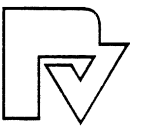
Essays on Microeconomics and Industrial Organisation



Contributions to Economics

www.springeronline.com/series/1262

Further volumes of this series can be found at our homepage.

Peter Meusburger/Heike Jöns (Eds.)
Transformations in Hungary
2001. ISBN 3-7908-1412-1

Claus Brand
Money Stock Control and Inflation Targeting in Germany
2001. ISBN 3-7908-1393-1

Erik Lüth
Private Intergenerational Transfers and Population Aging
2001. ISBN 3-7908-1402-4

Nicole Pohl
Mobility in Space and Time
2001. ISBN 3-7908-1380-X

Pablo Coto-Millán (Ed.)
Essays on Microeconomics and Industrial Organisation
2002. ISBN 3-7908-1390-7

Mario A. Maggioni
Clustering Dynamics and the Locations of High-Tech-Firms
2002. ISBN 3-7908-1431-8

Ludwig Schätzl/Javier Revilla Diez (Eds.)
Technological Change and Regional Development in Europe
2002. ISBN 3-7908-1460-1

Alberto Quadrio Curzio/Marco Fortis (Eds.)
Complexity and Industrial Clusters
2002. ISBN 3-7908-1471-7

Friedel Bolle/Marco Lehmann-Waffenschmidt (Eds.)
Surveys in Experimental Economics
2002. ISBN 3-7908-1472-5

Pablo Coto-Millán
General Equilibrium and Welfare
2002. ISBN 7908-1491-1

Wojciech W. Charemza/Krystyna Strzala (Eds.)
East European Transition and EU Enlargement
2002. ISBN 3-7908-1501-1

Natalja von Westernhagen
Systemic Transformation, Trade and Economic Growth
2002. ISBN 3-7908-1521-7

Josef Falkinger
A Theory of Employment in Firms
2002. ISBN 3-7908-1520-9

Engelbert Plassmann
Econometric Modelling of European Money Demand
2003. ISBN 3-7908-1522-5

Reginald Loyen/Erik Buyst/Greta Devos (Eds.)
Struggeling for Leadership: Antwerp-Rotterdam Port Competition between 1870–2000
2003. ISBN 3-7908-1524-1

Pablo Coto-Millán
Utility and Production, 2nd Edition
2003. ISBN 3-7908-1423-7

Emilio Colombo/John Driffill (Eds.)
The Role of Financial Markets in the Transition Process
2003. ISBN 3-7908-0004-X

Guido S. Merzoni
Strategic Delegation in Firms and in the Trade Union
2003. ISBN 3-7908-1432-6

Jan B. Kuné
On Global Aging
2003. ISBN 3-7908-0030-9

Sugata Marjit, Rajat Acharyya
International Trade, Wage Inequality and the Developing Economy
2003. ISBN 3-7908-0031-7

Francesco C. Billari/Alexia Prskawetz (Eds.)
Agent-Based Computational Demography
2003. ISBN 3-7908-1550-0

Georg Bol/Gholamreza Nakhaeizadeh/ Svetlozar T. Rachev/Thomas Ridder/ Karl-Heinz Vollmer (Eds.)
Credit Risk
2003. ISBN 3-7908-0054-6

Christian Müller
Money Demand in Europe
2003. ISBN 3-7908-0064-3

Cristina Nardi Spiller
The Dynamics of the Price Structure and the Business Cycle
2003. ISBN 3-7908-0063-5

Michael Bräuninger
Public Debt and Endogenous Growth
2003. ISBN 3-7908-0056-1

Brigitte Preissl/Laura Solimene
The Dynamics of Clusters and Innovation
2003. ISBN 3-7908-0077-5

Markus Gangl
Unemployment Dynamics in the United States and West Germany
2003. ISBN 3-7908-1533-0

Pablo Coto-Millán
Editor

Essays on Microeconomics and Industrial Organisation

Second, Revised and Enlarged Edition

With 48 Figures and 103 Tables

Springer-Verlag Berlin Heidelberg GmbH

Series Editors
Werner A. Müller
Martina Bihn

Editor

Professor Dr. Pablo Coto-Millán
University of Cantabria
Department of Economics
Avda. Los Castros s/n.
39005 Santander
Spain
cotop@unican.es

ISSN 1431-1933
ISBN 978-3-7908-0104-0 ISBN 978-3-7908-2670-8 (eBook)
DOI 10.1007/978-3-7908-2670-8

Cataloging-in-Publication Data applied for
A catalog record for this book is available from the Library of Congress.

Originally published by Physica-Verlag Heidelberg New York in 2002, 2004
Softcover reprint of the hardcover 1st edition 2004

Softcover Design: Erich Kirchner, Heidelberg

SPIN 10942383 88/3130/DK-5 4 3 2 1 0 – Printed on acid-free and non-aging paper

Table of Contents

Introduction

The aim of ***Essays on Microeconomics and Industrial Organisation (2nd edition)*** is to serve as a source and work of reference and consultation for the field of Microeconomics in general and of Industrial Organisation in particular.

Traditionally, Microeconomics is essentially taught as theory and, although handbooks illustrate the various microeconomic theories with examples and practical cases, they hardly ever offer an estimation of a demand, production and cost function. In fact, Microeconomics is explained with self-contained theories without empirical tests. The editor of these Essays has taught Microeconomics for twelve years in the traditional way; for the last ten years –in advanced courses and doctorates- he has offered a selection of empirical applications, which have complemented the traditional theoretical teaching. These applications have emerged from various research projects managed by the editor during the last ten years with the financial support of several institutions (DGICYT, DGESIT, CICYT, and R&D National Plan). The success in this type of teaching and the availability of recent original applications from authors usually collaborating with the editor has led the later to compose this text.

This text combines microeconomic theories with appropriate empirical tests. The standardised microeconomic analysis of demand, production and costs (supply) is set forth along with appropriate econometric techniques. Moreover, it should be pointed out that over the last two decades Microeconomics has greatly broadened its field of application. On the one hand, this has been due to the fact that the conditions required for existence, unicity and stability of the general competitive equilibrium have been met. This was the prevailing focus of Microeconomics in the Sixties and part of the Seventies. On the other hand, as big samples of inter- and intra-industrial data were increasingly available, a neo-classical Microeconomics branch emerged in the mid Seventies traditionally called in Britain New Microeconomics or Industrial Organisation, which on the structure-behaviour-results paradigm, binds together the earlier and new works on structural change and technical progress and applies new techniques -mainly panel data- which enable us to observe how the behaviour of the new agents affects industrial structure.

Demand, production and costs are parts of Microeconomics which are greatly active at present. Industrial structure and regulation, markets and failures of market constitute central nuclei of Industrial Organisation. Therefore, the second part of the title of these essays records this expression. Although I dare not give it a Roman numeral, a new volume will foreseeably emerge in the future recording new advances both in the expression of ideas and in the econometric cointegration and panel data techniques used here.

P. Coto-Millán (ed.), *Essays on Microeconomics and Industrial Organisation*

The text consists of four parts: Demand, Production and Costs (Supply), Market and Industrial Structure and Failures of Market and Industrial Regulation. Each part has three chapters.

Section I deals with demand and starts with a paper that studies industrial demand. Chapter 2 offers a study on air transport demand with respect to the remaining modes of transport. Chapter 3 presents the behaviour of the consumer's with respect to the introduction of a new product (High Speed Train AVE). To conclude, this section presents an approach to the hub-and-spoke systems from SVARs models, with a practical application to container traffic between the port of Bahía de Algeciras and other ports of the Spanish port system.

Section II deals with supply. In Chapter 5 a production function is estimated with panel data for the hospital industry and the results corresponding to asignative efficiency are presented. In Chapter 6 economic efficiency is analysed for the case of road transport firms. In Chapter 7 technical efficiency is analysed for the case of international air transport (1992-2000). Finally, Chapter 8 offers the productor's behaviour with technological innovation and structural change.

Section III studies the market and industrial structure. Chapter 9 covers the problems of measurement of inter- and intra-industry trade. Chapter 10 presents a categorical aggregation exercise with Spanish trade data. The following chapter presents a typical model of Industrial Organisation in which intra-industry activity is explained by industrial structure and the behaviour of the agents. To conclude, this section presents a paper on economic integration in which the link between structural adjustment and horizontal and vertical intra-industry trade is analysed.

Section IV starts with a chapter on organisation and regulation of the port industry with an application to Europe and Spain. Chapter 14 presents an application of positive theory of regulation to Spanish foreign trade. Chapter 15 studies the structure and regulation in electrical industry. In Chapter 16 the effects of a regulation of working time on labour market are studied. Chapter 17 studies transitional dynamics and endogenous growth for the case of public capital. Chapter 18 analyses the relationships among unions, wages and productivity. Chapter 19 includes a study on economic impact in Spanish ports. Chapter 20 studies economic impact in an airport. Chapter 21 includes a theoretical model which explains how the lack of co-ordination in the input and output of agents in industry may generate complex situations and presents an application. Chapter 22 describes a theoretical model that allows to calculate congestion effects on transport including an application. Finally, in Chapter 23 a cost-benefit analysis (CBA) of High Speed Train (AVE) Madrid-Barcelona is offered.

PART I. DEMAND

1 Modeling Seasonal Integrated Time Series: the Spanish Industrial Production Index

J. L. Gallego-Gómez
University of Cantabria (Spain)

In this paper the Box-Jenkins approach to the building of seasonal time series model is extended so that it is adequate to model seasonally integrated time series. To this end, the class of multiplicative ARIMA models is broadened in such a way that it allows to describe time series integrated at a few of the seasonal frequencies. Thus, tests for seasonal unit roots are not considered as a rival modeling approach, but can be used in the identification stage to decide the transformation inducing stationarity. The fit model is used to generate forecasts and to estimate unobservable components. The enhanced Box-Jenkins approach is illustrated modeling the Spanish Industrial Production Index.

1.1 Seasonal Time Series Models

The Box and Jenkins approach to the seasonal time series analysis can be sketched as follows. Firstly, the non-seasonal $1 - B$ and seasonal $1 - B^s$ differencing operators are used to convert a non-stationary series z_t into a stationary series w_t. It is usually necessary to use d-order non-seasonal and D-order seasonal differencing, that is,

$$w_t = (1-B)^d (1-B^s)^D$$

Then, the stationary series w_t is expressed, according the Wold decomposition theorem, as a weighted sum of current and past values of a white noise process

$$w_t = a_t + \sum_{j=1}^{\infty} \psi_j a_{t-j} = \psi(B) a_t$$

where

P. Coto-Millán (ed.), *Essays on Microeconomics and Industrial Organisation*

$$\psi(B) = 1 + \sum_{j=1}^{\infty} \psi_j B^j$$

Finally, to achieve parsimonious models the polynomial $\psi(B)$ is approximated by the rational polynomial

$$\psi(B) = \frac{\theta(B)\Theta(B^s)}{\phi(B)\Phi(B^s)}$$

where $\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$ and $\theta(B) = 1 - \theta_1 B - \ldots - \theta_q B^q$ are the non-seasonal autoregressive and moving-average polynomials which describe the dependence between consecutive data, and $\Phi(B) = 1 - \Phi_1 B - \ldots - \Phi_P B^P$ and $\Theta(B) = 1 - \Theta_1 B - \ldots - \Theta_Q B^Q$ are the seasonal autoregressive and moving-average polynomials describing the dependence between data which are s periods apart. Therefore, the nonstatinonary seasonal time series z_t is described by the general multiplicative model

$$\phi(B)\Phi(B^s)\nabla^d \nabla_s^D z_t = \theta(B)\Theta(B^s)a_t \tag{1.1}$$

The choice of the seasonal differencing to induce stationarity is based on the fact that seasonal time series show a cyclical behaviour with period s (s = 4 for quarterly data and s = 12 for monthly data). In the extreme case $z_t = z_{t-s}$, we have a homogeneous difference equation whose general solution is a linear combination of the s roots, $e^{-i2\pi j/s}$ (j = 0, 1, ..., s–1 and $i = \sqrt{-1}$), of the characteristic equation $1 - B^s = 0$,

$$z_t = \sum_{j=0}^{s-1} a_j e^{i2\pi j/s} = \alpha_0 + \sum_{k=1}^{[s/2]} (\alpha_k \cos(\frac{2\pi k}{s} t) + \beta_k \sin(\frac{2\pi k}{s} t))$$

where $\alpha_k \cos(\frac{2\pi k}{s} t) + \beta_k \sin(\frac{2\pi k}{s} t)$ is a harmonic oscillation with period $p_k = s/2\pi k$ and constant amplitude $A_k = (\alpha^2 + \beta^2)^{1/2}$, generated by the pair of conjugate complex roots $e^{-i2\pi k/s}$ and $e^{-i2\pi (k-s)/s}$; the harmonic with frequency $s/2$, $\alpha_{s/2}\cos(\pi t)$, only arises when s is even and is generated by the negative unit real root; $[s/2] = s/2$ if s is even and $[s/2] = (s-1)/2$ if s is odd. Hence, it is seen the key role played by the seasonal difference in representing the seasonal pattern.

Although the 1 - Bs filter has been commonly used in the Box-Jenkins approach, Abraham and Box (1978) pointed out that sometimes some of its factors could be sufficient to handle the seasonality. Such factors are found by expresing 1 - Bs in terms of its roots

$$1 - B^s = \prod_{j=0}^{s-1} (1 - e^{i2\pi j/s} B)$$

and joining each pair of complex roots into a second-order factor

$$(1 - e^{i2\pi k/s} B)(1 - e^{i2\pi (s-k)/s} B) = 1 - 2\cos(2\pi k/s)B + B^2$$

Thus, the seasonal difference can be factored as follows

$$1 - B^s = \prod_{k=0}^{[s/2]} S_k(B) \tag{1.2}$$

where

$$S_k(B) = \begin{cases} 1-B & k = 0 \\ 1 - 2\cos(2\pi k / s)B + B^2) & k = 1, \cdots, [(s-1)/2 \\ 1+B & k = [s/2] = s/2 \end{cases} \tag{1.3}$$

Hylleberg, Engle, Granger and Yoo [HEGY] (1990) developed unit root tests to identify the simplifying operators $S_k(B)$ in the underlying time series model. These authors extend the notion of integration to cover the seasonal frequencies. Thus, a time series z_t is integrated of order d at the frequency k, denoted by $z_t \sim I_k(d)$, if its autoregressive representation contains the factor $S_k(B)^d$. Empirical evidence based on the HEGY test reveals that most of the monthly and quarterly economic time series are integrated of order one at a few seasonal frequencies (see, e. g., Osborn 1990, Beaulieu and Miron 1993, Hylleberg, Jorgensen and Sorensen 1993).

In contrast, the empirical evidence found using the Box-Jenkins methodology has revealed that most of the monthly and quarterly seasonal time series can be adequately described by the IMA(0,1,1)(0,1,1)$_s$ model,

$$(1-B)(1-B^s)z_t = (1-\theta B)(1-\Theta B^s)a_t$$

implying that z_t is $I_0(2)$ and $I_k(1)$ for $k = 1, \ldots, [s/2]$.

Gallego and Treadway (1995) have shown why the empirical results obtained with the HEGY test cannot be found with the conventional multiplicative ARIMA class, and how to broaden such a family to allow for unit roots at a few seasonal frequencies. Consider for example the seasonal IMA(1,1)$_4$ model for quarterly time series

$$(1-B^4)z_t = (1-\Theta B^4)a_t \tag{1.4}$$

It is interesting to write (1.4) as

$$(1-B)(1+B^2)(1+B)z_t = (1-\Theta^{1/4}B)(1+\Theta^{1/2}B^2)(1+\Theta^{1/4}B)a_t \tag{1.5}$$

The three nonstationary factors on the left-hand side of (1.5) contribute to "spectral peaks" at the frequencies 0, 1 and 2, respectively. So, under the condition of invertibility $\Theta < 1$, z_t is integrated of order one, $z_t \sim I_k(1)$, at $k = 0$, 1 and 2. However, if the MA parameter Θ is positive and strictly noninvertible, $\Theta = 1$, then the model contains three common factors whose cancellation implies that z_t is integrated of order zero, $z_t \sim I_k(0)$, at $k = 0$, 1 and 2. In contrants, let the seasonal MA(1)$_4$ polynomial in model (1.4) be replaced with a nonseasonal MA(4) polynomial. Then, the resulting IMA(1,4)(1,0)$_4$ model

$$(1-B)(1+B^2)(1+B)z_t = (1-\theta_1 B - \theta_2 B^2 - \theta_3 B^3 - \theta_4 B^4)a_t \tag{1.6}$$

enables the description of a richer class of integrated series. Here, the case $z_t \sim I_k(0)$ arises when the model contains a common factor with frequency k, that is, when $\theta(B) = \theta^*(B)S_k(B)$. But, if the factor $S_k(B)$ is not included in the MA structure, then $z_t \sim I_k(1)$. In (1.6) the order of integration of the process is not necessarily the same at all frequencies. In fact, $z_t \sim I_k(d_k)$, being $d_k = 0$ or 1 for $k = 0$, 1, and 2.

When investigating the order of integration of quarterly time series, it is convenient to express the IMA(1,4)(0,1)$_4$ in such a way that it shows more clearly the presence of common factors. Thus, in the constrained IMA(1,4)(1,0)$_4$ model

$$(1-B)(1+B^2)(1+B)z_t = (1-\theta_0 B)(1+\theta_1 B^2)(1+\theta_1 B)a_t \tag{1.7}$$

where $0 < \theta_k \leq 1$, one can readily to see that $z_t \sim I_k(1)$ if $0 < \theta_k < 1$ and $z_t \sim I_k(0)$ if $\theta_k = 1$. The MA(4) polynomial has been constrained so that it has roots at the frequencies $k = 0$, 1 and 2. This constraint is especially relevant when modeling monthly time series because the replacement of a seasonal MA(1)$_{12}$ polynomial by a nonseasonal MA(12) polynomial leads to a nonparsimonious model, whose estimation could be problematic. In such a case, the constrained MA(12) polynomial can be expressed as the product of the following seven factors

$$(1 - \theta_0 B),\ (1 - \sqrt{3\theta_1}\, B + \theta_1 B^2),\ (1 - \sqrt{\theta_2}\, B + \theta_2 B^2),\ (1 + \theta_3 B^2),$$

$$(1 + \sqrt{\theta_4}\, B + \theta_4 B^2),\ (1 + \sqrt{3\theta_5}\, B + \theta_5 B^2),\ (1 + \theta_6 B)$$

where $0 < \theta k \leq 1$. In general, a MA(s) polynomial can be constrained so that it has s roots at the frequencies k = 0, 1, . . . , [s/2], which correspond to the following factors with real coefficients

$$\vartheta(B) = \prod_{k=0}^{[s/2]} \vartheta_k(B) \tag{1.8}$$

where

$$\vartheta_k(B) = \begin{cases} 1 - \vartheta_0 B & k = 0 \\ 1 - 2\cos(2\pi k / s)\sqrt{\vartheta_k}\, B + \vartheta_k B^2) & k = 1, \cdots, [(s-1)/2] \\ 1 + \vartheta_{s/2} B & k = s/2 \text{ and s even} \end{cases} \tag{1.9}$$

The subscript k of the coefficient ϑ_k indicates the number of cycles in a period of s time instants. Comparing (1.8)-(1.9) with (1.2)-(1.3) it can be clearly seen the one-to-one correspondence between the factors $\theta_k(B)$ of the constrained MA(s) polynomial and the simplifying factors $S_k(B)$ of the seasonal difference. Thus, each MA factor acts as an indicator for overdifferencing.

In summary, seasonal integrated time series can be described by the generalized multiplicative ARIMA model

$$\phi(B)\Phi(B^s)\nabla^d \prod_{k=1}^{[s/2]} S_k(B)^{d_k} z_t = \theta(B)\Theta(B^s)\vartheta(B)a_t \tag{1.10}$$

which is compactly written as $\varphi(B)z_t = \eta(B)a_t$.

1.2 Forecasting

In this section the implications that the generalisation of the conventional multiplicative ARIMA model has on the forecasting function are examined; in particular, those derived from the relaxation of the assumption that time series are integrated at all seasonal frequencies. To this end, two specifications for the model (1.10) are used: (i) the difference equation representation

$$z_{t+l} - \varphi_1 z_{t+l-1} - \cdots - \varphi_p z_{t+l-p} = a_{t+l} - \eta_1 a_{t+l-1} - \cdots - \eta_q a_{t+l-q} \tag{1.11}$$

which is useful to recursively compute point forecasts, and (ii) the specification in terms of innovations

$$z_{t+l} = a_{t+l} + \psi_1 a_{t+l-1} + \psi_2 a_{t+l-2} + \cdots \tag{1.12}$$

or

$$z_{t+l} = \psi(B)a_{t+l} = (1 + \psi_1 B + \psi_2 B^2 + \cdots)a_{t+l}$$

which is more convenient to calculate interval forecasts and to update forecasts as new data become available.

The minimum mean square error forecast of z_{t+l} at origin t, for lead time l, is the conditional expectation $\hat{z}_t(l) = E\left[z_{t+l} / z_t, z_{t-1}, \cdots\right]$. Thus, taking conditional expectations at time t in (1.11) we obtain for $l > q$ the difference equation

$$\hat{z}_t(l) - \varphi_1 \hat{z}_t(l-1) - \cdots - \varphi_p \hat{z}_t(l-p) = 0, \quad l > q \tag{1.13}$$

whose general solution is

$$\hat{z}_t(l) = b_1^{(t)} f_1(l) + b_2^{(t)} f_2(l) + \cdots + b_p^{(t)} f_p(l), \quad l > \max(q-p,0) \tag{1.14}$$

To compute the p coefficients $b_i^{(t)}$ we need p initial conditions or pivotal values $\hat{z}_t(r+1)$, $\hat{z}_t(r+2)$, ..., $\hat{z}_t(r+p)$, where $r = \max(q-p,0)$. Let the column vector $\mathbf{b}^{(t)}$ be the p coeficients $b_i^{(t)}$, $i = 1, \ldots, p$, then

$$\mathbf{b}^{(t)} = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\hat{\mathbf{z}}_t$$

where $\mathbf{F} = [f_i(j)]$ is a $p \times p$ matrix of fitting functions, $\hat{\mathbf{z}}_t = [\hat{z}_t(j)]$ is an $p \times 1$ column vector, and $j = r+1, \ldots, r+p$.

From (1.12), it follows that

$$\hat{z}_t(l) = \psi_l a_t + \psi_{l+1} a_{t-1} + \psi_{l+2} a_{t-2} + \cdots$$

Therefore, the forecast error l steps ahead

$$e_t(l) = a_{t+l} + \psi_1 a_{t+l-1} + \cdots + \psi_{l-1} a_{t+1}$$

has zero mean and variance

$$V[e_t(l)] = \sigma_a^2 (1 + \psi_1^2 + \cdots + \psi_{l-1}^2)$$

The coefficients ψ_j of the polynomial $\psi(B) = \eta(B)/\varphi(B)$ satisfy the difference equation

$$\psi_j - \varphi_1 \psi_{j-1} - \cdots - \varphi_p \psi_{j-p} = 0, \quad j > q$$

whose general solution is

$$\psi_j = c_1 f_1(j) + c_2 f_2(j) + \cdots + c_p f_p(j), \quad j > \max(0, q - p)$$

It is convenient to write

$$\boldsymbol{\psi} = \mathbf{F}\mathbf{c} \tag{1.15}$$

where $\boldsymbol{\psi} = (\psi_{r+1}, \ldots, \psi_{r+p})^T$ and $\mathbf{c} = (c_1, c_2, \ldots, c_p)^T$. The ψ_j weights are also used in updating the forecasts. As the forecasts of the future observation z_{t+l} made at origins t and $t-1$ are

$$\hat{z}_t(l) = \psi_l a_t + \psi_{l+1} a_{t-1} + \psi_{l+2} a_{t-2} + \cdots$$

$$\hat{z}_{t-1}(l+1) = \psi_{l+1} a_{t-1} + \psi_{l+2} a_{t-2} + \psi_{l+3} a_{t-3} + \cdots$$

the updating formula is

$$\hat{z}_t(l) = \hat{z}_{t-1}(l+1) + \psi_l a_t$$

Hence, it follows that

$$\mathbf{F}\mathbf{b}^{(t)} = \mathbf{G}\mathbf{b}^{(t-1)} + \boldsymbol{\psi} a_t$$

where $\mathbf{G} = [f_i(j+1)], j = r+1, \ldots, r+p$. Solving for $\mathbf{b}^{(t)}$ we have that

$$\mathbf{b}^{(t)} = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{G}\mathbf{b}^{(t-1)} + \mathbf{c} a_t$$

Now we can see the role played by the constrained MA operators. While the roots of the autoregressive operator determine the form of the eventual forecast function and the pattern of the ψ_j weights, the moving average coefficients are relevant to provide initical conditions. Since that the replacement of the seasonal $MA(1)_s$ operator by the nonseasonal $MA(s)$ operator does not change the orders of the extended ARMA model, neither the form of the forecast function nor the number of pivotal values are modified. Obviously to the extent the ψ_j weights changes, the pivotal values will change and also the position of the the eventual forecast function. However, the main implications arise when some of the MA roots are close to unit and cancel with some common simplifying operators, that is, when the series is $I_k(0)$ at a few k. In such a case, some coefficients c_i will be zero and, therefore, the forecasting function will contain non-updated or deterministic

components and the squares sum of ψ_j weights, $V[e_t(l)] = \sigma_a^2 c^T \mathbf{F}^T \mathbf{F} c$, will be smaller.

1.3 Unobservable Components

Several procedures have been proposed to estimate the trend T_t, seasonal S_t and irregular I_t components of a time series described by an ARIMA model (see, e. g. , Gallego 2000). I addopt an variant of the approach based on the eventual forecasting function. Equation (1.14) can be written as

$$\hat{z}_t(l) = \sum_{k=1}^{p_1} b_k^{(t)} f_k(l) + \sum_{k=p_1+1}^{p} b_k^{(t)} f_k(l) = \hat{T}_t + \hat{S}_t \tag{1.16}$$

where it has been assumed that the first p_1 fitting functions, $f_k(l)$ $(k = 1, \ldots, p_1)$ are nonseasonal, and $f_k(l)$ $(k = p_1+1, \ldots, p)$ are seasonal functions; $\hat{T}_t$ and $\hat{S}_t$ are the estimates for the trend and seasonal components.

Assuming that the unobservable components are described by ARIMA models, then it follows that

$$z_t = \hat{z}_{t-1}(1) + a_t,\ T_t = \hat{T}_{t-1}(1) + b_t,\ S_t = \hat{S}_{t-1}(1) + c_t,\ I_t = \hat{I}_{t-1}(1) + d_t$$

where b_t, c_t, and d_t are white noise innovations. Hence, it is clear that the estimation of each unobservable component involves the calculation of its one-step-ahead predictor and error. While the quantities $\hat{T}_{t-1}(1)$, $\hat{S}_{t-1}(1)$, and $\hat{I}_{t-1}(1)$ can be obtained easily by breaking down $\hat{z}_{t-1}(1)$, the remaining problem is how to estimate the shocks b_t, c_t, and d_t.

For the additive seasonal model

$$z_t = T_t + S_t + I_t$$

it follows that

$$\hat{z}_{t-1}(1) = \hat{T}_{t-1}(1) + \hat{S}_{t-1}(1) + \hat{I}_{t-1}(1)$$

and

$$a_t = b_t + c_t + d_t$$

It is now apparent that the difficulty inherent to the decomposition of seasonal time series is how to isolate the three component innovations of the residual series a_t. As in other approaches, the identification problem also arises here. Nonetheless, the class of observationally equivalent decompositions is readily obtainable from the different allocations of white noise series among the one-step-ahead predictors

$$T_t^* = \hat{T}_{t-1}(1) + e_{1t},\ S_t^* = \hat{S}_{t-1}(1) + e_{2t},\ I_t^* = \hat{I}_{t-1}(1) + a_t - e_{1t} - e_{2t}$$

where e_{1t} and e_{2t} are white noise process. Hence, the canonical decomposition is

$$T_t^C = \hat{T}_{t-1}(1), \quad S_t^C = \hat{S}_{t-1}(1), \quad I_t^C = \hat{I}_{t-1}(1) + a_t \tag{1.17}$$

The $\hat{I}_{t-1}(1)$ component will be zero for models with $p \leq q$, since none of the AR roots generates an irregular movement. In such a case, the irregular component is equal to the residuals of the model fit to the data. In contrast, for top-heavy models $p < q$ the eventual forecast function for $\hat{z}_{t-1}(1)$ includes moving average terms, which are allocated to the irregular component. This result is consistent with the observation by Burman (1980) that top heavy models lead to irregular components with moving average structure.

1.4 Empirical Analysis

In this section, I embed the seasonal integration notion into the Box-Jenkins iterative modeling strategy. Although this can be made in several ways I follow a "general to specific" approach by assuming that times series, when showing seasonality, are integrated of order one in all of the seasonal frequencies. This general assumption does not imply a big change in the Box-Jenkins practice, in which the airlines models arises often as a tentative representation for many seasonal time series. Then, a more specific model is arrived at by looking for overdifferencing at single frequencies. The specific model, which can be thought of as a mixed regression-ARIMA model or transfer function model, is used to generate forecast and to estimate unobservable components.

Fig. 1.1 displays several identification tools for the Spanish Industrial Production Index from January 1986 to January 2000, 169 monthly observations.

Following the Box-Jenkins approach it is concluded that $\nabla\nabla_{12} z_t$ seems to be a stationary series. On the other hand, its periodogram, column 3 in Fig. 1.1, peaks at the calendar frequency 0. 348 (close to $\pi/3$, indicating a trading day variation which can be modelled (see, e. g. , Bell and Hillmer 1983) as

$$TD_t = \sum_{i=0}^{6} \beta_i td_{it}$$

where td_{it} (i = 0, ..., 6) are, respectively, the number of Sundays, Mondays, . . . in month t. Since $lom_t = \sum_{i=0}^{6} td_{it}$, where lom_t is the length of month t, an alternative specification for the trading day variation is

$$TD_t = \beta_0^{'} lom_t + \sum_{i=1}^{6} \beta_i^* td_{it}^*$$

where $\beta_0^{'} = (1/7)\sum_{i=0}^{6} \beta_i$ $td_{it}^* = td_{it} - td_{0t}$, and $\beta_i^* = \beta_i - \beta_0^{'}$ measures the difference between the *i-th* day effect and the *lom* effect. Also it is expected that the production decreases due to holidays such as Easter. The Easter effect here is modeled as

$$E_t = \alpha \xi_t^{Easter}$$

where ξ_t^{Easter} is the proportion of Easter days falling in the month t. Here, it is assumed that the length of the Easter is four days.

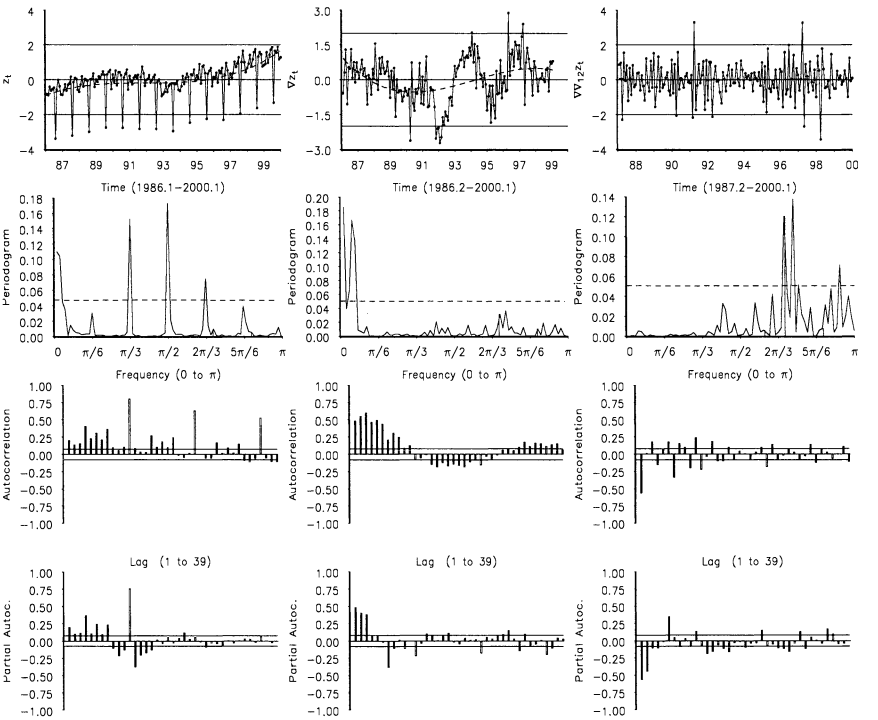


Fig. 1.1. Plot, Periodogram, Correlogram and Partial Autocorrelation Function for the series: z_t, ∇z_t y $\nabla\nabla_{12} z_t$

Since the sample autocorrelation function (SACF) and the sample partial autocorrelation function of $\nabla\nabla_{12} z_t$ can be distorted by the calendar and Easter effects, instead of identifying a tentative multiplicative ARMA structure, the airline model with deterministic variables

$$z_t = \beta_0' lom_t + \sum_{i=1}^{6} \beta_i^* td_{it}^* + \alpha \xi_t^{Easter} + N_t$$

$$\nabla\nabla_{12} N_t = (1 - \theta B)(1 - \Theta B^{12}) a_t$$

is estimated as starting point. The exact maximum likelihood estimates obtained with the AMB program (Gallego 2000b) are

$$z_t = \underset{(0.68)}{-6.93}\xi_t^{\text{Easter}} + \underset{(1.10)}{2.74}\text{lom}_t + \underset{(0.33)}{0.15}\text{td}_{1t}^* + \underset{(0.34)}{0.23}\text{td}_{2t}^* + \underset{(0.32)}{0.76}\text{td}_{3t}^*$$

$$+ \underset{(0.32)}{0.73}\text{td}_{4t}^* + \underset{(0.33)}{0.57}\text{td}_{5t}^* - \underset{(0.33)}{1.35}\text{td}_{6t}^* + N_t$$

$$\nabla\nabla_{12}N_t = \underset{(0.07)}{(1-0.47B)}\underset{(0.08)}{(1-0.83B^{12})}a_t$$

$$a = .054(.175), \sigma_a = 2.19, \max(\rho_{ij}) = -.58,$$

$$Q(39) = 58, \text{LLF} = -343.4, \text{AIC} = 1.69 \tag{1.18}$$

Some of the usual diagnostic checks reveal model inadequacies: (i) the Q-statistic for the first $n/4$ residual autocorrelations is greater than the 5% point for χ^2 with 37 degrees of freedom, 52. 192320; (ii) the SACF and SPACF take large values compared with the $\pm 2/\sqrt{n}$ bands at some lags. While it is not clear how to reformulate this model with conventional operators, some improvement is obtain by replacing the $\text{MA}(1)_{12}$ i by a constrained MA(12) polynomial and including an $\text{AR}(2)_{12}$ polynomial to describe the annual dependence structure indicated by lags 12, 24 and 36. The estimated model is

$$z_t = \underset{(0.53)}{-7.23}\xi_t^{\text{Easter}} + \underset{(0.91)}{2.21}\text{lom}_t + \underset{(0.26)}{0.25}\text{td}_{1t}^* - \underset{(0.27)}{0.10}\text{td}_{2t}^* + \underset{(0.26)}{1.26}\text{td}_{3t}^*$$

$$+ \underset{(0.26)}{0.28}\text{td}_{4t}^* + \underset{(0.26)}{1.35}\text{td}_{5t}^* - \underset{(0.26)}{1.95}\text{td}_{6t}^* + N_t$$

$$\underset{(.14)}{(1-.40B^{12}+.21B^{24})}\nabla\nabla_{12}N_t = \underset{(.09)}{(1-.34B)}\underset{(.06)}{(1-.99B)}\underset{(.04)}{(1-1.71B+.97B^2)}$$

$$\underset{(-.-)}{(1-1.0B+1.0B^2)}\underset{(.05)}{(1-.97B^2)}\underset{(.06)}{(1+.97B+.93B^2)}\underset{(.15)}{(1-1.73B+.99B^2)}\underset{(.03)}{(1+.96B)}a_t$$

$$a = .033(.174), \sigma_a = 2.18, \max(\rho_{ij}) = -.58,$$

$$Q(39) = 50, \text{LLF} = -342.8, \text{AIC} = 1.79 \tag{1.19}$$

where the presence of a common factor $1 - B + B^2$ on both sides of the model equation is discovered. This is indicative of the presence of a single deterministic seasonal component with a period of six months or a frequency of two cycles per

year. Thus, the common factor can be removed, but a pair of trigonometric variables c_{2t} and s_{2t} are added to the model (see,e. g. , Box, Jenkins and Reinsel 1994). Furthermore, the simplifying factors $1 - B$ and $1 - \sqrt{3}B + B^2$ of the seasonal difference are nearly cancelated by the associated MA operators, indicating that a constant term and a pair of trigonometric variables c_{1t} and c_{2t} must be included. Proceeding in this way we reach the estimated model

$$
\begin{aligned}
z_t = & -7.06\xi_t^{Easter} + 3.26lom_t + .06td_{1t}^* + .29td_{2t}^* + .66td_{3t}^* + .78td_{4t}^* + .60td_{5t}^* \\
& (0.54) \quad (0.91) \quad (.26) \quad (.27) \quad (.26) \quad (.26) \quad (.26) \\
& -1.40td_{6t}^* + 1.27c_{1t} + 5.90s_{1t} + 1.23c_{2t} - 7.73s_{2t} - 10.02c_{3t} - 2.03s_{3t} \\
& (.26) \quad (.31) \quad (.40) \quad (.29) \quad (.22) \quad (.46) \quad (.16) \\
& + 2.53c_{4t} + 5.15s_{4t} + 1.72c_{5t} - 7.66s_{2t} + N_t \\
& (.16) \quad (.40) \quad (.21) \quad (.74) \\
& (1-.28B^{12} + .23B^{24})[\nabla(1+B)N_t - .40] = (1-.36B)(1+.96B)at \\
& (.08) \quad (.09) \quad (.19) \quad (.09) \quad (.03) \\
& a = .033(.174), \sigma_a = 2.18, \max(\rho_{ij}) = -.58, \\
& Q(39) = 50, LLF = -342.8, AIC = 1.79
\end{aligned}
\tag{1.20}
$$

Now the main diagnosis checks do not reveal serious model inadequacies: the Q-statistics is less than the 50% point of the χ^2 with 40 degrees of freedom, and the residual plot, cumulative periodogram, correlogram and SPACF are consistent with the white noise hypothesis.

Several implications are derived from model (1.20). Firstly, the time series z_t is $I_k(1)$ at k = 0 and 6, but $I_k(0)$ at k = 1, ..., 5. Similar results are found with the Beaulieu and Miron test for seasonal unit roots showed in Table 1. Note that these results can not be indicated by the airlines model. Secondly, the trend component is described by a linear trend with stochastic ordinate, but deterministic slope. Thirdly, the seasonal component is the sum of five harmonic oscillations with deterministic amplitude and one oscillation with stochastic deterministic. Therefore, both the trend and seasonal component are a mixture of deterministic and stochastic terms. Finally, a damped cyclical component with period 4. 96 years is described by the annual $AR(2)_{12}$ polynomial. The estimates of three unobservable components, following the method described in the above section, are shown in Fig. 1.2.

Table 1.1. Test for seasonal unit roots

	a_0	a_1	b_1	a_2	b_2	a_3	b_3	a_4	b_4	a_5	b_5	a_6
t	-1. 49	-2. 69	-3. 80	-4. 32	-248	-4. 06	-1. 12	-3. 29	2. 21	-2. 93	. 20	-2. 68
F	.	10. 4	.	14. 4	.	9. 10	.	8. 48	.	4. 34	.	

Lags: 5, 6, 7, 12, 15, 32. Deterministic inputs: Linear trend, seasonal dummies and calendar effects. Diagnostic checks: $\hat{\sigma}_a$ = 1. 8301, $Q(31)$ = 33. 01, LLF = -252. 91, AIC = 1. 8327.

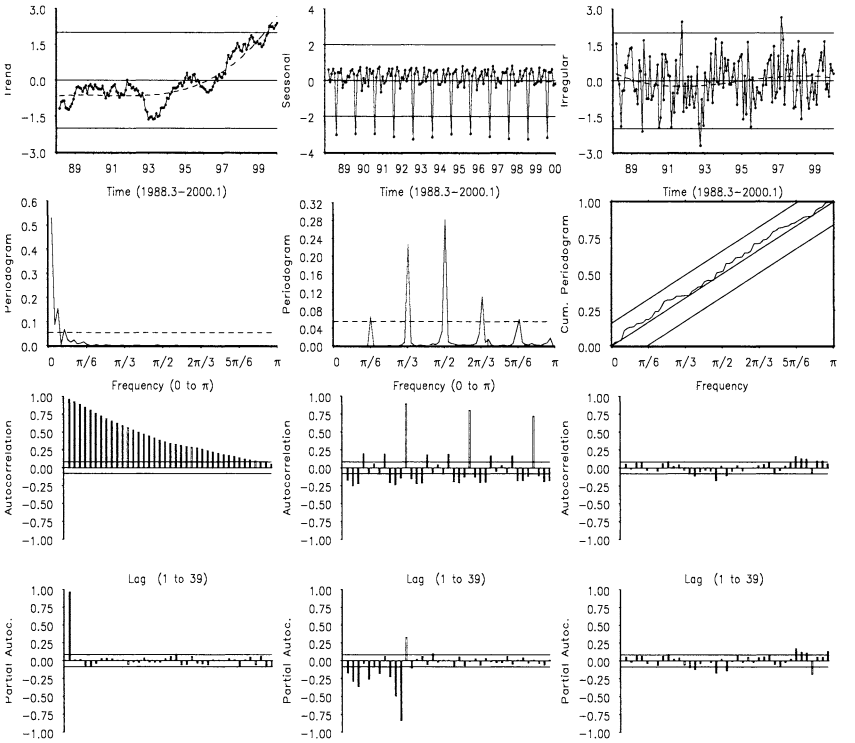


Fig. 1.2. Estimated trend, seasonal and irregular components for the Spanish IPI provided by model (1.20).

References

B. Abraham and G. E. P. Box. Deterministic and forecast-adaptive time-dependent models. Applied Statistics, 27:120 –130 (1978)

J. J. Beaulieu and J. A. Miron. Seasonal unit roots in aggregate U. S. data. Journal of Econometrics , 55:305 –328 (1993)

W. R. Bell and S. C. Hillmer. Modelling time series with calendar variation. Journal of the American Statistical Association, 78 (383):526 –311, September 1983

G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. Time series analysis, forecasting and control . Prentice Hall, Engle Woods, New Jersey, 3rd edition, 1994

J. P. Burman. Seasonal adjustment by signal extraction. Journal of the Royal Statistical Society A , 143:321 –337 (1980)

H. S. Lee E. Ghysels and J. Noh. Testing for unit roots in seasonal time series. Journal of Econometrics , 62:415 –442 (1994)

J. L. Gallego. User guide for the arima model building package. Departamento de Economía, Universidad de Cantabria (2000b)

J. L. Gallego. Measuring seasonal variations wit regarima models. Departamento de Economía, Universidad de Cantabria (2000a)

J. L. Gallego and A. B. Treadway. The general family of seasonal stochastic processes. Departamento de Economía (1995)

D. R. Osborn. A survey of seasonality in uk macroeconomic variables. International Journal of Forecasting , 6:327 –336 (1990)

C. Jorgensen S. Hylleberg and N. K. Sorensen. Seasonality in macroeconomic time series. Empirical Economics , 18:321, 325 (1993)

C. W. J. Granger S. Hylleberg, R. F. Engle and B. S. Yoo. Seasonal integration and cointegration. Journal of Econometrics , 44:215 –238 (1990)

G. W. Schwert. Tests for unit roots: a monte carlo investigation. Journal of Business and Economic Statistics , 7:147 –160 (1989)

2 Passenger's Choice of Air Transport under Road Competition: the Use of Cointegration Techniques*

J. Baños-Pino
University of Oviedo (Spain)

P. Coto-Millán
University of Cantabria (Spain)

V. Inglada-López de Sabando
University Carlos III (Spain)

In this study, a theoretical model for the passenger transport demands in Spain is presented. Quarterly data have been used for the 1980.I-1992.IV period.

Cointegration techniques, which are subject to a wide range of tests, are used to obtain short and long run equations. Moreover, the product, price and cross elasticities of each mode of transport are obtained. These estimations may be used to analyse the effects of transport tariffs on income changes, as well as to predict short- and long run traffics.

2.1 Introduction

The initial models of passenger transport demand were the aggregate "modal split models". In these models, there has been an attempt to determine the number of journeys by a given set of modes of transport for two towns, taking into account the characteristics of the passengers. Studies on modal split, such as Quandt and Baumol (1966), Boyer (1977), and Levin (1978), have been criticised by Oum

* This research was partly financed by the PS95-0095 research project from the DGICYT a body from the Spanish Ministry of Science and Education.

P. Coto-Millán (ed.), *Essays on Microeconomics and Industrial Organisation*

(1979) and Winston (1985), among others, for the few variables used to account for the motivation in the user's behaviour, and for using very simple linear patterns in their estimations.

Several models of aggregate passenger transport demand based on the user's behaviour have been carried out in order to improve the previous ones. The user's utility is optimised in these models in line with the classic theory of consumer's behaviour and demand. The work by Oum and Guillen (1979) based on the user's behaviour is a typical example in which the passenger demand in Canada is analysed.

Some disaggregate research based on the user's behaviour has also been done on passenger transport demand. The most significant work on these models is McFadden (1973, 1974). In these works, the user takes a discrete choice of some of the different modes of transport (railway, air, road transport, etc.) and it is assumed that the mode chosen optimises the utility for the user.

Spanish interurban passenger transport was first studied in the "Elasticidad de la Demanda del Transporte Público de Viajeros" (Elasticity of the Passenger Public Transport Demand) by the Instituto de Estudios de Transportes y Comunicaciones (Institute of Transport and Communication Studies) (1978). This was analysed by Vázquez (1985) in a work carried out by the Secretaría General del Ministerio de Transportes (General Secretariat of the Ministry of Transport). In addition, other studies such as that by Inglada (1992), Coto-Millán and Sarabia (1994) and Coto-Millán, Baños and Inglada (1997) have been carried out on this issue. The elasticities of the modes of transport in the main regions were studied in IETC (1978) and Vázquez (1985). Price elasticities have been studied in Inglada (1992) for monthly data between 1980.01 and 1988.12, with time series in which the residues have been modelled with the Box-Jenkins techniques. Uniequational models have been carried out in Coto-Millán and Sarabia (1994) in order to estimate income elasticities, using the Industrial Production Index and the Electric Power Consumption, for the 1980.01-1988.12 period, and monthly data have been used in the estimations. In these works, the series is also modelled by the Box-Jenkins methodology.

An original model is offered in this paper in order to estimate price income and cross elasticities for the 1980.I-1992.IV period, applying cointegration techniques and using monthly data. Such techniques allow the estimation of short run elasticities, which add immediate responses to price and income changing, and the estimation of long run elasticities which allow to see the effects of the price and/or income changing produced later on.

This research offers a model according to the second proposal above, being based on a microeconomic analysis, which can be considered as classic. Its structure is very simple. Point number 2.2 presents the theoretical model for Spanish passenger transport demand. In the next point, the data used are described. Point number 2.4 presents the estimations of the different demands. Finally, the main conclusions are offered in point number 2.5.

2.2 The Model

Suppose a typical user whose preferences are weakly separable, then modelling passenger transport demand constitutes the second stage of a two-stage budget process. Therefore, the user's spending firstly falls into two large categories: passenger transport services and all other goods and services; secondly, the user's income is assigned to the goods and services contained in each of these two categories. That is to say, the utility function of the representative user is as follows:

$$U = U(X_1, X_2, ..., X_k; X_{k+1}, ..., X_n)$$

where the vector $X_i = (X_1, X_2, ..., X_k)$ with i = 1, 2,..., k represents passenger transport services; the vector $X_j = (X_{k+1}, ..., X_n)$ with j = k+1,...n represents goods and services except for those corresponding to passenger transport, and U represents a utility function which is continuous and differentiable, monotonic, increasing, and strictly quasi-concave.

The consumer balance is reduced to:

$$\max U(X_i; X_j)$$

$$\text{subject to: } P_i \cdot X_i + P_j \cdot X_j = Y$$

where $P_i = (P_1, P_2, ..., P_k)$ and $P_j = (P_{k+1}, ..., P_n)$ are the prices, and where Y represents the user's level of income.

First order conditions allow obtaining the following typical user's Marshallian demands:

$$X_i = X_i(P_i, P_j, Y) \quad (2.1)$$

$$X_j = X_j(P_i, P_j, Y) \quad (2.2)$$

Of these individual demand functions, function (2.1) is interesting for us since it corresponds with passenger transport services.

Equation (2.1) still presents some problems. First, functions such as (2.1) should be valid for any income distribution among the different economic agents. If this were not the case, this function would provide as many values as income Y distributions among the possible users and, therefore, such a function would not exist. Another assumption would be that income is distributed under a specific rule. Once this rule has been established, integrability conditions are verified and the existence of the aggregate Marshallian demand functions is guaranteed (Varian, 1992). However, there are no data to go along these lines. In order to solve this problem in this study we can assume that all the users have the same level of income.

Function (2.1) is general enough to analyse the passenger transport service demands -long-distance railway, air and road transport- identifying the different subindexes for the amounts demanded in each service.

From 1980.I to 1992.IV, passenger transport services in Spain have been provided under different regulation conditions. The government company RENFE and Iberia have the monopoly of railway and air national transport in Spain

respectively, and road transport is provided by private companies which have exclusive routes under a system called "right of testing". It can be said that trump road passenger transport, which has a low incidence in quantitative terms, is the only mode of transport, which has not yet been regulated. However, given the impossibility to obtain quarterly statistical data on passenger road transport, and with the aim of adding interregional transport on the user's own vehicles, the premium petrol consumption variable has been used. The premium petrol consumption has also been regulated by the government during the period of this study. Under such regulatory conditions and with the aim of avoiding any problems arising from the supply-demand simultaneity, we assume that supply is exogenous in relation with prices and income and is determined by the decisions of the government.

2.3 The Data

Data on the series of passenger departures and arrivals of national fligths in Spanish airports (AERV) have been obtained from the series provided by the Informes de Coyuntura del Ministerio de Transportes, Turismo y Comunicaciones (Reports on the State of the Transport, Tourism and Communication Ministry) No data on road transport passengers are available and a "proxy" such as premium petrol consumption has been used in order to approximate the transport on the user's own vehicles. This variable, QGAS, has been obtained from the Dirección General de Previsión y Coyuntura del Ministerio de Economía y Hacienda (General Management of the Finance Ministry Forecast and State). The gas-oil consumption variable (QGLEO) has also been used with the aim of approximating the behaviour of regular and tramp passenger transport in public services. However, the results obtained are significantly anomalous and the reason for this may be that this variable shows the behaviour of road transport of goods (which is much more important in terms of consumption), rather than of passengers.

The data on the series of long distance railway prices (PF) and air transport fares (PA) have been obtained from the monthly series worked out from the fares of the Boletines Oficiales del Estado (Official State Reports), evaluated within the period in which each tariff is in force.

The data on the prices of premium petrol (PGAS) have been obtained from the Dirección General de Previsión y Coyuntura del Ministerio de Economía y Hacienda, as monthly data, also evaluated within the period in which each fare is in force.

Data on the prices of gas oil (PGLEO) have been obtained from CAMPSA until 1992. From then onwards, the data from the Compañía Logística de Hidrocarburos (Hydrocarbon Logistic Company) have been recorded for further studies.

The data on the income variable have been obtained considering the Spanish quarterly GDP as "proxy". The series used for the 1980.I-1989.IV period come from Contabilidad Nacional Española (Spanish National Accounting).

2.4
Marshallian or Non Compensated Demands of Interurban Passenger Transport: Air and Road Transport

In this paper we use cointegration analysis and error-correction modelling techniques to estimate transport demand functions. The analysis of cointegration offers a generic route to test the existence of long-term relationships, suggested by the economic theory, among various integrated variables, which has provided the most satisfactory results of the various approaches previously attempted (Inglada (1992), Coto-Millán and Sarabia (1994)). Engle and Granger (1987) formalize the equivalence between cointegration and error-correction.

Before an error-correction model can be constructed according to Engle and Granger's representation theorem, we need to establish that variables in demand equations are integrated, at most integrated of order 1, I(1), and are cointegrated. First of all, we examine the properties of individual time series by means of standard tests. The results of these tests suggest the presence of a single unit root in all series and this implies that the levels of these variables are nonstationary but the first difference is stationary.

The existence of a long run relationship among variables has been verified in this paper by applying the Engle and Granger (1987) technique and the Johansen (1988) and the Johansen and Juselius (1990) procedure. The Engle and Granger approach to test the existence of cointegration is characterized by a Dickey-Fuller statistic, which is used to test the existence of a unit root in the residual of static cointegration regressions. Johansen (1988) and Johansen and Juselius (1990) develop a maximum likelihood estimation procedure that has several advantages over the procedure suggested by Engle and Granger, because it eliminates the assumption that the cointegration vector is unique and it takes into account the error structure of the underlying process.

We have estimated some equations from the specifications in model (2.1) by adjusting the variables to each mode of transport. All variables headed by letter L are in natural logs and those headed by letter D are in differencess. The statistical "t" is presented within brackets under each coefficient.

2.4.1 Air Transport Demand

Long run

The estimated equation of long run balance cointegration has provided the following results:

$$LAERV_t = \underset{}{-1.88} \ \underset{(-2.13)}{-1.38} \ LPA_t \ + \underset{(21.14)}{1.48} \ LGDP_t$$

$$\underset{(-6.43)}{}$$

R^2 adjusted = 0.91 ; S.E. = 0.04 ; D.W. = 1.25 ; D.F.[a] = -4.14 ; D.W.[b] = 1.79

[a] Indicates statistical significance at the 5% level.

[b] Is the Durbin-Watson from equation used to computed the DF (Dickey-Fuller) statistic.

Table 2.1 presents the results of cointegration tests applying the Johansen methodology. To test the null hypothesis of at most r cointegrating vectors, Johansen (1988) proposed the trace test . The critical value for these statistics has been generated by simulation by Osterwald-Lenum (1992) and the Pantula principle has been employed to determine the cointegration rank. Then, if Johansen technique is applied to a VAR along with three lags and a restricted constant, it is concluded that there is only one cointegration vector.

Table 2.1

Number of cointegration vectors			
Under H_0	Under H_1	Trace test	95% critical value (a)
$r = 0$	$r \geq 1$	41.14	34.91
$r \leq 1$	$r \geq 2$	19.60	19.96
$r \leq 2$	$r = 3$	6.83	9.24

(a) Critical values are from Osterwald-Lenum (1992)

After normalisation, the following long run solution is obtained:

$$LAERV_t = -1.76 \ -1.41 \ LPA_t + 1.17 \ LGDP_t$$

In both estimated equations, the long run elasticity of air transport demand with respect to the GDP is close, somewhat higher than the unit ad takes 1.17 and 1.48 values as it would correspond to normal goods and particularly to "luxury" goods. The estimated long run own-price elasticity of goods is negative with values ranging from 1.38 to 1.41, which reflects a significant response of the demand to price changing.

Short run

All the demands of passenger transport have been specified, according to the Granger representation theorem, in the form of a model with error correction mechanism (ECM). This model incorporates the long run relationships, contained in the ECM, as well as the dynamics implied by the deviations from this equilibrium path and the adjustment process to recover it. The coefficients of the variables in differences represent short run elasticities. The joint non-linear estimation presents the following results:

$$DLAERV_t = \underset{(-3.92)}{-0.51} \ (LAERV_{t-1} + \underset{(1.96)}{2.24} + \underset{(2.06)}{1.12} \ LPA_{t-1} - \underset{(-9.8)}{1.47} \ LGDP_{t-1})$$

$$\underset{(-3.07)}{- 0.43} \ DLPAERV_{t-4} + \underset{(2.14)}{0.45} \ DLPGAS_t - \underset{(-2.36)}{0.78} \ DLPA_t$$

R^2 adjusted = 0.95 ; S.E. = 0.048; F = 147.13; D.W. = 2.10

Serial Correlation: Ljung-Box: Q(1) = 0.30;

Q(2) = 0.19;

Q(3) = 1.36;

Q(4) = 2.44

Residual Normality: Bera-Jarque: N(2) = 1.08

Heteroscedasticity: ARCH (1-4) = 1.27

The long run elasticities obtained for this and the previous model do not differ from each other significantly. Then, long run income elasticity is now 1.47 in comparison with the former values 1.17 and 1.48, as it corresponds to luxury goods or services. The negative value of the own-price elasticity of goods is 1.12 in comparison with the former 1.38 and 1.41 values.

Short and long run elasticities are once more slightly different. Short run elasticities clearly present the inelastic feature of the demand, and a substitution effect of road transport, which has never been revealed before, is detected. Gross and net substitution relationships between air and road transport result once more from these estimations.

2.4.2 Road Transport Demand

Long run

In the inter-city road passenger transport demand equation, the dependent variable is the amount of premium petrol, in logs, LQGAS, and the independent variables are log real rpices of premium petrol, LPGAS and LGDP already defined. The equation of the long-term balance cointegration estimated, yielded the following results:

$$LQGAS_t = \underset{(-3.21)}{-3.80} \underset{(-1.94)}{- 0.13\ LPGAS_t} + \underset{(8.29)}{1.11\ LGDP_t}$$

R^2 adjusted = 0.94 ; S.E. = 0.03 ; D.W. = 1.51; D.F.[a] = -5.52 ;D.W.[b] = 2.01

[a] Indicates statistical significance at the 5% level.

[b] Is the Durbin-Watson from equation used to computed the DF (Dickey-Fuller) statistic.

Table 2.2 presents the results of Johansen's cointegration test. This test strongly reject the null hypothesis of no cointegration (r = 0), but not the null hypothesis of at most one cointegrating vector, so there appears to be a single cointegrating vector, wich implies, after normalization, the following long run solution:

$$LQGAS_t = 2.85 - 0.47\ LPGAS_t + 0.3611\ LGDP_t$$

Table 2.2

Number of cointegration vectors			
Under H_0	Under H_1	Trace test	95 % critical value (a)
r = 0	r ≥ 1	41.87	53.12
r ≤ 1	r ≥ 2	19.19	34.91
r ≤ 2	r =3	7.90	19.96

(a) Critical values are from Osterwald-Lenum (1992)

The results obtained from the long run estimations provide elasticities of 0.361 and 1.11 with respect to the GDP, relationships that characterise these services as basic goods rather than as luxury goods, always within the context of normal goods. The own-price elasticities of the goods take the negative values 0.13 and 0.47, once more referring to essential goods with inelastic demand and slight demand variations as a response to tariff changes (if we consider such changes as proportional to premium petrol price changing).The gas-oil demand equation QGLEO presents very similar values with respect to its price and to the GDP variable.

Short run

Finally, we estimate an error-correction model to integrate short run dynamics with long run equilibrium, which presents the following results:

$$DLQGAS_t = \underset{(-4.85)}{-0.69}\,(LQGAS_{t-1} \underset{(2.20)}{+3.88} \underset{(1.68)}{+0.15}\,LPGAS_{t-1} \underset{(-6.19)}{-1.11}\,DLGDP_{t-1}) -$$

$$\underset{(-2.73)}{-0.36}\,DLPGAS_{t-4} \underset{(2.23)}{+0.34}\,DLPA_t$$

R^2 adjusted = 0.95 ; S.E. = 0.036 ; F = 212.45 ; D.W. = 2.13

Serial Correlation: Ljung-Box: Q(1) = 0.28;
Q(2) = 1.91;
Q(3) = 4.81;
Q(4) = 4.82

Residual Normality: Bera-Jarque: N(2) = 4.16

Heteroscedasticity: ARCH (1-4) = 1.17

The long run elasticities obtained for the ECM equation do not differ from the previous model. The value of the GDP long run demand elasticity is now 1.11, equal to the Engle and Granger approach, and the negative value of the long run own-price elasticity of goods is 0.15, while the former values were 0.13 and 0.47.

The estimated short run own-price elasticities of goods have the negative value 0.36 and a cross elasticity of 0.34 with respect to air transport price. In the short

run, it is possible to speak about gross substitution relationships between road and air transport. However, it is not possible to meet any conclusion with respect to the net substitution or complementary relationships of these transport services without any further assumption.

2.5 Conclusions

A theoretical model of air passenger transport demand has been presented in this paper. With quarterly aggregated Spanish data, equations of inter-city passenger air and road transport demand have been specified for 1980.I and 1992.IV.

Moreover, different demand function estimations have been carried out using cointegration techniques, and have been subject to a wide evaluation which allows us to check the adequacy of this method with respect to others used in earlier works by Inglada (1992) and Coto-Millán and Sarabia (1994).

Each specific demand may require more detailed studies, especially road transport. However, having carried out the estimations, it is possible to draw conclusions as regards income, the own-price elasticity of goods and cross price elasticities:

- Long-term income elasticities are all positive and all the services are normal goods. Income elasticities are very close to the unit for air transport, and slightly below the unit for road transport.
- The own-price elasticities of goods increase parallel to the quality of the service, since they increase with fares, and present values close to the unit for air transport. They are clearly inelastic for road transport.
- All cross elasticities present positive values and they are below the unit. Gross and net long run substitution relationships between air and road transport and gross substitution relationships between road and air transport can be guaranteed, but net substitution relationships between these cannot.

References

Boyer, K.D.: Minimum Rate Regulations, Modal Split Sensitives, and the Railroad Problem. Journal of Political Economy 85(3), 493-512 (1977)

Coto-Millán, P., Sarabia, J.M.: Intercity Public Transport in Spain 1980-1988: Elasticities, Prices, Income and Time series. Mimeo. University of Cantabria, Department of Economics (1994)

Coto-Millán, P., Baños-Pino, J., Inglada, V.: Marshallian Demands of Intercity Passenger Transport in Spain: 1980-1992. An Economic Analysis. Transportation Research-E 33-2, 79-96 (1997)

Engle, R.F. and Granger, C.W.: Cointegration and error Correction: representation, estimation and testing. Econometrica, 55, 251-276 (1987)

I.E.T.C.: Elasticidad de la Demanda del Transporte Público de Viajeros. Ministerio de Transporte (Spanish Ministry for Transport) (1978)

Inglada, V.: Intermodalidad y Elasticidades Precio en el Transporte Interurbano de Viajeros. Revista TTC Transportes y Comunicaciones 54, 3-14 (1992)

Johansen, S.: Statistical Analysis fo Cointegration Vectors. Journal of Economic Dynamics and Control 12, 231-254 (1988)

Johansen, S., Juselius, K.: Maximum Likelihood Estimation and Inference on Cointegration with Applications to the Demand for Money. Oxford Bulletin of Economics and Statistics 52, 169-210 (1990)

Levin, R.C.: Allocation in Surface Freight Transportation: Does Rate Regulation Matter?. Bell Journal of Economics 9(1), 18-45 (1978)

McFadden, D.: Conditional Logit Analysis of Qualitative Choice Behaviour, Frontiers in Econometrics. In: Zaremka, P. (ed.). NY & London: Academic Press 1973

McFadden, D.: The Measurement of Urban Travel Demand. Journal of Public Economics 3, 303-28 (1974)

Ministerio de Transportes (Ministry for Transport): Informes de Coyuntura de los Transportes y Comunicaciones. Madrid, 1980-1993 (1994)

Osterwald-Lenum, M.: A Note with Fractiles of the Asymptotic Distribution of the Maximum Likelihood Cointegration Rank Test Statistics: Four Cases. Oxford Bulletin of Economic and Statistics 54, 461- 472 (1992)

Oum, T.H.: A Warning on the Use Linear of Logits Models in Transport Mode Choice Studies. Bell Journal of Economics 10(1), 374-387 (1979)

Oum, T.H., Gillen, D.: The Structure of Intercity Travel Demands in Canada: Theory, Tests and Empirical Results. Working Paper No 79-18.Queen's University (1979)

Quandt, R., Baumol, W.: The Demand for Abstract Transport Modes: Theory and Measurement. Journal of Regional Science 6(2), 13-26 (1966)

Varian, H.: Microeconomic Analysis. 3rd ed. W.W. Norton & Company, Inc. (1992)

Vázquez, P.: Un Estudio Limitado sobre Elasticidades de Demanda al Precio en el Transporte Interubano. Revista del Ministerio de Transportes, Turismo y Comunicaciones 15, 21-33 (1985)

Winston, C.: Conceptual Developments in the Economics of Transportation: An Interpretative Survey. Journal of Economic Literature XXIII, 57-94 (1985)

3 Introduction of an Innovative Product: the High Speed Train (AVE)

V. Inglada-López de Sabando
University Carlos III (Spain)

P. Coto-Millán
University of Cantabria (Spain)

3.1 AVE: Characterisation

The high-speed train (AVE) is mainly characterised by a new infrastructure, which uses the adequate mobile material to be able to obtain highly operating speeds. Moreover, its way of management is generally different from that typical of the conventional train. In line with this, Plassard (1992) proposes the following defining characteristics of this mode of transport:

- A high speed of 250-300 km/h.
- A high frequency as the traffic requires it.
- A weak capacity (lower than 400 people per train).
- Slightly higher fares than those of the conventional train.

With these operation conditions, he states that the AVE must be the link between two big cities, excluding every intermediate service, since this is the only way to combine the two requirements of high speed and satisfactory rates of utilisation.

Bonnafous (1987) agrees with this assumption and points out that the French TGV resembles more a plane than a conventional train if we consider some factors such as the distance covered, the capacity and the speed. He finally concludes that its structural effects essentially affect the urban poles with the highest population, as the air transport does.

In relation with the Madrid-Seville corridor, which is the object of our research, we carry out a more detailed analysis, which enables us, to distinguish between

P. Coto-Millán (ed.), *Essays on Microeconomics and Industrial Organisation*

the following three clearly differentiated sub-products within the so-called high speed trains: the Shuttle, the Long Distance and the Talgo[1]. Chronologically, the first train used was the Long Distance. The Madrid-Seville route started to be operated on the 21st of April 1992 with intermediate stops in Ciudad Real, Puertollano and Cordoba. The demand for this sub-product had a great boom, especially the Madrid-Seville route due to the Universal Exhibition (EXPO) held at Seville in 1992. In October 1993, immediately after this event concluded, there was a fare reduction[2], which allowed this market to consolidate itself[3], thus compensating the decrease in the demand produced after the closure of EXPO. This also led to a rising tendency in the dynamics of the product to an aspect, which is always inherent to the stage of maturing of any new product.

Later, on the 18th of October 1992, a new train called "Shuttle"[4] was introduced. This train was characterised by the dramatic decrease in fares at rush hours, by the introduction of discounts and the addition of units with a greater number of tourist-class seats. The decrease in the average revenues of approximately 50% brought about an increase in demand of 275% and 115% in October and November of 1993 respectively. Therefore, this new offer, which had the initial aim of efficiently using the already existing units, was firmly consolidated.

Finally, the "Talgo" train started to operate in August 1992 from the introduction of the Cordoba gauge interchange service on the Madrid-Malaga route. This line was extended in the summer of 1993 by the introduction of the Majarabique (Seville) gauge interchange service, which offered the possibility of travelling from Madrid to Cadiz and Huelva on the new line without changing trains.

Table 3.1 presents the most significant features of each type of supply, which determine their differentiation from the High-Speed train. The primary differentiating factors between Shuttle and Long Distance are those of demand and fares. However, the Talgo differs essentially from the others for having to employ varied mobile material because it needs a track with different rail widths. This fact, along with the gauge interchange operation, leads to a lower average speed in this train and therefore, to a lower generalised cost reduction than that of the rest of the High Speed transport.

A more detailed analysis shows that the shuttle-service, which has low fares and in which discount tickets prevail, has generated a demand essentially for commuter journeys of day returns. This fact is mainly due to the reduction of the generalised cost of transport –especially in time[5] and fares -, which also leads to a reduction of the price of a supplementary commodity such as the housing.

---

1 The new High Speed line has been built with UIC rail width, which is different from the Spanish rail width. The great advantage of the Talgo, is that, due to its variable axle system, both types of rail can be used with the minimum waste of time when one infrastructure is changed into the other.

2 The average revenue fell from 19 ptas. to 14 ptas. per passenger-km.

3 This is shown by the level of the load factor which exceeds 80%.

4 In the Madrid-Ciudad Real, Ciudad Real-Puertollano and Madrid-Puertollano lines.

5 For instance, the 171-km Madrid-City takes slightly longer than 50 minutes, virtually the same time as any city center-outskirts journey takes in Madrid. Moreover, we must take into account that the average price of housing in Ciudad Real is almost half that of Madrid.

Table 3.1. Discriminating factors of the different sub-products

	SHUTTLE	LONG DISTANCE	TALGO
Routes	Madrid-C.Real C.Real-Puertollano Madrid-Puertollano	Madrid-Seville Madrid-Cordoba Other routes[6]	Madrid-Malaga Madrid-Cadiz Madrid-Huelva Others[7]
Material	Gec-Alsthom	Gec-Alsthom	Talgo 200
Infrastructure	New High Speed Track	New High Speed Track	New and Conventional Track
Fares (Average revenue/ Passenger-Km) (PTAs. 1993)	10.5	15.4	10.2
Occupancy rate	0.65	0.84	0.68
Type of demand	Local train with a high percentage or commuters journeys	Long distance	Long distance

Source: Own elaboration

3.2 Qualitative Analysis

Table 2 records the results obtained from the successive surveys to the high-speed train passengers, both in the Long Distance and in Shuttle trains. These results refer to qualitative characteristics. The following conclusions can be reached from these data.

The Long Distance Train

- Education and professional level: The Long Distance passenger has higher education and professional level: almost 63% have University education. Of these, 22% hold University diplomas while 41% hold higher University degrees. As regards professions, 40% are executives and businessmen and 26% are technicians.
- Reasons for travelling: 58% travel for professional reasons, this is an especially significant motivation among the business class passengers (75%). As regards other reasons for travelling, 22% travel for tourism and 20% for family reasons.
- Fares assessment: A survey carried out before the fares were increased in September, 1993 reveals that 64% passengers think that the AVE fares are

6 This includes the Cordoba-Seville, Ciudad Real-Cordoba, Ciudad Real-Seville, Puertollano-Cordoba, Puertollano-Seville lines.

7 This would include lines such as Barcelona-Malaga.

adequate. This fact led the company to raise the fares for the possibility, which such a measure offered of increasing returns.

- Level of and cause for satisfaction: There is an almost complete level of satisfaction (96%) and the main causes for satisfaction are: speed (29%), comfort (26%), good-quality service (11%), punctuality (8%), safety (4%) and good price (3%).

Table 3.2. Qualitative characterisation of the demand for high speed services

ATTRIBUTES	PERCENTAGES	
	Long Distance Train	Shuttle Train
Level of education		
Primary School	15	31
Secondary School	22	27
University Diplomas	22	16
Higher University Degrees	41	26
Professional occupation		
Executives	17	6
Businessmen	23	15
Technicians	26	23
Miscellaneous	14	19
Unemployed	19	37
Reasons for travelling		
Professional	58	37
Tourism	22	22
Family	40	41
Fares assessment		
Very high	5	6
High	30	25
Good	64	66
Low	1	2
Very low	0	1
Level of satisfaction		
Very satisfied	47	55
Satisfied	49	39
Indifferent	3	3
Unsatisfied	0	1
Very unsatisfied	1	2

Source: Own elaboration based on the RENFE surveys (1993)

The Shuttle Train

- Education and professional level: The passenger of this service has lower education in general, 42% of the passengers have University education. Of

these 16% hold University diplomas and 26% hold higher University degrees. As regards professions, 21% are executives, 21% are businessmen, and 23% are technicians, which constitutes lower percentages than in the Long Distance train.

- Reasons for travelling: For those passengers using this service, business (37%) has stopped being the main reason for travelling. However, it is worth remarking that 41% of passengers travel for family reasons, which is much higher percentage than that for the Long Distance train.
- Fares assessment: As in the Long Distance train, the survey carried out before fares were raised in September, 1993, shows that an elevated percentage of passengers (69%) think that the AVE fares are appropriate or low. This percentage is even higher than that obtained in the Long Distance service, due to the multiple discounts made in this service. Therefore, the rise margin of fares also seems higher for the Shuttle train.
- Satisfaction causes and level: Almost all the passengers (94%) show a high level of satisfaction by their answers. The main causes for it are the following: speed (38%), comfort (30%), punctuality (9%), safety (4%), good-quality service (3%) and good fares (2%). It is worth remarking in this service that the highest values are for speed and punctuality for the great importance given to time, in connection with the clear commuter typology, which characterises this demand.

3.3 The Concept of Generalised Cost

As far as transport is concerned, the traditional concept of a product price must be replaced by a broader concept called the generalised cost of the respective transport mode. This cost includes other components apart from the purely monetary ones. Among these components, which correspond to the various attributes which characterise the transport product, the outstanding features, according to García-Alvarez (1998), are the frequency, time, punctuality, safety, comfort, reliability, intermodality and price. The importance of each one of these components when the user chooses a particular transport depends on the type of product.

If we analyse these components in detail, we can define the frequency in railway transport as the number of trains, which travel between two points at a particular period of time. The demand elasticity with respect to the frequency is positive, a fact due not only to the existence of more alternative schedules when frequency increases but also to a lower waiting time. The value of this elasticity naturally depends on extent of frequency since the more elevated this magnitude is the lower the elasticity will be[8]. In addition, one of the most important attributes

[8] García-Alvarez et al. (1998) study the modal distribution and consider a coefficient due to the frequency of each operator measured in the number of services per day. This modal distribution coefficient increases when f increases, which reflects that the market share increases when the frequency increases. However, for high values of f, the additional growth of this coefficient is increasingly lower (for frequencies of more than 25 services per day the

of transport service is the travel time, especially in those journeys, which are not for leisure. Obviously, the time value varies depending on the individual and the reason for travelling. The total time includes not only the transport time but also the time of access and stay at departure or arrival terminals.

If we analyse fares or the monetary component within the framework of the generalised cost of transport, we observe a high demand elasticity with respect to price due to different reasons. Firstly, as in the case we are studying, we cannot consider the transport product as a first-necessity service, secondly, there are various substitutive products on the market we are analysing and finally, a large percentage of the population are car owners, a fact leading to the existence of a reference value of the marginal cost of car use, which enables a comparison of fares of the different modes of transport.

On the whole, the safety variable of a mode of transport does not significantly and directly conditions the decision of travelling or the choice of the mode of transport. In line with this, we only have to observe that the private car is the mode of transport with the lowest safety rate but it holds the highest passenger traffic rate all over the world. However, for certain individuals, this factor may be extremely important when choosing the mode of transport. Their decision is made from on a subjective point of view of the corresponding mode of transport rather than from an objective consideration. Thus, for example, in spite of the fact that the plane is much safer than the car, there are a great number of individuals who prefer any other means of transport to the air transport.

Along with the income level, comfort is increasingly important when choosing the mode of transport. This is a complex concept which includes, of course, a comfortable atmosphere, an adequate accommodation (dimensions and comfort of the seat) and the possibility of the minimum interruption in the passenger's day to day life during the journey. In connection with this, the availability of toilets, coffee and restaurant services, television or video significantly contributes to increase comfort.

The data appearing on table 3.3 refers to the reasons for the choice of the AVE in comparison with other Long Distance modes competing with it. We can reach several conclusions from these data about the importance of the various components of the AVE generalised cost. In short, we can state that the main causes for the choice of the AVE are, by order of importance, speed, comfort, price, novelty and safety.

A first result, certainly relevant, is the fact that the effect of the comfort component is virtually as important (29%) as that of time (30%) and it is even higher than the effect of price (11%). This result is especially significant within the group of air passengers, where the component of comfort is far more important (31%) than that of price (19%), a value which is even equal to the sum of the two "traditional" generalised cost components: price and time.

coefficient increases in a maximum 5%). For f=1 frequency, the coefficient value obtained is 1.187 while for f= infinite frequency, which may correspond for example to the car, the coefficient value is 4.55.

Table 3.3. Reasons for the choice of the AVE in comparison with other alternative modes of transport (% for the Long Distance Train)

	Plane	Car	Train	Bus	Total
Speed/Time	13	42	57	67	30
Punctuality	4	0	2	0	3
Comfort	31	35	19	13	29
Price	19	6	2	2	11
Novelty	11	3	9	3	9
Safety/Fear	6	10	0	0	5
City Centre	4	0	0	0	2
Schedules	6	0	8	10	5
Miscellaneous	6	4	3	5	6

Source: Own elaboration based on the RENFE surveys (1993)

In a more detailed disaggregated analysis of each mode of transport, we can state that, the main reasons for the choice of the AVE with respect to the plane are comfort (31%), price (19%), speed (13%), novelty (11%) and safety (6%). With respect to the car, time is the most important reason (42%), followed by comfort (35%) and safety (10%). In relation to the conventional train, the reasons to choose the AVE are speed (57%), comfort (19%) and novelty (9%). Finally, with respect to the bus, speed (67%) and comfort (13%) are the essential reasons for the choice of the new product.

3.4 Comparison Among Different Competing Products

When we compare the different components of the AVE generalised cost with those of competing models, we can observe the great advantages of the former. In line with this, we can emphasise the fact that the AVE travel time is much lower than that of the remaining modes of transport except for the air transport. This time reduction is due not only to its high speed but also to the less time employed in accessing to the service, as well as to a decrease in the distances travelled thanks to the new infrastructure.

Table 3.4 presents the time the main routes take. We must point out that in the routes in which the new infrastructure is completely used[9], the time employed by the AVE is less than half that used by the remaining modes of transport, except for the air transport. For example, in the Madrid-Seville route, the reduction of the travel time in the high-speed train is 2h 45' with respect to the car, 3h 55' with respect to the bus and 3h 20' with respect to the conventional train.

[9] In other lines such as Madrid-Cadiz, Madrid-Huelva and Madrid-Malaga the Talgo is employed with High-Speed and conventional track.

Table 3.4. Travel time for route

ROUTES	TRAVEL TIME				
	Air Transport	Car[10]	Bus	Conventional Train	AVE
Madrid-Seville	50'	5h 20'	6h 30'	5h 55'	2h 35'
Madrid-Malaga	55'	5h 40'	7h 15'	6h 50'	4h 45'
Madrid-Huelva	-	6h 20'	7h 50'	7h 40'	4h 40'
Madrid-Cadiz	55'	6h 30'	8h 10'	7h 45'	5h
Madrid-C. Real	-	2h 5'	2h 30'	1h 55'	55'
Madrid-Puertollano	-	2h 35'	3h	2h 45'	1h 15'
Madrid-Cordoba	-	3h 55'	5h	4h 25'	1h 45'

In addition, observing table 3.5, which shows the access time to the services of different modes of transport we see that the time used in the AVE is significantly lower than in that used in the conventional train and plane but it is equal to that used in the bus. The time of access used in the car, obviously null, is the only one, which is lower than that in the AVE. This is precisely one of the main advantages of this mode of transport since the user considers the access time more "lost time" than the travel time. Moreover, we must take into account that this concept includes delay penalties, safety, boarding time and access time to and from services. The advantageous aspects of the high-speed train vary depending on the mode of transport with which it is compared. For example, with respect to air travel, the new product has great advantages in relation to access - its stations are located in the city centre – and to delay since the AVE has a high punctuality rate. The advantage of this high punctuality rate is also apparent in the comparison of the AVE with conventional trains.

The use of a new infrastructure also adds a comparative advantage to the new product, which is the fact that the different distances travelled are quite lower as observed in table 3.6.

Finally, when comparing the prices of the different products appearing in table 3.7, we must remark that, on the whole, the AVE is – after air travel - the most expensive mode of transport. However, in some lines such as the short-distance ones, included in the segment of the Shuttle Train, differences with respect to the car and the conventional train are minimal. In every case the bus is the mode of transport with the lowest price.

[10] In order to obtain the V speed, the $V = 48 + 72\ (1\text{-}i/c)^{½}$ equation has been used for highways, while in the conventional road the $V = 90 - 22\ (i/c)$ alternative expression has been used, where i and c are the intensity and hour capacity respectively. Both expressions come from the "Highway Capacity Manual" from the Transportation Research Board (1985).

Table 3.5. Access time to service[11]

Plane	Car	Bus	Conventional Train	AVE
1h 55'	-	1h	1h 15'	1h

Table 3.6. Distances

	Road	AVE	Conventional Train
Madrid-C. Real	190	171	255
Madrid-Puertollano	228	210	293
Madrid-Cordoba	400	343	442
Madrid-Seville	538	471	565
Madrid-Malaga	544	528	627
Madrid-Cadiz	663	633	727
Madrid-Huelva	632	581	675

Table 3.7. Prices of different modes of transport (PTS. per passenger, 1993)

ROUTES	FARES						
	Plane[12]	Car[13]	Bus	Conventional Train		AVE	
				1st cl.	2nd cl.	Pref.	Tourist
Madrid-Seville	12650	7346	2210	7880	5450	10800	7900
Madrid-Malaga	13700	7655	2680	8750	6050	9200	6600
Madrid-Huelva	-	8630	2630	9450	6520	9700	6900
Madrid-Cadiz	13950[14]	9181	2600	10100	7250	10200	7300
Madrid-C. Real	-	2539	1440	3690	1425	2500	2000
Madrid-Puertollano	-	3080	1780	4280	1635	2900	2300
Madrid-Cordoba	-	5424	1650	6130	4480	7900	5800

11 For the Madrid-Seville route. For other routes with departure or arrival in Madrid or Seville, the access time is reduced by 10 minutes for the bus, the conventional train and AVE. Moreover, for the remaining lines, the access time is reduced by 20 minutes. In planes, this concept includes 1h 15' for access, 30' for safety and boarding and 10' for delay penalty. In the bus and the AVE, this access times includes 15' for safety and boarding and 45' for the transport to and from the station. Finally, a 15' delay penalty has been considered for the conventional train respect to the AVE.

12 Price for tourist class without weekend discounts.

13 The occupation for vehicle is of 1.8 individuals and half the depretiation is due to the number of Kilometers travelled.

14 Madrid-Jerez route.

3.5 Induction and Substitution Effects

The introduction of the AVE produces two clearly differentiated effects on transport demand viz. inducing and substitution effects. These effects correspond to those journeys which would have never been made if this new service had not existed and to those which would have been made but using a different mode of transport.

3.5.1 Induction Effect

Within the inducing or generation effect of new journeys, we must include not only those passengers who had never travelled on that route but also the component, which accounts for the increase in the travel frequency of those passengers who already used that route before the existence of the AVE. In table 3.8, we observe that the average journeys per year on the Madrid-Seville route have increased from 11 to 15 due to the introduction of the new product and that 28% of the users had never travelled on that route before. The inducting effect still continued, but at an increasingly lower rate, until 1996, which marked the end of the final period of maturing of the service. The results obtained show the importance of this effect, which, according to the studies made, once the period of maturing was completed, accounts for almost 45% of the AVE passengers[15].

Table 3.8. Travel frequency in the Madrid-Seville route

	Before the AVE	After the AVE
Twice of more a week	3%	6%
Once a week	6%	8%
Once a fortnight	7%	9%
Once a month	16%	16%
Once a quarter	18%	16%
Less	30%	17%
Did not travel	21%	-
Travelled on this route for the first time	-	28%
Average journeys per year	11.1	15.2

Source: Own elaboration based on the RENFE surveys (1993)

[15] A result similar to that indicated by Nash (1991) for the French Paris-Lyon line.

3.5.2 Substitution Effect

The reduction of the generalised cost in the new mode of transport in comparison with the substitutive modes[16] generates the change of part of the demand for these competitive modes of transport to the High Speed Train. Obviously, the intensity of this "substitution effect" is lower in the Talgo train because the reduction of the generalised cost is also lower.

The comparative analysis made in point 4 about the different components of transport generalised cost indicates that the high-speed train offers certain advantages in comparison with the remaining modes of transport. In line with this, as inferred from the reasons which lead to the choice of the AVE detailed in table 3, this new product is preferred to the air transport due to its lower price (19%) and higher comfort (31%) and safety (6%). Even a large number of passengers (13%) prefer this mode of transport for its reduced access time and waiting time, since these two aspects are considered a great time loss.

The large number of advantages presented by the AVE with respect to the conventional train, especially as regards time and comfort, causes such a high magnitude of the substitution effect that the demand for this mode of transport is virtually absorbed by the AVE. With respect to the car, the main causes for the choice of the AVE are the time (42%), comfort (35%) and safety (10%). Finally, in comparison with the bus, the main causes for substitution are time (67/%) and comfort (13%).

3.5.3 The Demand for the New Product

The demand for the AVE comes from the combination of the induction and substitution effects the above section. When we analyse the time evolution of these two effects, a period of maturing is observed during which the level of demand for that product adjusts and grows exponentially until it stabilises. This period of maturing is common to all new products. As a reference, we must remark that when the High Speed was introduced in the Paris-Lyon corridor, its period of maturing lasted 6 years. However, that period was lower – 4 years - in the Madrid-Seville case for various reasons[17]. A first reason for this time reduction was the demand shock generated by the EXPO, especially in the Madrid-Seville line. A second reason was the significant decrease in fares occurred in October 1992 at the end of the EXPO. This fall of prices, especially marked for the new shuttle-service, accelerated the rhythm of demand absorption by the new mode of transport. Consequently, the period of maturing was reduced[18].

---

16 This reduction is mainly due to the decrease in fares and the increase in comfort with respect to the plane as well as by the reduction of time and the increase in comfort with respect to the bus, the conventional train and the car.

17 This period lasted 6 years in the TALGO train.

18 However, the TALGO fares were kept almost fixed so we cannot consider them as a reduction of the period of maturing.

3.6 Impact on Demand

Due to the high magnitude of the substitution effect, the introduction of the high-speed train causes very strong effects on the demand for the remaining modes of transport competing with it in the Madrid-Seville route. A more detailed analysis enables us to observe the highly strong effect particularly on the conventional train since this lost almost 78% of its traffic prior to the introduction of the AVE, which virtually produces the disappearance of this transport for the route. The introduction of the High-Speed train has led to an approximately 50% fall of the air transport demand in the Madrid-Seville route. Therefore, we can remark that the operating company has acted efficiently when it adjusted its supply to the demand fall by reducing the number of flights[19] and trying to keep the occupation percentage almost stable[20]. As regards the car, the losses are lower than in the previously mentioned modes of transport, approximately 30% in the Madrid-Seville route. Finally, in bus transport it seems that the impact on long-distance routes is not great (11% of demand fall) since both products are hardly substitutive. However, in the Madrid-Ciudad Real route, the demand fall, which reached 34%, is significant due to the low fares of the high-speed train.

Therefore, we can conclude that the introduction of the high-speed train causes a dramatic change in the modal distribution of the demand, and we can speak of a pre AVE transport market and a post AVE market. This effect is especially important in the long-distance segment, where the railway is the predominant mode of transport in the Madrid-Seville route, and where its use even exceeds the use of the private car, a most unusual situation in the Spanish transport market.

References

Bonnafous, A.: The Regional Impact of the TGV. Transportation 14, 127-137 (1987)

Inglada, V.: Análisis Empírico del Impacto del AVE Sobre la Demanda de Transporte en el Corredor Madrid-Sevilla. Revista de Estudios de Transportes y Comunicaciones 62, January-March, 35-51 (1994)

García-Alvarez, A.; Cillero-Hernández, A., Rodríguez-Jericó, P.: Operación de Trenes de Viajeros. Fundación de los Ferrocarriles Españoles 1998

Nash, C.A.: The Case for High-Speed Rail. Investigaciones Económicas 15(2), 337-354 (1991)

Plassard, F.: El Impacto Espacial de los Trenes de Alta Velocidad en Europa. Transporte y Medio Ambiente. Ministry for Transport and Public Works (MOPT) 1992

RENFE : Surveys Carried Out to the AVE Passengers. 1993

Transportation Research Board : Highway Capacity Manual. Special Report 209. 1985

[19] Other adjustment measures such as the reduction of fares have not been considered in detail.

[20] In fact, this percentage has bee reduced less than 10% due to the introduction of the High-Speed train.

4 An Approach to the Hub-and-Spoke Systems from SVARs Models. A Practical Application to Container Traffic between the Port of Bahía de Algeciras and Other Ports of the Spanish Port System (Bahía de Cádiz and Las Palmas)

J. I. Castillo-Manzano
University of Sevilla (Spain)

P. Coto-Millán
University of Cantabria (Spain)

L. López-Valpuesta
University of Sevilla (Spain)

4.1 Introduction and Reasons for Analysis

The hub-and-spoke systems are frequently connected with maritime, air and road transport. The international container traffic in maritime transport is developed under this model. For a port to achieve a *"Hub Status"* in maritime container transport, its traffic should exceed 1,000,000 tons per year and be provided with the necessary equipment to use simultaneously three *post-panamax* cranes for a new-generation vessel. Huge container vessels dock only at Hub ports and international shipping companies transfer their containers from these ports to their *round the world* and *feeder*[1] lines.

[1] It is known that *feeder* vessels are used for the transport of containers from or to minor ports which are not served by the huge vessels pertaining to the *round the world* lines, and make

P. Coto-Millán (ed.), *Essays on Microeconomics and Industrial Organisation*

National air networks are also established as hub-and-spoke systems in connection with international transport. Small airports have barely any connections with foreign airports or even with other minor national airports, therefore, indirect flights through hub airports network are necessary in order to reach those airports. We can also find this system either in road transport (with the Logistics Area acting as hub) or in railways, whose stations facilitate the transfer between lines.

Given that we are dealing with an organisation system, in the hub-and-spoke system - which is constantly present in Transport Economy - it is necessary to find methodological formulae which enable us to tackle the study of this system in a common manner, notwithstanding the means of transport implied. In this paper, we deal with one of those possible common approaches from the econometric analysis of the time series by means of the SVARs (Structural Vector Autoregressions) models with Long-run Restrictions. Other alternatives can be found in the BVAR (Bayesian Vector Autoregressions) or in the VEC (Vector Error Correction Models) with restrictions. All these approaches allow us to introduce the information available[2] a priori on the hub-and-spoke system and make it compatible with the statistical information obtained from the evolution of traffics in the past.

4.2 VAR Models: Methodological Approach

During the last 20 years, we have witnessed important changes in the econometric modelling of time series, mainly due to the weaknesses of the economic theory when dealing with complex models of simultaneous equations. On rare occasions does the theory offer us a dynamic specification which clearly explains the relationships between variables, even less frequently in fields of more recent development, such as transport economics. This fact has led to the application of non-structural models which explain the relationships between the different economic variables. Perhaps the models most frequently used have been the VAR. These models are employed both to predict and analyse the impact over time of some variables in comparison with others, and even to demonstrate hypotheses of the variables' behaviour, although, as regards the latter use, their evidential valence is, without doubt, lower than that of the traditional simultaneous equations systems. VAR models also avoid having to formulate structural relationships between variables since they state that each endogenous variable of the model is a function of their lag values together with a set of exogenous variables. Below, we show an autoregressive representation of a VAR model:

---

only intermediate stops at the large transoceanic stations. Since *feeder* vessels operate mainly in small ports with less gauge, these usually have a lower capacity, between 500 and 700 tons, although some may reach 1,500 tons. In contrast with the huge container vessels and, in view of the limited provision of mobile facilities in the small ports where they operate, the vessels are indeed provided with their own loading and unloading means; they usually have two cranes located in the middle-line or on the port side of the vessel.

2 It is easy to have information available a priori on a hub-and-spoke system, even though it is only the structure of the lines and their commercial importance measured taking into account the volume of traffic.

$y_t = \Sigma_i A_i y_{t-i} + \Sigma_j B_j x_{t-j} + e_t = A_1 y_{t-1} + ... + A_p y_{t-p} + B_0 x_t + \ldots + B_q x_{t-q} + e_t$. $\forall i = 1,\ldots,p$ $\forall j = 0,\ldots,q$.

Being:

y_t = endogenous variables vector of k x 1 range.

x_t = exogenous variables vector of d x 1 range.

A_i = matrix of the endogenous variables ratios with i-lags and whose range is k x k.

Bj = matrix of the exogenous variables ratios with j-lags and whose range is d x d.

e_t = k x 1 dimension innovations vector. These innovations have zero mean and constant variance, and they are generally correlated between one another if they are contemporary, but incorrelated if they correspond to different time periods.

4.2.1 Impulse Response

As mentioned above, one of the main uses of the VAR is the analysis of the dynamic transmission mechanisms of the effects of changes undergone by the variables. A shock or change in the i variable will affect all the variables of the model through a structure of lags of the VAR model. A priori, the interpretation of the Impulse Response functions is not so simple since the innovations observed in the model (e_t) will generally present contemporary correlations between one another. In line with this, it can be stated that the innovations have a common component which cannot be associated to a specific variable.

In order to make the interpretation of the Impulse Response functions easier, the innovations must be incorrelated either over time or during the same time period. The incorrelation is obtained by applying a transformation to the innovations in the following manner:

$Bu_t = Ae_t$ being $u_t = B^{-1}Ae_t \sim (0,\Sigma u_t)$

Being Σu_t a diagonal covariance matrix[3]. The Impulse Response functions are generally estimated to show the response in the model to a shock of a standard deviation in the structural innovations. Due to this, it is necessary that a SVAR model be standardised by adjusting the variances of the structural innovations to 1, since, in this way, a shock of a standard deviation corresponds to a unit innovation in u_t. Naturally, the standardisation only affects the scale of the system formulated and does not alter any relationship or substantial feature of this system.

In order to orthogonalise the impulses, we have used in general the Cholesky factor of the residual covariance matrix. This system has important shortcomings since the results obtained depend on the ordering of the variables so that, the responses to the impulses will be affected according to this order. The Cholesky decomposition can only be employed when the order of the variables proposed has it foundations in strong theoretical assumptions. This problem has been traditionally overcome by ordering the variables from high to low exogeneity, for

[3] Technically speaking $P = B^{-1}$ A matrix is standardised in such a way that $P \Sigma e_t P' = \Sigma u_t = I$, I being the identity matrix.

which we should apply Granger's causality tests. Another solution is to employ the SVAR models, as set out in the section below.

4.3 Formulating Long-Run Restrictions under a SVAR Model Applied to a Hub-and-Spoke System

Prior to explaining how to formulate restrictions under the SVAR model, we should prove its utility. SVAR models have become an increasingly popular tool during the past decade. Among their main uses is the analysis of the economic cycle's sources, and they are also employed in the study of the transmission mechanisms of monetary policies. Thanks to short- and long-run restrictions, the error term can be orthogonalised in a non-recursive way for the analysis of the Impulse Response functions. In other words, it is a an alternative to Cholesky recursive orthogonalisation for which it is necessary to have sufficient restrictions to identify the orthogonal components of the error term.

The literature of the SVAR models makes a difference between short- and long-run restrictions. In this paper we focus on long-run restrictions mainly because we deal with annual series. The way in which short-run restrictions[4] are formulated usually implies that one of the variables cannot instantaneously respond to the changes or shocks produced in other variables. These changes or shocks are unpredictable so, in the analysis of the hub-and-spoke models, we would be stating that the managers of one of the transport infrastructures of the system - let us call it X -, need some time to evaluate the changes in the traffic statistics of the infrastructure which has undergone the shock, called Y. In line with this, the short-run restriction would mean that, during the following period, the response in the traffic of X infrastructure to a shock in the Y infrastructure of the system would be 0.

Given that transport infrastructures (ports, airports, railway stations...) operate with the newest information systems, these hypotheses can only be kept when we work with monthly, term or quarterly series rather than with annual series. This fact reduces the scope of application of short-run restrictions to port and railway traffic; although it is possible to find monthly statistics in port traffic, these are usually mere approximations. In spite of having traffic statistics available every six months – or even more regularly - it is difficult to keep short-run restrictions useful for the analysis of the hub-and-spoke transport systems. We cannot overlook the fact that a positive (or negative) shock in a spoke infrastructure immediately implies an increase (or decrease) in the hub infrastructure since the hub infrastructure is the natural entry and departure point of traffic in the spoke infrastructure.

On the contrary, there is a priori a wide scope for the employment of SVAR models with long-run restrictions. These restrictions can be stated under quite realistic assumptions, as can be seen below. Next, we account for the analytic formulation of the long-run restrictions included in a SVAR model[5], the

[4] For the analysis of short-run restrictions, see Amisano and Giannini (1997).
[5] For a deeper analysis see Blanchard and Quah (1989).

components of the C matrix being the long-run accumulated response to the shocks in the structural innovations[6]:

$$C = \hat{\Psi}_{\infty} A^{-1}B$$

$$\hat{\Psi}_{\infty} = (\mathbf{I} - \hat{\mathbf{A}}_1 - \ldots - \hat{\mathbf{A}}_p)^{-1}$$

being the estimation of the accumulated response to the shocks in the innovations observed.

The Impulse Response functions for a hub-and-spoke system will allow us to check how a positive or negative shock in one of the transport infrastructures will affect the remaining ones, mainly in the hub system. Within the frame of this general formulation we can distinguish between two main aims:

1. It is especially interesting to study how a variation in the traffic of an infrastructure which plays a "spoke" role will affect the "hub" infrastructure of the system. With this study, we will be able to contribute to a better distribution of the "hub" infrastructure's capacity thus avoiding the appearance of bottlenecks or reducing the appearance of idle time periods.
2. In our SVAR model, it may happen that the transport infrastructures considered share their hinterland or scope of influence, that is to say, the area where the traffic is generated because the industrial companies which depend on the existence of the infrastructures studied are located there. Therefore, we will have to take into account that a shock in an infrastructure will be composed by two effects with different signs, one positive, which implies the creation of commerce, and another one negative which means a trade deviation. If the hinterland is shared, the positive effect which the favourable shock of one of the spoke infrastructures would have on the hub, is reduced by the traffic loss produced by trade deviation. We can even find that a positive shock of a spoke infrastructure may have a negative accumulated response on its hub infrastructure. This will happen whenever we are dealing with a multiple hubs system. In the event of this, the market share of the spoke infrastructure which undergoes the shock, held by the hub infrastructure analysed, shall be quite small. In other words:

   $$Chs = k(CMhs*Css - DC) \qquad (4.1)$$

   Being:
   Chs = Accumulated response on the hub infrastructure to a shock in the spoke infrastructure.
   CMhs = Market share of the hub infrastructure over the set of traffics of the spoke infrastructure. This share is calculated as the traffic rate of the spoke infrastructure which has its origin or destiny at the hub infrastructure studied. Naturally, this rate shall depend on multiple factors, among which we can stand out the following: the number of hubs with which the spoke infrastructure

[6] Therefore, C12 = O is interpreted in such a way that the long-run effect on the variable 1 of a shock produced in variable 2 is nil. In other words, variable 2 will only have transitory effects on variable 1.

operates - and the geographic proximity between these and the spoke infrastructure -, or their relative importance within the transport category implied.

Css = Accumulated response on the spoke infrastructure to a shock produced in it.

DChs = This item includes that part of Css traffics which does not imply trade creation - in other words, new traffics - but they are the result of a trade deviation from the hub infrastructure towards the spoke infrastructure. As mentioned above, this effect can only be stated under the initial assumption that the hub and spoke infrastructures share a geographic area known as "shared hinterland".

k = Is a positive proportionality constant.

In order to implement a SVAR model and be able to measure the objectives exposed above (eq. (4.1)), it is necessary to establish a criterion under which long-run restrictions are to be established, based upon the economic behaviour of the hub-and-spoke systems. A priori, it is logical to assume that in a bivariate SVAR model Csh = 0, in other words, that the long-run effect of a shock in the hub variable on the spoke infrastructure is nil. Here we are assuming that, a change in the traffics of the hub infrastructure (a shock in the hub traffic) is due either to the increase in the demand of the hub's hinterland[7] or to a former shock in another spoke infrastructure in connection with it. A different view of this would be provided by deriving the following analytic expression from formula 1:

$$Csh = c(CMsh*Chh - DCsh) = 0 \quad (4.2)$$

This restriction is based upon the following initial assumptions:

CMsh = Market share of the spoke infrastructure over the set of traffics of the hub infrastructure. Analogously to CMhs, this share is calculated as the traffic rate of the hub infrastructure which has its origin or destiny at the spoke infrastructure studied. Also in this case, this rate will depend on a multiplicity of factors such as the number of spokes with which the hub infrastructure operates, the geographic proximity or the cost structure of both infrastructures, among others. Presumably, the more important the hub infrastructure[8] is, the lower this coefficient will be and therefore, CMsh*Chh will tend to 0 or be slightly over zero.

Chh = Accumulated response on the hub infrastructure to a shock in it.

DCsh = Trade deviation from the spoke infrastructure towards the hub infrastructure. Again, this effect shall be estimated if the hub and the spoke infrastructures share the hinterland. Obviously, this effect shall be less important than the DChs since the majority of the variations in the traffics of a hub infrastructure shall be due to changes in the demands of the hub or spokes infrastructures' hinterlands in relation with the hub infrastructure. In general, this

[7] We cannot miss the fact that the demand for transport services is a derived demand, as all demands for productive factors. Therefore, an increase (decrease) in the traffics of an infrastructure is a consequence of a former demand for these traffics generated by an increase (decrease) in the economic activity of a hinterland.

[8] This importance is measured either with respect to the hub's traffic volume or to the number of different connections established by this hub with the spoke infrastructures and with other hubs.

effect shall be nil, even when a small hinterland area is shared, or, at best, will be slightly negative, being compensated with the CMsh*Chh value. Under these assumptions, the CMsh*Chh – DC difference will tend to 0 and, therefore, it can be maintained, as a long-run restriction, that Csh = 0.

c = Is a positive proportionality constant.

Obviously, in order to support this reasoning we must previously assume that the hub infrastructure is not congested, in other words, that the spoke infrastructure can employ the hub for its transfers as much as necessary. Moreover, in order to analytically sustain this assumption, it is highly recommended to check that the Csh = 0 value is within the Csh confidence interval by applying any recursive orthogonalisation method[9] and elaborating the confidence intervals from the estimation of the accumulated responses' standard errors[10].

Under the above reasoning, in the event that the hub infrastructure were congested, it would not be of much sense to analyse the Impulse Response functions, and all conclusions would be altered. This analysis would only be useful as an attempt to simulate what would happen if the hub infrastructure extended its facilities, a fact that would explain the positive shock in its traffics. Even in this case, if there were congestion problems in the hub infrastructure, the above-mentioned restrictions could not be used, in other words, we could not state that Csh = 0.

Under this simple theoretical outline, assuming that no congestion problems exist in the hub infrastructure and applying a methodology of tested solvency to the dynamic analysis of time series (the SVAR models), a very interesting set of possibilities appears in all fields of transport economy. In order to illustrate the usefulness of this methodology, a practical application is set forth in the following section.

4.4 Practical Application to Container Traffic between the Bahía de Algeciras Port and other Ports of the Spanish Port System[11].

As mentioned at the beginning of this work, this methodology could be applied to any transport system developed under a hub-and-spoke structure. We should not miss the fact that many of the properties of the VAR models - and of the tests

---

9 For this work we have employed Cholesky factorization by ordering the variables from high to low exogeneity after applying the Granger's test and adjusting the number of degrees of freedom for the residual covariance matrix. In line with this, the (i,j) element of the residual

10 covariance matrix would be $\Sigma_t e_{i,t} e_{j,t}/(t-p)$, p being the number of parameters per equation. This does not imply any difficulty since standard econometric packages account for the possibility of estimating these standard errors either analytically (asymptotically) or by means of the Montecarlo method. For the applications in section 4 we have chosen the analytical

11 estimation.
All the estimations of this section have been carried out through the Eviews 4.1 statistical package.

applied to them - are for infinite samples[12]. This makes it necessary to seek infrastructures with broad time series to which this methodology can be applied, a fact which will in general restrict its scope of action to port and airport systems. For this practical application, we have chosen the container traffic between the Bahía de Algeciras port and other ports of the Spanish port system.

Within the Spanish port system, the most important hub system in the international container traffic is the Bahía de Algeciras port[13], together with the Valencia and Barcelona ports. Although the difference in the volume of containers moved in these three ports is not so big as to justify the selection of the Bahía de Algeciras port as the unquestioned hub[14] of the Spanish port system, this difference has to be measured due to the extension of their hinterlands. The Bahía de Algeciras port has the minor hinterland, which leads to the fact that, historically, an average of 90% of its container traffic does not depart from its hinterland. This 90% is dedicated to the international or national transfer by means of its round the world and feeder lines, and, therefore, this traffic is only due to its hub function within the Spanish port system. For all these reasons, we have chosen it as a reference of our practical application.

Since time series are not very long (1975-2001) and, in addition, they are not stationary, we have decided to work with bivariate VAR models in order to avoid losing all degrees of freedom. This would happen if we worked with a wide number of variables. Our objective is to demonstrate the suitability of the two scenarios depicted above, in other words, we will study the interdependency of these infrastructures in the system whether there be a shared hinterland or not. We have chosen a group of applications in which, a priori, we could certainly know which are the predicted results in such a way that any deviation from these results produced by using SVAR models would be difficult to justify and therefore, the utility of these models within the field of the hub-and-spoke systems economic analysis.

4.4.1 Scenario 1: Existence of a Shared Hinterland between the Infrastructures. Practical Applications: Bivariate SVAR between the Ports Bahía de Algeciras and Cádiz

In order to illustrate the first scenario, we have established a bivariate SVAR between the Bahía de Algeciras and Bahía de Cádiz ports. The distance between these two ports is less than 100 km; it takes one hour to a truck to travel by road between these two ports. In line with this approach, the Bahía de Cádiz port would be a spoke port with respect to the Bahía de Algeciras port, which would be a hub

[12] It is widely known that, in time series, the ordinary least squares (OLS) results are the same as in linear economy but for infinite series. Therefore, the estimations by OLS of the parameters of a VAR converge in their distribution towards the real values of the parameters, provided that the model's residues are white noise. In order to prove this assumption, it is necessary to apply a multivariate normality test to these values.

[13] The Bahía de Algeciras port, in its function of hub port in container traffic, has feeder lines within the Spanish port system with the Melilla, Alicante, Valencia, Barcelona, Las Palmas de Gran Canaria, Santa Cruz de Tenerife, Ceuta and Tarragona ports.

[14] In 2001, the Bahía de Algeciras port moved 1,414,390 tons in comparison with the 1,144,775 tons of Valencia and the 947,767 of Barcelona.

port. It is certain that the Bahía de Cádiz port is not a spoke port of Bahía de Algeciras since it does not have feeder lines with it.

This fact does not invalidate the application of the above-mentioned methodology, indeed, it has clearly favourable effects. On the one hand, it reinforces the theoretical justification of the long-run restriction (Csh = 0) for various reasons. Firstly, since there is not a feeder line between the Bahía de Algeciras and Bahía de Cádiz ports, the shock of the hub infrastructure would not have any favourable effect on the spoke port, in other words, CMsh*Chh = 0, since CMsh = 0 by definition. Moreover, we must assume that the effects of trade deviation derived from sharing a common hinterland (DCsh) will be very low as regards the accumulated response of the spoke infrastructure to the hub shock. This can be easily supported if we take into account the low magnitude of traffic in the spoke ports with respect to the hub ports; for example, the Bahía de Cádiz port would account for less than 3.5% of container traffic in the Bahía de Algeciras port, measured in tons.

Another possibility of analysis for this particular and exceptional case - where no hub-spoke relationship exists between the infrastructures covered in the analysis - is to consider that DCsh is different from zero. Therefore, we would be assuming that a shock in the hub infrastructure has a slightly negative effect on the spoke infrastructure.

Obviously, it is much more difficult to establish a priori by deduction a long-run restriction different from zero, either positive or negative. In this work, we have used the same analytical method to choose the value a priori of this new negative restriction as the one used to test that Csh = 0. To start with, we have established the confidence interval for Csh by - with a view to orthogonalising the errors observed - using the inverse of Cholesky factor, then adjusting the number of degrees of freedom and ordering the variables from high to low exogeneity under Granger's causality model[15]. The highest exogeneity of the variable as we expose it, coincides with the concept of causal priority as Granger understands it. Once the interval has been established, a slightly negative value within this interval must be chosen.

The first step to establish a bivariate SVAR model between the container traffics, expressed in tons, of the Bahía de Algeciras and Bahía de Cádiz ports, is to survey the series which are to be dealt with. Graphically, we observe a positive relationship between the mean and the typical deviation of the variables expressed in levels, especially for the case of the Bahía de Algeciras port (see figure 4.1). This evident hint of heterocedasticity suggests the employment of the logarithmic transformation in both series. Moreover, without having to resort to a unit-root test, we can conclude the nonstationariety in the mean of the level series, as seen in figure 4.2. In order to avoid these shortcomings, we will establish the SVAR model with the following ∇logALG and ∇logCAD series, in other words, we will explain the evolution of the growth rate of each one of the level series depending on the last growth rates. Table 4.1 shows the results of the unit-root tests for these

[15] We consider a variable y_{2t} to cause, in the Granger's meaning, another variable y_{1t} if the future projection of y_{1t}, with all the information available ($Y_{1t} = y_{1t} + y_{2t}$), and the projection with only the information of variable y_{1t}, are different.

series, in which the nil hypothesis is rejected and therefore, the original series (logALG and logCAD) are just I(1).

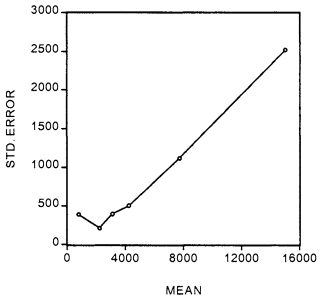


Fig. 4.1. Graph of the mean-typical deviation of the Bahía de Algeciras port series

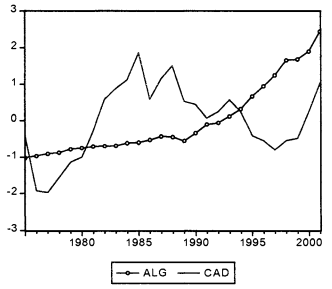


Fig. 4.2. Bahía de Cádiz and Bahía de Algeciras port series (standardised data)

Table 4.1. Unit-root test for ∇logALG and ∇logCAD

Test ∇logALG	t-stat	MacKinnon p-values
Augmented Dickey-Fuller based SIC	-4.8840	0.0007
Phillips-Perron (Newey-West using Bartlett kernel)	-9.0163	0.0000
Test ∇logCAD	t-stat	MacKinnon p-values
Augmented Dickey-Fuller based SIC	-5.6475	0.0001
Phillips-Perron (Newey-West using Bartlett kernel)	-5.5837	0.0001

Once the likelihood ratio tests have been applied to determine the optimal lag of the VAR[16], we obtain the following structure:

$$\nabla \text{logCAD} = 0.2176*\nabla \text{logCAD}(-1) + 0.1193*\nabla \text{logALG}(-1) + 0.0162.$$

$$\nabla \text{logALG} = -0.3264*\nabla \text{logCAD}(-1) + 0.1522*\nabla \text{logALG}(-1) + 0.1221$$

Table 4.2. Optimal lag in the VAR (∇logALG y ∇logCAD)

Lag	LogL	LR	FPE	AIC	SC	HQ
0	27.71421	NA	0.000438	-2.057137	-1.959627*	-2.030092
1	33.32200	9.869710*	0.000386*	-2.185760*	-1.893230	-2.104624*

* indicates lag order selected by the criterion
LR: sequential modified LR test statistic (each test at 5% level)
FPE: Final prediction error
AIC: Akaike information criterion
SC: Schwarz information criterion
HQ: Hannan-Quinn information criterion

One data necessary to specify the SVAR model is the confidence intervals of the long-run restriction which we wish to establish (Csh = 0). In this case, the ranges for an accumulated response of 10 periods are (0.1311,-0.0464), estimated from the Cholesky factor by adjusting the degrees of freedom and ordering them from high to low exogeneity. Then, it would be possible that Csh = 0 or slightly negative taking into account that only trade deviation effects exist. For example, an option would be to state that Csh = -0,010. Table 4.3 accounts for the results of the VAR model eventually selected.

Below appear the graphs (figures 4.3 and 4.4) of the Impulse Response functions under two choices Csh = 0 and Csh = -0,010. Csh would be Ccad/alg in this model. For the results to be coherent with the model exposed, the graphs must show that Ccad/cad and Calg/alg will be certainly positive while Calg/cad must be negative. The following graphs present the estimations of the accumulated functions Ccad/cad and Calg/alg under the two hypotheses considered (Csh = 0 and Csh = -0,010), which, as can be observed, satisfy the assumptions exposed for them.

After analysing the model's residues we have applied the Jarque-Bera standardisation test for multivariate series. With the orthogonalisation restrictions proposed Csh = 0 and Csh = -0,010, we have obtained p-values of 0,5028 and 0,4905 respectively. These values are more than sufficient to keep the residues' white noise hypothesis.

[16] Table 4.2 shows the results of the tests in order to determine the optimal lag in the model studied. The optimal lag for each election criterion is marked with an asterisk.

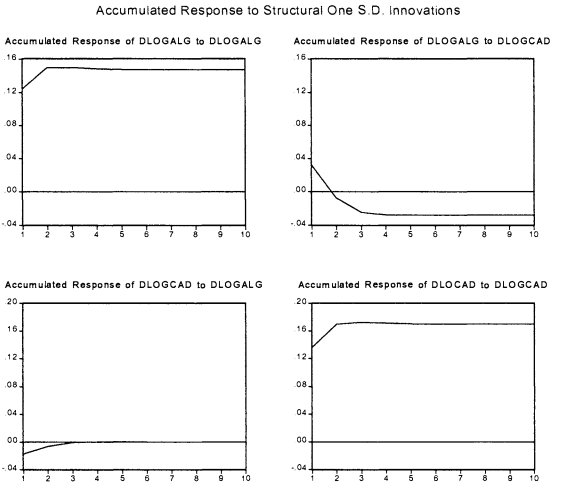


Fig. 4.3. Impulse Response functions accumulated under the Csh = 0 restriction

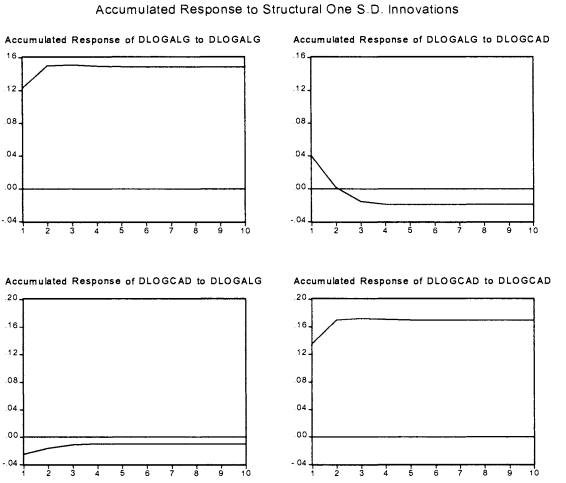


Fig. 4.4. Impulse Response functions accumulated under the Csh = -0,010 restriction

Table 4.3. VAR Bahía de Algeciras and Bahía de Cádiz ports

	∇LogALG	∇LogCAD
∇LogALG(-1)	0.152218	0.119393
	(0.14036)	(0.14995)
	[1.08449]	[0.79620]
∇LogCAD(-1)	-0.326435	0.217643
	(0.16384)	(0.17504)
	[-1.99238]	[1.24339]
C	0.122171	0.016222
	(0.03620)	(0.03868)
	[3.37482]	[0.41944]
R-squared	0.305305	0.067350
Adj. R-squared	0.242151	-0.017436
Sum sq. Resids	0.364097	0.415571
S.E. equation	0.128646	0.137439
F-statistic	4.834294	0.794353
Log likelihood	17.39166	15.73875
Akaike AIC	-1.151333	-1.019100
Schwarz SC	-1.005.068	-0.872835
Mean dependent	0.146757	0.038944
S.D. dependent	0.147777	0.136257
Determinant Residual Covariance		0.000308
Log Likelihood (d.f. adjusted)		30.12616
Akaike Information Criteria		-1.930093
Schwarz Criteria		-1.637563

4.4.2 Scenario 2: Lack of Shared Hinterland between Infrastructures. A Practical Application: Bivariate SVAR between the Bahía de Algeciras and Las Palmas Ports

In this scenario we have two transport infrastructures which are sufficiently far apart so that no hinterland is shared by them. This makes the analysis much simpler since the effects of trade deviation between the infrastructures shall be nil and we therefore have that DCsh = DChs = 0. Within this scenario our model's equations would be as follows:

Chs = kCMhs*Css.

The long-run restriction being:
Csh = cCMsh*Chh = 0, since, as explained above, CMsh will tend to 0 in a truly hub infrastructure with multiple national and international connections and much more global traffic than the one of the spoke infrastructure.

The current Spanish ports which have a feeder line in connection with the Bahía de Algeciras port are Alicante, Barcelona, Bilbao, Cartagena, Las Palmas de

Gran Canaria, Melilla, Santa Cruz de Tenerife, Tarragona, Valencia and Vigo. However, not all these ports are suitable for our analysis. Firstly, we must discard those ports which have hub infrastructures at the same time, since we cannot assume that Csh =0 when both ports are hub. Moreover, since these ports would compete for the same transoceanic lines within the same geographic area – in this case within the western Mediterranean – the effects of trade deviation between them will be strong and would have to be taken into account, even though the hinterland shared between the Bahía de Algeciras port and the rest of hub ports may be nil. For these reasons, we must exclude the analysis of the Valencia and Barcelona ports[17]. Finally, in this first approach, we have chosen the Las Palmas port since, being an insular port, it is to be expected that the hinterland shared with the port of Bahía de Algeciras is nil.

Following the same line of the previous model, the first step to specify this model is to survey the series which are to be dealt with. As was the case previously, we can observe on the graph a positive relationship between the mean and the standard deviation of the variables expressed in levels, especially in the port of Bahía de Algeciras[18] (see figures 4.1 and 4.5). This evidence of heterocedasticity suggests us again the employment of the logarithmic transformation in both series. In addition, again by merely observing the series, we can conclude logALG and logLasP are nonstationary series, as seen in figure 4.6. For all this, the SVAR model shall be specified with the following series ∇logALG and ∇logLasP.

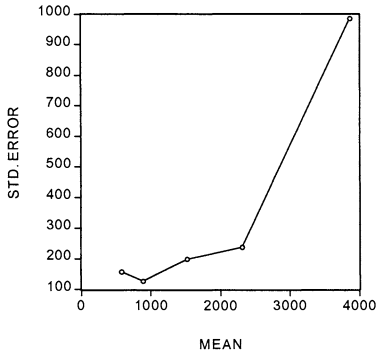


Fig. 4.5. Graph of the mean-typical deviation of Las Palmas port series

[17] In the case of Valencia, Barcelona ports and, to a lesser extent, Bilbao port, we must take into account that they really have a shared hinterland with the port of Bahía de Algeciras although indirectly, through the Dry Dock of Madrid. This shared hinterland is the central area of Spain.

[18] It is logical to assume that the heterocedasticity will be higher in a hub infrastructure than in a spoke one. The traffics of the former will undergone a stronger fluctuation depending on the phase of the cycle, while the spoke infrastructure will be more stable since it uniquely depends on its hinterland.

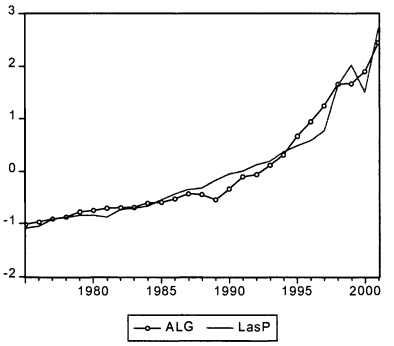


Fig. 4.6. Bahía de Algeciras and Las Palmas port series (standardised data)

By merely observing the ∇logALG and ∇logLasP series, we cannot conclude that they are stationary so we must apply the unit-root tests. Table 4.4 sets forth the results of these unit-root tests for these series, which show that the nil hypothesis is again rejected, and therefore, the original series (logALG y logLasP) are just I(1).

Table 4.4. Unit-root test for ∇logALG and ∇logLasP

Test ∇logALG	t-stat	MacKinnon p-values
Augmented Dickey-Fuller based SIC	-4.8840	0.0007
Phillips-Perron(Newey-West using Bartlett kernel)	-9.0163	0.0000
Test ∇logLasP	t-stat	MacKinnon p-values
Augmented Dickey-Fuller based SIC	-5.9147	0.0001
Phillips-Perron(Newey-West using Bartlett kernel)	-6.5017	0.0000

Once the likelihood ratio tests have been applied to determine the optimal lag of the VAR[19] and the lag exclusion tests, we have chosen the following specification:

$$\nabla logALG = - 0.0060*\nabla logALG(-2) + 0.4472*\nabla logLasP(-2) + 0.0819.$$

$$\nabla logLasP = - 0.0702*\nabla logALG(-2) - 0.3233*\nabla logLasP(-2) + 0.1448.$$

[19] Table 4.5 accounts for the results of the tests in order to determine the optimal lag in the model studied. The optimal lag selected by the criterion is marked with an asterisk.

Table 4.5. VAR optimal lag (∇logALG and ∇logLasP)

Lag	LogL	LR	FPE	AIC	SC	HQ
0	33.18294	NA	0.000201	-2.834813	-2.735627*	-2.811448
1	34.76988	2.741085	0.000252	-2.615444	-2.317887	-2.545349
2	42.51476	11.96935*	0.000181*	-2.955887*	-2.459958	-2.839061*

* indicates lag order selected by the criterion
LR: sequential modified LR test statistic (each test at 5% level)
FPE: Final prediction error
AIC: Akaike information criterion
SC: Schwarz information criterion
HQ: Hannan-Quinn information criterion

Table 4.6. VAR Bahía de Algeciras and Las Palmas

	∇LogALG	∇LogLasP
∇LogALG(-2)	-0.006029	-0.070259
	(0.12120)	(0.11066)
	[-0.04974]	[-0.63490]
∇LogLasP(-2)	0.447254	-0.323318
	(0.24665)	(0.22520)
	[1.81331]	[-1.43569]
C	0.081991	0.144847
	(0.03998)	(0.03650)
	[2.05073]	[3.96795]
R-squared	0.140674	0.127223
Adj. R-squared	0.058833	0.044101
Sum sq. Resids	0.321877	0.268328
S.E. equation	0.123804	0.113038
F-statistic	1.718877	1.530559
Log likelihood	17.68515	19.86865
Akaike AIC	-1.223.762	-1.405721
Schwarz SC	-1.076506	-1.258464
Mean dependent	0.130969	0.095923
S.D. dependent	0.127615	0.115616
Determinant Residual Covariance		0.000193
Log Likelihood (d.f. adjusted)		34.54087
Akaike Information Criteria		-2.378406
Schwarz Criteria		-2.083893

As can be observed, the first lag has been removed since the p-value associated with the Nil Hypothesis of TheWald's Exclusion Test of the coefficients is

0.6714[20]. The following step is to check that Csh = 0 is within our model's confidence intervals. In this case, the intervals for an accumulated response of 10 periods are (0.0135,-0.0261), so we can support this hypothesis.

Below are set out the graphs of the Impulse Response functions with the long-run restriction of Csh = 0, Csh being Clasp/alg in this model. The graphs correspond to the assumptions of the model proposed since, as can be observed, Clasp/lasp, Calg/lasp and Calg/alg are clearly positive.

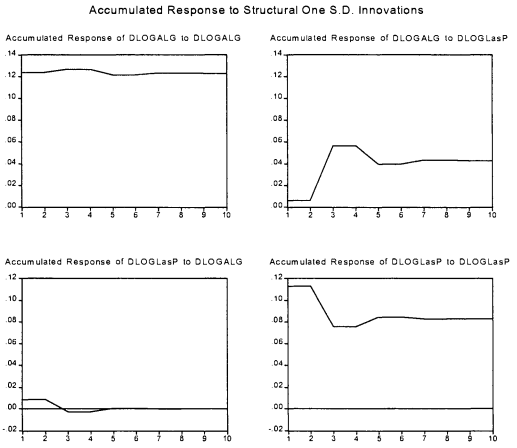


Fig. 4.7. Impulse Response functions accumulated under the Csh = 0 restriction

Finally, in this case, the analysis of residues again supports the white noise assumption for the residues. We have obtained a p-value of 0,7961 under the orthogonalisation restriction proposed.

4.5 Conclusions

The applications detailed in this paper make us be optimistic as regards the employment of the SVAR methodology in transport economy. There is clear evidence that this category of models may be a useful instrument not only for traffic predictions but also for planning infrastructures and reducing the congestion costs associated with them. We should not overlook that we are dealing with an economic sector – transport - in which , the infrastructure demand is growing at a higher annual rate than that of the supply, both public and private. This fact makes it necessary to have the widest number of complementary mathematical instruments which help to plan the sector.

[20] Table 4.6 accounts for the results of the VAR model specified.

Among the positive aspects of this methodology we can stand out its compatibility with other methods to make economic decisions such as the cost-social welfare analysis or the economic impact methodologies. As far as its negative aspects are concerned, we find that this methodology is initially restricted to hub-and-spoke transport structures, since we cannot establish long-run restrictions to carry out the SVAR for other ways of transport which do not have a central infrastructure to distribute the traffic. For this type of transport infrastructures which do not have a hub and operate with a common traffic, other possibilities within cointegration analyses can be employed.

The tests here exposed have shortcomings which force us to be cautious and wait for further works which should be orientated not only to the port activity but also to other sectors such as the airport system.

References

Amisano, G. and Giannini, C.: Topics in Structural VAR Econometrics. Springer 1997

Blanchad, O. J. and Quah, D. T.: The Dinamics Effects of Aggregate of Demand and Supply Disturbances. The American Economic Review 79(4), 655-673 (1989)

Breitung, J.: A convenient representation for structural vector autoregressions. Empirical Economics 26, 447-459 (2001)

Cochrane, J. H.: What do the VARs Mean? Measuring the Output Effects of Monetary Policy. Journal of Monetary Economics 41 (2), 277–300 (1998)

Eviews 4 User's Guide. U. S. A.: Quantitative Micro Software 2002

DeSerres, A. and Guay, A.: Estimating and Projecting Potential Output Using Structural Var Methodology. Working Paper 95-2. Bank of Canada 1995

Dwyer, M.: Impulse Response Priors for Discriminating Structural Vector Autoregresions. Working Paper. UCLA 1999

Faust, J.: The Robustness of Identified VAR Conclusions About Money. Carnegie-Rochester Conference on Public Policy 49, 207–244 (1998)

Faust, J., and Leeper, E. M.: When Do Long-Run Identifying Restrictions Give Reliable Results?. Journal of Business and Economic Statistics 15 (3), 345–353 (1997)

Gottshalk, J.: An Introduction into the SVAR Methodology: Identification, Interpretation and Limitations of SVAR models. Kiel Working Paper 1072. Germany 2001

Hamilton, J. D.: Time Series Analysis. Princeton University Press 1994

Jang, K.: Impulse Response Analysis with Long Run Restrictions on Error Models. Working Paper 01-04. Ohio University 2001

Phillips, P. C. B. and Perron, P.: Testing for a Unit Root in Time Series Regression. Biometrika 75, 335–346 (1988)

Rudebusch, G. D.: Do Measures of Policy in a VAR Make Sense?. International Economic Review 39 (4), 907–931 (1998)

Sims, C. A.: Comment on Glen Rudebusch's Do Measures of Monetary Policy in a VAR Make Sense?. International Economic Review 39 (4), 933–941 (1998)

Urzua, C. M. (1997). Omnibus Tests for Multivariate Normality Based on a Class of Maximum Entropy Distributions. In: Advances in Econometrics Volume 12, 341-358. JAI Press 1997

PART II. PRODUCTION AND COSTS (SUPPLY)

5 Technical and Allocative Inefficiency in Spanish Public Hospitals*

C. García-Prieto
University of Valladolid (Spain)

The OECD economies are undergoing an upward trend in health expenditure. The usual explanations for this tendency are based on such elements as the progressive ageing of population, the increasing use of technology in health procedures, and the fact that public health services become more extensive (cover more people) year after year.

All this has led governments to try to curb health spending by reducing the degree of inefficiency of public health services. Such inefficiency has been generally admitted by many studies carried out by experts in the area[1].

In Spain, health spending is mainly public and about 60% of this expenditure goes to hospitals. This draws attention to a sector that has continuously exceeded budgets in recent years, requiring extra money and generating a continuous pressure on costs.

There has been an attempt to deal with the situation from the point of view of prospective budgeting, through establishing *Program-Contracts*. This attempt, according to recent papers by González and Barber (1996) and Ventura and González (1999), has led to a significant decrease of inefficiency in hospitals directly managed by the national government. Ventura and González's study concentrates on technical efficiency by applying Data Envelopment Analysis. On the other hand, González and Barber follow a double approach: First, they analyse global inefficiency on the basis of the estimation of a parametric cost frontier.

* I wish to thank professors Guillem López-Casasnovas, Carlos Pérez Domínguez and José Miguel Sánchez Molinero for their comments, which I have found very helpful. However, I am responsible for any possible mistake.

1 Undoubtedly, the "Dunning report" (1991) from Holland is the most renowned. In Spain, the Congress entrusted a committee headed by Abril Martorell with a study, and their conclusions were recorded in a document known as "Abril's Report" (1991).

P. Coto-Millán (ed.), *Essays on Microeconomics and Industrial Organisation*

Second, they use Data Envelopment Analysis to quantify the technical component of hospital inefficiency.

However, the allocative component of hospital inefficiency has not been given much attention in the literature. This is probably due to information problems concerning input prices. In spite of such problems, there have been some relevant papers on this matter, such as Eakin and Kniesner (1988) and Eakin (1991). These two papers use the same sample of 133 hospitals in the USA and concentrate exclusively on allocative inefficiency. Other papers, such as Morey, Fine and Loree (1990) and Byrnes and Valdmanis (1994), quantify allocative inefficiency for a set of hospitals in California by means of Data Envelopment Analysis.

For the case of Spain, only Puig-Junoy (2000) has studied allocative efficiency in hospitals, using a sample of both public and private hospitals in the region of Catalonia. Puig-Junoy's approach uses the Data Envelopment Analysis technique. This author finds an interesting negative association between the proportion of public funds in the total budget of each hospital and de degree of allocative efficiency.

The present paper studies the degree of efficiency of the Spanish hospitals managed by the *INSALUD,*[2] by quantifying the extent to which each of the two inefficiency components (technical and allocative) increases cost. We have used an econometric analysis based on the estimation of a system consisting of a cost frontier and a set of input share equations.

This paper is organised as follows: section 5.1 presents the theoretical model to be used; section 5.2 describes the variables and data to be used in the estimation; section 5.3 analyses the results, and section 5.4 summarises the final conclusions.

5.1 The Model

Empirical studies on economic efficiency have usually followed the frontier methodology, ever since Farrell (1957) set the bases for this approach[3]. Farrell's approach basically starts from the assumption that firms do not follow optimal behaviour patterns (they do not maximise profits) due to inefficiencies in the management of the production process.

When calculating a cost frontier, it must be taken into account that firms may fail in their attempts to minimise costs for two different reasons: On the one hand, firms may not manage to use the minimum quantity possible of each input for the obtained level of output; this causes technical inefficiency. On the other hand, the factor combination used by the firm may not be the cheapest one, given the input prices; this would lead to allocative inefficiency. Both behaviours are shown in figure 5.1 for the case of two inputs.

---

2 *INSALUD* is the public agency which finances and administrates the Spanish hospitals in 11 out of the 17 autonomous regions existing in the country.

3 Forsund, Lovell and Schmidt (1980), Bauer (1990), Greene (1993) and Coelli, Rao and Battesse (1998) are good surveys on this technique.

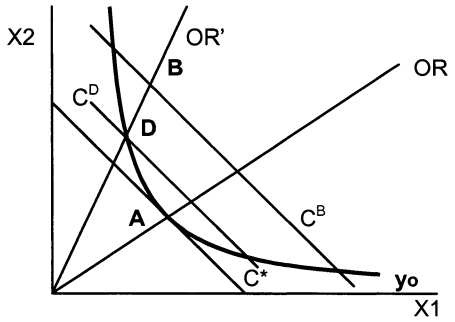


Fig. 5.1. Technical and allocative inefficiency

A firm operating at point B would present both types of inefficiency: allocative inefficiency, since the firm is using an input combination that is not optimal, and technical inefficiency, since both inputs are being used in quantities greater than those required by output level y_0, when production process OR' is chosen. The fact that OR' is chosen, instead of OR, results in allocative inefficiency, while the fact that input combination B is chosen, instead of D, results in technical inefficiency.

If the firm chose input combination D instead of B, there would be no technical inefficiency but there would be allocative inefficiency. This allows us to measure allocative inefficiency at point B by the ratio $AI_B = C^D/C^*$, where C^D stands for the total cost at point D and C* stands for the minimum cost, reachable only when both sources of inefficiency have been eliminated (point A). The measure of technical inefficiency is given by the ratio $TI_B = C^B/C^D$ where C^B stands for total cost at point B. The product of these two measures ($TI_B * AI_B = C^B/C^*$) shows the proportion in which the actual cost is higher than the minimum cost. Therefore, we can state that C^B/C^* is a global measure of economic inefficiency.

Economic efficiency can be measured by the inverse ratio, C^*/C^B. It shows the proportion in which the minimum cost is lower than actual cost. This ratio is the product of two measures: that of technical efficiency and that of allocative efficiency:

$$\frac{C^*}{C^B} = \frac{1}{TI_B AI_B} = \frac{1}{TI_B}\frac{1}{AI_B} = TE_B AE_B$$

According to Aigner, Lovell and Schmidt (1977), the cost incurred by a firm may be outside the minimum cost function due to economic inefficiency (either technical or allocative) as well as to purely random elements (thus, the term "stochastic frontier"). This is shown by the expression:

$$C(y, w) = C^*(y, w, \beta)\ e^v\ TI\ AI \qquad (5.1)$$

where C (y, w) is the effective cost of the firm and C* (y, w, β) is the minimum cost function which depends on production, y, input prices, w, and a number of

parameters β, that must be estimated; the term e^{v} is the random disturbance and the product TI AI, is the measure of economic inefficiency as stated earlier[4].
In logarithmic form:

$$\ln C(y, w) = \ln C^*(y, w, \beta) + \ln AI + \ln TI + v \tag{5.2}$$

While the random disturbance can make costs go up or down (υ may be either positive of negative), inefficiency always increases costs, as shown by the fact that lnAI and lnTI are both non-negative.

Expression (5.2) could be estimated directly. We would only need to postulate a particular functional form for the cost frontier and an adequate distribution for each of the error terms (v, lnAI and lnTI). Nevertheless, this procedure would be subject to strong objections. The basic problem here is that allocative inefficiency (the lnAI term) is clearly related to input prices and this implies that estimators cannot be expected to be consistent.

In order to avoid this problem, we could estimate (5.2) together with the input share equations. This will be the procedure followed in this paper. Next, we will proceed to specify the input share equations, taking into consideration the fact that there may be allocative inefficiency.

The effective share of the j^{th} input in total cost, $S_j(y, w)$, is given by the ratio:

$$S_j(y, w) = x_j w_j / C(y, w) \tag{5.3}$$

where x_j represents the amount of the j^{th} input used by the firm, and w_i the price of that particular input. Now, if we take into account that $C(y, w) = \Sigma_j x_j w_j$, it follows that $\partial C(y, w)/\partial w_j = x_j$, which allows us to write:

$$\frac{\partial \ln C(y,w)}{\partial \ln w_j} = \frac{x_j w_j}{C(y,w)} \tag{5.4}$$

Hence, we may define the effective share of the j^{th} input as:

$$S_j(y,w) = \frac{\partial \ln C(y,w)}{\partial \ln w_j} \tag{5.5}$$

and taking into account expression (5.2), it follows that:

$$S_j(y,w) = \frac{\partial \ln C^*(y,w,\beta)}{\partial \ln w_j} + \frac{\partial AI}{\partial \ln w_j} \tag{5.6}$$

since TI and υ do not depend on input prices.

Now, applying Shephard's lemma, we can identify $\partial \ln C^*(y,w,\beta)/\partial \ln w_j$ with the optimal share of the j^{th} input; that is:

$$S_j^*(y,w,\beta) = \partial \ln C^*(y,w,\beta)/\partial \ln w_j \tag{5.7}$$

This expression allows us to write:

$$S_j(y,w) = S_j^*(y,w,\beta) + \varepsilon_j \tag{5.8}$$

where the error term, ε_j , would include the effects of allocative inefficiency plus a purely random element, ξ_j; that is:

[4] Subscripts have been removed with the aim of simplifying the expression.

$$\varepsilon_j = \partial \ln AI / \partial \ln w_j + \xi_j \qquad (5.9)$$

ε_j can have both positive and negative values, showing over or under-utilization of the j^{th} input with respect to the optimum level.
In short, we will estimate the following system:

$$\ln C\,(y, w) = \ln C^*\,(y, w, \beta) + \ln AI + \ln TI + \nu$$

$$S_j\,(y,w) = S_j^*(y,w,\beta) + \varepsilon_j \qquad j = 1...n-1 \qquad (5.10)$$

It must be observed that the number of share equations is not n, but n-1. This is so in order to avoid the estimation problems derived from the fact that the sum of the n shares must be equal to one.

The relationship between the error terms of the input share equations and the lnIA term in the cost function is usually approximated[5] by a quadratic form. This quadratic expression accounts for the fact that both positive and negative deviations from optimality result in cost increases.

Schmidt (1984) proposed a specification in which the sum of error squares was weighed through a positive semi-definite matrix of parameters, F, such that lnAI = $\varepsilon'F\varepsilon$. The problem here is how to chose such parameters. On this subject, Kumbakhar (1991) showed that the expression lnAI = $\varepsilon'F\varepsilon$ cannot be taken without restrictions. More specifically, the integrability condition,

$$\partial \ln (\varepsilon'F\varepsilon) / \partial \ln w_j = \varepsilon_j \qquad j = 1...n \qquad (5.11)$$

must be fulfilled. As shown by Kumbhakar (1991), this integrability condition implies that the parameters in the F matrix are exact functions of the cost frontier parameters when a translog frontier is considered. To reach such a conclusion, he uses a rather restrictive specification of the share equations. In essence, he regards these equations as deterministic; that is to say, the error terms, ε_j's, are only due to allocative inefficiency and do not have any random component.

In a later paper, Melfi (1984) proposed simplifying F into an identity matrix. This would make allocative inefficiency take values systematically close to zero, since the absolute value of share deviations is lower than one. Faced with this problem, Bauer (1990) suggested weighing the identity matrix through a positive parameter, which, depending on the final value it takes in the estimation, enables us to avoid this systematic tendency. This paper will follow Bauer's suggestion. Hence, for us:

$$\ln AI = c \sum_{j=1}^{n} \varepsilon_j^2 \qquad (5.12)$$

In order to carry out the estimation of the system, some assumptions must be made about the distribution of the disturbances: lnTI, ν and ε_j, taking into account that they are independent of one another. It will be assumed that ν is independently and identically normal distributed with zero mean and variance σ^2_ν ; ε is distributed independently and identically as multivariate normal with constant mean μ and constant covariance Σ, where $\varepsilon = (\varepsilon_1...\varepsilon_{n-1})'$ and $\mu = (\mu_1,... \mu_{n-1})'$.

[5] The exact relationship for the restrictive Cobb-Douglas function has been proposed by Schmidt and Lovell (1979). It has also been obtained for the translog function –Kumbhakar (1997)- although, given its complexity, it has not yet been employed in any empirical study.

The fact that the mean of the errors in the share equations may be different from zero, as Schmidt and Lovell (1979) suggested, can be interpreted as evidence of systematic mistakes in the choice of the input combination. If the error mean is shown to be equal to zero, deviations with respect to optimal share levels will be merely random.

As far as technical inefficiency is concerned, we must propose a distribution over positive values. In this case, we have chosen an exponential distribution[6] with constant mean $1/\alpha$, although results show a high correlation with those obtained when the chosen distribution is a seminormal[7].

Once the errors of the input share equations have been estimated, lnAI can be obtained from expression (5.12). Therefore, the likelihood function of the system (5.10) is given by the product of two density functions: that of the composed error, $\ln TI + v$, and that of the errors of share equations, ε, that is:

$$f[(\ln TI + \ln AI + v), \varepsilon] = g(\ln TI + v)\, h(\varepsilon) \tag{5.13}$$

The maximum likelihood technique enables us to obtain a consistent and asymptotically efficient estimation of the cost frontier parameters and the allocative inefficiency of each firm. From the cost frontier residuals obtained, it is possible to estimate the mean technical inefficiency of all firms; particular estimates of technical inefficiency for individual firms can also be obtained using the procedure suggested by Jondrow et al. (1982). It must be observed, however, that the estimates of technical efficiency obtained by such a procedure are unbiased but non-consistent.

5.2 Definition of Variables and Functional Specification

Most of the data used in this estimation has been obtained from the EESRI, a survey of hospitals carried out by the INE (the Spanish Institute of Statistics). All data refer to 1994.

These data refer only to public hospitals managed by the *Insalud*. The very small hospitals (less than 80 beds) and the very large ones (more than 1000 beds) have been excluded from the sample. The reason for this is that small and large hospitals usually differ substantially in their endowments of technical equipment. They also differ widely in the type of cases attended. This fact could be a handicap for inefficiency estimation since the highest use of resources could be understood as inefficiency when, in fact, it could simply mean a more intensive attendance to very complex cases. A total of 67 hospitals have been studied.

Since we are dealing with a cross-sectional analysis, a short-run cost function has been specified, which depends on the output level, the input prices and the amount used of the fixed input, capital.

---

[6] The literature has mainly dealt with the following distributions: the seminormal, the truncated normal (only for positive values) and the exponential distributions; see, for instance, Greene (1993) for further details on this issue.

[7] In the case of a truncated normal distribution for positive values, the mean was not found different from zero.

5.2.1 Outputs

The first difficulty encountered when defining the variables involved in the cost function is how to measure the final output of hospital activity. Hospitals produce health and health improvement is difficult to measure. Nevertheless, since we do need a measure, we have to use some proxies. Such proxies could be the number of inpatient days or the number of cases attended. We have chosen the number of cases rather than the hospital stays. The problem with the number of stays is that many of them are presumably unnecessary and cost minimization may often lead the hospital to reduce the average length of the stay in each case by administering treatment in fewer days, all of which causes a cost increase per stay. To sum up, the accommodation services provided by hospitals may be considered as input in the "production of health" rather than an output.

When discussing hospital activity, it is convenient to distinguish between inpatient and outpatient attendance. This provides us with two basic output measures:

CASES : this variable accounts for inpatient hospital activity, and it is developed from the number of discharged patients adjusted by the casemix complexity. It is a weighed sum of discharges from the different hospital services. The weighs have been carried out using the coefficients defined by the *Insalud* in the UPA[8].

AMBU: this variable measures ambulatory hospital care: first and successive visits and emergencies with no hospitalization. In this case, weighing is also carried out with the coefficients fixed by the UPA.

5.2.2 Input Prices

Two inputs, labor and materials, have been considered. The variables that account for their respective prices are:

WAGE: The wage measure used in this paper is the average earnings of all hospital employees. *Insalud* wages are equal for all hospitals within each particular labor category (doctor, nurses, etc.), and wage data come from the BOE[9] (the Official State Bulletin).

According to the previous definition, differences in wages among hospitals basically show the existence of different mixes of labor inputs (that is, the fact that hospitals employ different proportions of doctors, nurses, etc.). In spite of everything, our wage variable can be regarded as a global measure of labor costs, which allows us to use it as a proxy for the price of labor. If instead of using this global measure, we were to use the "real" price of any particular labor category (doctors' salaries, for instance) we would observe identical values for all hospitals and that would make our estimation impossible.

---

[8] These coefficients have been determined by the *Insalud* in a study that estimated the relative consumption of resources of a one day stay in each service, outpatient visits and emergency attendance.

[9] 21/1993 Act of the *Presupuestos Generales del Estado* (National Budget Account) for 1994, December 29th.

PMAT: This variable accounts for the price of an aggregated input which includes the purchasing of varied health equipment, cleaning, food, energy supply..., all of which can be called materials. The diversity of components makes it difficult to estimate a particular price for this category. Hence, we have resorted to an approximation: the total spending in this category per day of stay[10].

The variability of PMAT may be related to both differences in prices and differences in the casemix of each hospital. This is indeed a drawback of the PMAT measure. Nevertheless, it can be argued that most elements included in this category (for instance: expenditure in food, clothes, cleaning, heating...) are unrelated to the casemix complexity.

5.2.3 Fixed Input

BEDS: The amount of fixed input (capital) used by each hospital is approximated by the number of beds. The justification for this is that the number of beds in the hospital reflects the stock of capital invested around them.

In order to use this proxy properly, it is necessary to have a homogeneous sample of hospitals. Therefore, we have focused the analysis on general hospitals[11], leaving out big hospitals with expensive modern technologies, and small hospitals with less sophisticated equipment. This leads us to consider that the hospitals in the sample have invested the same amount of capital per bed and, therefore, that the use of the bed variable seems adequate as an approximation to capital.

The statistical description of the variables defined above is summarised in table 5.1.

Table 5.1. Statistical description of the variables

VARIABLE	MEAN	STD. ERROR	MAX	MIN
COST	5,784,800	4,670,787	20,545,979	939,649
CASES	23,573	17,342	72,962	989
AMBU	119,038	92,075	533,145	7,530
WAGE	3,347,091	102,420	3,560,862	3,034,687
PMAT	19,870	12,925	95,672	12,088
BEDS	360	242	958	80

5.2.4 Functional Form

As regards the functional form, a translog cost function has been chosen for two reasons. On the one hand, for its flexibility, since it hardly places restrictions on technological features[12] from the beginning, but these restrictions can be tested

[10] Therefore, it is considered that the stays produced may be an adequate proxy of the amount of factor employed.

[11] Hospitals specialized in particular health services have not been included since they may show differences with respect to the rest for requiring either too much or too little capital investment.

[12] We should not forget that the translog function is the approximation to the real function in one point.

later. On the other hand, it enables us to introduce more than one output, thus accounting for the multiproduct character of hospital activity.
Then, the following system has been estimated:

$$\ln(\text{Cost}) = \alpha_0 + \alpha_1 \ln(\text{CASES}) + \alpha_2 \ln(\text{AMBU}) + \beta_1 \ln(\text{Wage}) +$$
$$\tfrac{1}{2}\delta_{11} \ln(\text{CASES})^2 + \tfrac{1}{2}\delta_{22} \ln(\text{AMBU})^2 + \delta_{12} \ln(\text{CASES}) \ln(\text{AMBU}) +$$
$$\tfrac{1}{2}\gamma_{11} \ln(\text{Wage})^2 + \rho_{11} \ln(\text{CASES}) \ln(\text{Wage}) + \rho_{21} \ln(\text{AMBU}) \ln(\text{Wage}) +$$
$$\eta_{11} \ln(\text{BEDS}) + \ln\text{AI} + \ln\text{TI} + \nu$$

$$S_1 = \beta_1 + \gamma_{11} \ln(\text{Wage}) + \rho_{11} \ln(\text{CASES}) + \rho_{21} \ln(\text{AMBU}) + \varepsilon_1 \qquad (5.14)$$

in which only the labor share equation[13] has been included; the usual symmetry restrictions have been imposed, and the theoretical requirement of homogeneity of degree one in input prices has been achieved by dividing costs and input prices by the price of the second input, which yields:

$$\text{Cost} = \text{COST} / \text{PMAT},$$
$$\text{Wage} = \text{WAGE} / \text{PMAT}.$$

5.3 Estimation and Results

Maximum likelihood estimation has been carried out on deviations with respect to the mean. Therefore, the parameters of the first-order terms represent the respective cost function elasticities for the average hospital. All of them are positive and, except for the case of the AMBU variable, significant at 1%, as observed in table 5.2, which summarises the results obtained.

On the whole, parameters are highly significant, except those associated with the AMBU variable and its cross products. This is due to the strong correlation between the AMBU and the CASES variables. However, the existing multicolinearity does not affect the precision of the parameters relevant for inefficiency; so, the conclusions still stand.

The significance of the exponential distribution parameter (α) confirms the existence of technical inefficiency, which explains 60% of the total variability of the composed error in the cost function[14].

As far as labor share is concerned, the mean error, μ_1, is significantly different from zero, indicating a systematic overutilization of labor. The estimated value of μ_1 means that hospitals, on the average, spend about 16.8 percentage points more on labor than required by minimum cost considerations.

---

[13] The share equation of the remaining factor has not been included because it does not provide any additional information (the sum of both equations is one).

[14] The composed error variability is explained by the variance of each one of its two components: lnTI and ν. When the exponential distribution parameter, α, approaches infinity, the technical inefficiency variance tends to zero. In such a case, the variability of lnTI+ν would be entirely explained by the random disturbance variance. Then, we would conclude that there is no technical inefficiency in that case.

Table 5.2. Estimation results

VARIABLE	COEFFICIENT	STD. ERROR
constant	-0.214891***	0.07496
ln (CASES)	0.596417***	0.11039
ln (AMBU)	0.052727	0.06816
½ ln (CASES)^2	-0.166184	0.14222
½ ln (AMBU)^2	-0.189554*	0.12799
ln (Wage)	0.549951***	0.07594
½ ln (Wage)^2	0.090459***	0.01210
ln (BEDS)	0.361394***	0.07057
ln (CASES) ln (AMBU)	0.143169	0.13046
ln (Wage) ln (CASES)	-0.034781***	0.00737
ln (Wage) ln (AMBU)	0.008727	0.00840
c	3.002640**	1.74996
α	12.97428***	4.59385
σ_v	0.063042***	0.01652
μ_1	0.168106**	0.07587
σ_ε	0.022380***	0.00246
Log likelihood		222.7361
Akaike info criterion		-6.171227

*** significant at 1%; ** significant at 5%; *significant at 10%.

The overutilization of labor implies underutilization of the other factor. Hence, we can state that the wrong input combination is been used, which is evidence of allocative inefficiency, and its overall impact on costs can be calculated through expression (5.12).

From these results, efficiency indexes, whose statistical description is summarized in table 5.3, are estimated for each hospital. In short, mean technical efficiency reached by the hospitals is considerably higher than mean allocative efficiency (with a difference of eight and a half points).

Table 5.3. Estimated economic, technical and allocative efficiency

	Economic ef.	Technical ef.	Allocative ef.
Mean	78.12	92.74	84.22
Std. Error	5.77	5.26	3.75
Min.	55.45	70.86	77.37
Max.	90.33	97.99	93.96

These indexes can be interpreted as follows: if *Insalud* hospitals were as efficient as the best ones, then their cost would be, on average, only 78.12% of their actual cost. If they used the least amount of resources possible, their cost would be around 92.7% of their actual cost, and if they chose factors in the right proportion (depending on prices), their cost would be 84.2% of their current cost.

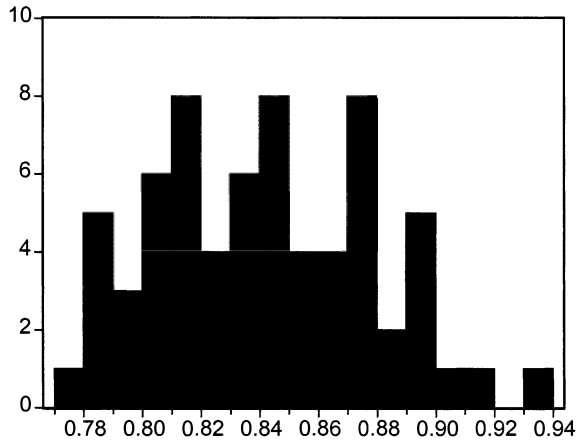


Fig. 5.2. Allocative efficiency distribution

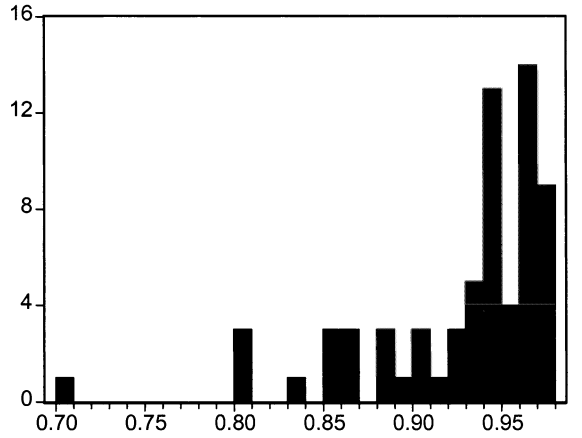


Fig. 5.3. Technical efficiency distribution

Figures 5.2 and 5.3 show the distribution of each of the two indexes among the hospitals. Both series are hardly correlative so it cannot be concluded that the most technically efficient hospitals also present the highest allocative efficiency.

Comparing the present paper with earlier works on Insalud hospitals, we may conclude the following: our technical efficiency index of 92.7% is slightly lower than the indexes estimated by González and Barber (1996). These authors use Data Envelopment Analysis and obtain estimates of technical efficiency for 1991, 1992 and 1993 of 95%, 96% and 97% respectively. However, our result is higher

than the one obtained by Ventura and González (1999), who also used a non-parametric analysis. Ventura and González studied Insalud hospitals between 1993 and 1996, and obtained values that go from 80.7% in 1993 up to 84.7% in 1996.

As regards allocative inefficiency, we agree with Puig-Junoy (1999) in that this inefficiency represents a problem quantitatively more important than that of technical inefficiency. Puig-Junoy (2000) uses Data Envelopment Analysis on a sample including both public and private hospitals in the region of Catalonia and obtains a measure of allocative efficiency with a higher dispersion (mean 89.1% and standard error 9.5).

Table 5.4 shows the cost increase caused by both types of inefficiency. According to our previous results, allocative inefficiency has a much higher impact on costs than technical inefficiency (18.96% and 8.2% respectively). Hence, any study focussing exclusively on technical inefficiency ignores the quantitatively more important aspect of hospital inefficiency.

Table 5.4. Cost increase due to economic, technical and allocative inefficiency

	Economic inef.	Technical inef.	Allocative inef.
Mean	28.76	8.2	18.96
Std. Error	10.52	6.9	5.25
Min.	10.69	2.05	6.43
Max.	80.31	41.12	29.25

As shown in table 5.4, total inefficiency (both technical and allocative) accounts for a 28.76% increase in hospital costs above minimum cost. Previous estimates of global inefficiency in Spanish hospitals have shown a wide variety of figures ranging from 13% to 58%, -see for instance: González and Barber (1996), Wagstaff and López (1996), López and Wagstaff (1993)-.

5.4 Conclusions

The aim of this paper has been to quantify separately technical and allocative inefficiency in the Spanish public hospitals. Given that a separate estimation of the hospitals' cost frontier is not adequate, we have resorted to a joint estimation of the cost frontier together with the input share equations. On the basis of this joint estimation, we have obtained two separate measures of technical and allocative efficiency.

These measures allow us to conclude that total inefficiency accounts for a 28.7% increase in hospital costs above minimum cost. Previous estimates of global inefficiency in Spanish hospitals have shown a wide variety of figures ranging from 13% to 58%, -see for instance: González and Barber (1996), Wagstaff and López (1996), López and Wagstaff (1993)-.

Perhaps the most important aspect of the present paper is that it provides an estimate of allocative inefficiency, which is missing in previous papers. The present paper also identifies allocative inefficiency as the main component of

economic inefficiency. More specifically, allocative inefficiency accounts for 65% of the total cost increase due to global inefficiency.

References

Aigner, D., Lovell C.A.K., Schmidt, P.: Formulation and Estimation of Stochastic Frontier Production Function Models. Journal of Econometrics 6, 21-37 (1977)

Bauer, P.W.: Recent Developments in the Econometric Estimation of Frontiers. Journal of Econometrics 46, 39-56 (1990)

Byrnes, P., Valdmanis, V.: Analysing Technical and Allocative Efficiency of Hospitals. In: Charnes, A., Cooper, W., Lewin, A.Y., Seiford, L.M., Data Envelopment Analysis: Theory Methodology and Application. Kluwer Academic Publisher 1994

Coelli, T., Rao, D.S.P., Battese, G.E.: An Introduction to Efficiency and Productivity Analysis. Kluwer Academic Publishers 1998

Comisión de Análisis y Evaluación del Sistema Nacional de Salud: Informe y Recomendaciones. Madrid 1991

Eakin, B.K.: Allocative Inefficiency in the Production of Hospital Services. Southern Economic Journal 58, 240-248 (1991)

Eakin, B.K., Kniesner, T.J.: Estimating a Non-Minimum Cost Function for Hospitals. Southern Economic Journal 54, 583-97 (1988)

Farrel, M.J.: The Measurement of Productive Efficiency. Journal of Royal Statistical Society 120, 253-281 (1957)

Forsund, F., Lovell, D.A.K., Schmidt, P.: A Survey of Frontier Production Functions and of Their Relationship to Efficiency Measurement. Journal of Econometrics 13, 5-25 (1980)

González, B., Barber, P.: Changes in the Efficiency of Spanish Public Hospitals after the Introduction of Program-Contracts. Investigaciones Económicas XX 3, 377-402 (1996)

Government Committee on Choices in Health Care: Report. Zoetermeer. The Nederlands 1991

Greene, W.H.: The Econometric Approach to Efficiency Analysis. In: Fried, H.O., Lovell, C.A.K., Schmidt, S.S., The Measurement of Productive Efficiency. Oxford University Press 1993

INE: Estadística de Establecimientos Sanitarios con Régimen de Internado. EESRI 1994

Jondrow, J., Lovell, C.A.K., Materov, I.S., Schmidt, P.: On the Estimation of Technical Inefficiency in the Stochastic Frontier Production Models. Journal of Econometrics 4, 23, 269-274 (1982)

Kumbhakar, S.C.: The Measurement and Decomposition of Cost-inefficiency: the Translog Cost System. Oxford Economic Papers 43, 667-683 (1991)

Kumbhakar, S.C.: Modelling Allocative Inefficiency in a Translog Cost Function and Cost Share Equations: An Exact Relationship. Journal of Econometrics 76, 351-356 (1997)

López, G., Saez, M.: Finance Versus Costs for Teaching Hospital in Spain. Working paper. CRES-UPF 1998

López, G., Wagstaff, A.: Eficiencia y Competitividad en los Servicios Públicos: Algunas Consideraciones Relativas a la Asistencia Sanitaria. Moneda y crédito 196, 181-231 (1993)

Melfi, C. A.: Estimation and Decomposition of Productive Efficiency in a Panel Data Model: An Application to Electric Utilities. Unpublished Doctoral Dissertation. University of North Carolina. Chapel Hill. NC 1984

Morey, R.C., Fine, D.J., Loree, S.W.: Comparing the Allocative Efficiencies of Hospitals. Omega 18, 71-83 (1990)

Puig-Junoy, J.: Partitioning Input Cost Efficiency Into Its Allocative and Technical Components. An Empirical DEA Application to Hospitals. Socio-Economic Planning Sciences, 34, 3, 199-218.

Schmidt, P.: An Error Structure for System of Translog Cost and Share Equations. Econometrics Workshop Paper 8309. Michigan State University 1984

Schmidt, P., Lovell. C.A.K.: Estimating Technical and Allocative Inefficiency Relative to Stochastic Production and Cost Frontiers. Journal of Econometrics 9, 343-366 (1979)

Shephard, R.W.: Cost and Production Functions. Princeton University Press 1953

Ventura, J., González, E.: Análisis de la Eficiencia Técnica Hospitalaria del Insalud GD en Castilla y León. Revista de Investigación Económica y Social de Castilla y León 1, 39-50 (1999)

Wagstaff, A., López, G.: Hospital Costs in Catalonia: a Stochastic Frontier Analysis. Applied Economics Letters 3, 471-474 (1996)

6 Technical Efficiency of Road Haulage Firms

J. Baños-Pino
University of Oviedo (Spain)

P. Coto-Millán
University of Cantabria (Spain)

A. Rodríguez-Álvarez
University of Oviedo (Spain)

V. Inglada-López de Sabando
University Carlos III (Spain)

This study justifies the contemporary importance of efficiency analysis. We put forward the theoretical concepts of technical, allocative, and economic efficiency. We then tackle problems of an empirical nature. In this way, different ways of measuring efficiency are presented, with their main disadvantages. To finish, a theoretical application is presented for haulage firms operating in Spanish roads in six different sub-sectors with panel data.

6.1 Introduction

The production theory of the firm, in most basic Microeconomics textbooks, starts from the premise that the manager behaves in an efficient manner, making the best possible use of the resources available. To begin with, there is the hypothesis that the firm, given the technology and the inputs available, produces the maximum output possible. This relationship is represented in the *production function*. Secondly, the first order conditions for cost minimisation, given the factor prices, are introduced through the *cost function*. Finally, there is the assumption that the manager chooses that quantity of output which maximises profits, as represented through the *profit function*. However, in reality it may happen that some or none

P. Coto-Millán (ed.), *Essays on Microeconomics and Industrial Organisation*

of these hypotheses are fulfilled. There are various reason for this, ranging from bad luck to the pursuit of objectives different to those outlined above. As a result, in the empirical estimation of these behavioural functions much research has been dedicated to testing firm behaviour by analysing the problem of inefficiency in the productive system.

6.2 Efficiency

We now distinguish between the different types of efficiency dealt with in the economics literature. We can distinguish between at least three types of efficiency: technical, allocative, and economic. We refer to *technical efficiency* when the optimality condition is defined by the production function. This means that the exact amount of inputs necessary are used in order to produce a vector of outputs, and hence there are no redundant (i.e. more than were strictly necessary) inputs. To estimate technical efficiency we need data about the quantities of inputs and outputs but it is not necessary to have data on prices.

Another concept is that of *allocative efficiency*, which occurs when the productive inputs are used in the proportions which minimise costs, that is, when the ratio of marginal products is equal to the ratio of their prices.

Finally, we talk about *economic efficiency* when both of the aforementioned kinds of efficiencies are achieved simultaneously. In Farrell (1957) this concept (he refers to it as productive or global efficiency) is defined as the product of technical efficiency and allocative efficiency (he refers to the latter as 'price efficiency').

We use Farrell's original ideas in order to provide a brief introduction to these concepts of efficiency. To begin, consider a firm employing two factors of production (x_1, x_2) to produce a single product. Under conditions of constant returns to scale, the technology of the company could be represented by a single isoquant, the unitary isoquant ($y = 1$), which can be drawn as in Figure 6.1.

According to the definition of technical efficiency used above, all the combinations of inputs lying on the curve $y = 1$ are technically efficient. The combinations below the frontier are unfeasible while the combinations which are above are technically inefficient. With reference to Figure 6.1, point A represents the situation of a company that employs (x_1^0, x_2^0) of inputs to produce the unit y. Now, according to the isoquant, $y = 1$ can be produced with $(OB/OA)x_1^0$ and $(OB/OA)x_2^0$. Alternatively, with (x_1^0, x_2^0) an efficient company can obtain $(OA/OB)y$ of the product. This means that technical efficiency (TE) can be measured using the ratio OB/OA. The maximum value of this index is one, which would mean that the firm is operating on the isoquant and is thus technically efficient. When a company is inefficient, then $OB \neq OA$ (as can be seen in Figure 6.1) and the index takes a value lower than one which informs us about the degree of technical efficiency achieved by the firm.

The definitions of allocative and economic efficiency can also be illustrated using Fig. 6.1. Allocative efficiency reflects the ability of a firm to use the inputs in the optimal proportions, given their respective prices and the production technology. Hence, if the relative prices of the factors of production are available

it is possible to know the slope of the isocost line PP'. A company that allocates resources efficiently will thus produce with a input combination given by the slope of the isocost where it is equal to that of the isoquant, represented at point C. At this point, the firm is producing the output at minimum cost, given the technology and the input prices. Hence, the measure of allocative efficiency (AE) is derived from the ratio OD/OB.

It is also possible to compare the total cost of producing a unit y using the input combination at C with that of producing y using the combination at A. We can use the ratio OD/OA to derive a measurement of economic or global efficiency (EE). A producer is economically efficient if it minimises the cost of producing the output given the input prices and the technology. A producer is economically efficient if and only if it is technically and allocatively efficient. In addition, we can see that:

$$OD/OA = (OB/OA)\ (OD/OB)$$

That is to say:

$$EE = (TE)\ (AE)$$

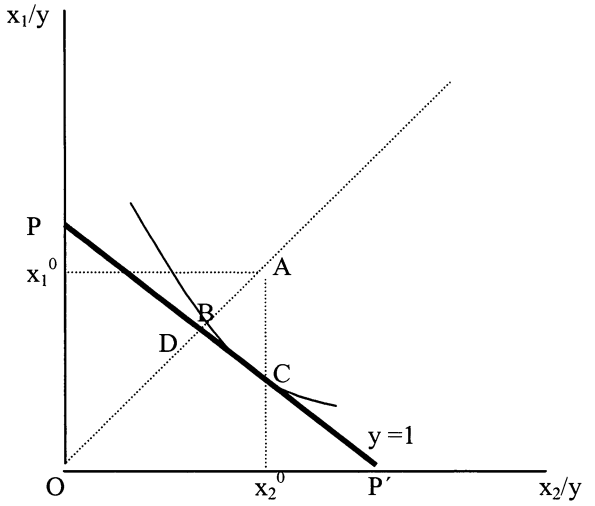


Fig. 6.1

Technical efficiency as described above could be expressed in different ways depending on whether we use the output or the inputs as a reference. For example, we could take the production quantity as given and calculate the efficient quantity of inputs (input-oriented technical efficiency). Alternatively, we could take the quantities of inputs as given and calculate the efficient quantity of output (output-oriented technical efficiency). Only if the technology exhibits constant returns scale will the input-orientated technical efficiency index be the same that the output-orientated technical efficiency index (Färe and Lovell, 1978). Although

constant returns to scale have been assumed in the explanation above, Forsund and Hjalmarsson (1979) generalises this for variable returns to scale.

It is important to point out that the concept of economic efficiency is very wide-ranging. For example, we could also define economic efficiency in terms of the objectives of the company or sector under study. If the objective is to maximise profits, the optimum is associated with a profit function, given the prices of outputs and inputs. If the objective is to minimise costs, the optimum is associated with the cost function, given the prices of inputs and the quantity of outputs. Economic efficiency may also be dealt with from other perspectives. For instance, when the price of production is not equal to the marginal cost of production, there will be economic inefficiency. Here we are referring to deviations with respect to the behaviour of a competitive industry, which generates inefficient allocations of resources in the long term.

6.3 Classification of Efficiency Frontiers

Measurements of companies' efficiency are obtained by comparing the observed values for each productive unit with the optimum defined by the estimated frontier.

The studies on functions of production frontiers (or those of costs or profits) can be classified in the following way:

i) Deterministic studies: the error term is composed of a single one-sided term, which captures inefficiency.
ii) Stochastic studies: the error term is composed of two parts - a component to capture "noise" and a one-sided component which captures inefficiency.

These give rise to two types of production frontier: deterministic or stochastic.

Either of these two production frontiers can be estimated using specifications that may be:

i) Parametric specifications: a concrete parametric function is specified (Cobb-Douglas, CES, Translog...)
ii) Non-parametric specifications: no function is specified.

In addition, to calculate efficiency measurements one can:

i) Use methods of mathematical programming to construct the frontier.
ii) Specify a statistical relationship to estimate the frontier.

From here eight possible combinations emerge in order to carry out empirical investigations. In the different empirical investigations stochastic frontiers are usually used with parametric specifications (for example, translog functions) and deterministic frontiers with non-parametric specifications (for example, Data Envelopment Analysis).

If we use *deterministic frontiers* (see for example Aigner and Chu, 1968), all the observed cases below the production frontier represent inefficiency. Then, if

we have cross-sectional data with N inputs used to produce a single output, it is possible to define the production function as:[1]

$$y_i = f(x_i)\exp(-u_i)$$

where y_i is the output of producer i, x_i is the input vector of N inputs used by producer i, $f(x)$ is the production frontier and $u_i \geq 0$ ($0 \leq e^{-u_i} \leq 1$). It is assumed that the observations on u_i are independently and identically distributed (iid) and that x is exogenous (independent of u). The technical efficiency index $TE_i = \exp(-u_i)$ will be:

$$TE_i = \frac{y_i}{f(x_i)}$$

In words, the technical efficiency index is the ratio of observed output to maximum output. However, the main disadvantage of this definition is that it attributes all of the deviation from the frontier to technical inefficiency.

Stochastic frontiers, proposed by Aigner, Lovell, Schmidt (1977) and Meusen and van den Broeck (1977), are also called frontiers of composed errors because to every frontier we add an error which is composed of two components: a random error, and another error which captures the degree of inefficiency in the company. That is:

$$y_i = f(x_i) \exp (v_i - u_i)$$

where v_i is a random error with some symmetric distribution to capture the random effects of measurement error and exogenous shocks, and $\{f(x_i) \exp (v_i) \}$ is the stochastic production frontier. In this way, technical inefficiency is captured by the one-sided error component $\exp (-u_i)$, $u_i \geq 0$. Hence, the technical efficiency index is defined as:

$$TE_i = \frac{y_i}{f(x_i)\exp(v_i)}$$

However, while this model is capable of revealing the average efficiency of the sample, it is unable to yield estimates of technical efficiecny for each observation.

A solution, proposed by Jondrow, Lovell, Materov and Schmidt (1982), was to specify the functional form of the distribution of the u_i component and to derive the conditional distribution $(u_i|v_i+u_i)$.

At first, empirical studies using frontier models used cross-sectional data. Schmidt and Sickles (1984) noted three difficulties with cross-sectional stochastic production frontier models:

i) It is necessary to make assumptions about the distribution of technical inefficiency (e.g., half-normal) and statistical noise (e.g., normal) in order to be able to decompose the error term. It is not clear how robust one's results are to these assumptions.
ii) It is necessary to assume that the degree of inefficiency is independent of the regressors, although it is easy to imagine that if the firm knows the technical inefficiency level, this should affect its input choices.

[1] Given the objectives of this study, the analysis will follow a production function focus.

iii) Technical inefficiency of a particular firm (observation) could be estimated, but not consistently.

Schmidt and Sickles (1984) formulated and estimated some methods with panel data that would overcome these disadvantages. If each producer is observed over a period of time, we can apply panel data techniques. A production frontier with time-invariant technical efficiency can be written as:

$$y_{it} = f(x_{it}) \exp (v_{it}-u_i)$$

where there are T time periods. We can then consider different panel-data models. The *fixed-effects model* treats the technical inefficiency component (u_i) as fixed effects, that is to say, as non-random. Thus, it is possible to allow for technical inefficiency to be correlated with regressors. However, if we assume that inefficiency indices are randomly distributed with a constant mean, we have the *random-effects model.* In this model it is necessary to assume that the technical inefficiency is uncorrelated with the regressors. To test this assumption, Hausman and Taylor (1981) created a test between individual effects (technical inefficiency) and explanatory variables (inputs vector in the case of the production function).

The development of the analysis of econometric models from the 1980s has led the way to applications to frontiers with panel data. The economic literature in the 1990s up to the present has gone in at least two directions: a) The first comes from the increasing use of data panels with a dynamic character. In models of efficiency there is an attempt to estimate indices of efficiency which vary with time. Studies by Cornwell, Schmidt and Sickles (1990), Kumbhakar (1990) and and Battesse and Coelli (1992) constitute and good example of the literature in this vein, b) The second direction has an empirical character. There have been efforts to compare research done on the same sector, using the same methods and concepts of efficiency, in different countries or groups of countries, and the most relevant conclusions are drawn. An excellent survey of stochastic frontier analysis since its beginnings was given by Kumbhakar and Lovell (2000).

6.4
A Theoretical Application to Goods Haulage Companies on Spanish Roads Using Panel Data

In this section, we calculate technical efficiency for goods haulage firms on Spanish Roads, considering six sectors (Full Load Transport, Grouppage-Service, Refrigerated Transport, Crane Transport, Special Transport and International Transport) over the period 1994-1997. We have chosen a translog functional form to estimate the production function. Thus, the econometric specification will be:

$$\text{Log GVA} = \beta_0 + \beta_1 \text{LogCI} + \beta_2 \text{LogK} + \beta_3 \text{LogL} + \beta_4 (\text{LogCI})^2 + \beta_5 \text{LogCI LogK} + \beta_6 \text{LogCI LogL} + \beta_7 (\text{LogK})^2 + \beta_8 \text{LogK LogL} + \beta_9 (\text{LogL})^2 + \beta_{10} \text{T} + \beta_{11} (\text{T})^2 + \beta_{12} \text{LogCI T} + \beta_{13} \text{LogK T} + \beta_{14} \text{LogL T}$$

where GVA is the output, measured as the Gross Value Added of the different companies calculated as the sum of the Gross Operating Surplus (GOS) and the compensation to employees, where the GOS is calculated as the difference

between the income and expenditures, excluding consumption of fixed capital; CI refers to the intermediate consumption, which in stock accounting is called provisions; K represents capital and is approximated by fixed assets dedicated to production; and L represents the number of employees.
All these variables except the variable L appear in millions of pesetas and are considered in constant pesetas. To achieve this, the variables VAB and K have been deflated by the Consumer Price Index (CPI), and the variable CI has been deflated by the Price of Energy Index to take into account that a very significant part of such consumption refers to the fuel of transport vehicles. In addition, all the variables are expressed in the form of deviations from their mean values, so that (at the mean) the first-order coefficients of the production function indicate how production varies with respect to the inputs.

In all the estimations a fixed-effects model was chosen, that is, a separate dummy is estimated for every firm and the constant term is suppressed. These effects will be interpreted as the indices of technical inefficiency for each firm (Schmidt and Sickles, 1984). According to this method, the most technically efficient firm takes the value one. To correct for endogeneity of inputs we have used an instrumental variables approach, using their lagged values as instruments. Though the generic expression of the production function takes into account a temporal trend to measure technical progress, it was finally decided not to include it since it either did not prove significant or, as in the case of international transport, it showed a coefficient of around 8%, which was not very credible.

In Tables 6.1 to 6.6, we summarise the results of the estimations. The estimations are quite satisfactory for most of the distinct sectors of haulage by road. In Table 6.7 the scale parameter and the technical efficiency parameter is measured for each sector. As can be observed, in haulage by road there is a sector such as Groupage-Service with increasing returns to scale, since the scale parameter is significantly higher than 1. In the other sectors there are decreasing returns to scale with an average scale parameter of 0.81, as in the case of Crane Transport and International Transport.

In Figures 6.2 to 6.7 the mean of the level of technical (in)efficiency and other statistics are summarised for each sector. The arithmetic mean of the indicators of technical efficiency relative to the sectoral mean is 0.80. The economic interpretation of this indicator is the following: in the haulage by road sectors, on average, the production could have been increased by 20% during the years 1994-1997 without increasing the quantity of inputs.

However, the average levels and the dispersion of the (in)efficiency show a great deal of sectoral heterogeneity. The average level varies from 92% in the sectors of Full Load Transport and Crane Transport to 44% in the sector of Groupage-Service. The sector in which the indicator of technical (in)efficiency presents the highest dispersion with respect to the average sectoral value is Groupage-Service. This sector also has an average level of efficiency that is quite low. The sectors with less dispersion in the grade of inefficiency of companies are International Transport, Full Load Transport and Crane Transport, which are mostly sectors with a high level of efficiency.

In developing this application we wanted to illustrate one of the many methods which have been developed and which are available for the empirical study of the relative technical efficiency of companies using the methodology of efficiency

frontiers. As has been seen, the topic of relative efficiency is a fundamental part of the Microeconomics of the Theory of Production and it is a subject whose contribution is important in terms of its applications. There remains another important economic topic, which is the analysis of the causes of the differences between efficiencies across companies of the same sector. This is needed in order to identify allocation problems and to guide sectoral policies. This is important because, as is obvious, inefficiency is not only costly for the productive unit in question, but for the economy as a whole.

Table 6.1. Full load transport

	Coefficient	t-Student
β_1	0.3364	4.3769
β_2	0.7789	6.9367
β_3	0.0761	2.3175
β_4	0.5735	2.4592
β_5	-2.4978	-2.4874
β_6	0.3555	1.3071
β_7	1.2029	1.9595
β_8	0.6404	1.4513
β_9	-0.1562	-1.4568

Table 6.2. Grouppage-service

	Coefficient	t-Student
β1	0.3854	6.7586
β2	0.8036	10.744
β3	0.0527	1.2349
β4	0.3198	2.9957
β5	0.2315	0.7376
β_6	-0.7004	-3.6098
β_7	0.2552	0.5308
β_8	0.2529	0.5032
β_9	-0.5415	-5.7813

Table 6.3. Refrigerated transport

	Coefficient	t-Student
β_1	0.1771	4.1177
β_2	0.1993	3.3153
β_3	0.7388	6.6673
β_4	0.2408	2.7204
β_5	-0.0363	-0.1296
β_6	-0.8211	-2.1997
β_7	-0.2916	-1.9447
β_8	0.4617	0.9723

Table 6.3. Continued

	Coefficient	t-Student
β_9	0.0022	0.0059

Table 6.4. Crane transport

	Coefficient	t-Student
β_1	0.0635	1.6573
β_2	-0.0367	-1.3134
β_3	0.8261	10.678
β_4	0.0432	1.6237
β_5	0.0293	0.4857
β_6	-1.0788	-3.1990
β_7	0.0988	2.4034
β_8	-0.4035	-2.0788
β_9	1.8919	4.9323

Table 6.5. Special transport

	Coefficient	t-Student
β_1	0.2041	2.2994
β_2	0.0079	0.2239
β_3	0.5063	3.7002
β_4	0.3246	2.2737
β_5	0.4331	0.7951
β_6	-2.0812	-3.4231
β_7	0.1933	1.6674
β_8	1.2879	2.1186
β_9	-1.0610	-1.8020

Table 6.6. International transport

	Coefficient	t-Student
β_1	0.2336	5.9920
β_2	0.0865	3.7535
β_3	0.4440	10.449
β_4	0.2874	1.2973
β_5	0.2846	1.3142
β_6	-1.1295	-2.6446
β_7	-0.2230	-3.6819
β_8	-0.0015	-0.0064
β_9	0.9369	2.5462

Table 6.7. Summary of results

Sector	Scale Parameter	χ_1^2	Average Technical Efficiency
Full Load Transport	1.19	2.75	0.92
Groupage-Service	1.24	8.92*	0.44
Refrig Transport	1.11	0.98	0.80
Crane Transport	0.85	3.85*	0.92
Special Transport	0.72	3.21	0.80
International Transport	0.76	25.44*	0.89

*Statistically significant different from one at 5% level.

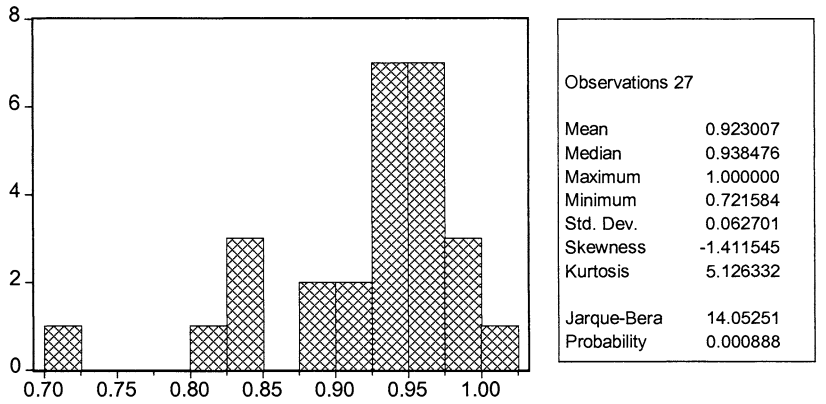


Fig. 6.2. Technical efficiency. Full load transport

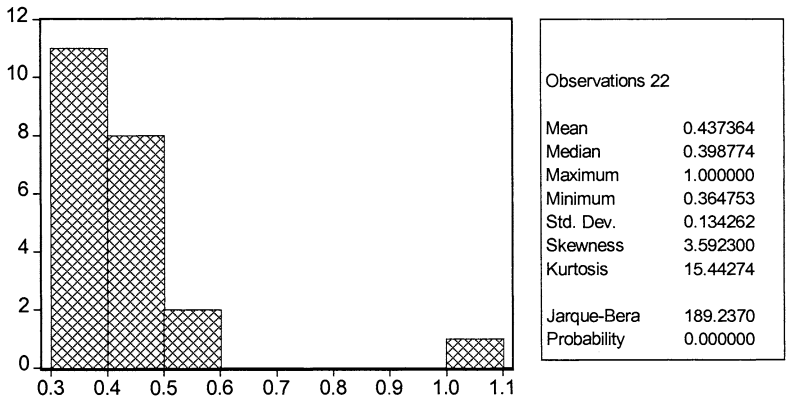


Fig. 6.3. Technical efficiency. Groupage-service

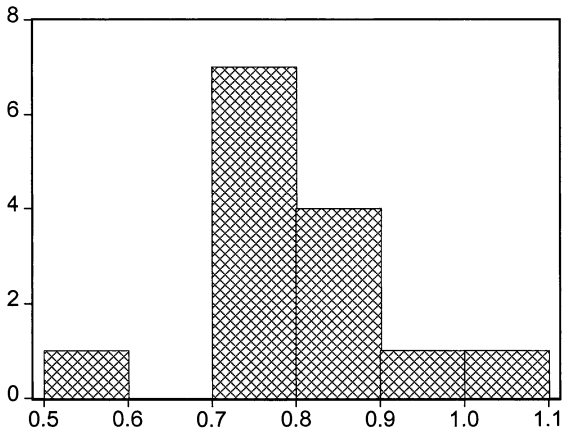


Fig. 6.4. Technical efficiency. Refrigerated transport

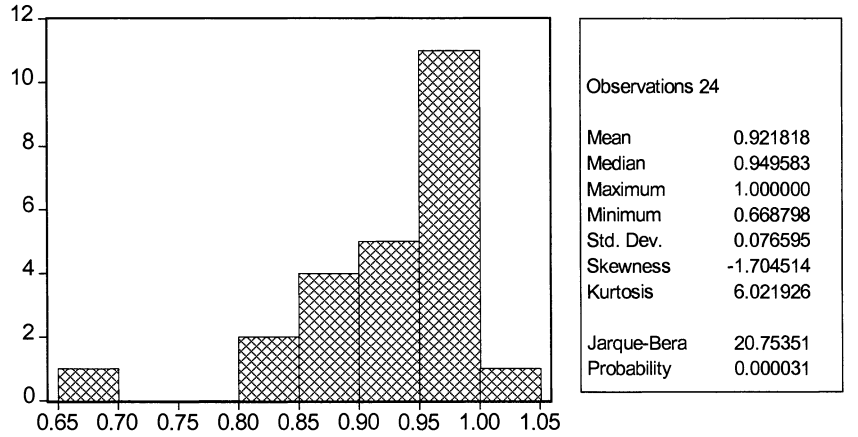


Fig. 6.5. Crane transport

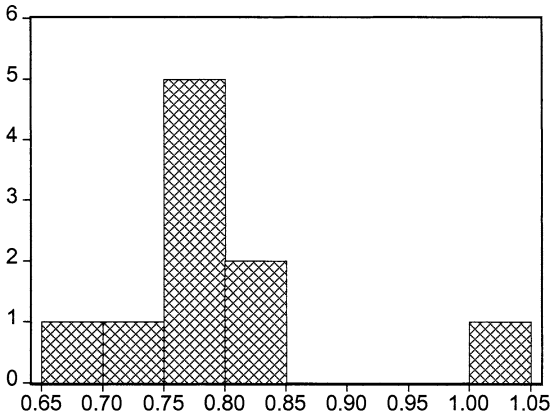


Fig. 6.6. Technical efficiency. Special transport

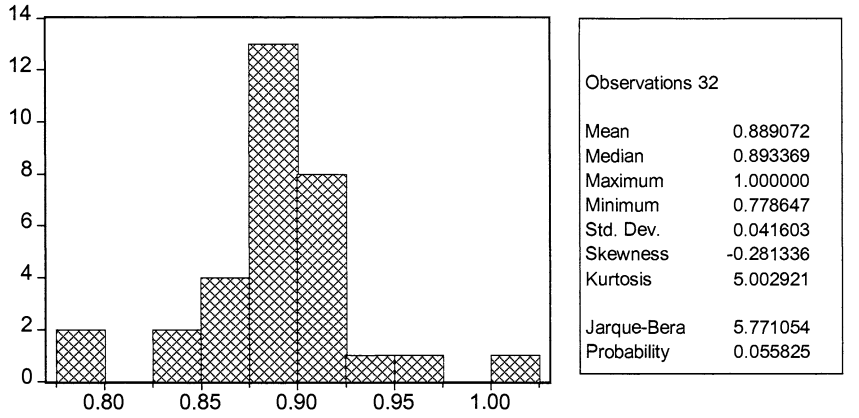


Fig. 6.7. Technical efficiency. International transport

References

Aigner, D. J. and Chu, S. F.: On Estimating the Industry Production Function. American Economic Review, 58, 826-839 (1968)

Aigner, D. J., Lovell, C. A. K. and Schmidt, P.: Formulation and Estimation of Stochastic Frontier Production Function Models. Journal of Econometrics 6, 21-37 (1977)

Battesse, G. and Coelli, T.: Frontier Production Functions, Technical Efficiency and Panel Data: with Application to Paddy Farmers in India. The Journal of Productivity Analysis 3, 153-169 (1992)

Cornwell, C., Schmidt, P. and Sickles, R. C.: Production Frontiers with Cross-Sectional and Times –Series Variations in Efficiency Levels. Journal of Econometrics 46,182-200 (1990)

Färe, R. and Lovell, C. A. K.: Measuring the Technical Efficiency of Production, Journal of Economic Theory, 19, 150-162 (1978)

Farrell, M. J.: The Measurement of Productive Efficiency. Journal of the Royal Statistic Society Series A 120, 253-281 (1957)

Forsund, F. R. and Hjalmarsson, L.: Generalized Farrell Measurement of Efficiency: An Application to Milk Processing in Swedish Dairy Plants. Economic Journal 89, 294-315 (1979)

Hausman, J. A. and Taylor, W. E.: Panel Data and Unobservable Individual Effects. Econometrica 49, 1377-1398 (1981)

Jondow, J., Lovell, C. A. K., Materov, I. S. and Schmidt, P.: On the Estimation of Technical Inefficiency in the Stochastic Frontier Production Function Model. Journal of Econometrics 19, 233-238 (1982)

Kumbhakar, S. C.: Production Frontiers, Panel Data and Time-Varing Technical Inefficiency. Journal of Econometrics 46, 201-211 (1990)

Kumbhakar, S. C. and Lovell, C. A. K.: Stochastic Frontier Analysis. Cambridge University Press (2000)

Meusen, W. and van den Broeck, J.: Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error. International Economic Review 18, 435-444 (1977)

Schmidt, P. and Sickles, R. C.: Production Frontiers and Panel Data. Journal of Business & Economic Statistics vol. 2, 4, 367-376 (1984)

7 Technical Efficiency and Liberalisation within International Air Transport (1992-2000)

P. Coto-Millán
University of Cantabria (Spain)

V. Inglada-López de Sabando
University Carlos III of Madrid (Spain)

B. Rey
University Complutense of Madrid (Spain)

A. Rodríguez-Álvarez
University of Oviedo (Spain)

7.1 Introduction

The main aim of this article is to analyse the technical efficiency of different international airline companies during the period 1992-2000. For a number of reasons this analysis is of particular value. Firstly, dealing with the degree of market competition, the main studies into the question of efficiency -Forsyth, Hill and Trengove (1986), Encaoua (1991)- use a period of analysis during which air transport was tightly regulated in many countries and the majority of national airline companies were under public ownership. More recent studies, such as those published by Good, Lars-Hendrik and Sickles (1999) or by Inglada, Coto-Millán and Rodríguez-Álvarez (1999), analyse efficiency by comparing regulated periods with other periods of partial liberalisation. In both studies, improvements in efficiency are documented as the liberalisation process advances. Nevertheless,

P. Coto-Millán (ed.), *Essays on Microeconomics and Industrial Organisation*

there are no studies available regarding the efficiency of air transport companies that were operational during a complete period of liberalisation, through which we might discover whether there are trends towards increases or reductions in efficiency.

In second place, when focussing on the ownership of the companies, most of the authors agree that the organisational system of large companies under public ownership diminishes incentives to increase efficiency. Economic models exist that consider the lack of inter-company competition to be one of the principal determinants for inefficiency and lack of productivity, since there are no incentives for them to behave efficiently (see, for example, Leibenstein (1966) or De Alessi (1983)). Indeed, at the international level, since the mid-1980s the trend has been towards the privatisation of big companies. Pioneers in this process, for instance, were British Airways, Singapur Airlines (SIA) and Japan Airlines (JAL). More recently, in Spain we have witnessed the privatisation of Iberia. In this article we intend to make a comparative behaviour study of the efficiency of these 19 airline companies, most of which are in private ownership.

In the light of all these changes that have taken place in the air transport industry, it seems logical to pose the question of how this liberalisation process has affected company efficiency. Consequently, this article undertakes an analysis of technical efficiency through the estimation of a production function for a data panel charting the development of 19 airline companies in the period 1992-2000. The methodology proposed is that of Battese and Coelli (1995), which makes it possible to make a model of the technical inefficiency of companies on the basis of a set of explanatory variables that vary over time, as well as to differentiate the components of that inefficiency from other factors that are purely random and beyond the control of airline companies.

The outline of the paper is as follows: in the first instance, we present a résumé of the most pertinent institutional changes in the air transport industry; we then go on to review the methodology we used to make the estimations, based on Battese and Coelli (1995); thirdly, a description of the data and variables is provided; in fourth place, the economic specification and functional form chosen are introduced and, lastly, we provide the results of the estimation and the conclusions drawn from them.

Main policy changes in international civil aviation

The international air transport industry, in the United States and Europe, to be specific, has been undergoing profound institutional changes since the beginning of the 1980s. The Asian countries, on the other hand, recently initiated moves to liberalise their markets and are close to an "open skies" policy. Although it is beyond the reach of this article to provide a detailed background description, it is nevertheless worthwhile running through the main transformations this industry has undergone, concentrating on the differences between the European countries and their American competitors.

From the early 1950s, the legal structure of civil aviation policy had been based on complex systems of bilateral agreements between countries. These varied from deals that introduced great flexibility (in relation to capacity and fares) to others of a far more restrictive nature, based on the designation of companies and as a consequence, on pre-established capacity and fares. The USA was at the head of the field in its efforts to adopt competitive bilateral agreements and this took on

concrete form in the International Air Transport Competition Act in 1979 (IATCA). In it, the aims of the United States were defined as the striking of bilateral agreements allowing for the multiple designation of companies, free access of charter companies, the elimination of restrictions on capacity and fares and single treatment regarding the access of national and foreign companies to airport services. So from that moment on the American air transport industry went into full liberalisation, although, dating from the early 1970s, serious efforts had already been taken by the administration to limit restrictions, resulting in important gains in efficiency in comparison with previous periods (Sickles, Good and Johnson (1986)).

In contrast, the European countries maintained a more protectionist approach until the decade after. Suffice it to say Article 84 of the Treaty of Rome excluded both air and maritime transport from the obligation of having to observe the general rules of policy concerning competition. On the other hand, most national companies in Europe were under public ownership; this form of ownership reduces the efficiency of airline companies (De Alessi (1983)) and, furthermore, encourages the subsidising of companies when they suffer losses.

The first effect in Europe of the application of the IATCA was the unilateral agreement signed between the United States and Holland, allowing each company to freely determine the capacity of their aeroplanes without any government intervention.

The changes in Europe, however, were slow during the 1980s. Great Britain and Holland were the first two countries to attempt to end restrictions on competition, negotiating a bilateral agreement in 1984 whereby any change that affected capacity or fares would be considered valid if neither of the countries expressed disagreement (principle of disapproval). But it was not until December 31, 1987, that the Official Journal of the European Communities published four Council provisions (comprising two Regulations concerning the defense of competition, a Directive on fares and a Decision on bilateral air transport agreements) that were to constitute the first package of measures in what was known as "EEC air transport deregulation". These provisions were later brought to completion with Commission Regulations governing the defense of competition in the field of air transport.[1].

Subsequent to that date, two more packages of liberalising measures were brought into force. To be specific, the period of 1993-1997 coincided with the application in Europe of the third package of air transport liberalisation measures, which was brought to completion in 1997. From that moment onwards, any company from any country in the European community was at liberty to run passenger transport services between any two destinations within the Community, with absolute freedom to establish their own air fares.

If we compare Asian countries with the USA and Europe, we can see that the markets are still very fragmented and Asian companies are currently still subject to very restrictive bilateral agreements (Oum and Lee (2002)). However, the threat posed by the creation of international alliances between American and European companies, is leading these countries towards liberalising their air transport. One of the first measures was the introduction of second tier airlines, such as Silk Air,

[1] A detailed description of the air transport liberalisation process may be found in Rey (2000).

Eva Airways, Japan Asia Airways, All Nippon Airways, Asiana, Sempati and Dragon Air which have increased market competitiveness, against the traditional national companies that had been operating as big monopolies. In this fashion, there has been a gradual introduction of a "limited or partial open skies" policy, begun by Thailand, the Philippines and Indonesia. And, parallel to this, the national companies of these countries have been privatised.

In 1991 Korea signed a revised version of its bilateral agreement with the USA, described as "open carrier designation, open route and double disapproval pricing", making it one of the most favourable bilateral deals struck in terms of the stimulus it provided to competition.

The bilateral deal signed between Japan and the United States in March 1998, as well as the agreement the USA and China came to in April 1999, bears the same liberalising hallmark. These constitute symptoms that indicate the advance of the liberalisation process. Nevertheless, the pressure exercised by the USA and Europe in their drive to go beyond the "fifth liberty" in the region is still encountering stumbling blocks in many Asian countries.

7.2 Methodology

To use Koopmans' (1951) definition, a company is technically efficient if it reaches the maximum output possible for the inputs and technology used. Farrell's pioneering work (1957) developed a methodology designed to measure the concept of relative efficiency. This methodology makes it possible to "construct" an efficient frontier with the best productive units observed, where the (in)efficiency of a unit is defined as an index that measures the distance between itself and that frontier. This pioneering contribution gave rise to a wealth of written literature devoted to the task of discovering the best way to construct an efficient frontier (employing both mathematical and statistical methods). The first contribution in the parametric estimation[2] of frontier functions came from Aigner and Chu (1968), who obtained a deterministic production frontier using mathematical programming methods as their starting point. Afriat (1972) and Richmond (1974) proposed their estimation using parametric techniques, involving the possibility of bringing statistical inference to bear on the results obtained.

The main drawback of the deterministic frontier is that any deviation from it is deemed to be inefficiency and, therefore, there is no consideration of the possibility that productivity may be affected by random exogenous shocks that, along with possible errors of measurement of variables or problems of poor model specification, make it inadvisable to interpret the term of error as an isolated measurement of inefficiency. From quite another angle, the parametric and stochastic frontier approach (Aigner, Lovell and Schmidt (1977) and Meeusen and Van den Broeck (1977)) starts out from the assumption that the deviation between the observed output level and the possible maximum is comprised of two

[2] Alternatively, non-parametric frontiers can be calculated via techniques for analysing data envelopment (DEA) and mathematical programming.

components: a symmetric term of error to capture the effect of variables that are not under the control of the productive unit under analysis, errors of measurement in variables and other statistic noise; and a second term that is supposed to capture the degree of inefficiency, positioning the production level below the frontier's maximum output, making it necessary, in consequence, to specify an asymmetric distribution for this second term of error. Schmidt and Sickles (1984) develop a production frontier estimation using panel data. Kumbhakar, Ghosh and McGukin (1991) and Reifschneider and Stevenson (1991) establish a stochastic frontier model that sets out to explain inefficiency through the use of a specific variables vector for different companies and a random error through a second stage analysis. Later, Battese and Coelli (1995) developed an alternative model to get round the inconsistencies in the second stage analysis. [3]

In this research, the Battese and Coelli model (1995) has been specified, as it provides the possibility of analysing the determinants of the development of technical inefficiency of a productive unit in terms of a set of explanatory variables that, in addition, may vary in time. [4] To be specific, delimiting the analysis to the estimation of a production frontier using a data panel, the model can be expressed in the following form:

$$Y_{it} = X_{it}\beta + (V_{it} - U_{it}); \; i=1,...,N; \; t=1,...,T. \tag{7.1}$$

where:

Y_i is the production of the i-th firm;

X_{it} is a $k\times1$ vector of input quantities of the i-th firm in the t-th time period;

β is a vector of unknown parameters;

the V_{it} are random variables which are assumed to be iid. $N(0,\sigma_V^2)$, and independent of the

U_{it} which are non-negative random variables which are assumed to account for technical inefficiency in production and are assumed to be independently distributed as truncations at zero of the $N(m_{it},\sigma_U^2)$ distribution; where:

$$m_{it} = z_{it}\delta, \tag{7.2}$$

where z_{it} is a $p\times1$ vector of variables which may influence the efficiency of a firm; and δ is an $1\times p$ vector of parameters to be estimated.

So, the special contribution of this methodological approach lies in the fact that, without resorting to a second stage analysis, it enables us to specify the time trajectory of the technical inefficiency of a company in relation to a set of explanatory variables that may change over time.

---

[3] For a review of this literature, see Kumbhakar and Lovell (2000).

[4] For possible regressors in the equation of inefficiency, Battese y Coelli (1995) propose both the explanatory variables of the production function and any variable able to determine changes in inefficiency.

7.3 Description of the Data and the Variables Used

The database used in this article is constituted by a data panel formed by 19 international airline companies, of which 6 are European (Lufthansa, SAS, Finnair, Spannair, Iberia, British Airways), 6 North American (American Airlines, United, Delta, Northwest, USAir, Continental) 1 Canadian (Canadian), 2 Mexican (Aeroméxico and Mexicana) and 4 Asian (JAL, Korean Air, Cathay Pacific and SIA). They are all large-scale companies operating international flights. The observations are annual and cover the period 1992-2000. The basic reference is provided by ICAO statistics (Digest of Statistics from the International Civil Aviation Organisation) and World Air Transport Statistics, published by the International Air Transport Association (IATA). Using this source of information, the database was constructed with the aim of estimating a production function for the air transport industry, utilising the methodology described in the previous section. The production function to be considered is as follows:

$$Y = F(L, K, N) \tag{7.3}$$

where: Y is the production that has been calculated for the number of tons-kilometre available[5]; L represents the total number of workers in the airline company; K represents the capital estimated for the aeroplane capacity available (expressed as tons available per plane); and, finally, N indicates the average distance covered in a flight stage (total kilometres flown / number of departures) and gives us an approximation of a measurement of the routes that make up the network of each airline company. Table 7.1 shows a descriptive analysis of the data.

Table 7.1. Descriptive data analysis

Variable	Average	Typical Deviation	Minimum	Maximum
Production (Y)	14,297,400	10,623,000	757,895	40,237,300
Labor (L)	34,109.21	25,613.07	1,837	87,586.56
Capital (K)	6,639.49	4,699.45	693.799	17,290.34
Average Distance (N)	1,607.36	799.55	702.15	4,764.49
Charter (%)	3.59	10.53	0.01	69.86

7.4 Econometric Specification

The functional form of the production frontier adopted is a transcendental logarithmic production function. The choice was based on the flexibility this function displays in adapting to all kinds of productive technology without it being necessary to impose a priori restrictions on scale capacity (see, for instance, Oum

[5] The reason for using tons-kilometre available as a measure of output is that this combines passengers-kilometre and cargo/tons-kilometre.

and Yu (1996)). So the function representing the production of international airline companies is represented as:

$$\ln Y_{it} = \beta_0 + \sum_{j=1}^{3} \beta_j \ln(X_{jit}) + \sum_{j=1}^{3}\sum_{h=1}^{3} \beta_{jh} \ln(X_{jit}) \ln(X_{hit}) + \\ + \sum_{t=1}^{T} \gamma_T D_t + v_{it} - u_{it} \tag{7.4}$$

i = 1, ..., 19 countries; t = 1, ..., 9 years

where Y_{it} is output and X_{it} is a vector referring to the inputs under consideration. Technical progress is incorporated via time dummies (D_t) that register the effect of variables and, varying over time, have an equal influence on all the companies. Lastly, v_{it} is the random error and u_{it} represents the term of inefficiency and they are distributed following the assumptions defined in the previous section. Component u_{it} is defined by means of the equation:

$$u_{it} = \delta_0 + \delta_1 (Trend_{it}) + \delta_2 (Trend^2{}_{it}) + W_{it} \tag{7.5}$$

i = 1, ..., 19 countries; t = 1, ..., 9 years

where W_{it} is a random error.

In equation (7.5), factors have been included that are assumed to have the capacity to influence technical inefficiency. In the first place, we introduced a trend which registers the effects of the time period used and which, as we mentioned above, corresponds in time with the total liberalisation of air transport in the countries selected. In addition, with the aim of lending flexibility to the effect of time on inefficiency, the variable trend was also included in the table.

7.5 Results of the Estimation

Equations (7.4)-(7.5) were simultaneously estimated for Maximum Likelihood, using the Frontier 4.1. program (Coelli (1996)). The variables were taken on deviations from their averages (approximate functional form). Thus, the estimated translog function is a second order approximation to the real function in the data average and the first order coefficients are the production elasticities of each input, evaluated on the sample average.

The results of the estimation are recorded in table 7.2. The first order coefficients of the inputs bear the expected positive sign and are significant. The time dummy variables introduced into the model, as already explained, aim to register the effect of the technical change that has taken place in the air transport sector.

Table 7.2. Estimation of the production function (Battese and Coelli, 1995)

Variable	Coefficient	t-Statistic
Production Function (equation 4)		
Constant	16.2507	392.7113
L(L)	0.0811	2.5937
L(K)	0.9172	32.0844
L(N)	0.4259	19.0189
L(L) L(L)	-0.1335	-1.3777
L(K) L(K)	0.4362	3.7914
L(N) L(N)	-0.2383	-3.1769
L(L) L(K)	-0.0919	-0.8810
L(L) L(N)	-0.1725	-2.3922
L(K) L(N)	-0.0445	-0.7103
D93	-0.0267	-0.7170
D94	-0.0688	-1.7424
D95	-0.1092	-2.6007
D96	-0.1242	-2.9545
D97	-0.1106	-2.4538
D98	-0.0492	-1.2264
D99	-0.0814	-1.8071
D00	-0.0398	-0.8571
Model of Inefficiency (equation 7.5)		
Constant	0.3966	5.1966
T	-0.1520	-3.9147
T^2	0.0128	3.3112
σ^2	0.0173	7.1838
γ	0.9931	50.0521
Log. F. Likelihood	168.75	

Using expression (7.6) it is possible to understand how these time effects have affected the production frontier from one period to another:

$$TC_{T+1,T} = \gamma_{T+1} - \gamma_T \tag{7.6}$$

So, a positive TC value would imply technical progress, that is to say, an upward shift of the production frontier, making it possible to obtain a higher output quantity without incurring a greater consumption of resources, and vice versa. The values obtained from eq. (7.6) are presented in table 7.3 and figure 7.1.

Table 7.3. Technological change

Period	TC
92-93	-0.067
93-94	-0.042
94-95	-0.040
95-96	-0.015
96-97	0.013
97-98	0.061
98-99	-0.032
99-00	0.041

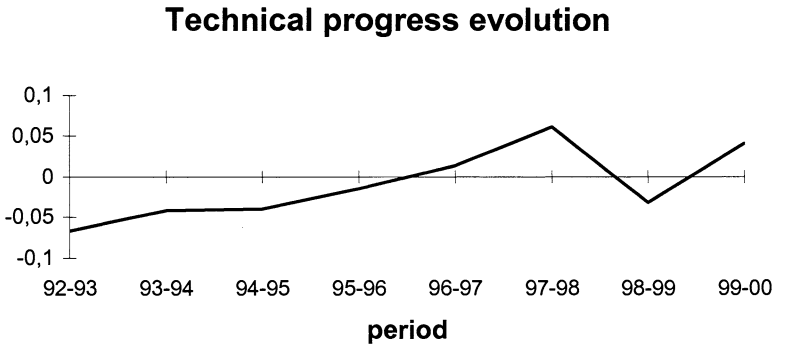


Fig. 7.1. Development of technological change for the period of 1992-2000

The coefficients obtained reveal that, during a first period (1992-1996), the passing of time had a negative influence on air company technology, even though development was favourable. This improvement was particularly evident after the period of 1997-1996 where the coefficients indicated discrete technical progress (with the exception of the period 1999-1998). These results bear out our initial hypothesis that a direct relationship exists between air market liberalisation and the productive improvement registered in the companies that make up the sector.

Moreover, the interpretation of the coefficients obtained for the inefficiency model is particularly pertinent to the objectives of this study. The coefficient linked with the time variable is negative and statistically significant, suggesting that air company efficiency improved during the time period under consideration and that this coincided, as previously suggested, with the process of deregulation in the sector. The squared coefficient of this variable is also significant and bears an opposite sign, indicating the suitability of the model proposed, since it makes it possible to introduce flexibility into the effect of time on inefficiency.

The estimated value of parameter γ indicates that the proportion of variance of u_{it} on total compound error is 99.3% and serves to warn us that it is a mistake to use average production functions in which differences in inefficiency are not taken

into account. Furthermore, in order to justify the methodology adopted here, a series of specification contrasts were carried out, based on the ratio of likelihood. Firstly, we tested out the suitability of the functional translog form chosen. In this first contrast (see table 7.4), there is a rejection of the null hypothesis that the Cobb-Douglas functional form is preferable to the translog function. We then went on to contrast and reject the hypothesis of non-existence of technical inefficiency in the term of error. The third null hypothesis, considering inefficiency not to be a function of the regressors used, was also rejected.

Table 7.4. Specification contrasts

Null Hypothesis	Likelihood Ratio	Number of Restrictions	Decision (95%)
$H_0 : \beta_{jh} = 0$	**45.54**	**6**	**REJECTION**
$H_0 : \gamma = \delta_0 = \delta_1 = \delta_2 = 0$	**27.49**	**4**	**REJECTION**
$H_0 : \delta_1 = \delta_2 = 0$	**9.28**	**2**	**REJECTION**

Finally, the estimation made allows us to calculate the indices of technical efficiency for each company and year via the following expression:

$$ET_{it} = \exp(-u_{it}) = \exp\left[-\left(\delta_0 + \delta_1(trend_{it}) + \delta_2(trend^2_{it} / Y_{it})\right) - W_{it}\right] \quad (7.7)$$

In this way, technical efficiency is calculated as the ratio of the production level obtained in relation to the achievable maximum (that is, when $u_{it} = 0$) given the amount of inputs and the technology. Its value fluctuates between 0 and 1, the latter being the most favourable case. A more detailed analysis of these indicators is found in figure 7.2 and tables 7.5 and 7.6. In figure 7.2 we see the favourable development of the indices of technical efficiency within the time period analysed and observe that they stabilise at the end of the period in values close to the level of efficiency. Tables 7.5 and 7.6 show the average indices of technical efficiency for each company, classified, respectively, by area. It can be seen in table 7.5 that the German company (LUFTHANSA) reaches the highest level of efficiency (0.952), closely followed by the American company DELTA, both with an efficiency index of 0.839. IBERIA (0.767) and USAIR are among the most inefficient companies (0.792). By area, the results displayed in table 7.6 show that the Asian and North American companies are the most efficient, followed by their European counterparts. One possible explanation of these results, in the case of the Asian companies in particular, would be found in the growing competition introduced into these markets from the end of the 1980s onwards. That was when the monopoly of the national companies was broken and the market was opened up to tier companies, whose interests have had a striking impact on air transport policy in Asia (Knibb (1993)). Also, Hooper (1996) describes how a number of competing companies sprang up in various Asian countries and the effect this had on international traffic in Korea, the Philippines, China, India y Japan.

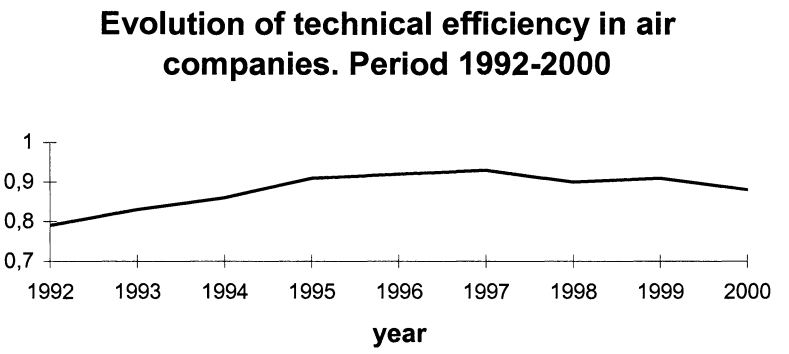


Fig. 7.2. Development of technical efficiency

For these reasons, the Asian air market is powerful and experiencing very rapid growth. But, beyond factors due to the introduction of competition, some Asian companies display features that make them more efficient than American and European companies. If we look at SIA or Cathay Pacific, for instance, both companies have an established reputation for the passenger service they provide, for the efficient collection and delivery of luggage, for keeping their planes clean, and so on.

From a cost perspective, their labour costs are lower, and their aircraft are of the latest generation.

Last, but not at all least, we must mention their computer reservation system. In Asia, Abacus, the joint system established by six of the regional airlines, including Cathay Pacific and SIA, is already up and running in both Hong Kong and Singapore. Of equal significance, Abacus has come to an agreement to establish a joint company with All Nippon in Japan, which will give the partner access to the dominant market in Asia.

The European CRS systems, although established prior to Abacus, are still struggling to create a viable product, and appear to be a long way from having something concrete to offer the market. Again, there is a clear danger that the lead established by Asian airlines in this field might give them a competitive advantage that could ultimately enable them to control the market.

Table 7.5. Indices of technical efficiency

COMPANIES	Average	Minimum	Maximum
LUFTHANSA (GERMANY)	0.9522	0.8728	0.9916
SAS (SACNDINAVIA)	0.8631	0.7305	0.9498
FINNAIR (FINLAND)	0.8022	0.5831	0.9903
SPANNAIR (SPAIN)	0.9183	0.8634	0.9795
IBERIA (SPAIN)	0.7679	0.6863	0.9952
BRITISH AIRWAYS (UK)	0.8873	0.7515	0.9737
AMERICAN (UNITED STATES)	0.8997	0.8177	0.9607
UNITED (UNITED STATES)	0.9309	0.8099	0.9900
DELTA (UNITED STATES)	0.9435	0.8932	0.9856
NORTHWEST (UNITED STATES)	0.8564	0.8009	0.8948
USAIR (UNITED STATES)	0.7928	0.7466	0.8141
CONTINENTAL (UNITED STATES)	0.9309	0.8531	0.9828
JAL (JAPAN)	0.8951	0.7655	0.9908
CANADIAN (CANADA)	0.8623	0.7239	0.9786
AEROMEXICO (MÉXICO)	0.9533	0.8957	0.9930
MEXICANA (MÉXICO)	0.8194	0.8046	0.8395
KOREAN AIR (KOREAN REPUBLIC)	0.8843	0.8034	0.9765
CATHAY PACIFIC (CHINA)	0.8998	0.7183	0.9903
SIA (SINGAPORE)	0.9147	0.8776	0.9719

Table 7.6. Technical efficiency per company

Companies	Average	Minimum	Maximum
Total sample	0.8838	0.5831	0.9952
European Companies	0.8641	0.5831	0.9952
North American Companies	0.8916	0.7239	0.9930
Asian Companies	0.8975	0.7183	0.9908

7.6 Conclusions

It is customary in economic theory to link processes of deregulation and liberalisation with incentives directed at improving the productivity and efficiency of companies. In this context, the objective of our research was to explore the relationship between the liberalisation of the air traffic market that took place in the period under study (1992-2000) and the productive change that was brought

about in companies in the sector. To this end, a study was made of the behaviour of 19 international airline companies (European, Asian and American), utilising the stochastic frontier model designed by Battese y Coelli (1995), which makes it possible to estimate the development of technical progress and technical inefficiency, in addition to introducing variables related to liberalisation as determinants of technical inefficiency.

The results obtained in this investigation bear out the hypothesis we started out from, that a direct relationship exists between market liberalisation of airlines and the productive improvement that has been seen in companies in this sector. Moreover, the time period in which this deregulation process was carried out had a significant influence on improvements in the levels of technical efficiency, American and Asian companies being most efficient of all. Nevertheless, in the case of the latter there were factors that had a positive effect on efficiency, such as the good reputation and quality approach of some Asian companies, coupled with advantages in labour costs and the fact that their SCR is much more competitive than that of the Europeans. Consequently, within the international market, the Asian air market is amongst those with the greatest potential for future growth, while American and European companies are naturally investigating ways to increase their share in these markets.

References

Afriat, S.: Efficiency Estimation of Production Functions. International Economic Review 13 (3), 568-598 (1972)

Aigner, D. and Chu, S.: On Estimating the Industry Production Function. American Economic Review 58, 826-839 (1968)

Aigner, D. J., Lovell, C. A. K. and Schmidt, P.: Formulation and Estimation of Stochastic Frontier Production Models. Journal of Econometrics 6, 21-37 (1977)

Battese, G. E. and Coelli, T. J.: A Model for Technical Inefficiency Effects in a Stochastic Frontier Production Function for Panel Data. Empirical Economics 20, 325-332 (1995)

Coelli, T. J.: A Guide to FRONTIER Version 4.1: A Computer Program for Stochastic Frontier Production and Cost Function Estimation. Department of Econometrics. University of New England. Armidale, Australia 1996

De Alessi, L.: Property Rights, Transaction Costs and X-Efficiency: An Essay in Economic Theory. American Economic Review 73, 64-81 (1983)

Encaoua, D.: Liberalizing European Airlines: Cost and Factor Productivity Evidence. International Journal of Industrial Organization 9, 109-124 (1991)

Farrell, M. J.: The Measurement of Productive Efficiency. Journal of the Royal Statistical Society Series A General 120, 253-281 (1957)

Forsyth. P. J., Hill, R. D. and Trengrove, C. D.: Measuring Airline Efficiency. Fiscal Studies 7, 61-81 (1986)

Good, D., Lars-Hendrik, R. and Sickles, R.: Airline Efficiency Differences Between Europe and the US: Implications for the Pace of EC Integration and Domestic Regulation. European Journal of Operational Research 80, 508-518 (1995)

Hooper, P.: Airline markets in Asia: the domestic/ international regulatory interface. In: Dick, H. (eds): Towards "open skies"; Airline deregulation in Asia-Pacific. Institute of Transport Studies. University of Sidney 1996

Inglada, V., Coto-Millán, P. and Rodríguez-Álvarez, A.: Economic and Technical Efficiency in the World Air Industry. International Journal of Transport Economics, 219-236 (1999)

Knibb, D.: Asia's Little Tigers: an expanding group of regional carriers is taking a greater share of intra-Asia passenger traffic as markets matures. Airline Business, October (1993)

Koopmans, T. C.: An Analysis of Production as an Efficient Combination of Activities. in T. C. Koopmans, ed., Activity Analysis of Production and Allocation. Cowls Commission for Research in economics, Monograph no.13. New York 1951

Kumbhakar, S. C. and Lovell, C. A. K.: Stochastic Frontier Analysis. Cambridge University Press. New York 2000

Kumbhakar, S. C., Ghosh, S. and Mc. Gukin, T.: A Generalized Production Frontier Approach for Estimating Determinants of Inefficiency in U.S. Dairy Farms. Journal of Business and Economic Statistics 9(3), 279-86 (1991)

Leibenstein, H.: Allocative Efficiency vs. X-Efficiency. The American Economic Review 56, 392-415 (1966)

Meeusen, W. and Van den Broeck, J.: Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error. International Economic Review 18, 435-444 (1977)

Oum, T. and Yu, C.: A Productivity Comparison of the World's Major Airlines. Journal of Air Transport Management 2, 181-195 (1996)

Reifschneider, D. and Stevenson, R.: Systematic Departures from the Frontier: A Framework for the Analysis of Firm Inefficiency. International Economic Review 32, 715-723 (1991)

Rey, M. B.: Efectos de la liberalización del transporte aéreo sobre el mercado español de vuelos regulares(1989-1997). Ph. D. Dissertation. University Complutense 2000

Richmond, J.: Estimating the Efficiency of Production. International Economic Review 15 (2), 515-521 (1974)

Sickles, R. C., Good, D. H. and Johnson, R.: Allocative Distortions and the Regulatory Transition of the US Airline Industry. Journal of Econometrics 33, 143-163 (1986)

Smitdt, P. and Sickles, R. C.: Production Frontiers and Panel. Journal of Business and Economic Statistics 2, 367-376 (1984)

8 Technological Innovation and Employment: Intersectoral Appraisals of Structural Change in the Service Economy

D. Díaz-Fuentes
University of Cantabria (Spain)

8.1 Introduction

One of the main structural changes in all OECD countries has been a shift in employment and total output towards services. This trend indicates that services are a dynamic part of the economy and make an increasing contribution to employment and economic growth. Thus it could be argued that the "service or tertiary society" is the stage towards which all industrialised countries are moving.

Although the share of services in the economy has increased, there is a lack of understanding about their industrial performance. Traditional views considered that employment in services was low-skilled and unproductive, while tertiary activities were considered neither dynamic nor innovative. In fact, several of the service activities require highly skilled jobs, show increasing productivity and are highly innovative and dynamic. Although services have been classified as non-tradable, they are increasingly exposed to competition and are becoming more tradable. The increasing importance of services in the so-called "industrialised countries" and the poor understanding of the process of services expansion require a review of the interpretations of the expansion of services.

The primary interpretation of the growth in services was the "theory of stages" (Kindleberger, 1958, Rostow, 1960) whose explanation was based mainly on the patterns of final consumption (Petty's Law and, specifically, Engel's Law: as income per capita increases, final demand shifts towards superior goods including services. This reflects a shift in consumer demand due to a high income elasticity of services). Furthermore, the growth of the service sector generated diverse

P. Coto-Millán (ed.), *Essays on Microeconomics and Industrial Organisation*

interpretations about the "post-industrial societies" (Bell, 1974), according to which the service sector is gradually taking the place of industry as the new engine of growth. However, a different set of questions were raised about the growth in services with Baumol's (1967) "theory of unbalanced growth". Amongst the most important questions posed were those relating to the definition of the role of the pattern of consumption and the role of productivity differentials in relation to service output and employment growth. Fuchs (1968) showed that the pattern of consumption had a less important role than that of productivity differentials (relatively slow productivity growth in some services). As a consequence of the lower productivity growth of the service sector as a whole compared with the manufacturing sector, plus the low skilled labour intensive characteristics of many services ("cost disease of personal service"), a secondary set of questions appeared when the intersectoral comparisons are made in constant or current prices (Gershuny & Miles, 1983 and Kravis *et al.*, 1983, Baumol *et al* 1989). In general, the increase in the contribution of services to GDP during the last five decades has stemmed more from changes in relative prices than from an increase in output. In order to understand this trend, however, it is necessary to distinguish between different services.

One of the problems encountered when examining the "services" is how they can be most accurately defined. Since the service sector is highly heterogeneous, it must be broken down into different categories according to the functions performed, the transformation processes and to the market served: producer or consumer services; distributive, social, personal or business services; market or non-market services; and physical, person-centred or information services. Most of these branches have been incorporated gradually into the Systems of National Accounts to improve the classification of services for analytical purposes (CEC-EUROSTAT, IMF, OECD, UN & World Bank 1994 and OECD 1995a & B). In order to understand the shift towards the service economy, it is necessary to analyse *what* is being produced in the economic system and *how* it is being produced.

Classification of services is not the only difficulty, traditional analysis is too biased on manufacturing performance. The process that drives services activities in many cases is different to that of manufacturing. In the same way, the key factors in services have a different relevance to those in manufacturing (R&D, innovation, organisational change or human capital). Notwithstanding, several services activities are becoming similar to manufacturing and vice versa (standard process and mass production), while the differences among services are as varied as those among manufacturing. In fact, the focus on the different categories between manufacturing and services is becoming less interesting and it seems more relevant to analyse the interaction between sectors and activities (Tomlinson 1997).

Standard indicators of technology intensity show that services make a contribution to total R&D expenditures which is relatively limited compared with the size of the sector in total employment. Certainly there is a problem of measurement, since most countries have only recently covered services in R&D and innovation surveys because it was assumed that manufacturing was the source of technological change and traditional measures do not usually capture key factors of innovation in services (such as patent registration).

One advantage of macroeconomic analysis based on Input Output techniques is that it enables overall production to be disaggregated by sector and by sub-system. It then becomes possible to examine the growth of different services and industries in relation to the process of structural change of the economic system. Furthermore, since the evolution of a sector or sub-system is not independent of the rest of the economy, it is necessary to evaluate the links between sectors (intersectoral relations) in terms of changes of final demand and technical change (Diaz Fuentes1993).

With these general premises in mind, and with the specific methodological approach, the following sections of this paper aim to explore a number of key questions:

- What has been the extent of structural change in employment in services in the main "industrialised countries" over the past four decades?
- Which technological trends explain the directions of structural change in employment in manufacturing and services? Was there a relationship between direct technology intensity and employment growth by sector?
- What is the embodied contribution of R&D and innovation expenditures to total technology intensity in manufacturing and services? What is the sectoral importance of the acquisition of technology embodied in inputs generated by other industries?
- Is the growth of employment in services explained solely by the relative increase in the final demand of these services, or is it also necessary to consider the growth of intermediate demand of services? Has the growth of services been caused by a greater demand for their use in the production of manufactured goods?

While accepting that there is a positive correlation between economic growth and employment in services, this research departs from an optimistic point of view about services. This seeks to analyse the increasing integration between manufacturing and services in technological innovation. This integration includes the outsourcing of manufacturing to specialised services activities and services that were not performed previously by other firms.

The following sections discuss these questions and demonstrate the importance of detailed analysis of inter-industry performance. Part 8.2 considers the trends in employment in the main "industrialised countries" taking a global measure of the extent of structural change of the major sectors share from 1960 to 1980, and from 1980 to 1997 (OECD, 1992a & annual a). Part 8.3 discusses the relationship between conventional measurements of technology intensity and the trend's directions in employment in manufacturing and services since 1980 (OECD, annual b and 1998). Part 8.4 goes beyond examining the extent and direction of structural change in services and provides an inter-industry analysis of innovation which, on the basis of the most recently published input-output table, estimates technology flows and the acquisition of innovation by sector in 1994 (INE 1997a and 2000, EUROSTAT 1997). Part 8.5 presents the results of the changing composition of employment in services due to final and intermediate demand (Garcia *et al.* 1994, EUROSTAT 1987, 1992 and OCDE 1995c). These results have been obtained through the use of inter-industry analysis. This analysis is founded on the notion of vertical integration of the sub-systems and provides an

enhanced view of the intersectoral relationships. Finally, some implications of the results are drawn in part six.

8.2
Extent of Structural Change in Services Employment

During the last four decades the world economy has exhibited extensive economic structural changes. At the same time, economic performance has varied significantly; the real increase in OECD GDP and productivity during the period from 1980 to 1997 was half that experienced between 1960 and 1980. The slow-down of economic growth in OECD countries since the end of the 1970s has been accompanied by three significant recessions, two oil shocks, growth of international trade, globalisation of financial markets and the diffusion of a set of new technologies (Freeman & Soete, 1994). As a consequence of these changes, the economic structures of these countries have been transformed markedly, reflecting structural, as opposed to cyclical, shifts in the composition of employment and production.[1] Notable economists consider that the reasons for this slow-down and higher unemployment are structural and are either caused by restrictions in markets,[2] or by the lack of technological innovative capabilities.[3]
Certainly, despite the considerable economic growth in the postwar period (GDP growth rates for OECD countries averaged around 4 per cent between 1960 and 1980 and 2 per cent between 1980 and 1990), labour absorption has been limited throughout the three decades, since employment average growth rates were around 1 per cent in both periods considered. This trend has been even more significant for EU countries, whose GDP growth averaged 4 per cent (1960-1980) and 2 per cent (1980-1997) while total employment growth rates were only 0.2 and 0.5 per cent for the respective periods. Considering the problems of labour absorption during the "golden age", it is interesting to explore employment trends since the turning point of the middle of the 1970s (OECD, annual a).
Taking a global measurement of the extent of structural change of the major sectors shares for the main industrialised countries including Spain (Table 8.1), it is clear that there are basic similarities in the patterns of structural change among

1 "Whilst economic theory has pointed to compensation mechanism generating new employment to replace jobs which are lost through technical change, no one has claimed that this process is instantaneous or painless. Economists differ however on the extent to which they would rely on self-adjusting market-clearing mechanism or on active public investment and labour market policies". See Freeman & Soete (1994, pp. 17-38).

2 Over the last decade the European unemployment rate has averaged 10 per cent, which is a much more serious matter than the fluctuations around the average. Conventional business cycles account for relatively little of the history of unemployment. Most of the annual variations in unemployment come from the long-frequency fluctuations between half decades rather than from the short-frequency fluctuations within half decades. This is because there are long term changes in social institutions, and the shocks (wars, oil or financial crisis) have long-lasting effects, see Layard *et al.* (1994, pp. 91-109).

3 These Information and Communication Technologies, although they have a vast range of present and future applications, do not yet easily match the inherited previous skill profile, management organisation, industrial structure or institutional framework. See Freeman & Soete (1994, pp. 47-66).

countries during the whole period: agriculture is declining while services are increasing as a share of overall employment and in all the three areas the sectoral changes were shifting towards services. However, in the period 1960- 80 the rising share in employment services was mainly correlated to the declining share of agriculture rather than in manufacturing or industry as in the period 1980-97. Additionally, countries differ widely in the sectoral composition of employment, in the proportion of structural adjustment, and in the degree of flexibility which work organisation displays in response to changes. On the one hand, from 1960 to 1980, some countries showed an increase in structural changes and industrial employment (Japan, Spain and Italy) and GDP shares (Japan and Spain), implying a significant catching-up in comparison with the leader country. On the other hand, from 1980 to 1997, the contribution of services to GDP has declined in the two countries with higher productivity growth (Japan and Germany). Furthermore, structural change can be considered as a source of growth. This applies, in particular, to countries in which employment is high in agriculture and productivity is low, since labour can be reallocated to other sectors of higher productivity.[4]

Table 8.1. Structural trends in employment 1960-97

	Agriculture		Industry		Services		Wholesale & retail trade, restaurants & hotels	Transport, storage & commun - ications	Finance, insurance, real estate, business services	Community, Social & Personal services
	1960-80	1980-97	1960-80	1980-97	1960-80	1980-97	1980-97	1980-97	1980-97	1980-97
USA	-4.9	-0.9	-4.0	-7.4	8.9	8.3	1.2	-1.3	7.5	1.1
Canada	-7.8	-1.5	-4.2	-5.3	12.0	6.8	1.1	-0.9	3.4	3.5
UK	-2.1	-0.7	-10.1	-10.7	12.2	11.4	0.3	-0.7	4.6	1.9
Germany	-8.4	-2.4	-2.9	-7.6	11.3	10.0	2.9	-0.6	1.4	3.0
France	-13.8	-4.2	-1.7	-10.3	15.5	14.5	1.7	-0.7	4.2	10.1
Italy	-18.3	-7.5	4.0	-5.9	14.3	13.4	3.3	0.6	0.8	11.5
Spain	-19.5	-10.8	5.9	-6.2	13.7	16.9	5.7	-0.1	1.9	7.6
Japan	-19.8	-5.1	6.8	-2.3	13.0	7.4	-1.1	-0.6	1.7	6.9
Australia	-4.5	-1.3	-8.0	-8.8	12.5	10.1	1.8	-0.6	5.5	1.7

Source: OECD (annual a & b).

Within services, the share of finance, insurance, real estate and business services (FIREBS) and community, social and personal services (CSPS) increases proportionally in all cases while that of transport, storage and communications (TSC) and, in certain cases, wholesale and retail trade, restaurants and hotels

[4] There is some evidence, based on 30 countries for the period 1960-90, that the GDP growth rates correspond negatively to increasing services shares. See Chenery (1986).

(WRTRH) lose their share in total service employment. Breaking down these figures shows that, from 63 to 85 per cent of growth was due to service sectors and that among these activities FIREBS and CSPS have the greatest impact. The increasing share of business and other intermediary services makes it gradually more difficult to measure compositional change, since it becomes necessary to look at structural links among sectors instead of showing only change in employment or GDP shares by sector.

8.3 Innovation and Employment Trends in Structural Change

The classification of the main sectors into three parts is a conventional but limited method of measuring the extent of structural change. A more accurate approach would be to disaggregate the sectoral evolution of GDP and employment in relation to the different trends in each individual sector, so it becomes possible to distinguish growing, medium and declining growth activities.[5] A complementary measurement of structural change in terms of direction could be obtained by classifying the branches according to their technology intensity.[6]

Contemporary theories of economic growth and international trade have stressed the role of innovation as a fundamental source of growth, employment and productivity, the capacity to innovate depends on multiple factors (Grossman, G. & Helpman, E. 1994). Technology investments are developed in a few manufacturing industries, however, the overall performance of the economic system depends on putting technology to work by using ideas and products developed in other activities. However, most firms and industries, in particular services such as FIRBS, acquire technology by purchasing and assimilating capital embodied technology machinery. This fact has changed the EU attitude to science, technology and innovation: "Support to innovation should be broadened from "mission-orientated" projects with specific research outcomes, such as a new

[5] In OECD (annual b) the industries were classified according to their annual growth rate (1974-90) in the main industrialised countries as: **High-Growth**: 1. Computers and office machinery, 2. Aerospace, 3. Communications, 4. Finance and insurance, 5. Business service, 6. Government, 7. Rubber plastic, 8. Pharmaceutical, 9. Social and personal service, 10. Instruments. **Medium-Growth**: 11. Chemical, 12. Trade, 13. Transport, 14. Agriculture, 15. Electrical machinery, 16. Paper and printing, 17. Electricity, gas and water, 18. Non-ferrous metals, 19. Food, drink and tobacco, 20. Motor vehicles, 21. Hotels and restaurants. **Low-Growth**: 22. Mining, 23. Non-electrical machinery, 24. Construction, 25. Fabricated metals, 26. Stone, clay and glass, 27. Textiles, 28. Petroleum refining, 29. Wood and furnitures, 30. Ferrous metals, 31. Shipbuilding.

[6] In OECD (annual b) the 21 manufacturing branches across 11 industrialised countries are ranked according to the R&D expenditures to gross output as a rough estimate of technological sophistication, with the following scheme: **High-Tech**: 1. Aerospace, 2. Computer and office machinery, 3. Communication equipment, 4. Pharmaceutical, 5. Instruments, 6. Electrical machinery. **Medium-Tech**: 7. Motor vehicles, 8. Chemical, 9. Non-electrical machinery, 10. Rubber and plastic, 11. Non-ferrous metals, 12. Other transports. **Low-Tech**: 13. Stone, clay and glass, 14. Food, drink and tobacco, 15. Shipbuilding, 16. Petroleum refining, 17. Ferrous metals, 18. Fabricated metals, 19. Paper and printing, 20. Wood and furnitures, 21. Textiles, footwear and leather.

combat aircraft, to "diffusion-orientated" programmes, such as educating small firms about new products and process" (Working Group on Innovation and Technology Policy 1999).

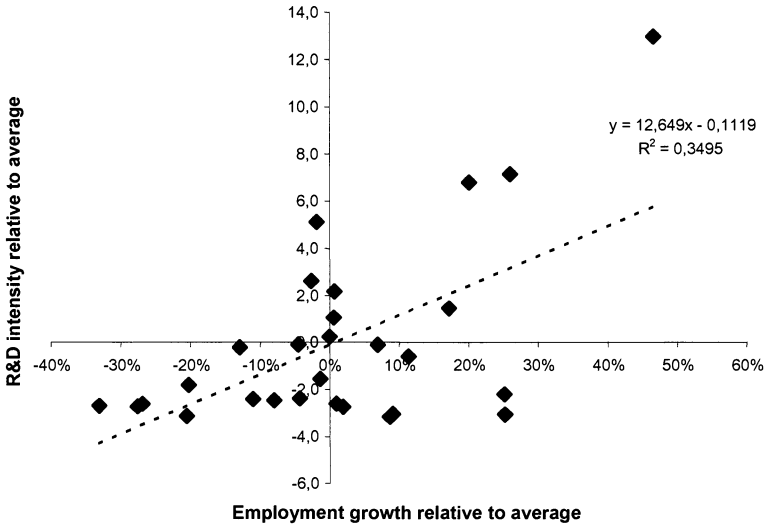


Fig. 8.1. R&D intensity and employment growth 1980-95

Source: OECD (annual b & c, 1998)

Technology intensity concerns the degree to which technology is produced and used within different industries. Activities with relatively high R&D or innovative expenditures per unit of output or value added are classified as high technology industries or technology intensive industries. However, technology generation and use are hard to measure, the roughest estimation is R&D intensity that measures the expenditure directly incurred by an industry, while technology use is estimated by R&D expenditures incurred by the acquisition of intermediate and capital goods. While the shares of manufacturing in employment and output have declined over the last two decades, the share of high technology industries in manufacturing output has increased steadily. However, employment shares of high technology industries have risen less than output shares, which means that labour productivity has risen in these industries. Fig. 8.1 provides some evidence of R&D effort and relative employment growth in the manufacturing industries for the period 1980-1995 for 15 OECD countries. The graph shows a correlation between high technology industries and employment. This indicates that technological change has accompanied structural change, favouring the emergence of

employment in high technology sub-sectors such as Aircraft, Office and computing machinery, Drugs and medicines, and Professional goods.
In the services, the evidence shows that the most rapidly growing sectors in both output and employment terms are FIREBS and CSPS; in certain cases WRTRH has noted a significant expansion (in particular in countries specialised in tourism). Although the service sectors are increasingly recognised as being important innovators, the existing indicators of technological intensity do not reveal their significance (Andersen & Howells 1998). Innovation in services is less based on R&D and more linked to acquired technology and changes in process, markets and organisation. In the past many countries focused R&D surveys on manufacturing because it was assumed that this was the source of technological change and innovation (Young 1996).

The latter approach would be useful in helping to evaluate the direction of structural change, but would fail to identify the transformations taking place between industries below the aggregate levels. The economies are also undergoing a different structural change, as the firms change the organisation of their production and source inputs. These changes affect the linkage between and within firms, industries and sectors. These sorts of structural changes can be analysed by looking at the structure of production in each industry and sector. These can be done applying input-output techniques, which provide a picture of inter-industry relations and linkages. Additionally, it could not connect the change in the structure to other factors such as shifts in domestic demand, foreign trade, technical change or input productivity.

8.4 Embodied Technology and Technology Diffusion in Services

Input-output techniques make it possible to analyse the economic system based on intersectoral relations of innovation expenditures that represent the technology flow of embodied R&D or innovation. This methodology applied to technology diffusion allows the measurement of total innovation intensity by sectors (Sakurai *et al.*1996).

In order to understand the shift towards the service economy it is important to examine the principal factors affecting both what is produced in the whole economy and how it is produced. This means that, when examining the service sector, it is not enough to look at this sector alone, since structural changes in the patterns of final demand, the intermediate demand and technical change must also be considered.

Services form a network through which economic activity takes place. The process of structural change can influence the degree to which services are required throughout the whole economic system. The supply of services, in turn, makes it possible to attain greater specialisation and division of labour. These factors are essential in reinforcing the observed shift towards services employment.

In order to present the methodology, a sector will be defined as a cluster of branches (firms) producing commodities in agreement with a standard classification (NACE), and a sub-system as a group of different activities which are required in the economic system to produce a specific product or service. In the context of inter-industry relations, a sub-system consists of activities of different branches, all of which directly and indirectly contribute to the production of a specific final output.

The balance equation of output in an input-output table can be defined as:

$$\mathbf{x} = \mathbf{A}\,\mathbf{x} + \mathbf{f} \tag{8.1}$$

$$\mathbf{X} = \mathbf{L}\,\mathbf{F} \tag{8.2}$$

where **x** is the vector of output by branch, **A** is the matrix of technical coefficients whose typical elements $a_{ij} = x_{ij} / \Sigma\, x_{ij}$ represent the value of the ith input needed to produce one unit of industry j`s output; $\mathbf{L} = (\mathbf{I}-\mathbf{A})^{-1}$, known as the Leontief inverse, represents the total requirements per unit of final output in terms of gross output; **X** is the diagonalised vector **x**; and **F** is the diagonalised vector **f** of the final demand by industry. The intersectoral relationships presented in equation (8.1) for an open static economic system can be written as:

$$\mathbf{X} = \mathbf{A}^{\mathbf{d}}\,\mathbf{x} + \mathbf{f}^{\mathbf{d}} + \mathbf{e}$$

$$\mathbf{X} = \mathbf{L}^{\mathbf{d}}\,\mathbf{F} \tag{8.3}$$

where $\mathbf{A}^{\mathbf{d}}$ is the matrix of domestic coefficients, $\mathbf{f}^{\mathbf{d}}$ is the domestic final demand vector for domestic outputs, **e** is the foreign demand or exports vector and the inverse $\mathbf{L}^{\mathbf{d}} = (\mathbf{I}-\mathbf{A}^{\mathbf{d}})^{-1}$ represents the total requirements per unit of final domestic output whose typical element is $\mathbf{l}^{\mathbf{d}}_{\mathbf{ij}}$.

Technology intensity for each industry can be defined as the R&D or innovation directly incurred expenditure per output for industry i : $\mathbf{r}_i = r_i / \mathbf{x}_I$, whose diagonalised vector is **R**. Thus, the vector of technology or innovation embodiment **ti** can be expressed in diagonalised form as:

$$\mathbf{TI}^{\mathbf{d}} = \mathbf{R}\,\mathbf{L}^{\mathbf{d}}\,\mathbf{F} \tag{8.4}$$

This equation connects innovations to final domestic and foreign demand and the multiplier of total innovation embodiment per unit of final demand is $\boldsymbol{mi} = \Sigma\, \mathbf{r}_i\, \mathbf{l}^{\mathbf{d}}_{\mathbf{ij}}$. Finally, in an open economy, imports are a source of technology diffusion, while production for exports induces demand for imported inputs.

$$\mathbf{m} = \mathbf{A}^{\mathbf{m}}\,\mathbf{x} + \mathbf{f}^{\mathbf{m}} = \mathbf{A}^{\mathbf{m}}\,\mathbf{L}^{\mathbf{d}}\,\mathbf{F} + \mathbf{F}^{\mathbf{m}} = \mathbf{M(4)}$$

$$\mathbf{TI}^{\mathbf{m}} = \mathbf{R}^{*}\,[\mathbf{A}^{\mathbf{m}}\,\mathbf{L}^{\mathbf{d}}\,\mathbf{F} + \mathbf{F}^{\mathbf{m}}] \tag{8.5}$$

As was noted in the third section, conventional indicators of R&D or innovation intensity such as direct R&D per unit of output or value added, distort the measurement of total R&D and innovation of industries. Indicators that estimate total innovation embodiment per unit of output or (such as $\boldsymbol{mi} = \Sigma\, \mathbf{r}_i\, \mathbf{l}^{\mathbf{d}}_{\mathbf{ij}}$ and $\mathbf{R}^{\mathbf{m}}\,\mathbf{A}^{\mathbf{m}}$) are more appropriate measures.

Table 8.2. Re-estimation of embodied technology intensity indicators by sector 1994

	R&D direct	Total R&D intensity	Innovation direct	Total Innovation intensity
High technology sectors				
28 Communications	***15,9***	***16,5***	***80,6***	***83,1***
13 Other transport n.e.c.	36,2	40,8	55,3	64,8
11 Electrical Machinery and electronic equipment	24,4	27,1	37,3	43,5
12 Motor vehicles	8,4	12,5	31,2	43,5
Medium technology sectors				
21 Rubber & Plastic Products	5,1	9,1	15,5	25,7
20 Publishing, Printing & Reproduction or recording media	0,4	3,6	12,7	24,5
30 Finance and insurance intermediation	***0,0***	***3,1***	***0,0***	***13,6***
10 Office and Computing Machinery	5,3	8,5	7,0	12,5
24 Recovery and repair	***0,0***	***4,7***	***0,0***	***12,4***
17 Leather products and footwear	0,6	2,9	3,4	11,5
29 Electricity, gas & water supply.	***2,8***	***4,2***	***5,2***	***10,5***
Lower technology sectors				
1 Agriculture, fishing	0,0	2,3	0,0	8,0
23 Construction	***0,0***	***2,2***	***0,0***	***7,3***
26 Hotels & restaurants	***0,0***	***1,4***	***0,0***	***7,0***
15 Tobacco products	2,0	2,6	4,5	6,9
27 Transport	***0,0***	***2,3***	***0,0***	***6,4***
33 Public services	0,0	1,9	0,0	5,0
32 Private social and personal services	***0,0***	***1,3***	***0,0***	***3,4***
31 Business services & real state	***0,0***	***0,6***	***0,0***	***2,8***
25 Wholesale & retail trade	***0,0***	***0,7***	***0,0***	***2,7***

Sources: Elaborated by the author based on R&D and innovation data (INE 1997 & 1999) and Input-Output tables for 1994 (INE, 2000).

Considering that R&D and innovation direct expenditures are reduced in most services with the exceptions of communications and electricity, gas and water distribution, the classification by sectors is blurred when considering the purchase of R&D and innovation inputs. The change is of particular significance for R&D users such as financial and insurance intermediaries and recovery and repair, and to a lesser extent for the rest of the services. In terms of innovation, table 8.2 shows similar results but the contribution of acquired innovation is relatively larger in sectors such as construction, hotels and restaurants and transport.

R&D expenditures are widely considered as indicators of technology intensity and the main determinant of economic growth and productivity. In Spain as well as in other "industrialised countries" these expenditures mainly originate in a few

manufacturing industries such as: electronic machinery and electronic equipment, chemical products (including pharmaceutical) and other transport (including aerospace) (Papaconstantinou *et al.* 1996). A large share of the outputs of these "high technology" manufacturing sectors are demanded as intermediate inputs into the production process by different sectors and are also sold to final demand. In this way the direct R&D expenditures of the provider industries become embodied in products, process and services across the economic system.

The estimation of performed and acquired technology is estimated using input-output techniques and R&D and innovation expenditures. The shares of R&D and innovation are expressed in relative terms to the total indirect expenditures embodied in output using the methodology explained above (equations 8.4 and 8.5 for **TI**).

Table 8.3. Technology clusters of R&D and innovation. The 7 largest sectors in Spain in 1994

R&D providers	*share*	***R&D users***	*share*
11 Electrical machinery and electronic eq	*18,05*	***24 Recovery and repair***	***6,8***
7 Chemical products	*15,92*	*21 Rubber & Plastic Products*	*6,4*
13 Other transport n.e.c.	*12,09*	*12 Motor vehicles*	*6,1*
12 Motor vehicles	*11,85*	*7 Chemical products*	*4,5*
28 ***Communications***	*11,01*	***32 Private social and personal services***	***4,1***
9 Machinery n.e.c.	*6,90*	*4 Basic Metal Ferrous*	*4,1*
29 ***Electricity, gas & water supply.***	*3,45*	***30 Finance and insurance int.***	***3,4***
	79,28		35,3
		Total services	***28,3***

Innovation providers	*share*	***Innovation users***	*share*
28 ***Communications***	*17,31*	*1 Agriculture, fishing*	*6,2*
12 Motor vehicles	*13,76*	*24* ***Recovery and repair***	*6,0*
14 Food & beverages	*13,23*	*12 Motor vehicles*	*5,7*
7 Chemical products	*10,43*	*26* ***Hotel & restaurants***	*5,6*
11 Electrical Machinery and electronic eq.	*8,59*	*21 Rubber & Plastic Products*	*4,8*
13 Other transport n.e.c.	*5,74*	*30* ***Finance and insurance int.***	***4,6***
9 Machinery n.e.c.	*4,14*	*32* ***Private social and personal services***	***4,5***
	73,21		*37,4*
		Total services	***35,1***

The picture that emerges in terms of R&D is, on the one hand, a concentrated cluster of "high technology" industries providing most of the R&D in the manufacturing sector and in certain services such as communication and electricity, gas and water distribution. On the other hand, the cluster of users is more disperse and includes different services and manufacturing activities, such as

recovery and repair services, private and personal services and finance and insurance institutions.

In terms of innovation expenditures the picture that emerges is slightly different. The cluster of providers is still concentrated and includes some of the same industries than in R&D but also others such as food and beverages. The bulk of innovation is in communication and motor vehicles rather than in "high technology" manufacturing providers. Last but not least, the cluster of innovative users includes services that are not usually considered "high or medium technology intensive" such as: agriculture, forestry and fishery; recovery and repair services; hotels and restaurants; finance and insurance; and private social and personal services. Thus, many services industries act as the main users of technology and innovation and constitute significant R&D and innovative clusters.

8.5 Compositional Structural Change in Employment (Final and Intermediate Demand for Services)

In the second section the extent of the structural change has been examined and, in the third section, the direction of these changes in the face of technology intensities. A more precise definition of compositional structural change considers changes in the sectoral integration of an economy: output and employment shares reported for different sectors, and the changes in the inputs used by them. The advantage of this method is that it provides a detailed image of how the structure of an economic system and its linkages are at one moment, and how they have unfolded over time. The evaluation of both the extent and direction is connected to the broad sources of change for each sector, namely: final demand, import substitution and pattern of inter-industry linkages in the economy (referred to as technical change), and this represents the path of change followed to reach a specific sectoral structure.

The Input-Output (IO) technique enables changes in output and employment to be estimated, and it is also useful in helping to evaluate the relationship between employment and technology. The temporal variations of IO intermediate coefficients themselves reveal significant information about the technical change that operates in an economic system. In so doing, they represent an extension of previous measurements of structural change. A complementary IO analysis of employment and innovation is based on the concept of the sub-system and the notion of a vertically integrated sector, which was introduced by Sraffa (1960) and Pasinetti 1981) for theoretical purposes, but can also be used in applied terms (Sakurai 1993, Diaz Fuentes 1999).

The intersectoral relationships between branches and sub-systems from an IOT will be represented in a single table as matrix operator **B**. The last two terms on the right hand side of (8.2) $\mathbf{L}^d \mathbf{F}$ correspond to the actual amount of all domestic input that is directly and indirectly required for the production of a final commodity (column $_j$). When **X** is multiplied by the matrix $\mathbf{L}^d \mathbf{F}$ the operator **B** is obtained:

$$\mathbf{B} = \mathbf{X}\,\mathbf{L}^d\mathbf{F} \tag{8.6}$$

This is the matrix of shares of production or **B** operators, the elements of which (b_{ij}) shows the share of total output x_i which is required in the sub-system $_j$. The sum of these elements is one. The results of matrix operator **B** can be utilised to re-analyse variables associated with the production by branch such as employment, R&D and innovation. This can be disaggregated to the highest level (Terleckyj 1974, Barker 1990 and Sakurai 1997). In this case, the diagonalised vector of employment $\hat{\mathbf{u}}$ by branch has been used to calculate the matrix of employment **U**:

$$\mathbf{U} = \hat{\mathbf{u}}\,\mathbf{B}$$

U shows, by rows, the amount of employment that each branch contributes to each sub-system (and of which the sums are the same total of **u**), and the columns of **U** show the employment of each sub-system. These matrices **B** and **U** disclose the direct and indirect shares of output and employment by sector and sub-system, and the indirect shares of output can be separated by replacing the Leontief inverse in (8.3) and (8.6) by **L** - **I**. With this methodology, employment in services can be separated by final and intermediate requirements.

Table 8.4

Employment in intermediate and final demand market services (thousands of employees and percentage of total employment in services)					
	1975	**1980**	**1985**	**1989**	**1994**
Intermediate	1.089	1.213	1.408	1.506	1.643
	26,0	*30,0*	*33,4*	*31,1*	*31,5*
Final	3.103	2.836	2.810	3.329	3.566
	74,0	*70,0*	*66,6*	*68,9*	*68,5*
	4.192	4.049	4.218	4.835	5.209
Employment in intermediate market services in manufacturing as a sub-system (thousands of employees and percentage of manufacturing sub-system)					
Intermediate market service in Manufacturing	402	390	378	407	418
% of manufacturing subsystem	***16,1***	***18,0***	***19,1***	***22,1***	***28,1***

The general results that have been extracted can be summarised by the following two main points. First, table 8.4 (top part) shows that the growth of market service employment in Spain was due mainly to an increase in intermediate demand; and this cannot be explained exclusively by the "stages of growth theory". This corresponds to the trends exhibited by the main advanced European countries that experienced constant growth in the share of employment due to intermediate market services since the 1960s. Therefore the growth of market service employment in the European economies is not directly, nor mainly, due to an increase in final demand, but rather to an increase in the intermediate demand of services. Second, table 8.4 (bottom part) shows that the expansion in intermediate demand for services is accompanied by their increasing use in the production of manufactured goods in Spain, and this corresponds with the trends exhibited by

the principal European economies throughout the considered period. This implies that the production of manufacturing goods goes beyond the industrial sector and requires increasing services.

8.6 Conclusions

This chapter has examined the expansion of employment in services that constitutes one of the most significant features of long term structural change in "industrialised economies". Four issues regarding structural change in employment in terms of innovation have been presented: structural change extension; direct technological intensity effects; embodied innovation; and the inter-industrial dimension in terms of the changes in the final and intermediate demand.

The most rapidly growing sectors in terms of employment in most "industrialised economies" have been services and in particular: FIREBS and CSPS. However, the increase in employment in WRTSRH has been significant in Spain, France and Italy, which have been at the same time the countries with the largest increases in service employment between 1980 and 1997. The observed patterns of the extension of structural change in employment do not indicate whether the growing role of services reflects a change in final demand, business demand or outsourcing to specialised services sectors.

Assuming the importance of technological innovation in international economic growth and productivity, and the fact that a few manufacturing industries concentrate most of the R&D surveyed expenditures, the correlation between direct technology intensity and employment was evaluated. It was observed that technological change determined structural change in employment in sub-sectors such as aerospace, office and computing machinery, telecommunications equipment, drugs and medicines and professional goods. Moreover, some services sectors such as FIREBS and CSPS acquire technology by purchasing and assimilating inputs embodied in innovation related to the previously mentioned high technology intensity manufacturing sub-sectors. However, it is notable that, although services are increasingly innovators, the existing indicators of R&D and technological intensity do not reflect the whole scope of innovation in services because traditional surveys have been designed for manufacturing.

One important element of innovation is the acquisition and not only the direct expenditures on R&D. To identify the inter-industrial flows of innovation taking place between sector below the aggregate levels input-output techniques were applied, in first instance, to analyse the embodied R&D and innovation technological flows with the purpose of measuring "total technology intensities" by sectors and sub-systems. Following a defined methodology it was noticed that the acquisition of technology was an important component of the innovation expenditures in services. Given the low level of direct technology intensity in services, the total innovation and R&D embodied intensity increased significantly in all services categories, but in particular in: finance and insurance, recovery and repair, electricity, gas and water distribution, hotels and restaurants and transport, which correspond with the aggregated trends envisaged in the second and third

section. Moreover, given that a limited number of manufacturing sectors are the main providers of technology, the estimation of performed and acquired technology based on input-output techniques shows a concentrated cluster of high technology industries including communication services and a cluster of R&D and innovation users that includes recovery and repair, private and personal services, finance and insurance, and hotels and restaurants. Thus, many services act as the main users of technology and constitute a technology cluster which importance should be reconsidered in national survey on innovation (OECD 1995c).

This result points to the importance of policies of innovation diffusion in services. Although new technologies are concentrated in a small number of manufacturing industries that spend directly in R&D, the new process, products and services created in that industries generate benefits that become widespread through diffusion and use. The performance of an economic system depends on applying technology by using and adapting products, process and services generated elsewhere. This ability of the firms and industries is critical for the economic system's productivity and growth.

Finally, a more precise analysis of structural change was presented in the fifth section. This considered not only the extension and the technological direction component but also the broad sources of structural change from the intermediate and final demand side. Growth in service employment must be explained by considering the increasing integration between sectors (services and industry). This relationship can be explained by the following four factors. First, the increasing specialisation among sectors, which requires a complex network of services such as communications, transport, banks and insurance, recovery and repair, and after-sales services that link the different sub-systems. Second, the expansion in the foreign trade of goods and services, which is another perspective of specialisation. Third, the augmenting regulations (standards of quality or environment) which require specialised services (such as legal, tax, engineering, publicity, training, accounting, or finance and insurance). Fourth, the emergence of new economies of scale in the production of services, which induce a process of externalisation or outsourcing to specialised service sectors.

The application of intersectoral analysis in this research on Spain and other European countries has generated a different conclusion from that derived by the aggregated analysis, since the methodology allows the interpretation of the links between manufacturing and services, and the sets of innovation expenditures and employment that are directly and indirectly utilised in the productive systems.

References

Andersen, B.and Howells, J.: Innovation Dynamics in Services: Intellectual Property Rights as Indicators and Shaping Systems in Innovation. CRIC Discussion Paper No 8. Manchester, University of Manchester 1998

Barker, T.: Sources of Structural Change of the UK Service Industries 1979-1984. Economic System Research 2, 173-183 (1990)

Baumol, W.J.: The Macroeconomic of Unbalanced Growth. The Anatomy of the Urban Crisis. American Economic Review 57, 415-426 (1967)

Baumol, W. J., Blackman, S. A. B. and Wolf, E.: Productivity and American Leadership: The Long View. Cambridge MA, MIT Press 1989

Bell, D.: The Coming of Post-Industrial Society: A Venture in Social Forecasting. London, Heinemann 1974

CEC (Commission of the European Communities EUROSTAT), IMF (International Monetary Fund), OECD, UN (United Nations) and World Bank: System of National Accounts 1993. Brussels/Luxembourg, New York, EUROSTAT, IMF, OECD, UN and World Bank 1994

Chenery, H. B.: Growth and Transformation. In: H. Chenery, Robinson, S. and Syrquin, M. (Eds.), Industrialization and Growth. New York, World Bank-Oxford University Press 1986

Diaz-Fuentes, D.: Relaciones entre Cambio Tecnológico y Empleo a partir del Análisis Input-output: España 1980-1985. Revista de Economía y Sociología del Trabajo 19-20, 21-33 (1993)

Diaz-Fuentes, D.: On the Limits of the Post-Industrial Society: Structural Change and Service Sector Employment in Spain. International Review of Applied Economics 13(1), 111-24 (1993)

EUROSTAT (Statistical Office of the European Communities): European System of Integrated Economic Accounts, ESA. Luxembourg, EUROSTAT 1985

EUROSTAT (Statistical Office of the European Communities): Input-Output Tables, 1975-1980, (magnetic tape CEE IO TAB). Luxembourg, EUROSTAT 1987

EUROSTAT (Statistical Office of the European Communities): Coding System of the Input-Output Tables Database of Eurostat used on Magnetic Support (National Accounts Tables). Luxembourg, EUROSTAT 1992

EUROSTAT: The First European Innovation Survey. Luxembourg, EUROSTAT 1997

Freeman, C. and Soete, L.: Work for all or Mass Unemployment: Computerised Technical Change into the 21st Century. London, Pinter 1997

Fuchs, V. and Leveson, I.: The Service Economy. New York, National Bureau of Economic Research 1968

García-Perea, P. and Gomez, R.: Elaboración de Series Históricas de Empleo a partir de la Encuesta de Población Activa, D. 38668. Madrid, Banco de España 1994

Gershuny, J. and Miles, I.: The New Service Economy: the Transformation of Employment in Industrial Societies. London, Pinter 1983

Greenhalgh, C.; Gregory, M. and Ray, A.: Employment and Structural Change in Britain. Working Paper 44. Institute of Economics and Statistics. University of Oxford 1988

Grossman, G. and Helpman, E.: Endogenous Innovation and the Theory of Growth. Journal of Economic Perspectives 8(1), (1994)

INE (Instituto Nacional de Estadística): Encuesta sobre Innovación Tecnológica de las Empresas. INE, Madrid 1997a

INE: Estadistica sobre las Actividades en Investigacion Científica y Desarrollo Tecnológico (I+D) Indicadores Básicos 1994. INE, Madrid 1997b-annual

Kindleberger, C.: Economic Development. New York, McGraw-Hill 1958

Kravis, I. B., Heston, A. and Summer, R.: The Share of Service in Economic Growth. In: Adam, F. G. and Hickman, B. G. (Eds.) Global Econometrics. Essays in Honour of Lawrence R. Klein. Cambridge MA 1983

Layard, R., Nickell, S. and Jackman, R.: The Unemployment Crisis. Oxford, Oxford University Press 1994

OECD (Organisation for Economic Cooperation and Development): Historical Statistics 1960. Paris, OECD 1960 (annual a)

OECD (Organisation for Economic Cooperation and Development): Basic Science and Technology Statistics 1993. Paris, OECD 1960 (annual b)

OECD (Organisation for Economic Cooperation and Development): The OECD STAN database for Industrial Analysis OECD. Paris, OECD 1960 (annual c)

OECD (Organisation for Economic Cooperation and Development): Structural Change and Industrial Performance. Paris, OECD 1992

OECD (Organisation for Economic Cooperation and Development): The Measurement of Scientific and Technologic Activities: Proposed Standard Practices for Surveys of Research and Experimental Development (Frascati Manual 1993). Paris, OECD 1994

OECD (Organisation for Economic Cooperation and Development): ISDB Version 1995 - International Sector Data-base. Paris, OECD 1995a

OECD (Organisation for Economic Cooperation and Development): Services Innovation: Statistical and Conceptual Issues. Paris, OECD 1995b

OECD (Organisation for Economic Cooperation and Development): The Input-Ouput Database. Paris, OECD 1995c

OECD (Organisation for Economic Cooperation and Development): Oslo Manual, Proposed Guidelines for Collecting and Interpreting Technological Innovation Data. Paris, OECD 1997

OECD (Organisation for Economic Cooperation and Development): Science, Technology and Industry Outlook. Paris, OECD 1998

OECD (Organisation for Economic Cooperation and Development): Strategic Business Services. Paris, OECD 1999

Pasinetti, L.: Structural Change and Economic Growth. A Theoretical Essay on the Dynamic of Wealth of Nations. Cambridge, Cambridge University Press 1981

Papaconstantinou, G., Sakurai, N. and Wyckoff, A.: Embodied Technology Diffusion: an Emprical Analysis for ten OECD Countries. STI Working Paper 1966/1/OCDE/GD(96)26. OCDE 1996

Rostow, W. W.: The Stages of Economic Growth: a Non-communist Manifesto. Cambridge, Cambridge University Press 1960

Sakurai, N.: Structural Change and Employment: Empirical Evidence for Eight OECD Countries. Paper of the Helsinki Conference on Technology, Innovation Policy and Employment 7-9, October. Paris, OECD 1993

Sakurai, N., Ioannidis, E. and Papaconstantinou, G.: Impact of R&D and Technology Diffusion on Productivity Growth: Empirical Evidence for 10 OECD Countries. Economic System Research 9(1), 81-110 (1996)

Sraffa, P.: Production of Commodities by Means of Commodities. Cambridge, Cambridge University Press 1956

Tomlinson, M.: The Contribution of Services to Manufacturing Industry: Beyond the De-industrialisation. Debate, CRIC Discussion Paper 5. Manchester, University of Manchester 1997

Working Group on Innovation and Technology Policy: Promoting Innovation and Growth in Services. DSTI/STP/TIP(99)4. Paris, OECD 1999

Young, A.: Measuring R&D in the Services. STI Working Paper 1996/7. Paris, OECD 1996

PART III. MARKET AND INDUSTRIAL STRUCTURE

9 The Measurement of Intra-industry Trade and Specialisation: a Review

G. Carrera-Gómez
University of Cantabria (Spain)

The assessment of the relevance of the new theories of trade as opposed to the traditional theoretical approach, when it comes to explain real trade patterns, is essentially an empirical question. Nevertheless, there are two problems (which remain unsolved) which influence the results obtained when measuring intra-industry trade and specialisation. The first problem refers to the very existence of the phenomenon and is related to the definition of "industry" and the selection of the level of data dissagregation more appropriate to study such a phenomenon. The second problem, closely connected to the former, comes from the objective difficulty of finding a convenient quantitative measure.

This chapter is devoted to the analysis of the obstacles faced by empirical treatment of intra-industry trade. With this purpose, we present first in 9.1 a general critical overview of the main measurement indices proposed by the literature. We deal next in 9.2 with two important problems of measurement: the adjustment of global trade imbalance and the question of categorical aggregation. Finally, in 9.3 we offer a summary of the main conclusions reached regarding these aspects of intra-industry trade.

9.1 Measures of Intra-industry Trade and Specialisation

Intra-industry trade may be defined as simultaneously importing and exporting products belonging to a particular country and the same industry. This phenomenon comes usually accompanied by intra-industry specialisation, which is the concentration of production factors in particular groups of products inside an industry, at the expense of other production lines. Although both expressions are often indiscriminately used, they refer to two different aspects.

P. Coto-Millán (ed.), *Essays on Microeconomics and Industrial Organisation*

Therefore, at a conceptual level, it is possible to establish a distinction between intra-industry trade and intra-industry specialisation. According to that distinction and following the classification provided by Greenaway and Milner (1986) and Kol and Mennes (1983, 1986), we can differentiate two broad categories of indices regarding empirical measurement of the aforementioned concepts:

i) Indices determining the extension of intra-industry trade by measuring the level of overlap existing in trade flows.
ii) Indices of intra-industry specialisation, which are measuring the degree of similarity of trade patterns (relative structure of imports and exports).

We next present separately the main indices used for the measurement of both aspects. Later, a comparison is made among the diverse measures and a suggestion is proposed as to which may be more appropriate according to the purpose of the analysis to be performed. Finally, we summarise in subsection 9.1.4 the later developments on the measurement of intra-industry trade and mention the current unresolved issues on the subject.

9.1.1 Intra-industry Trade Indices

The question of intra-industry trade measurement and the inherent problems was first explicitly analysed in early works of Grubel and Lloyd (1971, 1975). Nevertheless, some previous papers (Verdoon, 1960; Michaely, 1962; Kojima, 1964; Balassa, 1966) already provided several procedures which could be applied to the measurement of intra-industry trade; although, that was not the main purpose of these works.

Verdoorn (1960) used the following indicator to examine changes in trade patterns in the Benelux:

$$S_j = \frac{X_j}{M_j}$$

where X_j and M_j are exports and imports of commodity j and S_j takes values inside the interval [0 , $+\infty$). When the value of the index comes closer to unity along time, it would be an indicator of increasing intra-industry specialisation.

The disadvantage of this measure, as Grubel and Lloyd (1975) already pointed out, is that any fraction and its inverse measure the same level of intra-industry trade, what makes it complicated to compare among industries. It can be seen that, except for the case where $S_j = 1$, this indicator does not measure directly the degree in which imports and exports overlap in a specific group of products.

Kojima (1964) solved this problem by defining the degree of "horizontal trade" in a group of products j between two countries A and B as follows:

$$D_j = \frac{X_j}{M_j} \text{ if } M_j > X_j$$

and

$$D_j = \frac{M_j}{X_j} \text{ if } M_j < X_j$$

D_j takes values inside the interval [0, 1], making it easier comparison among industries. The value of the index approaches unity as the degree of "horizontal trade" increases[1]. Although the index D_j represents an improvement compared to the Verdoorn (1960) measure, both indices share a problem: the use of ratios between trade flows is not giving a direct measure of intra-industry trade as a proportion of total trade.

Kojima (1964) also provided an aggregated indicator of "horizontal trade" between two countries. The aggregated index was a weighted sum of indices D_j using as weight the share of trade of commodity group *j* in total trade. This aggregated indicator was defined as follows:

$$\overline{D} = \sum_{j=1}^{n} \left(\frac{X_j}{M_j} \cdot \frac{X_j + M_j}{\sum_{j=1}^{n} (X_j + M_j)} \right) = \sum_{j=1}^{n} D_j \cdot w_j \text{ , if } M_j > X_j$$

$$\overline{D} = \sum_{j=1}^{n} \left(\frac{M_j}{X_j} \cdot \frac{X_j + M_j}{\sum_{j=1}^{n} (X_j + M_j)} \right) = \sum_{j=1}^{n} D_j \cdot w_j \text{ , if } M_j < X_j$$

where *n* is the number of groups considered and $w_j = \frac{X_j + M_j}{\sum_{j=1}^{n} (X_j + M_j)}$ [2].

In a work aimed at analysing the effects of the completion of the Common Market on international specialisation in EC countries, Balassa (1966) proposed the following indices to measure the degree of overlapping between exports and imports:

---

1 Note that D_j is equal to S_j when $M_j > X_j$ and it is equal to S_j^{-1} when $M_j < X_j$.

2 Note that it is in fact a weighted mean since the sum of weights is unity.

$$\sum_{j=1}^{n} w_j = \sum_{j=1}^{n} \left(\frac{X_j + M_j}{\sum_{j=1}^{n} (X_j + M_j)} \right) = 1.$$

For this reason and from now on, we will call mean the type of weighted sums in which the sum of weights is 1.

$$A_j = \frac{|X_j - M_j|}{X_j + M_j} \text{ for commodity group } j,$$

and

$$\overline{A}_j = \frac{1}{n}\sum_{j=1}^{n} A_j \text{ , as an aggregate index for } n \text{ groups of products.}$$

where $\overline{A}_j$ is an arithmetic mean of indices A_j for a specific level of disaggregation. Therefore, this procedure of aggregation has the following shortcoming: it gives the same weight (1/n) to all industries independent of their share in total trade.

The A_j index is inversely related to the degree of intra-industry trade. It takes values from 1 (when all trade is inter-industrial) to 0 (when all trade is intra-industrial). It has the advantage compared to previous measures of giving information on the proportion of total trade that is of the intra-industry type.

Another important point to mention is that, as a result of expressing net trade as a proportion of total trade in a specific group of products, absolute values of imports and exports that may be quite different may produce the same value of the index[3].

Moreover, suppose that commodity group j is formed by a number of subgroups m (which will be denoted by the subindex i). In this case, A_j may be expressed as follows:

$$A_j = \frac{\left|\sum_{i=1}^{m}(X_{ij} - M_{ij})\right|}{\sum_{i=1}^{m}(X_{ij} + M_{ij})}$$

Assuming that either $(X_{ij}\text{-}M_{ij}) < 0, \forall\, i = 1, 2, .., m$, or $(X_{ij}\text{-}M_{ij}) > 0, \forall\, i = 1, 2, .., m$, then:

$$\left|\sum_{i=1}^{m}(X_{ij} - M_{ij})\right| = \sum_{i=1}^{m}|X_{ij} - M_{ij}|$$

and we can write:

$$A_j = \frac{\sum_{i=1}^{m}|X_{ij} - M_{ij}|}{\sum_{i=1}^{m}(X_{ij} + M_{ij})} = \sum_{i=1}^{m}\frac{|X_{ij} - M_{ij}|}{(X_j + M_j)}$$

Multiplying and dividing by $(X_{ij} + M_{ij})$:

$$A_j = \sum_{i=1}^{m} \frac{|X_{ij} - M_{ij}|}{(X_{ij} + M_{ij})} \cdot \frac{(X_{ij} + M_{ij})}{(X_j + M_j)} = \sum_{i=1}^{m} A_{ij} \cdot w_{ij}$$

where $A_{ij} = \dfrac{|X_{ij} - M_{ij}|}{(X_{ij} + M_{ij})}$ is the correspondent index for subgroup *i* and where

$w_{ij} = \dfrac{(X_{ij} + M_{ij})}{(X_j + M_j)} = \dfrac{(X_{ij} + M_{ij})}{\sum_{i=1}^{m} (X_{ij} + M_{ij})}$ is the weight for that subgroup *i*.

Therefore, A_j is a weighted average of the indices of the subgroups that form a group *j* only when the sign of trade imbalance is the same for all the subgroups. This weighting effect is lost, nevertheless, when there are trade imbalances of opposite sign.

In general, a weighted average may be more suitable when we want to obtain a *compendium* measure, which reflects the relative importance of intra-industry trade at a particular level of aggregation[4]. This is so specifically when such a *compendium* index refers to an economy in a wide sense, including all trade. On the other hand, if the level of aggregation that has been selected to calculate individual indices closely corresponds to the researcher's view of homogeneity inside a group or industry, weighting individual indices may be inappropriate.

Greenaway and Milner (1986) propose the following procedure to guarantee the weighting effect of the index:

$$A'_j = \frac{\sum_{i=1}^{m} |X_{ij} - M_{ij}|}{X_j + M_j}$$

Note that multiplying and dividing by $(X_{ij} + M_{ij})$ we have:

$$A'_j = \sum_{i=1}^{m} \frac{|X_{ij} - M_{ij}|}{(X_{ij} + M_{ij})} \cdot \frac{(X_{ij} + M_{ij})}{(X_j + M_j)} = \sum_{i=1}^{m} A_{ij} \cdot w_{ij}$$

where $w_{ij} = \dfrac{(X_{ij} + M_{ij})}{(X_j + M_j)}$.

---

3 Note, for instance, that A_j is always 0 when X_j equals M_j, independent of the actual value of imports and exports.

4 For instance, if we want to show the relevance of intra-industry trade in subgroups i, we would want A_j to be a weighted average of indices A_{ij}.

Therefore, when all the subgroups trade imbalances have the same sign, the relationship $A_j' = A_j$ holds.

Using the same procedure it is possible to obtain *global* weighted averages of all indices A_j, that is:

$$\overline{A}_j' = \frac{\sum_{j=1}^{n}|X_j - M_j|}{\sum_{j=1}^{n}(X_j - M_j)} = \sum_{j=1}^{n} A_j \cdot w_j$$

where $w_j = \dfrac{(X_j + M_j)}{\sum_{j=1}^{n} X_j + \sum_{j=1}^{n} M_j}$.

That is to say, the weight employed is the ratio volume of trade in industry *j* to total trade.

Grubel and Lloyd (1975) proposed an index according to which the share of intra-industry trade in total trade flows of an industry *j* can be computed by the following expression:

$$B_j = \frac{(X_j + M_j) - |X_j - M_j|}{(X_j + M_j)} = 1 - \frac{|X_j - M_j|}{(X_j + M_j)}$$

where $0 \leq B_j \leq 1$.

The average of indices B_j for a set of industries (j=1, 2, 3, ..., n) may be computed as follows[5]:

$$\overline{B}_j = \sum_{j=1}^{n} w_j \cdot B_j = \sum_{j=1}^{n} \frac{(X_j + M_j)}{\sum_{j=1}^{n}(X_j + M_j)} \cdot \frac{(X_j + M_j) - |X_j - M_j|}{(X_j + M_j)} =$$

$$= \frac{\sum_{j=1}^{n}(X_j + M_j) - \sum_{j=1}^{n}|X_j - M_j|}{\sum_{j=1}^{n}(X_j + M_j)} = 1 - \frac{\sum_{j=1}^{n}|X_j - M_j|}{\sum_{j=1}^{n}(X_j + M_j)},$$

where $\sum_{j=1}^{n}|X_j - M_j| \neq \left|\sum_{j=1}^{n}(X_j - M_j)\right|$ when the sign of trade imbalances is not the same for all industries.

---

5 Note that the same weighting procedure as in index $\overline{A}_j'$ is used.

The advantage of index $\overline{B}_j$ when compared to Balassa's $\overline{A}_j$ is that the latter gives the same weight to all industries, whereas the former takes into account the share of industry j trade in total trade.

A potential shortcoming of the aggregated G-L index $\overline{B}_j$ is that it introduces a downward bias in the measurement of intra-industry trade when total trade is not balanced, that is when $\sum_{j=1}^{n} X_j \neq \sum_{j=1}^{n} M_j$. In this case imports and exports cannot match exactly in all industries and therefore the index can never reach its maximum value 1. Grubel and Lloyd propose a correction procedure to deal with this problem, which we will discuss in section 9.2, totally devoted to trade imbalance adjustments.

On the other hand, it can be noted that index B_j is a modification of Balassa's A_j[6], sharing the same properties concerning the weighting effect. However B_j is directly related to the level of intra-industry trade. It reaches its maximum value when total trade is of the intra-industry type and its value is zero when there is no matching at all between exports and imports in a specific industry.

The same comments we have pointed out for the Balassa (1966) index concerning the trade imbalances opposite sign effects hold for the B_j index. If we consider the industry *j* composed of *m* subgroups *i*, we can rewrite the B_j index as follows:

$$B_j = \frac{\sum_{i=1}^{m}(X_{ij} + M_{ij}) - \left|\sum_{i=1}^{m}(X_{ij} - M_{ij})\right|}{\sum_{i=1}^{m}(X_{ij} + M_{ij})}$$

There is, however, an alternative aggregating procedure, which will provide the following index:

$$B_j^* = \sum_{i=1}^{m} w_{ij} \cdot B_{ij} = \frac{\sum_{i=1}^{m}(X_{ij} + M_{ij}) - \sum_{i=1}^{m}|X_{ij} - M_{ij}|}{\sum_{i=1}^{m}(X_{ij} + M_{ij})}$$

where $w_{ij} = \frac{(X_{ij} + M_{ij})}{(X_j + M_j)} = \frac{(X_{ij} + M_{ij})}{\sum_{i=1}^{m}(X_{ij} + M_{ij})}$ and $B_{ij} = 1 - \frac{|X_{ij} - M_{ij}|}{(X_{ij} + M_{ij})}$.

As $\sum_{i=1}^{m}|X_{ij} - M_{ij}| \geq \left|\sum_{i=1}^{m}(X_{ij} - M_{ij})\right|$, we have that $B_j \geq B_j^*$

---

6 $B_j = 1 - A_j$.

As we commented before, the question to consider is which procedure is more suitable. If the trade flows to aggregate are not homogeneous from the intra-industry trade point of view (that is, if they are not belonging to the same *industry*) it doesn't seem desirable trade imbalance to cancel among themselves. In this case, B_j^* would be a more convenient measure to use. On the other hand, if trade flows to aggregate are homogeneous, the use of B_j would be preferable. Nevertheless, the problem in practice consists of finding out if a group of products included under a concrete level of aggregation by a classification system can be considered as homogeneous from the perspective of intra-industry trade. That is to say, if this commodity group constitutes an *industry*. International trade flows classification systems may include under the same group products with very different factor requirements. This gives rises to the denominated categorical aggregation problem, which will be treated in section 9.3.

Moreover, as B_j measures the degree of trade overlapping in a commodity group or industry *j* related to total trade in the industry, the behaviour of the index is affected by changes in total trade among industries and/or over time. For example, a high degree of trade overlapping may be registered in a specific industry in which the volume of trade is relatively small compared to total trade in manufactures. Besides, the volume of trade in an industry may or may not vary directly with the level of production or the level of sales.

Other features of index B_j are its simetry regarding X_j and M_j and its non-linearity[7].

A different way of addressing the question of intra-industry trade measurement is provided by Vona (1991). This author takes as elementary units the 5 digit SITC items[8] and points out that, for this high level of disaggregation, all trade is to be considered as intra-industry trade or inter-industry trade depending on the existence of two way flows[9]. Vona proposes the following intra-industry trade index for bilateral exports between two countries A and B in a commodity group *i*:

$$I_{A,B,i} = X_{A,B,i} + X_{B,A,i}, \text{ if } X_{A,B,i} \quad 0 \text{ and } X_{B,A,i} \quad 0$$

$$I_{A,B,i} = 0, \text{ if } X_{A,B,i} = 0 \text{ or } X_{B,A,i} = 0$$

where

$X_{A,B,i}$ = exports from country A to country B in commodity group *i* (5 digit SITC item)

$X_{B,A,i}$ = exports from country B to country A in commodity group *i* (5 digit SITC item)

---

7 For an illustration of these features of B_j see Carrera (1996).

8 Standard International Trade Classification.

9 According to this approach industries may be classified under two categories. A first group of industries is characterised by scale economies, product differentiation and imperfect competition, giving rise to intra-industry trade. A second group is formed by industries under perfect competition producing homogeneous goods (according to the H-O-S model), which leads to inter-industry trade.

Vona calculates an aggregated index for 3 digit SITC groups *j* from 5 digit SITC items *i* as follows:

$$I_{A,B,j} = \frac{\sum_{i=1}^{n} I_{A,B,i}}{X_{A,B,j} + X_{B,A,j}} \cdot 100$$

where

$X_{A,B,j}$ = exports from country A to country B in commodity group *j* (3 digit SITC)

$X_{B,A,j}$ = exports from country B to country A in commodity group *j* (3 digit SITC)

Vona's indicator takes values in the interval [0, 100]. It is 0 when all $I_{A,B,i} = 0$ and there is no intra-industry trade and it is 100 when all $I_{A,B,i}$ 0 and all trade is intra-industry trade.

As a feature of this indicator Vona mentions that it is not affected by the existence of trade imbalances, as all trade (imports and exports) is considered intra-industry trade or inter-industry trade independent of the exact coincidence of the volumes of imports and exports[10]. Nevertheless, the problem of sensitivity of the index to the number of items included in every industry still remains unsolved. That is, the intra-industry index tends to decrease as the number of subgroups included in an industry increases.

9.1.2 Intra-industry Specialisation Indices

Michaely (1962) proposed the following indicator to measure the degree of overlapping between the share of exports and the share of imports in a commodity group *j*:

$$\overline{H} = 1 - \frac{1}{2} \cdot \sum_{j=1}^{n} \left| \frac{X_j}{\sum_{j=1}^{n} X_j} - \frac{M_j}{\sum_{j=1}^{n} M_j} \right| .$$

The indices ranges from 0 to 1[11], higher values suggesting a higher degree of similarity in the composition of exports and imports. A value of 0 would imply

---

[10] According to this point of view it is not the degree of overlapping between imports and exports what matters, but the mere existence of exchange of products in both directions.

[11] Note that it is feasible for $\overline{H}$ to reach unity without global trade balance ($\sum_{j=1}^{n} X_j = \sum_{j=1}^{n} M_j$) or without trade balance in industry *j* ($X_j = M_j$). Actually $\overline{H}$ =1 when the following condition is satisfied:

perfect inter-industry trade, with non-existence of simultaneous exports and imports in a specific commodity group for a country during the time period considered[12].

Balassa (1966) introduces the concept of "revealed" comparative advantage and proposes the following indicator to measure it:

$$RCA = \frac{X_j}{X_{gj}} / \frac{X}{X_g}$$

where $\frac{X_j}{X_{gj}}$ indicates the share of a country in world exports of commodity *j* and $\frac{X}{X_g}$ is the share of a country in world exports of manufactured products.

When the value of the index tends to 0, it would be indicating a "revealed" comparative disadvantage for industry *j*, while a value above 1 would identify a comparative advantage in the correspondent industry[13]. From a general point of view, this index is measuring similarity in trade patterns and thus it can be employed to measure intra-industry specialisation[14]. As we will show later, the index is, in fact, equivalent to the measure proposed by Glejser, Goossens and Vanden Eede (1982) aimed at measuring intra-industry specialisation in foreign trade.

Finger and Kreinin (1979) develop an indicator that uses the share of every industry's exports in a country's total exports to measure the degree of similarity of export patterns from two countries (*a* and *b*) to a third market (*c*). The indicator was computed as follows:

$$\frac{X_j}{\sum_{j=1}^{n} X_j} = \frac{M_j}{\sum_{j=1}^{n} M_j}.$$

12 The index originally proposed by Michaely was actually the following:

$$D = \sum_{j=1}^{n} \left| \frac{X_j}{\sum_{j=1}^{n} X_j} - \frac{M_j}{\sum_{j=1}^{n} M_j} \right|.$$

This indicator ranges from 0 to 2. Index $\overline{H}$ (which ranges from 0 to 1) is just an adaptation of D in order to make it easier comparison to other measures.

13 A value of RCA = 1.1, for example, would indicate that the share of a country in exports of commodity *j* is 10% higher than its share in total exports of manufactures.

14 Aquino (1978) is using the standard deviation of Balassas's index as a measure of the intensity of inter-industry specialisation in a country. A deviation of 0 would indicate the absence of inter-industry specialisation while higher dispersion of the values of the index would point at higher inter-industry specialisation.

$$S\,(ab, c) = \sum_{j=1}^{n} \min\,[X_j\,(ac), X_j\,(bc)],$$

where X_j (ac) is the share of industry *j* in the exports from country *a* to country *c* and X_j (bc) is the share of industry *j* in the exports from country *b* to country *c*. The indicator takes the same values than the index of Michaely (1962) and, in fact, both measures are equivalent, as it is shown in Kol (1988). For this reason, Michaely's index can be used to compare trade patterns from many points of view. It can be employed, for instance, to compare imports and exports patterns in a country, exports (or imports) patterns for two countries or a group of countries, among them or related to a third market.

Glejser, Goossens and Vanden Eede (1979) mark a break with previous methodology and propose a conceptually different approach to quantify the magnitude and variations of intra-industry trade. Unlike other measures that consider simultaneously imports and exports, these authors make a distinction between supply specialisation (exports) and demand specialisation (imports). The indicators are built on the assumption that a country is specialised in a specific industry when it exports (or imports) relatively more than a group of conveniently selected countries.

The comparative supply specialisation for a given time period can be computed as follows:

$$\xi = \frac{1}{n} \sum_{j=1}^{n} \log\left(\frac{X_j}{X} \Big/ \frac{X_{gj}}{X_g}\right) = \frac{1}{n} \sum_{j=1}^{n} \xi_j$$

where:

n = number of industries considered
X_j = exports from industry *j* made from a country to the group of countries considered
X = total exports from a country to the group of countries considered
X_{gj} = total exports from industry *j* made by the group of countries considered (except for the country that is being analysed)
X_g = total exports from the group of countries considered (except for the country that is being analysed)

The index of comparative demand specialisation is similarly computed as follows:

$$\mu = \frac{1}{n} \sum_{j=1}^{n} \log\left(\frac{M_j}{M} \Big/ \frac{M_{gj}}{M_g}\right) = \frac{1}{n} \sum_{j=1}^{n} \mu_j$$

where:

n = number of industries considered
M_j = imports from industry *j* made from a country to the group of countries considered
M = total imports from a country to the group of countries considered

M_{gj} = total imports from industry *j* made by the group of countries considered (except for the country that is being analysed)
M_g = total imports from the group of countries considered (except for the country that is being analysed)

A higher divergence between X_j/X and X_{gj}/X_g and between M_j/M and M_{gj}/M_g is to be expected with a higher degree of inter-industry specialisation. On the other hand, if intra-industry specialisation is dominating, the quotiens $(X_j/X)/(X_{gj}/X_g)$ and $(M_j/M)/(M_{gj}/M_g)$ approach unity in every industry *j* and, consequently, the unweighted means $\xi = \frac{1}{n}\sum_{j=1}^{n}\xi_j$ and $\mu = \frac{1}{n}\sum_{j=1}^{n}\mu_j$ approach zero.

The variability among industries, that is, the variances of ξ and μ can be computed as follows:

$$S_\xi^2 = \frac{1}{n}\sum_{j=1}^{n}(\xi_j - \xi)^2$$

$$S_\mu^2 = \frac{1}{n}\sum_{j=1}^{n}(\mu_j - \mu)^2$$

Kol and Mennes (1986) and Kol (1988) establish an equivalence between this indicator and the one introduced by Balassa (1966) to measure "revealed" comparative advantage based on the share of exports from every industry in total world exports[15]. Although the purpose of those authors was not the same, both indicators are able to measure "revealed" comparative advantage and intra-industry specialisation in the form the authors define such concepts.

9.1.3 Comparison of Measures

In previous subsections we have examined two groups or families of measures and evaluated the most commonly used indices of intra-industry trade and specialisation. The first group of indicators is mainly based on the degree of overlapping of imports and exports in every industry[16]. The second group of measures takes into account the degree of similarity in relative structure of imports and exports.

Concerning the first family of measures, the following considerations can be mentioned:

- The index S_j introduced by Verdoorn (1960) has the following shortcomings:
  a) Every fraction and its inverse measure the same degree of intra-industry

[15] Operating in RCA we obtain the following relationship with the index of Glejser, Goossens and Vanden Eede:
$\xi_j = \log(\text{RCA})$.

[16] Except for the index of Vona (1991).

trade, making it complicated to compare among industries.

b) The index is not giving a direct measure of intra-industry trade as a proportion of total trade.

- The index D_j introduced by Kojima (1964) is solving the former problem in a) but the problem in b) remains unsolved.
- The measure A_j introduced by Balassa (1966) has, compared to the previous ones, the advantage that it gives information on the proportion of total trade that is of the intra-industry type. However, it has the unattractive feature of being inversely related to the degree of intra-industry trade. Moreover, the aggregated measure $\overline{A}_j$ is giving the same weight to all industries, independent on their share in total trade.
- Grubel and Lloyd (1975) introduce the index B_j, which is the most widely used in the empirical measurement of intra-industry trade. The B_j index shares some features of the A_j index of Balassa (1966) but has the advantage of being directly related to the level of intra-industry trade. Moreover, the aggregated measure $\overline{B}_j$ is preferable to $\overline{A}_j$ because it takes into account the share of every industry in total trade unlike the later, which gives the same weight to every industry.
- The index $I_{A,\,B,\,i}$ of Vona (1991) gives a different approach to the problem of intra-industry trade measurement. It has the advantage of not being influenced by trade imbalance. However, its use at the empirical level is complicated because the information on the high level of data disaggregation required[17] is not easily available and it is difficult to handle.

From the previous analysis it can be concluded that the indicators proposed by Grubel and Lloyd (1975) are the more convenient among those indices that form the first group of measures.

Concerning the second family of measures a distinction can be made among those that are using differences in relative shares of imports and exports in total respective flows (Michaely, 1962; Finger-Kreinin, 1979) and those, which are using quotients among such shares (Balassa, 1966; Glejser, Goossens y Vanden Eede, 1979). Following Kol (1988) and with the purpose of making it easier to compare both groups of measures we will first express the index of Balassa (1966) in terms of the indicator of Michaely (1962):

$$RCA = \frac{X_j}{X} / \frac{M_j}{M}$$

This indicator has some shortcomings. First, it has to be corrected in case $M_j = 0$, a case not uncommon at high levels of data disaggregation. Second, while a total similarity of trade patterns gives a value zero for the index, in case of increasing dissimilarity the value of the indicator tends to zero or ∞. This unlimited range of values complicates the interpretation of results concerning the degree of dissimilarity of trade patterns. A third problem is the following: when changing numerator and denominator we obtain completely different values of the index

[17] Vona proposes the 5-digit SITC level of aggregation.

(and the variance of the sample). The degree of similarity is however the same[18]. Finally, the use of quotients to compare trade patterns gives rise to a fourth problem. Imports and exports with a comparatively small value but very different shares in total respective flows may influence the value of the index in an undesirable way, by increasing it in case of high similarity of exports and imports patterns[19].

On the other hand, the index provided by Michaely (1962) has none of these disadvantages. Values of $X_j = 0$ or $M_j = 0$ are not a problem. The interpretation of results is easier because the index range from 0 (complete dissimilarity) to 1 (complete similarity). The third and fourth problems do not apply here because differences are used rather than quotients. Because of all these reasons it can be concluded that, when trying to measure the degree of similarity of trade patterns, it is preferable to use Michaely's index. As stated before, this indicator has a wide range of applications in numerous situations and it is suitable to compare trade patterns from diverse points of view.

Two criteria have been thus basically used when it comes to measuring intra-industry trade's intensity: the magnitude of trade flows and the similarity in relative structure of imports and exports. These two criteria give rise to two families of indices, which are not strictly comparable. However, Silber and Broll (1990) show that both types of measures[20] can be alternatively expressed as measures of distance or measures of similarity, deriving the corresponding family of indices. The empirical section of their work shows that, although the two groups of indices measure different types of distances or equalities, both are highly correlated.

9.1.4 Later Developments

In later years a great deal of work on IIT measurement has concentrated in two refinements of the traditional Grubel-Lloyd index. One of them is aimed at measuring what has been called marginal IIT, an issue related to the analysis of adjustment costs associated to trade liberalisation. The other one is targeted at adjusting IIT measures to disentangle vertical and horizontal IIT.

9.1.4.1 Measuring Marginal Intra-industry Trade

It is generally assumed that the costs of the adjustment process following trade liberalisation may differ depending on whether the new trade generated is classified as inter-industry or intra-industry trade. If emerging trade has an intra-industry nature the adjustment costs will be probably lower, because in this case reallocation is to take place within every industry rather than among different

[18] Using the logarithm of the indicator may solve this third problem. This is, in fact, what Glejser, Goossens and Vanden Eede are doing. In this case, when changing numerator and denominator, the indicator changes its sign but not its value. However, the aggregated index still ranges from 0 to . The second problem remains thus unsolved.

[19] Kol (1988) is giving an illustrative example of this feature.

[20] That is, those that are based on the degree of overlapping of trade flows and those based on the similarity of trade patterns.

industries. When analysing the adjustment consequences of trade expansion, what is relevant is not the current level of IIT but the composition of changes in exports and imports. In other words, how IIT changes at the margin.

To capture this important issue, Hamilton and Kniest (1991) developed a measure, which calculated the share of IIT in new trade flows as follows:

$$\text{MIIT} = \begin{cases} \dfrac{X_t - X_{t-n}}{M_t - M_{t-n}} \text{ for } M_t - M_{t-n} > X_t - X_{t-n} > 0 \\ \dfrac{M_t - M_{t-n}}{X_t - X_{t-n}} \text{ for } X_t - X_{t-n} > M_t - M_{t-n} > 0 \end{cases}$$

where X_j and M_j refer to exports and imports of commodity group *j* and *t* and *t-n* refer to the two points in time taken into consideration. The index is calculating the proportion of the increase in imports or exports that is matched. When all new trade is matched (the increase in imports totally matches the increase in exports for a particular industry) the index will be unity. When all new trade is inter-industry trade (there is no matching at all between new imports and new exports for a concrete industry) the index will be zero.

The Hamilton and Kniest measure, although being a considerable improvement for the analysis of adjustment issues, has some important shortcomings, as several authors have pointed out [see, for example, Brülhart (1994) and Greenaway, Hine, Milner and Elliot (1994)]. First of all, the index cannot be calculated when the change in exports or imports is negative. Second, the measure is unscaled, that is to say, it is not related to total amount of new trade or to the initial level of trade or to the value of production in the industry considered. Third, the index measures the changes in nominal terms rather than in real ones. To address these shortcomings the aforementioned authors have proposed several alternatives providing a menu of indices (which deal with the scaling problem and are always defined) to choice depending on the purpose of the job at hand.

The Brülhart (1994) index has been widely used in later works on adjustment issues. This *dynamic* measure can be expressed as follows:

$$A_i = 1 - \frac{|\Delta X_i - \Delta M_i|}{|\Delta X_i| + |\Delta M_i|} = 1 - \frac{|(X_t - X_{t-n}) - (M_t - M_{t-n})|}{|X_t - X_{t-n}| + |M_t - M_{t-n}|},$$

where X_j and M_j refer to exports and imports of commodity *i* and *t* and *t-n* refer to the two points in time taken into consideration. The index has values ranging from 0 to 1, a value of 0 indicating that new trade is entirely of the inter-industry type and a value of 1 showing complete matching between new exports and imports in industry *j* and so lower transitional adjustment costs. The measure is dynamic in the sense that it provides information on the proportion of changes in total trade flows that are of an intra-industry type.

Recent work by Thom and McDowell (1999) shows that the proposed marginal intra-industry trade indicators [such as Brülhart (1994)] may be underestimating the extent of intra-industry trade as they cannot distinguish between inter-industry trade and vertical intra-industry trade. These authors propose an alternative

procedure of classifying marginal trade flows based on the joint application of the Brülhart's index and the aggregate measure defined as follows:

$$A_j = 1 - \frac{\left|\Delta X_j - \Delta M_j\right|}{\sum_{i=1}^{n}\left|\Delta X_i\right| + \sum_{i=1}^{n}\left|\Delta M_i\right|}$$

where A_j measures the extent of total intra-industry trade in industry *j*, which has *n* subindustries, aggregating vertical and horizontal intra-industry trade. The difference between A_j and Brülhart's index would give vertical trade and inter-industry trade would be given by the residual.

9.1.4.2 Disentangling Vertical and Horizontal Intra-industry Trade

The distinction between horizontal and vertical differentiation of product is an important one when dealing with adjustment issues. As several authors point out [see Greenaway, Hine and Milner (1994, 1995)] horizontal IIT is likely to lead to lower adjustment pressures than vertical IIT. This is so since different industry and country characteristics can be associated with the exchange of products depending on whether such products involve similar or different qualities (horizontal or vertical differentiation).

Greenaway, Hine and Milner (1994, 1995) propose a methodology to identify vertical and horizontal IIT built upon previous work of Abd-el-Rahman (1991). This approach assumes that quality is reflected in price and price can be proxied by unit values. The share of vertical and horizontal IIT in total trade is computed as follows:

$$IIT_j^h = \left[1 - \frac{\sum_{i=1}^{n}\left|X_{ij}^h - M_{ij}^h\right|}{\sum_{i=1}^{n}\left(X_{ij}^h + M_{ij}^h\right)}\right] \cdot \frac{\sum_{i=1}^{n}\left(X_{ij}^h + M_{ij}^h\right)}{\left(X_j + M_j\right)}, \tag{9.1}$$

$$IIT_j^v = \left[1 - \frac{\sum_{i=1}^{n}\left|X_{ij}^v - M_{ij}^v\right|}{\sum_{i=1}^{n}\left(X_{ij}^v + M_{ij}^v\right)}\right] \cdot \frac{\sum_{i=1}^{n}\left(X_{ij}^v + M_{ij}^v\right)}{\left(X_j + M_j\right)}, \tag{9.2}$$

where *i* refers to the 5th digit SITC products in a commodity group or industry *j*.

Total IIT is then decomposed in vertical IIT and horizontal IIT as follows:

$$IIT_j^* = IIT_j^v + IIT_j^h$$

$$IIT_j^* = 1 - \frac{\sum_{i=1}^{n} \left| X_{ij} - M_{ij} \right|}{\sum_{i=1}^{n} \left(X_{ij} + M_{ij} \right)}$$

Horizontal intra-industry trade (IIT_j^h) is given by (9.1) for the items i in commodity group j where the following condition holds:

$$1 - \alpha \leq \frac{UV_{ij}^x}{UV_{ij}^m} \leq 1 + \alpha ,$$

where UV^x and UV^m refers to unit values of exports and imports and α is a given dispersion factor.

Vertical intra-industry trade (IIT_j^v) is given by (9.2) for the items i in commodity group j where the following condition holds:

$$\frac{UV_{ij}^x}{UV_{ij}^m} < 1 - \alpha \quad \text{or} \quad \frac{UV_{ij}^x}{UV_{ij}^m} > 1 + \alpha .$$

The dispersion factor α has been given several values in the diverse works where this procedure is employed[21] giving different price wedges within which IIT is considered as horizontal and outside of which it is defined as vertical. However, even when large price wedges have been used, vertical IIT has been found to be very significant in empirical studies. The results obtained suggest that the distinction between horizontal and vertical IIT should be taken into account in econometric models, as the determinants of both types of IIT may differ[22].

9.2 Trade Imbalance Adjustment

9.2.1 Grubel and Lloyd (1975)

One of the shortcomings of the aggregated index $\overline{B}_j$ is that, in case of global trade imbalance (that is, $\sum_{j=1}^{n} X_j \quad \sum_{j=1}^{n} M_j$), it can introduce a downward bias in intra-

21 Abd-el-Rahman (1991) uses $\alpha = 0.15$ and Greenaway, Hine and Milner (1994, 1995) employ both $\alpha = 0.15$ and $\alpha = 0.25$.

22 Some work on this subject can be found in Menon, Greenaway and Milner (1999).

industry trade measurement. This is due to the fact that, under this assumption, $\sum_{j=1}^{n}|X_j - M_j| > 0$ and, therefore, $\overline{B}_j$ can never reach its maximum value unity.

To deal with this problem, Grubel and Lloyd (1975) proposed an alternative adjusted measure, which expresses intra-industry trade as a proportion of total trade minus the value of trade imbalance (in absolute terms), as follows:

$$\overline{C}_j = \frac{\sum_{j=1}^{n}(X_j + M_j) - \sum_{j=1}^{n}|X_j - M_j|}{\sum_{j=1}^{n}(X_j + M_j) - \left|\sum_{j=1}^{n}(X_j - M_j)\right|} = \frac{\overline{B}_j}{1-K}$$

where $K = \dfrac{\left|\sum_{j=1}^{n}(X_j - M_j)\right|}{\sum_{j=1}^{n}(X_j + M_j)}$

and where $0 \leq \overline{C}_j \leq 1$[23].

As it can be observed, $\overline{C}_j$ increases as the value of K (value of trade imbalance as a proportion of total trade) gets higher. On the other hand, the index refers to aggregated trade flows and it has no counterpart for individual industries. Moreover, when trade imbalance has the same sign in all industries, $\overline{C}_j$ reaches the value 1, independent of the size of such trade imbalances[24].

In short, the adjusted measure $\overline{C}_j$ intends to indicate what the level of intra-industry trade would be when no trade imbalance would exist. Nevertheless, Greenaway and Milner (1986), among others, argue that the need for the index to be adjusted when that is not the case should be based on wider theoretical grounds than the mentioned functional restriction of $\overline{B}_j$. Specifically, it has to be

---

[23] Note that $\sum_{j=1}^{n}(X_j + M_j) \geq \sum_{j=1}^{n}|X_j - M_j| \geq \left|\sum_{j=1}^{n}(X_j - M_j)\right|$.

[24] Note that, in this case, $\left|\sum_{j=1}^{n}(X_j - M_j)\right| = \sum_{j=1}^{n}|X_j - M_j|$.

The index reaches thus its maximum value unity even when not all trade is intra-industry trade in the analysed country. Vona (1991) points out that this problem is not, nevertheless, very important at the empirical level. The probability for this problem to appear is inversely related to the number of elementary items used to compute $\overline{B}_j$ or $\overline{C}_j$, what makes it difficult to happen when using a high level of data disaggregation.

considered that a situation of global macroeconomic equilibrium may be compatible with the existence of trade imbalance for a group of industries. Moreover, in case we actually have a situation of disequilibrium, the direction of the adjustment may be different from the one proposed by Grubel and Lloyd. The restoring forces could imply an increase in trade balance and, in this case, $\overline{B}_j$ would be upwards biased.

Finally, it is worthy to mention that Grubel and Lloyd's argumentation on the need to correct in the presence of global trade imbalance implies the classification of trade flows in three categories: inter-industry trade flows, intra-industry trade flows and trade imbalances. The third of these categories is considered as a "disturbing" factor that has to be excluded of the analysis. Nevertheless, and if we take into account that most empirical works on intra-industry trade refer to the manufacturing sector, it is difficult to find a theoretical reason why trade in manufactured products should be balanced. A country may have, by instance, a deficit in this sector while having a superavit in other sectors. This matter becomes more evident if we consider bilateral trade flows, because it is not clear which bilateral position would be consistent with a situation of multilateral equilibrium.[25].

The critics to the Grubel and Lloyd's adjustment refer not only to the suitability of the proposed procedure [Aquino (1978)], but also to the very need to adjust global trade imbalance [Greenaway and Milner (1981, 1983), Kol and Mennes, 1983). Other contributions to the debate may be found in Greenaway (1984), Pomfret (1985), Greenaway and Milner (1986, 1987), Kol (1988), Kol and Mennes (1989) and Vona (1990, 1991).

9.2.2 Aquino (1978)

Aquino (1978) argues that, if $\overline{B}_j$ then there is a downwards biased when there exists a trade imbalance. It is precisely because individual indices B_j are themselves downward biased measures of intra-industry trade at the industry level. Consequently, this author proposes to adjust individual indices B_j under the assumption of equiproportional distribution of trade imbalance effects among all industries. The adjustment procedure begins by estimating the value of exports and imports in case no global trade imbalance would exist. For this purpose, the following expressions are computed:

$$X_{je} = X_j \cdot \frac{1}{2} \cdot \frac{\sum_{j=1}^{n} (X_j + M_j)}{\sum_{j=1}^{n} X_j} \tag{9.3}$$

[25] The equilibrium condition could be thus imposed only when measuring intra-industry trade including total exchange of goods and services between a country and the rest of the world (excluding, to simplify, capital flows).

$$M_{je} = M_j \cdot \frac{1}{2} \cdot \frac{\sum_{j=1}^{n}(X_j + M_j)}{\sum_{j=1}^{n} M_j} \tag{9.4}$$

Operating in (9.3) and (9.4) we have:

$$\sum_{j=1}^{n} X_{je} = \sum_{j=1}^{n} M_{je} = \frac{1}{2} \cdot \sum_{j=1}^{n}(X_j + M_j)$$

Using these "artificial" values the following intra-industry trade index for commodity group *j* can be computed:

$$Q_j = 1 - \frac{|X_{je} - M_{je}|}{(X_{je} + M_{je})} = 1 - \frac{\left| \frac{X_j}{\sum_{j=1}^{n} X_j} - \frac{M_j}{\sum_{j=1}^{n} M_j} \right|}{\left(\frac{X_j}{\sum_{j=1}^{n} X_j} + \frac{M_j}{\sum_{j=1}^{n} M_j} \right)}$$

Finally, based on individual indices Q_j the following aggregated index is computed:

$$\overline{Q}_j = \frac{\sum_{j=1}^{n}(X_j + M_j) - \sum_{j=1}^{n}|X_{je} - M_{je}|}{\sum_{j=1}^{n}(X_j + M_j)} \tag{9.5}$$

Examining (9.5) it can be deduced that this indicator uses as a criterion for measurement the degree of similarity of relative shares of imports and exports. Taking into account this feature, an equivalence between $\overline{Q}_j$ and the Michaely's indicator $\overline{H}$ can be obtained as follows:

$$\overline{Q}_j = \frac{\sum_{j=1}^{n}(X_j + M_j) - \sum_{j=1}^{n}|X_{je} - M_{je}|}{\sum_{j=1}^{n}(X_j + M_j)} =$$

$$= \frac{\sum_{j=1}^{n}(X_j + M_j) - \frac{1}{2} \cdot \sum_{j=1}^{n}(X_j + M_j) \cdot \sum_{j=1}^{n} \left| \frac{X_j}{\sum_{j=1}^{n} X_j} - \frac{M_j}{\sum_{j=1}^{n} M_j} \right|}{\sum_{j=1}^{n}(X_j + M_j)} =$$

$$= 1 - \frac{1}{2} \cdot \sum_{j=1}^{n} \left| \frac{X_j}{\sum_{j=1}^{n} X_j} - \frac{M_j}{\sum_{j=1}^{n} M_j} \right| = \overline{H}.$$

The use of this adjustment procedure does not imply a variation in total trade flows as the following holds:

$$\sum_{j=1}^{n}(X_j + M_j) = \sum_{j=1}^{n}(X_{je} + M_{je}) \tag{9.6}$$

Exports and imports patterns are also preserved after adjustment, that is:

$$\frac{X_{je}}{\sum_{j=1}^{n} X_{je}} = \frac{X_j}{\sum_{j=1}^{n} X_j}$$

and

$$\frac{M_{je}}{\sum_{j=1}^{n} M_{je}} = \frac{M_j}{\sum_{j=1}^{n} M_j}.$$

The above equality (9.6), however, does not hold necessarily at the industry *j* level. In fact, when total trade is not balanced, the fulfilment of $(X_{je}+M_{je} = X_j+M_j)$ implies that:

$$\frac{M_j}{\sum_{j=1}^{n} M_j} = \frac{X_j}{\sum_{j=1}^{n} X_j}$$

On the other hand, as Greenaway and Milner (1986) point out, the direction of the adjustment of B_j for every industry depends on the relationship between the sign of trade imbalance in the industry and total trade imbalance.

$$Xj > M_j \text{ and } \sum_{j=1}^{n} X_j > \sum_{j=1}^{n} M_j \text{ imply that } Q_j > B_j;$$

$$Xj > M_j \text{ and } \sum_{j=1}^{n} X_j < \sum_{j=1}^{n} M_j \text{ imply that } Q_j < B_j;$$

$$Xj < M_j \text{ and } \sum_{j=1}^{n} X_j > \sum_{j=1}^{n} M_j \text{ imply that } Q_j > B_j;$$

$$Xj < M_j \text{ and } \sum_{j=1}^{n} X_j < \sum_{j=1}^{n} M_j \text{ imply that } Q_j < B_j.$$

This variability in the direction of the correction affects the ranking of industries according to their intra-industry trade index. The election of the group of transactions, which will be the basis for the adjustment of trade imbalance, becomes thus critical, especially when the adjusted index is to be used in econometric analysis. Other critics to this procedure are the following: first, the fact that a trade imbalance in a group of industries is considered as a disequilibrium situation and, second, the assumption that the effects of restoring forces are equiproportionally distributed among industries[26]. For these reasons, it would be advisable, when analysing the determinants of intra-industry trade, a sensible selection of the time periods included in the sample rather than the use of adjusted indices. Another alternative proposed by the aforementioned authors is the use of an average measure of the indices corresponding to a carefully selected period of time.

9.2.3 Balassa (1979)

Balassa (1979, 1986) utilises a correcting procedure similar to the one used by Aquino (1978). However, while the latter is using as a basis for adjustment the group of transactions analysed (manufactured products, total trade or other sets of transactions), Balassa only takes into account total trade imbalance[27].

Technically, the Balassa adjustment is identical to Aquino's correction:

---

[26] Aquino (1981) argues that such a feature can be acceptable in absence of other elements that help to identify an alternative criterion.

[27] Balassa (1986) justifies this selection "in order to allow for inter-industry specialisation between primary and manufactured products".

$$X_j^b = X_j \cdot \frac{1}{2} \cdot \frac{\sum_{j=1}^{n}(X_j + M_j)}{\sum_{j=1}^{n} X_j}$$

and

$$M_j^b = M_j \cdot \frac{1}{2} \cdot \frac{\sum_{j=1}^{n}(X_j + M_j)}{\sum_{j=1}^{n} M_j}$$

Like in the previous case, the correction is equiproportionally distributed among industries[28]. However, the election of a specific trade imbalance as the basis for the adjustment represents an improvement regarding the problem of ambiguity commented for Aquino (1978).

9.2.4 Loertscher and Wolter (1980)

Loertscher and Wolter (1980) use the same correcting procedure as Aquino (1978). However, they apply it to bilateral trade flows of manufactured products instead of total trade of a country with the rest of the world. We have already mentioned, nevertheless, the lack of theoretical support for the existence of bilateral equilibrium with specific countries in a concrete set of transactions. Greenaway and Milner (1981) comment that, particularly, bilateral disequilibrium may be the result of factors leading to inter-industrial and intra-industrial specialisation. It may be thus inconvenient to adjust on the basis of bilateral disequilibria because it can hide the influences that one pretends to measure.

9.2.5 Bergstrand (1982)

Finally, Bergstrand (1982) considers total trade imbalance as the basis for adjustment, as in Balassa (1979), but, opposite to the later, he uses bilateral trade flows to measure intra-industry trade[29].

The indicator proposed for the measurement of bilateral intra-industry trade between countries *i* and *j* in products of industry *k*[30] is the following:

---

28 In case of a deficit (superavit), all exports are increased (diminished) in the same proportion. The same applies for imports.

29 Bergstrand justifies this decision arguing that in a model with multiple countries, multiple goods, two factors of production and non equality of factor prices, the Heckscher-Ohlin theorem is always fulfilled for bilateral trade flows between pairs of countries and thus the focus of interest should be the presence of bilateral intra-industry trade, which is "non expected" according to the mentioned theorem.

30 Adjusted on the basis of each country's multilateral trade imbalance.

$$IIT_{ij}^{k*} = 1 - \frac{\left|X_{ij}^{k*} - X_{ji}^{k*}\right|}{\left(X_{ij}^{k*} - X_{ji}^{k*}\right)}$$

where

$$X_{ij}^{k*} = \frac{1}{2} \cdot \left[\frac{(X_i + M_i)}{2X_i} + \frac{(X_j + M_j)}{2M_j}\right] \cdot X_{ij}^{k}$$

$$X_{ji}^{k*} = \frac{1}{2} \cdot \left[\frac{(X_j + M_j)}{2X_j} + \frac{(X_i + M_i)}{2M_i}\right] \cdot X_{ji}^{k}$$

X_{ij}^{k} = value of exports from country *i* to country *j* of products in industry *k*

$X_i = \sum_{k=1}^{n} \sum_{j=1}^{m} X_{ij}^{k}$ = total exports of country *i*

$M_i = \sum_{k=1}^{n} \sum_{j=1}^{m} X_{ji}^{k}$ = total imports of country *I*

Finally, X_{ji}^{k}, X_j and M_j are similarly defined.

Taking, for example, the case of exports of industry *k* from country *i* to country *j* it can be noted that the correcting factor $\frac{1}{2} \cdot \left[\frac{(X_i + M_i)}{2X_i} + \frac{(X_j + M_j)}{2M_j}\right]$ is an arithmetic mean of two elements. The first of them, $\frac{(X_i + M_i)}{2X_i}$, tends to increase exports from *i* when such country presents a global trade deficit ($X_i+M_i > 2X_i$) and to reduced them in the opposite case. Regarding the second element, $\frac{(X_j + M_j)}{2M_j}$, it tends to increase exports from *i* to *j* when country *j* has global trade superavit ($X_j+M_j > 2M_j$) and to reduce them in the opposite case. A similar analysis applies to X_{ji}^{k}.

In short, both elements of the correcting factor act in the same direction when one of the countries has a deficit and the other one has a superavit. However, both elements tend to offset when the sign of trade imbalance is the same in the two countries.

On the other hand, it can be noted that the correcting factors for X_{ij}^{k} and X_{ji}^{k} are independent on the industry considered k[31]. This means that the correction is equiproportionaly applied to all industries, the same as in Aquino's and Balassa's procedures.

The correcting procedure stops when countries i and j present a multilateral equilibrium (independent on the existence of bilateral disequilibria). It is thus an iterative procedure to compute X_{ij}^{k*} and X_{ji}^{k*} in which bilateral trade flows are changed until all countries reach a multilateral equilibrium. The amount of information required to apply this method is much higher than the one needed by the other indicators commented.

9.2.6 Comments and Conclusions

We have revised in previous subsections different methods, which can be applied to correct for trade imbalance, either in global trade or in specific sectors, for bilateral trade flows or multilateral exchanges.

The correction of the index $\overline{B}_j$ was introduced to allow the indicator to reach its maximum value unity, even in the presence of trade imbalance. However, this argument has no theoretical foundations. Moreover, the adjusted index $\overline{C}_j$ has no counterpart at the industry level, as opposite to indices B_j and $\overline{B}_j$. Finally, we have to mention that, in case all trade imbalances have the same sign in all industries, the index $\overline{C}_j$ reaches its maximum value 1 independent on the size of such trade imbalances[32].

Concerning the Aquino (1978) index, although it is an improving compared to the previous one, it suffers from several shortcomings. In concrete, we have already commented that the ranking of industries according to their level of intra-industry trade experiments significant variations, depending on the trade balance of the set of transactions considered as a basis for the adjustment. On the other hand, the correction is equiproportionaly distributed among industries, what may imply the lost of important information on specific industry characteristics related to intra-industry trade. Moreover, the Aquino (1978) index always takes the same value when the share of imports and exports in every industry in total imports and exports, respectively, remains constant.

In Aquino (1978) and Loertscher and Wolter (1980) the adjustment is made on the basis of trade imbalance in manufactured products but, again, there is *a priori*

---

31 The exports of every industry k are multiplied by the same factor and the same applies for imports.

32 This is due to the following:

$$\left|\sum_{j=1}^{n}(X_j - M_j)\right| = \sum_{j=1}^{n}\left|X_j - M_j\right|.$$

no theoretical foundations for trade in a specific group of transactions to be balanced.

Loertscher and Wolter (1980) correct in terms of bilateral trade flows but the situations of bilateral disequilibrium may be consistent with multilateral equilibrium and, in fact, they may be reflecting a trend towards inter or intra-industry specialisation. For this reason, it is even more hazardous to use and adjustment procedure of the Aquino (1978) type on a bilateral basis because it may hide the very influences that one pretends to measure.

In our opinion, a convenient method to correct for trade imbalance should satisfy, at least, the following conditions: first, it should be applicable to several levels of data aggregation; second, it has to use total trade as a basis for the correction (and not only a particular set of transactions) and third, it has to simulate a situation of multilateral equilibrium. Taking all this into account, the methods proposed by Balassa (1979) and Bergstrand (1982) seem to be more suitable than the other procedures commented. Balassa's adjustment has, compared to Bergstrand, the advantage of its simplicity.

Nevertheless, none of the exposed procedures considers the causes underlying the disequilibrium situations considered, conveniently adjusting the procedure[33]. In fact, these methods could create more distortions than the ones they pretend to eliminate and have no clear relationship with theoretical foundations. However, in some cases, particularly when it comes to comparing among countries with different trade imbalances or for a country with very different trade imbalances along the time period considered, the use of adjusted indices may be convenient. Some alternative methods may be used [Greenaway and Milner (1981)]. A sensible selection of the years considered could be made, which avoids periods of evident global trade imbalance. On the other hand, an average of intra-industry trade indices may be used for a carefully selected time period.

In any case, the distortions that trade imbalance may introduce in the measurement of intra-industry trade are probably insignificant when compared with those brought about by the definition of "industry" and the consequent effects of categorical aggregation. We deal with this question in next section 9.3.

9.3 Categorical Aggregation

9.3.1 Definition of the Problem

One of the most important problems of intra-industry trade's measurement is the unknown influence of categorical aggregation. The identification and measurement of intra-industry trade clearly depends on the degree of homogeneity of the products classified under the same group in trade statistics and the nature of such homogeneity. The bias introduced by categorical aggregation on intra-industry trade measurement emerges when essentially heterogeneous products

[33] The equiproportional distribution of the correction among industries reflects the lack of flexibility of the proposed procedures.

(that is, products than cannot be considered belonging to the same industry) are classified under the same group in trade statistics. In general, the existence of bi-directional trade flows for a commodity group will identify "genuine" intra-industry trade and statistical aggregation (simultaneous exchange of products classified under the same category but with different factor requirements).

The most convenient procedure to eliminate this problem seems to be, at first sight, the reclassification of data in a way that all resulting categories correspond as close as possible with the theoretical construction of an "industry". The problem is that there are multiple criteria of regrouping the data and, even is a specific criterion is systematically used, reclassification is an extremely arduous job[34].

International trade data may be classified according to different criteria. The most commonly used system is the SITC nomenclature[35], in which products are grouped at various levels of aggregation identified by digits. The alternative to reclassification would be the election of a specific level of data disaggregation in official classifications as the best approximation to the concept of industry. In this case, a double question emerges. On the one hand, the most convenient level of data disaggregation has to be identified (that is, the level that more narrowly corresponds to the concept of "industry"). On the other hand, the influence of categorical aggregation for the selected level of data disaggregation has to be determined.

9.3.2 Categorical Aggregation Tests

Due to the difficulties of product reclassification under a systematic and coherent with the economic criterion procedure, several alternatives have been proposed. Particularly, three alternative methods to assess the influence of categorical aggregation on intra-industry trade indices have been suggested[36]:

i) To perform measurement at higher levels of statistical disaggregation.
ii) To execute diverse measurements based on different classification systems.
iii) To compute an adjusted intra-industry trade index.

The first of these procedures has been the most widely employed. It is based on the consideration that a substantial fall in average intra-industry indices when descending from a specific level of aggregation to a higher one may be indicating the presence of categorical aggregation problem.

However, an important point has to be taken into account. There is no specific standard pattern to evaluate if the fall in the value of the index is significant and so the results of this test are not at all conclusive but have an indicative character[37].

---

[34] An additional problem would be the classification of pieces and components.

[35] Standard International Trade Classification.

[36] Greenaway and Milner (1983).

[37] Usually, a fall in the value of the index of about 20% when passing from the 3-digit SITC to the 5-digit SITC level of aggregation, while maintaining a reasonably high volume of intra-industry trade for the 5-digit level, is considered to indicate that the main component of recorded intra-industry trade is not categorical aggregation.

Concerning the second type of categorical aggregation test mentioned, it is also instructive to observe the sensibility of indices B_j to classification systems based on different criteria (like processing characteristics or product features). The stability in the ranging of indices computed using different classifications may be illustrative but, like in the previous case, no conclusions can be obtained regarding the absolute significance of the categorical aggregation error. Moreover, the need to select a specific level of data aggregation for the indices' computation still remains.

The third type of categorical aggregation test involves a more systematic way of evaluating the categorical aggregation problem associated with trade imbalances of opposite sign in subgroups of products with different factor *ratios* and low degree of substitution[38]. If we considered commodity group *j* formed by *m* subgroups denoted by *i*, this procedure consists of computing an adjusted intra-industry trade index according to the following expression:

$$B_j^* = \sum_{i=1}^{m} w_{ij} \cdot B_{ij} = \frac{\sum_{i=1}^{m}(X_{ij} + M_{ij}) - \sum_{i=1}^{m}|X_{ij} - M_{ij}|}{\sum_{i=1}^{m}(X_{ij} + M_{ij})}$$

where $w_{ij} = \dfrac{X_{ij} + M_{ij}}{\sum_{i=1}^{m}(X_{ij} + M_{ij})} = \dfrac{X_{ij} + M_{ij}}{X_j + M_j}$ and $B_{ij} = 1 - \dfrac{|X_{ij} - M_{ij}|}{X_{ij} + M_{ij}}$.

The aggregation procedure to compute the intra-industry trade index for commodity group *j* consists of computing B_{ij} indices for the subgroups *i* and averaging them according to the share of every subgroup in total trade of *j*.

As $\sum_{i=1}^{m}|X_{ij} - M_{ij}| \geq \left|\sum_{i=1}^{m}(X_{ij} - M_{ij})\right|$ holds, we have $1 \geq B_j \geq B_j^* \geq 0$.

This means that, the more trade imbalances of opposite sign cancel between subgroups, the higher B_j will be related to B_j^* and, in case all trade imbalances in the subgroups have the same sign, we have $B_j = B_j^*$.

The adjustment is thus based on the assumption that categorical aggregation is associated with opposite sign trade imbalances in the subgroups *i* of which a commodity group *j* is composed. When this situation exists, a measurement of intra-industry trade at a higher level of data disaggregation generates a lower value of intra-industry trade and helps to correct for the influence of statistical aggregation.

---

[38] See subsection 9.1.1.

9.3.3 Critics and Conclusions

The available evidence collected in the literature devoted to examine the influence of statistical aggregation on intra-industry trade measurement and, specifically, the existence of high volumes of intra-industry trade even at high levels of data disaggregation, shows that this is a phenomenon that cannot be explained solely by statistical aggregation.

There is no doubt that the interpretation of intra-industry trade measurement becomes more complicated with the existence of categorical aggregation. The more convenient way of facing this problem could be the reclassification of data building product categories as homogeneous as possible, (that is, according to what can be considered an "industry"). However, this regrouping is problematic due to the lack of a unique reclassification criterion, among other factors.

The election of a concrete level of data disaggregation to measure intra-industry trade is not either exempt of problems. If we select a too low level of data disaggregation most intra-industry trade registered may be due to categorical aggregation. On the other hand, when the selected level of disaggregation is too high, we run the risk of losing the economic meaning of the subgroups employed. Some products with similar factor requirements may appear in different groups and part of the recorded inter-industrial trade would be, in fact, intra-industry trade. In most empirical works on intra-industry trade it is considered that the 3-digit SITC level is a reasonable approximation to the economic concept of "industry".

With the purpose of assessing the influence of categorical aggregation on intra-industry trade measurement at a particular level of data disaggregation, several procedures have been proposed. The first of them consists of performing measurement at higher levels of statistical disaggregation. Among other shortcomings, it has the problem of dealing with the elevated volume of information needed to compute indices at high levels of data disaggregation. The second procedure, consisting of computing indices based on different classification systems, is similarly arduous. Moreover, the problem of selecting a concrete level of disaggregation to perform the measurements still persists. Finally, the third procedure involves the computation of an adjusted intra-industry trade index: B_j^*. This adjusted index is giving a more convenient measure of intra-industry trade than index B_j in the presence of categorical aggregation. Yet, if categorical aggregation is not associated with opposite sign trade imbalances in the subgroups, then B_j^* would be a downward biased intra-industry trade measure. In the absence of information about the existence of categorical aggregation, it is not possible to know which of both measures is more suitable.

Finally, it is worthy to mention that none of the mentioned procedures gives conclusive results regarding the determination of the aggregation error's absolute relevance. However, it is convenient to employ at least one (or more) of these methods for indicative purposes, in order to reduce, in part, the arbitrary indiscriminate use of a specific level of data disaggregation to compute intra-industry trade indices.

9.4 Summary and Conclusions

In this chapter we have tried to carry out a critical review of the main indicators used for the purpose of intra-industry trade and specialisation measurement. Moreover, we have discussed the two main problems related to such measurement: trade imbalance adjustment and categorical aggregation.

The analysis performed may be summarised in the following points:

i) Two different criteria have been identified concerning intra-industry exchange, giving rise to two families of indices (not extrictly comparable): the degree of overlapping in trade flows and the degree of similarity in trade patterns (or relative structure of imports and exports). Among the indicators belonging to the first group, the indices proposed by Grubel and Lloyd (1975)[39] seem to be the more suitable. Concerning the second group of indicators, it can be concluded that the index of Michaely (1962)[40] is the most convenient measure when it comes to measuring the degree of similarity in trade patterns.

ii) Regarding the procedures aimed at correcting the distortions introduced by trade imbalance in intra-industry trade measurement, a common feature to all of them is the lack of theoretical support to justify the need for adjustment. It is thus preferable to use unadjusted indices for carefully selected periods of time.

iii) The problem of categorical aggregation remains unsolved. Although there is no definitive solution, several methods have been developed to test for the influence of categorical aggregation. With the possible exception of the product regrouping method, these tests are easy to perform and, even if not conclusive, they have an indicative character, which allows us to eliminate in part the arbitrariness of the use of a specific level of data disaggregation when measuring intra-industry trade.

iv) Two issues have dominated the literature on intra-industry trade measurement in the last years: marginal intra-industry trade and disentanglement of vertical and horizontal intra-industry trade.

The measurement of the so-called marginal intra-industry trade is related to the analysis of adjustment costs associated to trade liberalisation. The indices developed are based on the assumption that when assessing the relevance of intra-industry trade during an adjustment process, it is the proportion of intra-

---

[39] $B_j = 1 - \frac{|X_j - M_j|}{(X_j + M_j)}$ and $\overline{B}_j = 1 - \frac{\sum_{j=1}^{n} |X_j - M_j|}{\sum_{j=1}^{n} (X_j + M_j)}$.

[40] $\overline{H}_j = 1 - \frac{1}{2} \cdot \sum_{j=1}^{n} \left| \frac{X_j}{\sum_{j=1}^{n} X_j} - \frac{M_j}{\sum_{j=1}^{n} M_j} \right|$.

industry trade in new generated trade what matters, rather than the level of intra-industry trade.

On the other hand, the distinction between vertical and horizontal intra-industry trade is an important one as the determinants of both may differ. Diverse indices, which disentangle horizontal and vertical intra-industry trade, have been developed and used in empirical analysis.

Finally, it is important to mention that recent work on this subjects addresses the issue that the indicators of marginal intra-industry trade so far developed may be underestimating the extent of intra-industry trade, as they identify intra-industry trade with horizontal trade only. In other words, they cannot distinguish between inter-industry trade and vertical intra-industry trade. A method to deal with this question can be found in Thom and McDowell (1999).

References

Abd-el-Rahman, K.: Firms' Competitive and National Comparative Advantages as Joint Determinants of Trade Composition. Weltwirtschaftliches Archiv 127, 83-97 (1991)

Aquino, A.: The Measurement of Intra-industry Trade when Overall Trade is Imbalanced. Weltwirtschaftliches Archiv 117, 763-766 (1981)

Balassa, B.: Tariff Reductions and Trade in Manufactures among the Industrial Countries. The American Economic Review 56, 466-73 (1966)

Balassa, B.: Intra-industry Trade and the Integration of Developing Countries in the World Economy. In: Giersch, H. (ed.): On the Economics of Intra-industry Trade. Symposium, Tübingen 1979

Balassa, B.: The Determinants of Intra-industry Specialisation in United States Trade. Oxford Economic Papers 38, 220-233 (1986)

Bergstrand, H. J.: The Scope, Growth and Causes of Intra-Industry International Trade. New England Economic Review (Sept/Oct), 45-61 (1982)

Brülhart, M.: Marginal Intra-industry Trade: Measurement and Relevance for the Pattern of Industrial Adjustment. Weltwirtschaftliches Archiv 130-3, 600-613 (1994)

Carrera, M. G.: Comercio Intra-Industrial: Análisis del Caso Español. Ph. D. Dissertation. University of Cantabria. Department of Economics 1996

Finger, J. M. and Kreinin, M. E.: A measure of Export Similarity and its Possible Uses. The Economic Journal 89, 905-912 (1979)

Glejser, H., Goossens, K. and Vanden Eede, M.: Inter-industry versus Intra-industry specialisation in Exports and Imports (1959-1970-1973). Journal of International Economics 12, 363-369 (1982)

Greenaway, D.: The Measurement of Product Differentiation in Empirical Studies of Trade Flows. In: Kierzkowski, H. (ed.): Monopolistic Competition and International Trade. Oxford University Press 1984

Greenaway, D., Hine, R. C. and Milner, C.: Country-Specific Factors and the Pattern of Horizontal and Vertical Intra-industry Trade in the UK. Weltwirtschaftliches Archiv 130, 77-100 (1994)

Greenaway, D., Hine, R. C. and Milner, C.: Vertical and Horizontal Intra-industry Trade: a Cross-Industry Analysis For The United Kingdom. The Economic Journal 105, 1505-1518 (1995)

Greenaway, D., Hine, R. C., Milner, C. and Elliot, R.: Adjustment and the Measurement of Marginal Intra-industry Trade. Weltwirtschaftliches Archiv 130-2, 418-427 (1994)

Greenaway, D. and Milner, C.: Trade Imbalance Effects in the Measurement of Intra-industry Trade. Weltwirtschaftliches Archiv 117, 756-762 (1981)

Greenaway, D. and Milner, C.: On the Measurement of Intra-industry Trade. The Economic Journal 93, 900-908 (1983)

Greenaway, D. and Milner, C.: The Economics of Intra-industry Trade. Basil Blackwell 1986

Greenaway, D. and Milner, C.: Intra-industry Trade: Current Perspectives and Unresolved Issues. Weltwirtschaftliches Archiv 123, 39-57 (1987)

Grubel, H. G., Lloyd, P. J.: The Empirical Measurement of Intra-industry Trade. Economic Record 47, 494-517 (1971)

Grubel, H. G., Lloyd, P. J.: Intra-industry Trade: the Theory and Measurement of International Trade in Differentiated Products. London: Macmillan 1975

Hamilton, C. and Kniest, P.: Trade Liberalisation, Structural Adjustment and Intra-industry Trade: a Note. Weltwirtschaftliches Archiv 127-2, 356-367 (1991)

Kojima, K.: The Pattern of International Trade among Advanced Countries. Hitotsubashi Journal of Economics (June), 16-36 (1964)

Kol, J.: The Measurement of Intra-industry Trade. Erasmus University. Rotterdam 1988

Kol, J. and Mennes, L. B. M.: On Concepts and Measurement of Intra-industry Trade. Discussion Paper 66. Centre for Development Planning. Erasmus University, Rotterdam 1983

Kol, J. and Mennes, L. B. M.: Intra-industry specialisation: some Observations on Concepts and Measurement. Journal of International Economics 21. 173-181 (1986)

Kol, J. and Mennes, L. B. M.: Corrections for Trade Imbalance: a Survey. Weltwirtschaftliches Archiv 125, 703-717

Loertscher, R. and Wolter, F.: Determinants of Intra-industry Trade among Countries and across Sectors. Weltwirtschaftliches Archiv 116, 280-292

Menon, J., Greenaway, D. and Milner C.: Industrial Structure and Australia-UK Intra-industry Trade. The Economic Record 75-228, 19-27 (1999)

Michaely, M.: Multilateral Balancing in International Trade. The American Economic Review 52, 685-702 (1962)

Silber, J. and Broll, U.: Trade Overlap and Trade Pattern Indices of Intra-industry Trade: Theoretical Distinctions versus Empirical Similarities. Sonderforschungsbereich 178, Internationalisierung der Wirtschaft. Diskussionsbeiträge Serie II, N. 107. Fakultät für Wirtschaftswissenschafter und Statistik 1990

Thom, R. and McDowell, M.: Measuring Marginal Intra-industry Trade. Weltwirtschaftliches Archiv 135 (1), 48-61 (1999)

Verdoorn, P. J.: The Intra-block Trade of Benelux. In: Robinson, E. A. G. (ed.): Economic Consequences of the Size of Nations, 291-329. Macmillan 1960

Vona, S.: Intra-industry Trade: a Statistical Artefact or a Real Phenomenon?. Banca Nazionale del Lavoro Quarterly Review 175 (1990)

Vona, S.: On the Measurement of Intra-industry Trade: some further Thoughts. Weltwirtschaftliches Archiv 127, 678-700 (1991)

10 Measurement of Intra-industry Trade: a Categorical Aggregation Exercise with Spanish Trade Data

G. Carrera-Gómez
University of Cantabria (Spain)

10.1 Introduction

One of the features which characterises recent international trade flows is the prevalence of intra-industry trade. This phenomenon may be defined as simultaneous imports and exports of products belonging to the same industry and constitutes an important part of total world trade. The lack of concordance between the predictions of the traditional trade theory and the empirical evidence found has given rise to the appearance of new theories, which try to explain the structure and patterns of international trade by introducing imperfect competitive markets where the existence of scale economies and diversity of consumer preferences constitute essential ingredients.

The assessment of the relevance of the new theories of trade as opposed to the traditional theoretical approach, when it comes to explain real trade patterns, is essentially an empirical question. Nevertheless, there are two problems (which remain unsolved) which influence the results obtained when measuring intra-industry trade and specialisation. The first problem refers to the very existence of the phenomenon and is related to the definition of "industry" and the selection of the level of data disaggregation more appropriate to study such a phenomenon. The second problem, closely connected to the former, comes from the objective difficulty of finding a convenient quantitative measure.

This chapter is devoted to the analysis of one of the main obstacles faced by empirical treatment of intra-industry trade: the problem of categorical aggregation. The aim of this study is to effectuate a measurement of Spanish intra-industry

P. Coto-Millán (ed.), *Essays on Microeconomics and Industrial Organisation*

trade for several groups of products and country areas and to evaluate the influence of categorical aggregation applying several tests. With this purpose, we introduce first in section 10.2 the methodology employed. We offer next in section 10.3 the main results obtained. Finally, in section 10.4 we present a summary and the main conclusions reached regarding this aspect of intra-industry trade measurement.

10.2 Methodology

10.2.1 Measurement of Intra-Industry Trade

Intra-industry trade may be defined as simultaneous imports and exports of products belonging to the same industry. As stated in previous chapter, a variety of intra-industry trade indicators has been introduced in the literature. Among them, we have employed the most widely used, which is the Grubel and Lloyd (1975) index. This index measures the share of intra-industry trade in total trade flows of an industry *j*, which is computed by the following expression:

$$B_j = \frac{(X_j + M_j) - |X_j - M_j|}{(X_j + M_j)} = 1 - \frac{|X_j - M_j|}{(X_j + M_j)}$$

where $0 \leq B_j \leq 1$.

The average of indices B_j for a set of industries (j=1, 2, 3, ..., n) is computed as follows:

$$\overline{B}_j = \sum_{j=1}^{n} w_j \cdot B_j = \sum_{j=1}^{n} \frac{(X_j + M_j)}{\sum_{j=1}^{n}(X_j + M_j)} \cdot \frac{(X_j + M_j) - |X_j - M_j|}{(X_j + M_j)} =$$

$$= \frac{\sum_{j=1}^{n}(X_j + M_j) - \sum_{j=1}^{n}|X_j - M_j|}{\sum_{j=1}^{n}(X_j + M_j)} = 1 - \frac{\sum_{j=1}^{n}|X_j - M_j|}{\sum_{j=1}^{n}(X_j + M_j)}.$$

The problem in practice lies in finding out if a group of products included under a concrete level of aggregation by a trade classification system can be considered as homogeneous from the perspective of intra-industry trade. That is to say, if this

commodity group constitutes an *industry*[1]. This gives rises to the denominated categorical aggregation problem, which will be treated in next section.

10.2.2 Categorical Aggregation: Definition of the Problem and Assessment Methods

One of the most important problems of intra-industry trade's measurement is the unknown influence of categorical aggregation. The identification and the measured levels of intra-industry trade depend on the degree of homogeneity of the products classified under the same group in trade statistics and the nature of such homogeneity. The problem brought about by categorical aggregation on intra-industry trade measurement emerges when heterogeneous products[2] are classified under the same group in trade statistics. The existence of bi-directional trade flows for a commodity group will identify "genuine" intra-industry trade and "spurious" intra-industry trade or statistical aggregation[3].

While the most convenient method to eliminate this problem seems to be the regrouping of data in a way that all resulting groups correspond as close as possible with the theoretical construction of an "industry", the problem is that there are multiple criteria of accomplishing that regrouping. And even if a specific criterion is systematically used, reclassification is an extremely arduous task[4].

International trade data may be classified according to different criteria. One of the most extended systems is the Standard International Trade Classification (SITC), in which products are grouped at various levels of aggregation identified by digits. The alternative to reclassification has been the election of a specific level of data disaggregation in official nomenclatures as the best approximation to the concept of industry. In this case, two problems have to be solved. First, the most convenient level of data disaggregation has to be identified[5]. Second, the influence of categorical aggregation for the selected level of data disaggregation has to be determined.

Taking into account the difficulties of product reclassification using a systematic and coherent with the economic criterion procedure, several alternative methods have been proposed. Particularly, three alternative methods to assess the influence of categorical aggregation on intra-industry trade measures have been suggested[6]: first, performing measurement at higher levels of statistical disaggregation; second, executing diverse measurements based on different classification systems; and third, computing an adjusted intra-industry trade index. In this work, we have begun by performing intra-industry trade measurement

---

1 Most studies made on this subject consider that the Standard International Trade Classification 3-digit groups are a reasonable approximation to the economic concept of industry, existing a certain degree of consensus in this respect.

2 That is, products which do not belong to the same industry.

3 Simultaneous exchange of products classified under the same group but with different factor requirements.

4 An additional problem would be the classification of pieces and components.

5 That is, the level that more narrowly corresponds to the concept of "industry".

6 See Greenaway and Milner (1983).

using the 3-digit groups of the SITC. Next, we have used the first and the third methods commented above to assess the influence of categorical aggregation on the obtained measures.

The first of these procedures[7] is based on the consideration that a substantial fall in average intra-industry indices when descending from a specific level of aggregation to a higher one may be indicating the presence of categorical aggregation problem. There is no specific standard pattern to evaluate if the fall in the value of the index is significant and so the results of this test are not conclusive but have an indicative character. Usually, a fall in the value of the index of about 20% when passing from the 3-digit SITC to the 5-digit SITC level of aggregation, while maintaining a reasonably high volume of intra-industry trade for the 5-digit level, is considered to indicate that the main component of recorded intra-industry trade is not categorical aggregation. Consequently, we have performed measurement at the 4-digits and 5-digits levels of data disaggregation and evaluated the fall in the 3-digits indexes of intra-industry trade.

The other type of categorical aggregation test performed involves a more systematic way of evaluating the categorical aggregation problem associated with trade imbalances of opposite sign in subgroups of products with different factor *ratios* and low degree of substitution. If we considered commodity group *j* formed by *m* subgroups denoted by *i*, this procedure consists of computing an adjusted intra-industry trade index according to the following expression:

$$B_j^* = \sum_{i=1}^{m} w_{ij} \cdot B_{ij} = \frac{\sum_{i=1}^{m}(X_{ij}+M_{ij}) - \sum_{i=1}^{m}\left|X_{ij}-M_{ij}\right|}{\sum_{i=1}^{m}(X_{ij}+M_{ij})}$$

where $w_{ij} = \dfrac{X_{ij}+M_{ij}}{\sum_{i=1}^{m}(X_{ij}+M_{ij})} = \dfrac{X_{ij}+M_{ij}}{X_j+M_j}$ and $B_{ij} = 1 - \dfrac{\left|X_{ij}-M_{ij}\right|}{X_{ij}+M_{ij}}$.

So, the aggregation procedure to compute the intra-industry trade index for commodity group *j* consists of computing B_{ij} indices for the subgroups *i* and averaging them according to the share of every subgroup in total trade of *j*[8].

As commented in previous chapter, the adjustment is thus based on the assumption that categorical aggregation is associated with opposite sign trade imbalances in the subgroups *i* of which a commodity group *j* is composed. When this situation exists, a measurement of intra-industry trade at a higher level of data

---

[7] Which has been the most widely employed.

[8] As $\sum_{i=1}^{m}\left|X_{ij}-M_{ij}\right| \geq \left|\sum_{i=1}^{m}(X_{ij}-M_{ij})\right|$ holds, we have $1 \geq B_j \geq B_j^* \geq 0$.

This means that, the more trade imbalances of opposite sign cancel between subgroups, the higher B_j will be related to B_j^* and, in case all trade imbalances in the subgroups have the same sign, we have $B_j = B_j^*$.

disaggregation generates a lower value of intra-industry trade and helps to correct for the influence of statistical aggregation.

Applying this method, adjusted indexes have been calculated for every SITC 3-digits group *j* consisting of *m* subgroups *i* (4-digits).

10.3 Main Results Obtained

Table 10.1 shows the Grubel and Lloyd intra-industry trade indexes calculated for several levels of data aggregation for the 10 sections of SITC and for the aggregates of primary products, manufactured products and total products. The indexes have been calculated for several country aggregates: European Union, rest of world and total world.

Looking at the 3-digits indexes we can underline the importance of intra-industry trade in Spanish trade flows, especially if we consider the exchange of manufactured products with the EU countries. According to these indexes, more than 60 per cent of total trade is of the intra-industry type. The share of intra-industry trade is particularly high for some sections, including beverages and tobacco, manufactured goods classified chiefly by material, machinery and transport equipment and chemicals and related products. Additionally, we can observe that intra-industry trade seems to be more significant in the exchange of manufactured products than in primary products for all the country aggregates considered and that the proportion of this trade is higher for trade with the EU countries than for the rest of the world.

As can be seen from table 10.1, the influence of categorical aggregation varies depending on the group of products and the country aggregates considered. Considering the narrowest groups of products (5-digits), we can indicate that the share of intra-industry trade is still relatively high in the exchange of manufactured products (58.0 per cent for total trade and 59.8 per cent for trade with the EU). These figures are high enough to sustain that observed intra-industry trade in this case is not only the result of categorical aggregation[9].

The proportional fall experimented by intra-industry trade index when passing from 3-digits to 5-digits reaches a value of 9.9 per cent for the case of exchanges with EU countries and 10.3 per cent for total trade, being lower, in general, than the one observed for other countries. This fact, together with the maintenance of high proportions of intra-industry trade even for the highest disaggregation levels, confirms the existence of "genuine" intra-industry trade and not only categorical aggregation for the 3-digits groups of SITC:

On the other hand, comparing the effects of aggregation on the groups of primary products and manufactured products, we observe that the former shows the influence of this problem with higher intensity when we consider the exchanges with EU countries, while the opposite holds for trade with the rest of the world. The less intense falls in the indexes are to be found for the exchange of manufactured products with the EU.

---

[9] That is, the heterogeneity of products classified under the same industry.

Table 10.1. $\overline{B}_j$ intra-industry trade indexes for several data aggregation levels. Spanish SITC trade data 1992

	EU-12			Rest of World			Total World		
SITC Rev. 3 Sections	**3 dig.**	**4 dig.**	**5 dig.**	**3 dig.**	**4 dig.**	**5 dig.**	**3 dig.**	**4 dig.**	**5 dig.**
0 Food and live animals	0.357	0.283 (7.4)*	0.243 (4.0)** (11.4)***	0.377	0.242 (13.5)	0.169 (7.3) (20.8)	0.415	0.317 (9.8)	0.265 (5.2) (15.0)
1 Beverages and tobacco	0.848	0.164 (68.4)	0.115 (4.9) (73.3)	0.272	0.247 (2.5)	0.116 (13.1) (15.6)	0.781	0.229 (55.2)	0.152 (7.7) (62.9)
2 Crude materials, inedible, except fuels	0.523	0.447 (7.6)	0.350 (9.7) (17.3)	0.207	0.165 (4.2)	0.127 (3.8) (8.0)	0.439	0.374 (6.5)	0.281 (9.3) (15.8)
3 Mineral fuels, lubricants and related materials	0.667	0.457 (21.0)	0.406 (5.1) (26.1)	0.132	0.117 (1.5)	0.082 (3.5) (5.0)	0.240	0.193 (4.7)	0.181 (1.2) (5.9)
4 Animal and vegetable oils, fats and waxes	0.539	0.497 (4.2)	0.388 (10.9) (15.1)	0.251	0.209 (4.2)	0.207 (0.2) (4.4)	0.401	0.363 (3.8)	0.325 (3.8) (7.6)
5 Chemicals and related products, n.e.s.	0.560	0.537 (2.3)	0.473 (6.4) (8.7)	0.718	0.557 (16.1)	0.429 (12.8) (28.9)	0.640	0.588 (5.2)	0.507 (8.1) (13.3)
6 Manufactured goods classified chiefly by material	0.750	0.651 (9.9)	0.566 (8.5) (18.4)	0.580	0.450 (13.0)	0.370 (8.0) (21.0)	0.732	0.639 (9.3)	0.561 (7.8) (17.1)
7 Machinery and transport equipment	0.707	0.684 (2.3)	0.665 (1.9) (4.2)	0.572	0.503 (6.9)	0.456 (4.7) (11.6)	0.696	0.671 (2.5)	0.648 (2.3) (4.8)
8 Miscellaneous manufactured articles	0.550	0.519 (3.1)	0.478 (4.1) (7.2)	0.426	0.347 (7.9)	0.304 (4.3) (12.2)	0.525	0.483 (4.2)	0.446 (3.7) (7.9)
9 Commodities and transactions not classified elsewhere in the SITC	0.032	0.032 (0)	0.032 (0) (0)	0.030	0.030 (0)	0.030 (0) (0)	0.034	0.034 (0)	0.034 (0) (0)
Total primary products (0 to 4)	0.465	0.326 (13.9)	0.272 (5.4) (19.3)	0.228	0.171 (5.7)	0.121 (5.0) (10.7)	0.390	0.286 (10.4)	0.238 (4.8) (15.2)
Total manufactured products (5 to 9)	0.667	0.638 (2.9)	0.598 (4.0) (6.9)	0.555	0.460 (9.5)	0.398 (6.2) (15.7)	0.668	0.623 (4.5)	0.580 (4.3) (8.8)
Total products (0 to 9)	0.639	0.583 (5.6)	0.540 (4.3) (9.9)	0.441	0.359 (8.2)	0.301 (5.8) (14.0)	0.602	0.544 (5.8)	0.499 (4.5) (10.3)

* Fall in intra-industry trade index 3-digits to 4-digits
** Fall in intra-industry trade index 4-digits to 5-digits
*** Fall in intra-industry trade index 3-digits to 5-digits

Source: Own elaboration with trade data provided by Hamburg Institute for Economic Research

Another fact to be mentioned is that the measures obtained for trade in manufactured products with non-communitarian countries are more affected (that is, they suffer higher falls) when passing to higher disaggregation levels than those corresponding to trade with EU countries. On the other hand, the influence of categorical aggregation is more significant in the measurement of trade in primary products with communitarian countries than with the rest of the world except for

section 0 (food and live animals). This fact suggests that trade with these geographical areas involves different types of products.

Focusing on the effects that categorical aggregation has on the different sections of SITC, we observe that this problem does not affect equally to all of them.

Within the category of manufactured products, section 7 (machinery and transport equipment) registers the lowest falls in intra-industry trade index when passing to higher disaggregation level.

A case to be equally mentioned is section 5 (chemicals and related products). When measuring intra-industry trade at the 3-digits level we obtain for this section, contrary to the general trend, an index of intra-industry trade for non-communitarian countries higher than the one obtained for trade with EU countries. However, when we carry out the measurement at a higher disaggregation level (5-digits), this characteristic stops being and exception. While in the case of EU trade this section is less affected by categorical aggregation, the fall experimented by the share of intra-industry trade with non-communitarian countries reaches 28.9 per cent. This suggests, on the other hand, that intra-industry exchanges within this section affect different types of products depending on the group of countries considered.

Concerning primary products two cases deserve attention. On the one hand, section 0 (food and live animals), for which a higher intra-industry trade was obtained for trade with non-EU countries, loses this character when carrying out measurement at higher levels of disaggregation. Contrary to what can be observed for the rest of primary products, this section is more affected by categorical aggregation in the exchanges with non-communitarian countries than with EU. On the other hand, section 1 (beverages and tobacco), which showed an unusually high level of intra-industry trade when measured at the 3-digits level, seems to be dramatically affected by categorical aggregation, experimenting falls in the indexes of 73.3 per cent (for EU countries) and 62.9 per cent (for total trade) when measured at the 5-digits level.

With regard to adjusted intra-industry trade indexes test, table 10.2 shows the percentage of industries which present falls in intra-industry trade index lower than those mentioned when using adjusted index B_j^* instead of index B_j[10].

The results of the calculations accomplished allow to state that categorical aggregation does not have a big influence in the measurement carried out, according to this test.

From the information contained in table 10.2, we can indicate that more than 93 per cent of industries present decreases in intra-industry trade lower than 25 per cent when using adjusted indexes for the case of total trade. This figure is of 92.66 per cent for EU trade and more than 85 per cent for exchanges with rest of the world. Approximately 70 per cent of SITC groups (63 per cent for non-communitarian countries) present a decrease lower than 5 per cent in the share of intra-industry trade when using adjusted indexes. Even if we observe the number of industries in which intra-industry trade falls less than 1 per cent when using

[10] The full series of indexes are available at Carrera-Gómez (1996).

adjusted indexes, it accounts for more than 60 per cent (52 per cent for non-EU countries).

Table 10.2. Percentage of industries with intra-industry trade decrease after using adjusted index B_j^*

Reduction of intra-industry trade measured by B_j	Percentage of industries		
	EU-12	Rest of World	Total World
less than 25%	92.66	85.33	93.44
less than 20%	89.19	81.08	89.58
less than 15%	84.94	77.61	83.78
less than 10%	80.31	70.27	78.76
less than 5%	70.66	62.55	69.50
less than 1%	60.23	51.74	62.16

Source: Own elaboration

Finally, we can mention that correlation coefficients between the two series of indexes (unadjusted and adjusted) have been calculated for all the country groups considered. The high value of the coefficients (higher than 0.9 in all cases) indicates that the order of industries under the criterion of intra-industry trade intensity does not vary substantially when using adjusted indexes. This suggests that they are likely to produce similar results to those corresponding to unadjusted indexes when being used as dependent variable in cross-section analyses.

10.4 Summary and Conclusions

This work is devoted to one of the main controversies concerning intra-industry trade: the problem of categorical aggregation. This problem is caused by the fact that essentially heterogeneous products are grouped together in trade statistics introducing a bias in intra-industry measurement. The more convenient way of facing this problem could be the reclassification of data, building product categories as homogeneous as possible, (that is, according to what can be considered an "industry"). However, this regrouping is problematic due to the lack of a unique reclassification criterion, among other factors.

In this study we have first performed a measurement of intra-industry trade at the 3-digits level of the SITC, considering that this level is a reasonable approximation to the economic concept of "industry".

Next, with the purpose of assessing the influence of categorical aggregation on the measurement carried out, several procedures have been employed. The first of them consisted of performing measurement at higher levels of statistical disaggregation (4-digits and 5-digits levels of the SITC). The results obtained confirm the maintenance of an elevated proportion of intra-industry trade even

when measured at the highest disaggregation levels for the case of manufactured products. The share of intra-industry trade in primary products is considerably lower, being worthy to mention that, in this case, measurement is not highly affected by categorical aggregation either except for particular cases. A featured to be highlighted is the fact that manufactured products show a higher influence of categorical aggregation in trade with non-EU countries than in exchanges with the rest of the world, while the opposite holds for trade in primary products. On the other hand, intra-industry trade in manufactured products seems to be less affected by categorical aggregation than intra-industry trade in primary products for trade with communitarian countries, as opposed to the situation presented by trade with the rest of the world. This suggests that, as can be expected, trade with these geographical areas involves different kinds of products.

The second procedure employed involved the computation of adjusted intra-industry trade indexes B_j^*, which may give a more convenient measure of intra-industry trade than index B_j in the presence of categorical aggregation associated with opposite sign trade imbalances in the subgroups. The results of the calculations accomplished allow to state that, according to this test, categorical aggregation does not have a big influence in the measurement performed.

In conclusion, the evidence shown by the data and the results of the tests performed and, specifically, the existence of high volumes of intra-industry trade even at high levels of data disaggregation, demonstrates that this is a phenomenon that cannot be explained solely by statistical aggregation.

Finally, we have to mention that the problem of categorical aggregation remains unsolved. Although there is no definitive solution, several methods have been developed to test for the influence of this problem on intra-industry trade measurement. While none of them gives conclusive results regarding the determination of the aggregation error's absolute relevance, it is convenient to employ at least one (or more) of these methods for indicative purposes, in order to reduce, in part, the arbitrary indiscriminate use of a specific level of data disaggregation to compute intra-industry trade indices. With the possible exception of the product regrouping procedure, these tests are not difficult to perform[11] and, even if not conclusive, they have an indicative character, which allows us to eliminate in part the arbitrariness of using a specific level of data disaggregation when measuring intra-industry trade.

References

Carrera-Gómez, G.: Comercio Intra-Industrial: Análisis del Caso Español. Ph. D. Dissertation. University of Cantabria. Department of Economics 1996

Greenaway, D. and Milner, C.: On the Measurement of Intra-industry Trade. The Economic Journal 93, 900-908 (1983)

Grubel, H. G., Lloyd, P. J.: Intra-industry Trade: the Theory and Measurement of International Trade in Differentiated Products. London: Macmillan 1975

[11] Appart from the shortcoming of dealing with the elevated volume of information needed to compute indexes at high levels of data disaggregation.

11 The Determinants of Intra-industry Trade in Spanish Manufacturing Sectors: a Cross-section Analysis

G. Carrera-Gómez
University of Cantabria (Spain)

A wide range of studies devoted to recent evolution of international trade highlights the fact that a significant and increasing proportion of flows consists of bi-directional change of products belonging to the same industry. This phenomenon is known as intra-industry trade.

Among the several interesting aspects arising from the study of this type of trade we can mention the following. On the one hand, it is difficult to find a satisfactory explanation for this phenomenon inside the traditional theoretical frame. On the other hand, it is important to analyse the factors determining this kind of trade as they differ from the conventional determinants of international trade and, consequently, the effects of trade policies and structural adjustments may be different to those obtained under the traditional assumptions. The former question has given rise to the appearance of a wide range of explanatory models, what gives an idea of the diversity of conditions under which this phenomenon may arise[1]. On the empirical side we find, on the one hand, a set of studies aimed at determining the relative importance of intra-industry trade through the use of several methods to measure it. On the other hand, we have a series of papers whose purpose is to identify the industry and country characteristics leading to intra-industry trade[2].

This paper is among the latter and it is aimed at establishing the factors explaining intra-industry trade in the Spanish economy, from an industry perspective, given the increasing relevance that this type of flows has acquired in

[1] A comprehensive review of this literature can be found in Greenaway and Milner (1986), Tharakan and Kol (1989) and Carrera (1996).

[2] See, for example, Greenaway and Milner (1984), Balassa and Bauwens (1988), Fariñas and Martin (1988), Bano (1991), Hugues (1993) and Montaner and Orts (1995).

P. Coto-Millán (ed.), *Essays on Microeconomics and Industrial Organisation*

Spanish foreign trade[3]. With this purpose, in section 11.1 we deal whith the measurement of Spanish intra-industry trade and calculate intra-industry trade indexes for several industrial sectors. Section 11.2 contains an analysis of the determinant factors of Spanish intro-industry trade including empirical contrast of a series of hypotheses related to industry characteristics which may influence the extent of intra-industry trade in Spanish manufacturing industrial sectors. Finally, section 11.3 synthesises the main conclusions of the paper.

11.1
Spanish Intra-industry Trade Measurement

A variety of intra-industry trade indicators have been introduced in the literature, the index proposed by Grubel and Lloyd (1975) being the most widely used for this purpose. The G-L indicator expresses the proportion of intra-industry trade in a given industry, which will be referred as *i*, as follows:

$$B_i = 1 - \frac{|X_i - M_i|}{(X_i + M_i)}$$

where $0 \leq B_i \leq 1$.

For a set of *m* industries an aggregated index can be constructed according to the following expression:

$$\overline{B} = 1 - \frac{\sum_{i=1}^{m} |X_i - M_i|}{\sum_{i=1}^{m} (X_i + M_i)} = \sum_{i=1}^{m} B_i \cdot \frac{(X_i + M_i)}{\sum_{i=1}^{m} (X_i + M_i)}, \qquad (11.1)$$

where $0 \leq \overline{B} \leq 1$.

From a theoretical point of view, we can consider an industry as a group of products with similar factor requirements and a high degree of substitution. Nevertheless, this is not easy to take into practice and the available international trade classification systems contain aggregations of products that can be heterogeneous from an intra-industry trade perspective[4]. As a consequence, the main problem we face when trying to measure intra-industry trade is the difficulty of defining the industry from an operative point of view. The usual way of treating this problem consists of choosing a particular level of data aggregation, measuring and evaluating the influence of statistical aggregation, often calculating intra-industry trade indexes for higher levels of data disaggregation. Most works made on this subject use the SITC 3digit groups[5] as a reasonable approximation to the economic concept of industry, being possible to talk of a certain degree of consensus in this respect.

---

[3] See Martin (1992) and Carrera (1996).

[4] This fact causes the denominated categorical aggregation problem.

[5] Standard International Trade Classification.

Table 11.1 shows the aggregated intra-industry trade indexes for Spain for a number of sectors which have been constructed from SITC 3-digit groups. Additionally, intra-industry trade indexes have been calculated for SITC 5-digit items. The range correlation coefficient between both series of indexes is 0,96, indicating that the ranking of sectors according to their intra-industry trade intensity has practically no variations between both cases.

Table 11.1. Intra-industry trade indexes for Spain (1992)

Sector	Denomination	IIT index
1	Solid fuel, coke, hydrocarbons, radioactive minerals and petroleum refineries	0.226
2	Electric power, water supply and gas	0.430
3	Metallic mineral products	0.630
4	Non-metallic mineral products	0.675
5	Chemicals	0.616
6	Fabricated metal products (except machinery and transport material)	0.806
7	Agricultural and industrial machinery	0.632
8	Office and computing machinery, optical accuracy instruments and similar tools	0.438
9	Electrical and electronic machinery and material (except computers)	0.585
10	Automobiles, pieces and accessories	0.800
11	Other transport material	0.648
12	Food, drinks and tobacco	0.519
13	Textiles, leather and footwear	0.570
14	Wood and cork	0.541
15	Paper, graphic arts and publishing	0.716
16	Rubber and plastic products	0.734
17	Other manufacturing industries	0.563

Source: Own elaboration with foreign trade data provided by Hamburg Institute for Economic Research

11.2 Analysis of the Determinants of Intra-industry Trade in Spain

11.2.1 Hypotheses

The bulk of intra-industry trade explanatory theories conforms an analytical outline from which we can extract some basic hypotheses concerning industry characteristics influencing the degree and extent of intra-industry trade.

One of the elements the literature is emphasising as responsible of intra-industry trade share of total trade in a sector is the differentiation of product. What we could call "genuine" intra-industry trade refers to the exchange of horizontally differentiated products, where the diversity of tastes and preferences of the consumers plays a significant role[6]. On the other hand, technological differentiation of the product may lead to a specialisation of an inter-industry nature, where industries better endowed with technological knowledge produce a higher number of new products and enjoy a comparative advantage in this sense. Nevertheless, if the new products are included in international trade classifications together with the old ones, the exchange of both will appear as intra-industry trade.

Consequently, the product differentiation hypothesis may be formulated on the following terms:

- Hypothesis 1: The higher the degree of product differentiation in a particular production sector the greater the proportion of intra-industry trade in the sector.

On the other hand, the presence of scale economies is considered as an important condition for the emergence of intra-industry trade in differentiated products. It is precisely this element that causes a firm finding more profitable to produce only one or a small set of varieties of the product to export, importing meanwhile the other varieties. This way the consumers are allowed to enjoy the benefits inherent to the increase in their range of election and a certain degree of intra-industry trade is generated. These issues allow us to state the following hypothesis:

- Hypothesis 2: The share of intra-industry trade is higher in those industries with larger economies of scale.

In addition to the aforementioned, several theoretical foundations may be found that induce us to expect a positive association between the participation of multinationals in a productive sector and the degree of intra-industry trade in that sector. The diverse cost minimisation strategies employed by this type of firms under certain assumptions may generate intra-industry exchange. This makes it possible to formulate the following hypothesis:

- Hypothesis 3: The proportion of intra-industry trade in a particular industrial sector is larger the more intense the activity of multinational firms operating in that sector is.

Finally, it has been frequently noted in the literature on intra-industry trade that this type of trade is more sensitive to trade barriers than traditional inter-industry trade, what allows us to suggest the following hypothesis:

- Hypothesis 4: The share of intra-industry trade is higher in those industries facing a lower degree of trade barriers.

[6] On the other hand, it is often impossible to separate vertical and horizontal differentiation in empirical work (mainly because both types of differentiation can be simultaneously found in products). In any case, a positive association between both forms of product differentiation and intra-industry trade can be considered.

11.2.2 Data, Variables and Econometric Model

The set of trade and industry data used to construct the variables employed in the analysis were recorded in various classification systems or nomenclatures[7]. Therefore, and due to the lack of official tables of correspondence among such classifications, it was necessary to carry out a previous task of conversion among the different nomenclatures with the aim of having the data homogeneity required for carrying out the work. The data were finally grouped together into the industrial sectors listed in table 11.1 (see section 11.1) and the test carried out is referred to a cross-section of these industrial sectors for the year 1992.

The construction of the variables employed was made as follows:

The dependent variable used in our analysis is the level of intra-industry trade. For every sector an aggregated index has been calculated according to expression (11.1) as it is shown in table 11.2.

Table 11.2. Dependent variable

Dependent variable	Proxy
Intra-industry trade share in total sector j trade	Grubel-Lloyd (1975) aggregated index: $\bar{B}_j = 1 - \frac{\sum_{i=1}^{m} \lvert X_{ij} - M_{ij} \rvert}{\sum_{i=1}^{m} (X_{ij} + M_{ij})}$.

Notes:
j = 1, 2, 3, ..., 17.
i = 3-digt-SITC groups for every sector j

As regards the explanatory variables, the following have been used: horizontal product differentiation, technological product differentiation, scale economies, involvement of multinational firms in the sector and level of tariff protection. The construction of the independent variables is summarised in table 11.3.

[7] The following classification systems have been used:
-Stantandar International Trade Classification, Rev. 3.
-Industrial Survey 1989-1992 sectors (National Institute of Statistics).
-National Classification of Economic Activities (CNAE-74).
-Tariff sector used in Melo and Mones (1982).
-Industrial Situation Survey (Ministery of Industry and Energy).

Table 11.3. Explanatory variables

Explanatory variable	Proxy	
Horizontal product differentiation	Advertising expenditures (AE) to sales (S) ratio	$AES_j = \frac{\sum_{l=1}^{t} AE_l}{\sum_{l=1}^{t} S_l} = \frac{AE_j}{S_j} = \sum_{l=1}^{t}\left(\frac{AE_l}{S_l}\cdot\frac{S_l}{S_j}\right)$
	Relative AE	$RAE_j = \frac{AE_j}{\sum_{j=1}^{17} AE_j}$.
Technological product differentiation	Research and development expenditure (RD) to value added (VA) ratio	$RDR_j = \frac{\sum_{r=1}^{s} RD_r}{\sum_{r=1}^{m} VA_r} = \frac{RD_j}{VA_j} = \sum_{r=1}^{s}\left(\frac{RD_r}{VA_r}\cdot\frac{VA_r}{VA_j}\right)$
	Technological development expenditure (TD) to VA ratio	$TDR_j = \frac{\sum_{r=1}^{s} TD_r}{\sum_{r=1}^{s} VA_r} = \frac{TD_j}{VA_j} = \sum_{r=1}^{s}\left(\frac{TD_r}{VA_r}\cdot\frac{VA_r}{VA_j}\right)$
Scale economies	Minimum efficient size (MES)* to cost disadvantage (CDR)** ratio	$SE_l = \frac{MES_l}{CDR_l}$; $SE_j = \sum_{l=1}^{t} SE_l \cdot \frac{P_l}{P_j}$.
Involvement of multinational firms	Foreign direct investment (FDI)	$FDI_j = \sum_{r=1}^{s} FDI_r$
	FDI to VA ratio	$FDIVA_j = \frac{\sum_{r=1}^{s} FDI_r}{\sum_{r=1}^{s} VA_r} = \frac{FDI_j}{VA_j}$
	FDI to total direct investment (TDI) ratio	$FDIT_j = \frac{FDI_j}{TDI_j}$
Level of tariff protection	Tariff protection (T)	$T_j = \frac{\frac{1}{m}\sum_{k=1}^{m} T_k \cdot M_j^R}{M_j^R + M_j^{EU}} = \frac{\frac{1}{m}\sum_{k=1}^{m} T_k \cdot M_j^R}{M_j}$

Notes:
j = 1, 2, 3, ..., 17
i = 3-digt-SITC groups for every sector j
l = INE Industrial Survey sectors for every sector j
r = CNAE-74 sectors (2 digits) for every sector j. INE Research and Development Survey
k = Melo and Mones (1982) tariff sectors for every sector j
T_k = Average Common External Tariff (Melo y Monés, 1982)
M_j^{EU} = Imports from the European Union
M_j^R = Imports from the rest of the world
M_j = Total imports for sector j
* MES is computed as the size (value of production) of the median establishment divided by the value of the sector's total production. By median establishment we mean the establishment which corresponds to the median of the accumulated distribution of the sector's production.
** RDC is computed as the quotient whose numerator is the value added per employee in establishments smaller than the MES and whose denominator is the value added per employee in the remaining establishments.

Once the variables have been constructed, we have to associate the intra-industry trade indexes with the set of industry characteristics in table 11.3 using regression analysis. But one of the difficulties we face when doing so is the limited range of

the dependent variable, which takes values inside the interval [0, 1]. Taking into account this feature, there are no guaranties that the values generated from a linear regression equation or a linear-logarithmic equation fall inside the range of the dependent variable. A Logit function will not present this problem but has the disadvantage that it cannot capture the extreme values (0 and 1), which provide relevant information. A reasonable approximation to the problem is the use of a Tobit model, which enable to incorporate this limited range characteristic of the dependent variable.

Because of the above mentioned reasons, the procedure we have employed consists of estimating a standard Tobit model, which can be expressed as follows[8]:

$$y_i^* = \beta' x_i + \varepsilon_i \; ; \; \varepsilon_i \sim N(0,\sigma^2) \; ; \text{ with } \begin{cases} y_i = 0 & \text{if } y_i^* \le 0, \\ y_i = y_i^* & \text{if } 0 < y_i^* < 1, \\ y_i = 1 & \text{if } y_i^* \ge 1. \end{cases}$$

where β is a vector of unknown parameters, x_i is the vector of explanatory variables containing industry characteristics and where the limited dependent variable, y_i, is the aggregated Grubel and Lloyd (1975) intra-industry trade index.

The procedure employed for estimation is the maximum likelihood method. As it is known, the estimators obtained from this method are sensitive to the assumptions of homoscedasticity and normality of the residuals and to deal with this problem several tests have been applied to check the compliance of these conditions[9]. For the assumption of homoscedasticity the BP statistic (Breusch y Pagan, 1980) has been used. With respect to the assumption of normality of the residuals, the BJL statistic (Bera, Jarque y Lee, 1984) has been employed. Additionally, the Shapiro-Wilk statistic (SW) for normality contrast has been calculated.

11.2.3 Results

Table 11.4 shows the main results of the regressions made. As it can be seen in the table the estimated coefficients have the expected sign and most of them are significant from a statistical point of view. Furthermore, the homoscedasticity and normality tests performed do not suggest evidence of the presence of these problems in any of the cases. Moreover, the contrast of the likelihood ratio. (LR) allows us to accept the hypothesis of joint significance of the parameters.

The results obtained provide, in general, support for the hypotheses formulated in 11.2.1 concerning the determinant factors of intra-industry trade.

The coefficient of the variable related to the degree of horizontal differentiation of the product, approximated as relative advertising expenditure has the expected sign and is highly significant in all the regressions, suggesting that this type of differentiation has a positive influence on the intensity of intra-industry trade.

On the other hand, the coefficient of the variable referred to technological differentiation is statistically significant in every case, revealing a positive

[8] See Maddala (1988) and Greene (1993).
[9] See Maddala (1993) and Greene (1993).

association between this feature of the product and the extent of intra-industry trade. Therefore, the sectors in which these activities are more intense seem to enjoy a comparative advantage provided by their specific technological knowledge, leading to inter-industry exchange of products.

The coefficient of the variable related to scale economies is also highly significant in all the regressions and points at a positive association between this characteristic and the degree of intra-industry trade. This result confirms that the existence of scale economies is an important factor in the emergence of intra-industry trade, as it is stated from a theoretical perspective.

Table 11.4. Determinants of Spanish intra-industry trade: Tobit model estimation results

Variables	(1)	(2)	(3)	(4)	(5)
Const.	0.66824***	0.68151***	0.66091***	0.27330*	0.30013*
	(12.900)	(21.623)	(18.332)	(1.872)	(1.828)
RDR	0.23468E-02				
	(1.136)				
RAE		0.66637**	0.77782**	0.73087***	0.76309**
		(2.316)	(2.730)	(3.032)	(2.796)
RDR	-0.13025E-02***	-0.10736E-02**	-0.11453E-02***	-0.13255E-02***	
	(-3.395)	(-2.973)	(-3.259)	(-4.283)	
TDR					-0.18161E-02***
					(-3.338)
SE	1.8369**	1.4700**	1.6065***	1.1030**	1.0888*
	(2.686)	(2.837)	(3.105)	(2.448)	(2.097)
FDI	0.35176E-06	0.24436E-06			
	(1.574)	(1.262)			
FDIVA			0.27065		
			(1.530)		
FDIT				0.44029**	0.40696**
				(2.914)	(2.399)
T	-0.30198E-01***	-0.35881E-01***	-0.35389E-01***	-0.22287E-01***	-0.23631E-01***
	(-4.464)	(-5.571)	(-5.702)	(-3.344)	(-3.155)
Log-L	22.15955	23.83386	24.16422	26.44177	24.61722
BP (χ^2_5)	3.208191	1.261381	1.883921	2.327728	2.519485
BJL (χ^2_2)	0.537062	0.922578	0.001491	1.132723	1.110650
SW	0.967266	0.957539	0.948222	0.964331	0.944375
LR (χ^2_6)	17.60450	20.95310	21.61384	26.16894	22.51984
$\overline{R}^2$	0.519061	0.615289	0.631865	0.728282	0.653444

Notes:
t statistic in parenthesis
* 10% level of significance
** 5% level of significance
*** 1% level of significance

The participation of multinational firms, approximated by the share of foreign direct investment in total investment flows is again statistically significant in all the regressions in which it has been included. Furthermore, the association between this variable and the share of intra-industry trade is positive. These results give support to the hypothesis 3 in 11.2.1. The proportion of intra-industry trade in a particular industrial sector is larger when the activity of multinational firms operating in that sector is more intense.

Finally, we have to mention that tariff protection seems to play also an important role, in this case of a negative sign, in the intensity of intra-industry trade. It is true that, in principle, the existence of tariff barriers is acting as an obstacle to every type of trade. Nevertheless, we have to stress that, in the case of trade agreements among countries with similar characteristics, it may be justified an increase in the share of intra-industry trade. This is due to factors like demand similarity or cultural and geographical proximity, as well as a higher probability for the countries to belong to an economic integration scheme.

11.3 Summary and Conclusions

In this paper we have analysed the connection between sectoral intensity of intra-industry trade and a series of variables related to industry characteristics including product differentiation, scale economies, technological factors, participation of multinational firms and tariff protection. Using an econometric Tobit model a set of hypotheses concerning the determinants of the degree of intra-industry trade in an industrial sector has been tested.

Empirical evidence found support the existence of a positive association between the level of intra-industry trade and the degree of horizontal product differentiation, the existence of scale economies, and the participation of multinational firms. On the other hand, technological product differentiation and tariff protection seem to have a negative influence on the share of intra-industry trade in total trade of a sector.

References

Amemiya, T.: Advanced Econometrics. Basil Blackwell 1985

Balassa, B. and Bauwens, L.: Changing Trade Patterrns in Manufactured Goods. North Holland Publishing Company 1988

Bano, S. S. : Intra-Industry Trade. The Canadian Experience. Avebury 1991

Bera, A. K., Jarque, C.M. and Lee, L. F.: Testing the Normality Assumption in Limited Dependent Variable Models. International Economic Review 25-3, 563-578 (1984)

Breusch, T. S. and Pagan, A. R.: The Langrange Multiplier Test and its Applications to Model Specification in Econometrics. Review Of Economic Studies 47, 239-254 (1980)

Carrera, M. G.: Comercio Intra-Industrial: Análisis Del Caso Español. Ph. D. Dissertation. University of Cantabria. Department of Economics 1996

Fariñas, J.C. and Martín, C.: Determinantes Del Comercio Intra-Industrial En España. In: El Sector Exterior de la Economía Española. Colegio de Economistas de Madrid 1988

Greenaway, D. and Milner, C. R.: A cross-section Analysis of Intra-Industry Trade in the U.K.. European Economic Review 25, 319-344 (1984)

Greenaway, D. and Milner, C. R.: The Economics of Intra-Industry Trade. Basil Blackwell 1986

Greene, W.: Econometric Analysis. Prentice Hall 1993

Grubel, H. G. and Lloyd, P. J.: Intra-industry Trade: the Theory and Measurement of International Trade in Differentiated Products. John Wiley &Sons 1975

Hughes, K. S.: Intra-Industry Trade in the 1980s: a Panel Study. Weltwirtschaftliches Archiv 129-3, 561-572 (1993)

Maddala, G. S. (1988).- Limited-Dependent and Qualitative Variables In Econometrics. Cambridge University Press.

Maddala, G. S.: Contrastes de Especificación en Modelos de Variable Dependiente Limitada. Cuadernos Económicos del ICE 55, 185-223 (1993)

Martín, C.: El Comercio Industrial Español ante el Mercado Único Europeo. In: Viñals, J. (ed.): La Economía Española ante el Mercado Único Europeo. Alianza Editorial 1992

Melo, F. and Mones, M. A.: La Integración de España en el Mercado Común. Un Estudio de Protección Arancelaria Efectiva. Instituto de Estudios Económicos 1982

Montaner, J. M. and Orts, V.: Comercio Intra-Industrial en España: Determinantes Nacionales y Sectoriales. Revista de Economía Aplicada III, 45-62 (1995)

12 Economic Integration, Vertical and Horizontal Intra-industry Trade and Structural Adjustment: the Spanish Experience

G. Carrera-Gómez
University of Cantabria (Spain)

Current research in international trade suggests that intra-industry trade[1] (simultaneous imports and exports of similar products) has acquired an increasing importance in the last years, particularly in the case of transactions of manufactured products among industrialised countries.

One focus of interest to which economic literature on this subject has been addressed concentrates on the process of structural adjustment following trade liberalization, a matter involving important implications related to economic policy. The costs of such a process may be lower when new trade flows are of an intra-industry nature. This statement is owed to the fact that, in this case, the adjustments are to take place within every industry rather than among different ones. This perception has given wide support to several integration projects, including the implementation of the Common Market or the European Community.

Furthermore, there are good reasons to believe that the adjustment implications of a given trade expansion will differ depending on the nature of IIT itself[2]. Horizontal IIT, defined as trade in different varieties of a product with similar quality but diverse attributes, is expected to give rise to less adjustment problems than vertical IIT, considered as trade in different varieties of a product that offer different levels of service[3].

Despite of the great interest arising from these questions, due to the increasing trade liberalization, such issues are empirically underresearched. This is so, in part, because of the problems related to appropriately measure horizontal and

---

1 From now on IIT.
2 See Greenaway, Hine and Milner (1994, 1995).
3 That is to say, different varieties of a product involving different qualities.

P. Coto-Millán (ed.), *Essays on Microeconomics and Industrial Organisation*

vertical IIT (as well as IIT itself) and the difficulty to incorporate to the measurement the dynamic nature of the adjustment process.

In this paper, we examine the possible connections between the nature of the newly generated trade following economic integration and the extent of structural adjustment therefore produced. For this purpose, we use Spanish trade and industry data for the time period between Spanish entrance in the European Union and the completion of the Single Market. For this purpose, we use first a marginal IIT index to measure in a *dynamic* sense changes in IIT, while changes in horizontal and vertical IIT are measured according to a methodology proposed by Greenaway, Hine and Milner (1994, 1995). Next, we measure the changes came about on several industrial characteristics during the considered time period. Finally, we test for two propositions concerning Spanish manufacturing industries. The first one assumes that structural adjustment has been lower in industries with high levels of IIT and/or where new trade was largely IIT than in industries where this trade was mainly inter-industrial. The second one says that structural adjustment has been lower in industries with high levels of horizontal IIT and/or where horizontal IIT rather than vertical IIT prevailed in newly generated trade flows. The results obtained give some support for both propositions suggesting, at the same time, the importance of disentangling vertical and horizontal IIT when dealing with structural adjustment issues.

The rest of the paper is organised as follows. Section 12.1 deals with data issues and sets out the methodology employed. Section 12.2 presents the results of the analysis. Finally, Section 12.3 provides some conclusions.

12.1 Data and Methodology

The current study covers the time period from 1985 to 1991. The first point of reference refers to the year previous to Spain's entry to the current European Union, while the second one corresponds to the year immediately previous to the completion of the European Single Market. The time period chosen is accordingly of particular relevance for the purpose at hand.

Trade and industry data are assembled for a number of 74 industries to the 4th digit of the ISIC[4]-Rev. 2 Classification. Since trade figures are published according to the SITC[5] classification, an allocation of 5th digit SITC categories into 4th digit ISIC product groups had to be performed. This reallocation was based on the preliminary table of correspondence between ISIC-based codes and SITC-based codes provided by United Nations (Department for Economic and Social Information and Policy Analysis, Statistical Division)[6]. Industrial statistics are reported by UNIDO[7]. All the statistical information used in the analysis was supplied by *HWWA-Institut für Wirtschaftsforschung*[8].

4 International Standard Industrial Classification of All Economic Activities.
5 Standard International Trade Classification.
6 Details of the concordance are available on request from the author.
7 United Nations Industrial Development Organization.
8 Hamburg Institute for Economic Research (Germany).

In order to study the connections between IIT and changes in the structure of Spanish industries, it has been applied a recent methodological advance in the empirical analysis of IIT, which tends to incorporate the *dynamic* nature of the adjustment process. As it is widely agreed, the assumption that an increase in the standard Grubel-Lloyd IIT index between two points in time reflects a predominantly intra-industry change in trade flows (and thus lower adjustment costs) can be misleading and lead to erroneous conclusions[9]. It is the nature of new generated trade flows between two points in time what has to be taken into consideration when analysing the advantages of IIT associated with reduced adjustment costs rather than the final level or IIT. For this reason, a number of measures using changes in trade flows to compute the so-called marginal IIT have been proposed up to date[10].

In this study, together with the traditional Grubel-Lloyd indicator of IIT level at a particular point in time (IIT_j), we use a measure of marginal IIT, which focuses on the proportion of matched trade change relative to total trade change ($MIIT_j$)[11]. Indexes definitions are as follows:

$$IIT_j = 1 - \frac{\left| X_j - M_j \right|}{\left(X_j + M_j \right)},$$

$$MIIT_j = 1 - \frac{\left| \left(X_t - X_{t-n} \right) - \left(M_t - M_{t-n} \right) \right|}{\left| X_t - X_{t-n} \right| + \left| M_t - M_{t-n} \right|},$$

where j denotes the 4th digit ISIC industry, X and M denote exports and imports, and t and t-n refer to the two points in time taken into consideration.

The former measure is the standard Grubel-Lloyd index, widely used in empirical IIT studies, while the later, proposed by Brülhart (1994) refers to two points in time (t and t-n) and shows the proportion of new trade which is of the intra-industry type[12].

On the other hand, as several authors point out[13], horizontal IIT is likely to lead to lower adjustment pressures than vertical IIT, since different industry and country characteristics can be associated with trade in products involving similar or different qualities (horizontal or vertical differentiation).

In this study vertical and horizontal IIT are identified using a methodology proposed by Greenaway, Hine and Milner (1994, 1995) built upon previous work of Abd-el-Rahman (1991). As these authors point out, the purpose of measuring quality differences in trade has been mainly addressed with the use of unit value

9 The example provided by Shelburne (1993) illustrates this point.

10 See Hamilton and Kniest (1991), Shelburne (1993), Brülhart (1994), Greenaway, Hine, Milner and Elliot (1994) and Thom and McDowell (1999).

11 See Brülhart (1994) for a discussion.

12 The $MIIT_j$ measure (like the IIT_j index) has values ranging from 0 to 1, a value of 0 indicating that new trade is entirely of the inter-industry type, and a value of 1 showing complete matching between new exports and new imports.

13 See Greenaway, Hine and Milner (1994 and 1995).

indexes which measure the average price of a bundle of items from the same general product grouping. This approach assumes that quality is reflected in price and price can be proxied by unit values. Intra-industry trade can be thus divided into horizontal and vertical components using relative unit values of exports and imports[14].

The share of vertical and horizontal IIT in total trade is computed as follows:

$$IIT_j^h = \left[1 - \frac{\sum_{i=1}^{n} \left| X_{ij}^h - M_{ij}^h \right|}{\sum_{i=1}^{n} \left(X_{ij}^h + M_{ij}^h \right)}\right] \cdot \frac{\sum_{i=1}^{n} \left(X_{ij}^h + M_{ij}^h \right)}{\left(X_j + M_j \right)}, \tag{12.1}$$

$$IIT_j^v = \left[1 - \frac{\sum_{i=1}^{n} \left| X_{ij}^v - M_{ij}^v \right|}{\sum_{i=1}^{n} \left(X_{ij}^v + M_{ij}^v \right)}\right] \cdot \frac{\sum_{i=1}^{n} \left(X_{ij}^v + M_{ij}^v \right)}{\left(X_j + M_j \right)}, \tag{12.2}$$

where i refers to the 5th digit SITC products in a given 4th digit ISIC industry j.

Total IIT is then decomposed in vertical IIT and horizontal IIT as follows:

$$IIT_j^* = IIT_j^v + IIT_j^h$$

$$IIT_j^* = 1 - \frac{\sum_{i=1}^{n} \left| X_{ij} - M_{ij} \right|}{\sum_{i=1}^{n} \left(X_{ij} + M_{ij} \right)}.$$

Horizontal intra-industry trade (IIT_j^h) is given by (12.1) for the items i in commodity group j where the following condition holds:

$$1 - \alpha \leq \frac{UV_{ij}^x}{UV_{ij}^m} \leq 1 + \alpha$$

where UV^x and UV^m refers to unit values of exports and imports and α is a given dispersion factor.

Vertical intra-industry trade (IIT_j^v) is given by (12.2) for the items i in commodity group j where the following condition holds:

14 Unit values are computed by tonne. Although there are well known problems associated with the different ways of computing unit values (see Greenaway, Hine and Milner, 1994), the use of unit value per tonne tends to be the more widely spread in trade studies.

$$\frac{UV_{ij}^{x}}{UV_{ij}^{m}} < 1-\alpha \qquad \text{or} \qquad \frac{UV_{ij}^{x}}{UV_{ij}^{m}} > 1+\alpha$$

The dispersion factor α can be given several values giving rise to different price wedges within which IIT is considered as horizontal and outside of which it is defined as vertical. Following Greenaway, Hine and Milner (1994, 1995) we have employed $\alpha = 0.15$ and $\alpha = 0.25$. The first value gives a price wedge of ± 15% meaning that when the unit value of exports related to the unit value of imports falls within a range of 0.85 to 1.15 intra-industry trade is classified as horizontal and as vertical otherwise. The second value gives a price wedge of ± 25% within which intra-industry trade is considered as horizontal and as vertical other wise.

We test for two propositions concerning Spanish manufacturing industries. The first one assumes that structural adjustment has been lower in industries with high levels of IIT and/or where new trade was largely IIT than in industries where this trade was mainly inter-industrial. For this purpose, industries are classified into two groups according to the level and increase of IIT. The second proposition assumes that structural adjustment has been lower in industries with high levels of horizontal IIT and/or where horizontal IIT rather than vertical IIT prevailed in newly generated trade flows. To test for this proposition, industries are classified into two groups depending on the level and increase of horizontal and vertical IIT.

In both cases, averages are computed for a set of industry structural characteristics for the period 1985-1991. The industry characteristics analysed are listed below:

- Number of establishments.
- Number of employees.
- Wages and salaries paid to employees.
- Output in factor values.
- Value added in factor values.
- Output per employee.
- Value added per employee.
- Wages and salaries per employee.
- Share of value added in output.
- Share of wages and salaries in value added.

Mean differences are computed next and tested for significance from a statistical point of view. The results obtained are shown in the following section.

12.2 Results

Table 12.1 shows the results obtained concerning changes in several structural characteristics for Spanish manufacturing industries according to the level and increase of IIT.

Table 12.1. Changes in structural characteristics for Spanish manufacturing industries according to the level and increase of intra-industry trade, 1985 to 1991.

Industry characteristics (Percentage changes) [a]	IIT level and increase in marginal IIT [b]		Mean difference
	Low	High	
Number of establishments	59.90	16.54	43.36 ***
	(36.20)	(-2.41)	(38.61) ***
Number of employees	15.29	15.25	0.04
	(3.07)	(5.21)	(-2.14)
Wages and salaries paid to employees	72.03	78.36	-6.33
	(72.03)	(78.36)	(-6.33)
Output in factor values	82.53	67.16	15.37
	(82.53)	(64.88)	(17.65) *
Value added in factor values	71.23	63.01	8.22
	(71.23)	(61.81)	(9.42)
Output per employee	80.44	59.68	20.76 **
	(79.70)	(57.53)	(22.17) **
Value added per employee	68.41	54.81	13.60 *
	(68.41)	(54.41)	(14.00) *
Wages and salaries per employee	66.61	70.08	-3.47
	(66.61)	(70.08)	(-3.47)
Share of value added in output	12.80	11.92	0.88
	(-4.40)	(0.33)	(-4.73)
Share of wages and salaries in value added	17.39	17.38	0.01
	(4.12)	(13.92)	(-9.80) *

Notes:

[a] Table shows averages of absolute values of percentage changes. Values in parentheses are averages of changes without taking absolute values.

[b] An industry is classified under 'high' when the share of intra-industry trade in total trade equalled or exceeded 50% in 1985 and/or the marginal intra-industry trade between 1985 and 1991 equalled or exceeded 50 %.

*** Significant at the 97,5 % level of confidence

** Significant at the 95 % level of confidence

* Significant at the 90 % level of confidence

The data shown in table 12.1 give some support for the hypothesis that structural adjustment has been lower in industries with high levels of IIT and/or where new

trade was largely IIT than in industries where this trade was mainly inter-industrial. The mean difference between the industries classified under "low" and those ones classified under "high" is positive and statistically significant for the number of establishments, the output per employee and the value added per employee. When using averages of changes without taking absolute values, the mean difference is significant also for the output in factor values and the share of wages and salaries in value added. The intensity of changes after trade liberalization in all these industry characteristics seems thus to have been higher for the group of industries with low levels of IIT and/or when the proportion of inter-industry trade in the changes in trade flows between the two time periods is higher.

Tables 12.2 and 12.3 show changes in structural characteristics for Spanish manufacturing industries according to the level and increase of horizontal and vertical intra-industry trade for the values of the dispersion factor $\alpha = 0.15$ and $\alpha = 0.25$. For both price wedges the results obtained give some support for the hypothesis that structural changes have been more intensive in industries with low levels of horizontal IIT and/or when horizontal IIT has grown more than vertical IIT for the period 1985-1991.

For the price wedge of ± 15%, the difference between the changes in the two groups of industries is statistically significant for the number of establishments, the number of employees, the wages and salaries paid to employees, the output in factor values and the share of value added in output.

When we take the price wedge of ± 25% to classify industries under low or high horizontal IIT, all the industry characteristics considered (except for the share of wages and salaries in value added) seem to have changed more intensely for the group of industries in which vertical horizontal IIT is less important.

According to these results, we can say that the structural adjustment brought about by trade liberalization in the Spanish manufacturing sectors seems to have been less intense for those industries in which horizontal IIT predominates than for those with high level and increase of vertical IIT.

Table 12.2. Changes in structural characteristics for Spanish manufacturing industries according to the level and increase of horizontal and vertical intra-industry trade, 1985 to 1991.

Industry characteristics (Percentage changes) [a]	Horizontal intra-industry trade [b]		Mean difference
	Low	High	
Number of establishments	47.30	15.27	32.03 *
	(22.24)	(-1.67)	(23.91)
Number of employees	17.81	13.52	4.29
	(9.95)	(1.01)	(8.94) **
Wages and salaries paid to employees	87.60	69.19	18.41 **
	(87.60)	(69.19)	(18.41) **
Output in factor values	82.11	63.95	18.16 *
	(82.11)	(61.15)	(20.96) **
Value added in factor values	71.76	60.68	10.98
	(71.11)	(59.75)	(11.36)
Output per employee	70.75	61.56	9.19
	(68.44)	(60.16)	(8.28)
Value added per employee	57.84	58.93	-1.09
	(57.53)	(58.64)	(-1.11)
Wages and salaries per employee	70.65	68.11	2.54
	(70.65)	(68.12)	(2.54)
Share of value added in output	12.55	11.89	0.66
	(-4.17)	(1.26)	(-5.43) *
Share of wages and salaries in value added	17.73	17.14	0.59
	(12.70)	(10.30)	(2.40)

Notes:

[a] Table shows averages of absolute values of percentage changes. Values in parentheses are averages of changes without taking absolute values.

[b] Intra-industry trade is classified as horizontal where the unit value of exports related to the unit value of imports (at the 5-digit level of SITC) falls within a range of 0.85 to 1.15. Outside this range, intra-industry trade is classified as vertical. An industry is classified under 'high' when the share of horizontal intra-industry trade in total trade equalled or exceeded 40% in 1985 and/or the growth of horizontal intra-industry trade was higher than the growth of vertical intra-industry trade for the period 1985 to 1991.

*** Significant at the 97,5 % level of confidence

** Significant at the 95 % level of confidence

* Significant at the 90 % level of confidence

Table 12.3. Changes in structural characteristics for Spanish manufacturing industries according to the level and increase of horizontal and vertical intra-industry trade, 1985 to 1991

Industry characteristics (Percentage changes) [a]	Horizontal intra-industry trade [b]		Mean difference
	Low	High	
Number of establishments	49.72	14.43	35.29 **
	(21.08)	(-0.39)	(21.47)
Number of employees	18.41	13.23	5.19 *
	(8.44)	(2.18)	(6.26)
Wages and salaries paid to employees	90.06	68.01	22.05 ***
	(90.06)	(68.01)	(22.05) ***
Output in factor values	87.79	60.69	27.10 ***
	(87.79)	(57.95)	(29.84) ***
Value added in factor values	77.37	57.41	19.96 ***
	(76.70)	(56.40)	(20.30) ***
Output per employee	79.64	56.04	23.61 ***
	(77.26)	(54.67)	(22.59) **
Value added per employee	66.68	53.21	13.47 *
	(66.36)	(52.93)	(13.43) *
Wages and salaries per employee	75.78	64.87	10.91 ***
	(75.78)	(64.87)	(10.91) ***
Share of value added in output	14.15	10.87	3.28 *
	(-3.48)	(0.69)	(-4.17)
Share of wages and salaries in value added	19.60	15.94	3.66
	(11.29)	(11.26)	(0.23)

Notes:

[a] Table shows averages of absolute values of percentage changes. Values in parentheses are averages of changes without taking absolute values.

[b] Intra-industry trade is classified as horizontal where the unit value of exports related to the unit value of imports (at the 5-digit level of SITC) falls within a range of 0.75 to 1.25. Outside this range, intra-industry trade is classified as vertical. An industry is classified under 'high' when the share of horizontal intra-industry trade in total trade equalled or exceeded 40% in 1985 and/or the growth of horizontal intra-industry trade was higher than the growth of vertical intra-industry trade for the period 1985 to 1991.

*** Significant at the 97,5 % level of confidence

** Significant at the 95 % level of confidence

* Significant at the 90 % level of confidence

12.3 Summary and Conclusions

It is generally assumed that the adjustment costs brought about by trade liberalization may be different depending on the nature of changes in trade flows. Specifically, there is a greater potential for lower adjustment costs when new trade can be classified as intra-industry trade rather than inter-industry trade. These is so because in the former case reallocation is to take place within industries while the latter implies a reallocation between industries. On the other hand, it has been suggested the importance of distinguishing between horizontal IIT and vertical IIT as the determinants of both may differ. While the former affects diverse varieties of a product with similar characteristics, the latter refers to different varieties of a product involving different qualities. Reallocation of factors may differ in both cases, possibly implying less adjustment problems in the former case.

In this paper, we have examined the relationship between IIT and the adjustment consequences of trade expansion. We have tested the hypotheses that changes in industrial structural characteristics following trade expansion have been less intense for those industries with high level of IIT and/or where new trade was mainly intra-industrial and for those industries with high levels and increase of horizontal IIT.

The results obtained give some support for both propositions suggesting, on the one hand, that changes in industry structure brought about by trade liberalization are less intense when existing and emerging trade is mainly intra-industrial and, on the other hand, that adjustment costs tend to be lower for those industries in which horizontal IIT is predominant. These results indicate the importance of intra-industry trade for the costs of the adjustment process and suggest, at the same time, the relevance of disentangling vertical and horizontal IIT when dealing with structural adjustment issues.

References

Abd-El-Rahman, K.: Firms' Competitive and National Comparative Advantages as Joint Determinants of Trade Composition. Weltwirtschaftliches Archiv 127, 83-97 (1991).

Brülhart, M.: Marginal Intra-industry Trade: Measurement and Relevance for the Pattern of Industrial Adjustment. Weltwirtschaftliches Archiv 130-3, 600-613 (1994)

Brülhart, M. and Mcaleese, D.: Intra-Industry Trade and Industrial Adjustment: The Irish Experience. The Economic and Social Review 26-2, 107-129 (1995)

Greenaway, D., Hine, R. C. and Milner, C.: Country-Specific Factors and the Pattern of Horizontal and Vertical Intra-industry Trade in the UK. Weltwirtschaftliches Archiv 130, 77-100 (1994)

Greenaway, D., Hine, R. C. and Milner, C.: Vertical and Horizontal Intra-industry Trade: a Cross Industry Analysis For The United Kingdom. The Economic Journal 105, 1505-1518 (1995)

Greenaway, D., Hine, R. C., Milner, C. and Elliot, R.: Adjustment and the Measurement of Marginal Intra-industry Trade. Weltwirtschaftliches Archiv 130-2, 418-427 (1994)

Grubel, H. G., Lloyd, P. J.: Intra-industry Trade: the Theory and Measurement of International Trade in Differentiated Products. London: Macmillan 1975

Hamilton, C. and Kniest, P.: Trade Liberalisation, Structural Adjustment and Intra-industry Trade: a Note. Weltwirtschaftliches Archiv 127-2, 356-367 (1991)

Shelburne, R. L.: Changing Trade Patterns and The Intra-Industry Trade Index: a Note. Weltwirtschaftliches Archiv 129, 829-833 (1993)

Thom, R. and McDowell, M.: Measuring Marginal Intra-industry Trade. Weltwirtschaftliches Archiv 135 (1), 48-61 (1999)

PART IV. FAILURES OF MARKET AND INDUSTRIAL REGULATION

13 Organisation and Regulation of the Port Industry: Europe and Spain

B. Tovar
University of Las Palmas de Gran Canaria (Spain)

L. Trujillo
University of Las Palmas de Gran Canaria (Spain)

S. Jara-Díaz*
University of Chile (Chile)

13.1 Introduction

In most countries in the world, practically all imports and exports are done by sea. Maritime transport requires port facilities to facilitate the interchange with land transport or interior navigation. Thus, efficient ports are needed to help the economy in terms of input provision and output distribution.

Sea ports have developed as an answer to the economic demands of their *hinterland*[1] or market area. Their creation and development have been influenced by historical, geographical, political and economic factors, which have translated into different political objectives, management models, property structures and regulatory rules throughout the world.

Traditionally, the port management model has been characterised by the presence of a centralised public agent in charge of the long run planning and

* Corresponding author.

1 The *hinterland* can be defined as that space for which the generalised cost of a port operation is lower than the similar cost using an alternative port.

P. Coto-Millán (ed.), *Essays on Microeconomics and Industrial Organisation*

provision of most of the port services. Nowadays, in many cases the need to adjust public expenditure has motivated the search for the active participation of the private sector in many countries, not only for the provision of port services but also to construct and develop port facilities.

In parallel, maritime transport has undergone important technological innovations within the last decades, which has increased the demand for port facilities that should be able to cope with last generation vessels and the different forms of packing cargo. This has stimulated the competition among ports in order to attract modern ships and modern freight forms. Besides, intermodality has intensified this competition process, forcing the ports to expand their activities beyond the usual frontiers of the port yards. Thus, the traditional warehouse role played by the ports is slowly loosing importance in favour of better and more integrated logistical and physical distribution. These trends have had an impact on the organisation and regulation of ports, letting private participation grow in time, deregulating the activities and sometimes leading to complete privatisation.

Technological changes have also stressed the importance of specific terminals within the port areas (e.g. multi-purpose[2], containers, liquid and solid bulk). Therefore, terminal facilities are becoming heavily capital intensive and, depending on port size, more specialised, playing a key role in the choice of port. The private sector is becoming increasingly interested in this type of activities, which has moved the focus of the competitive strategy from the port to the terminals, making them the most important elements within the port industry. This change of focus is the main element to explain the increase of competition within the sector (Heaver (1995)).

The rest of the chapter is organised as follows. The different models of property and management within the port industry are summarised in section 13.2. Section 13.3 analyses the introduction of private participation in ports operations as well as the role of the public sector. Economic regulation of port activities is presented in section 13.4 and continued in section 13.5 with emphasis on cargo handling. The attempts to design and apply a common policy on ports within the European Union are analysed in section 13.6, along with a description of the present regulatory scheme including the Spanish case. Section 13.7 contains a final discussion.

13.2 Models of Port Property and Management

A first important characteristic of a port as on economic organisation is that it can not be considered as an entity producing a single service. A diversity of activities takes place within the boundaries of a port area. Thus, it is quite important to take into account the diverse characteristics of each particular service that may lead to different regulatory schemes, as some present natural monopoly properties while others could be better produced under competition. By the same token, and given

[2] A multiple-purpose (MP) terminal is designed to serve heterogeneous traffic, including non-containerised and containerised cargo. It can be transformed into a specialised one (e.g. containers only) by changing equipment.

that all services have to be produced within a limited area, it is important to analyse the ways and means of inducing coordination and to identify the role of *port authorities* as institutions in charge of the regulation of all facilities and activities that take place within the port.

In general, port authorities (PA) are local or provincial public entities (although some examples of private institutions do exist). Public administration of a port is present in a variety of forms around the world. In some countries, operation management and planning of the port capacity are very much centralised, as in Singapore, while in others port authorities are highly autonomous as in the USA. There are intermediate situations where port administration involves both the regional and national governments, as in Australia. It should be stressed that the public nature of a PA does not necessarily imply that the provision of port services is developed by the public sector as well.

A variety of port organisation models exist around the world. They differ according the degree of direct intervention of the PA on the provision of services. On one extreme, the port authority acts as a *landlord,* i.e., leaving as many activities as possible in the hands of the private sector. In this model, the PA owns the facilities and either rents or gives in concession these facilities to private operators. Examples of this type can be found in the USA, Canada, Australia and Europe. On the other extreme lies the *comprehensive*[3] PA model, where the authority is directly in charge of all (or nearly all) responsibilities for the activities within the port area. This case is characterised by a trend to keep the management monopoly, sometimes including cargo handling. Examples of this type can be found in Singapore and many African ports (Goss (1990), Heaver (1995) and De Monie (1994)). The usual case regarding property is that of state owned PA and private operators.

When it comes to analyse infrastructure investment, a variety of cases can be found throughout the world. There is a model of local (*municipal*) funding that is used in northern Europe (Holland, Belgium, Germany), where the responsibility on port policy is directly in the hands of the local administrative body. There is a second model where the *State* does the planning and financing of all investments in the principal port network, although the general trend is towards self financing the port system, which can be found in countries in southern Europe and Latin America. A third model is that of *self financing*, where investment funds are provided by private firms or the AP directly using own resources generated by the customers' payments. This model dominates within countries of an Anglo-Saxon tradition (Great Britain and the United States).

13.3 Private Participation in Ports

The brief description of the different types of port organisation shows clearly that the public sector and the private initiative usually coexist. In fact, the general trend

[3] Juhel (1997) distinguishes two sub-groups within this category: *service ports* and *tool ports*. In both cases the PA owns all the assets, but private firms provide the services in the second case.

is towards the *landlord* scheme, which implies an increasing involvement of private participation.

As in all economic activities, private firms search for the maximum profit. The public sector in general intends to maximise some measure of social benefit. Jeffery (1994) suggests that the public role is mostly that of providing an environmental, economic and social structure that permits progress in port activities, not necessarily through direct involvement in port operations.

Goss (1992) studied the existence of a boundary between public and private sectors in port activities and found a large variety of practices, which suggested that such a boundary was extremely fuzzy or non-existent. As an antecedent, the UNCTAD (1975) report at a world level revealed the presence of important differences between the management models in developed and developing countries, although a trend towards an increasing degree of port autonomy was already detected.

There are important theoretical reasons given to justify public involvement in both the development and management of ports, mostly based on either natural monopoly characteristics of some of the services provided, or on market failures, mostly externalities as security or environmental concerns. On the other hand, some of the activities are considered as mandatory "public service" in some countries, as is the case of cargo handling or pilotage and towage services in Spain, which does not imply, however, that they have to be provided by a public firm. Nevertheless, after analysing the roles of both public and private sectors in port activities, Jeffrey (1994) concludes that, depending on a variety of factors, these roles can vary quite importantly. For the public sector, it can go from setting operational standards to the direct implementation of activities, or to provide financial or physical resources to make it possible the operation of the port. Thus, there is no universally accepted division of responsibilities, although, as stated earlier, an increasing role for the private firms has been observed (Harris (1989), Ra'anan (1992), UNCTAD (1993) and World Bank (2001)).

Private participation has been caused by a variety of elements: the need for new financial sources for infrastructure or equipment, the need to reduce public deficit, or the idea to increase efficiency in some activities introducing flexibility for adaptation to varying conditions (Thompson and Budin (1997)). Other elements are the contribution to commerce growth and the acquisition of port management experience (Baird (2002)). The case is that the governments of many countries have decided to deregulate or privatise the main ports. Although some processes of complete port privatisation have been observed (e.g. Malasia, New Zealand and Great Britain), the novelty in recent years has been the presence of private firms building new facilities (terminals) through concession schemes. Bennett (1992) argues that total privatisation is not the only way to increase efficiency and management, stating that the relevant point is to let the organisation in charge exploit the port on a commercial basis, providing enough flexibility and the necessary tools to be able to do it successfully. On the other hand, Eyre (1990) argues that the ports which are managed entirely by state dependent organisations are more expensive and less efficient, but this has not been supported by the empirical evidence collected by Liu (1995) for British ports.

According to Cullinane *et al.* (2001), the empirical evidence does not permit a simple inference regarding property and port efficiency. Nevertheless, the

international experience seems to suggest that private involvement in some port services has indeed improved the output (Estache *et al.* (2002)). There is a general world trend towards the *landlord* scheme, which means that the PA keep property on infrastructure in order to avoid private monopoly power on some essential assets (land or space), letting port operations and investment in both maintenance and equipment be done by private firms. This is in line with the pioneering analysis and recommendations by Bayley and Friedlaender (1981), who suggested that separation between transport activities that present clear scale economies (infrastructure provision) and those that do not (operations) was cost efficient, letting competition rule in these latter. In the great majority of countries, the public sector keeps a key role in planning, investment, development and regulation of ports (Baird (2002)).

For synthesis, the introduction of private participation in the port industry seems like a feasible and potentially desirable road to adapt this sector of the economy to a new and more competitive environment, in which ports require modern management and equipment in order to satisfy the maritime transport demands. The collaboration between the public and private sectors seems convenient, such that the former moves from being a direct operator to a regulatory role and the latter takes advantage of the adaptability to market conditions, increasing efficiency induced by competition with others.

13.4 Regulation of Port Activities

The active introduction of private participation generates in some cases a need for regulation in the provision of some services, in order to prevent potential actions leading to inefficiencies due to local monopoly power. Such a case is more likely to occur in small ports with captive traffic, because of the inadequacy of competition within the port and the difficulty of competition between ports. Thus, regulation of port activities is a key aspect within the new strategic trend, not necessarily in the hands of port authorities themselves.

A usual way to introduce private participation in ports is through contracts between the private and the public entities. Thus, the contract is the obvious tool for the regulator. Its form will depend on the initial conditions prevailing at the port, on its size, and on the specific activity under consideration. Contracts present a wide variety, ranking from concessions (*Building, Operation and Transfer*, *BOT*), where the private firm is given temporarily the port site for constructions and operations returning the facilities afterwards, to licenses for the provision of a given type of service[4]. The choice of the most adequate alternative will depend on the objectives of the regulator and the restrictions faced. An important element to consider is the condition of asymmetric information of the parties involved, as the firms usually know their costs and demand conditions better than the regulator.

On the other hand, the most usual regulatory systems to prevent abuse from a dominant monopolist is the application of maximum prices (*price cap*) and the limitation of firm profits through the rate of return. Hybrid regulatory systems

[4] See Trujillo and Nombela (2000) for a full description of contract types.

contain pricing elements as well as profit limits. The advantage of *price cap* is the incentive towards efficiency, as a cost reduction makes profit grow under given prices. The disadvantage is that captivity of demand combined with price limits provokes an incentive to diminish quality and to increase environmental damage as part of a pseudo cost reduction strategy. The rate of return regulation diminishes capital risk and its cost because of a guaranteed profit rate, but has no incentives towards efficiency. The hybrid systems aim at combining the advantages of both systems while preserving product and environmental quality.

Thus, contract design and price regulations seem to be the most appropriate tools to introduce private participation in port activities preserving quality and inducing efficiency. See Guash (2003) for a review of empirical evidence regarding concession contracts.

Of course there are cases in which regulation might not be needed or play a minor role. This is when competition is feasible, as it has agreed advantages as an instrument to induce discipline on economic agents intervening in a given market. Whether competition is both feasible and desirable will depend on the traffic volume moved within a port. This has been analysed by Kent and Hochstein (1998), who established traffic thresholds to determine the type of competition that is feasible. Even when no competition is feasible within the port, the need to regulate prices would be subject to possible competition between ports. If this is the case, the role of the regulator can be reduced to a periodical control on prices in order to prevent potential collusion among competitors providing similar services within the port or within alternative port sites. It is worth noting that, in general, competition has increased within the port industry as a whole, but this does not have equal impact on all ports or all activities. It depends on many aspects, as location, type, level and structure of traffic served, and so on. Within the European Union, an open debate is going on regarding public subsidies to ports, as they could be an undesired limit to competition.

13.5 Port Services and Terminals

Economic activity within a port is somewhat complex. Running it successfully requires a set of agents and operations that are integrated in what can be generically called port services[5]. They encompass from the administration provided by the PA to pilotage, towage, supply of utilities such as water and power, cargo handling, catering, ship repair, and so on.

From an economic viewpoint, services are quite heterogeneous. The analysis of each one requires different approaches in general, taking into account their specific characteristics. Cargo handling requires special attention, as it means more than 80% of the bill of a vessel that arrives to a port for loading and unloading. Increasing homogenisation of the cargo unit and technical change in the equipment industry has induced an increasing number of specifically designed facilities for loading and unloading of, for example, containers or bulk. In a relatively short period of time, container terminals have acquired a prominent role

[5] See De Rus *et al.* (1994) for a full description of port services.

in large and medium size ports, where the large volume of this type of cargo makes it economically efficient.

The cargo handling service is usually viewed as one that has to be provided directly by the public sector or by private firms through concession contracts. The large investment usually required for this type of services has been used as an argument to justify private participation. This is why in a large proportion of ports throughout the world container terminals are private although they use public land and have to pay cannon. Because of the private provision of the service, firms operate under regulatory mechanisms regarding prices and profits. As stated earlier, the enforcement rules are stronger depending on the presence of competition within the port or between ports. When the traffic volume in the port makes it efficient to have only one terminal but there are other ports in the neighbourhood that exercise real competition, fare regulation could not be necessary as the market mechanism keeps the price levels within reasonable ranges because of the fear of loosing demand in favour of other ports. If traffic through the port is captive, price regulation is necessary. As an example, users of the Mexican ports of Veracruz and Manzanillo have complained because there is no regulation of the single terminals that operate at each location, which seem to be taking advantage of their local monopoly power through large prices.

Size of the port and type of service are two key elements when deciding whether competition is feasible, and how to promote it. Analysing this requires a profound knowledge of the cost structure of the activity involved. This means not only knowing total costs for different volumes of aggregated traffic, but also the behaviour of costs when part of the bundle is produced, i.e. when the mix changes. Cargo handling usually involves moving different things (containers, rolling stock, bulk, non-containerised cargo, to mention some). Whether it is convenient to have one terminal for all needs in a port or many specialised terminals, or many multiple-purpose ones is not a simple matter to investigate. Yet it is necessary for an adequate design of a port policy. Equally important is to know the specific marginal costs associated to the different type of cargo movements, in order to provide a basis for price regulations.

The difficulty that means knowing the cost structure of firms that obtain concession contracts has been usually approached by comparison[6]. However, the direct estimation of marginal costs by product, scale economies (global and specific) and economies of scope, are quite useful in order to determine the number and type of terminals that should be allowed at a port for a given forecast of demand (traffic mix and volume).

From this viewpoint, the work by Tovar (2002) is particularly relevant, as it is the first attempt at looking at the cargo handling service in port terminals by means of the estimation of a (multioutput) cost function using Spanish data. Previous research had focused on port activity as a whole or had concentrated on other services. It should be noted that Spain follows the *landlord* model. According to the prevailing law, in order to set the conditions for potential private participation, the PA should base on pursuing efficiency, productivity and reliability. To do this task properly, the PA needs a detailed knowledge of the cost

[6] This means looking at the price structure of the same services offered by other ports within the region with similar characteristics as the one that is being regulated.

structure of the different activities. In Tovar (2002), the activity of multi-purpose terminals moving mostly containers is analysed, using data on cost and production of three concessionaire firms in the port of La Luz and Las Palmas. The emphasis is put on the calculation of quantities that are meaningful when dealing with regulation, namely marginal costs by type of cargo moved, economies of scale and scope. These are pieces of information that, although necessary and important, are not readily available to the regulators but have to be calculated from data specifically collected, which gives the analysts an important specialised role.

13.5.1 Cargo Unitisation

Starting some decades ago, new technologies for cargo handling and vessel design have been developed such that productivity has increased due to mechanisation and work reduction that has translated into shorter stays of the ships at the port. This new technology can be described as "Unitisation", whose general idea is that of repackaging various cargo items of relatively small size into larger units of a standard size that can be moved using specifically designed machines and accommodated into specifically designed ships, speeding up the service. There are different techniques for unitisation. There are *pallets*, which can be handled by fork-lift; wheeled platforms manoeuvred by truck; cargoes that can be "rolled on" the vessel in the loading port, and "rolled off" the vessel in the destination port (e.g. roll-on/roll-off trucks and trailers); containers; and even barges which are loaded into the LASH vessel[7]. In each of these cases, the cargo handling process is associated with specific machines (cranes and vehicles) making the type of standardised or compact unit used more important than the type of cargo itself. This might cause that the same type of goods can receive different handling treatment depending on the repackaging: bags, pallets, containers, and so on.

One of the key aspects of cargo unitisation is the correlation between handling capacity and the weight of the standard unit. This is due to the large amount of time that takes the manipulation of small size packages, particularly the process of cargo handling within the vessel's hold. Thus, for a given cargo volume, the larger the standard unit the lower the number of units. The use of containers and rolling units has made it unnecessary the operations of cargo handling on vessel's hold.

13.5.2 Factors of Production and Their Regulation

In the production of cargo handling services the following three groups of factors are required: basic infrastructure, superstructure, machines and mobile equipment, and labour. The provision of these factors is affected by the type of organisation prevailing in each port to manage cargo handling.

Basic Infrastructure

An interesting characterisation of ports is given by the European Parliament (EU, 1993). The port area is defined as a complex of water basins and land areas where

[7] LASH means "Lighter aboard ship". This means that lash ships carry barges.

services to ships and cargo are provided. To get to this port area maritime access, defence infrastructure[8] and land access[9] are needed.

Civil works within the port area defined above are needed for the supply of services to ships and cargoes. These are the port infrastructures, including wharf, shipyard, road and railway network inside the port, and so on. The boundary of this port infrastructure can be set at the extremes of reinforced-concrete structures. Thus, canalisations would be part of them, whereas pavements and maintenance would be excluded.

Two kinds of customers use port infrastructures. On one hand are the ships that moor a buoy or anchor in the port's waters. On the other hand are those firms that work inside the port area and perform services to ships (pilotage, towage, stevedore, terminals, ship repair, and so on). Among the latter are the terminals that operate through concession contracts where concessionaire's obligations and payments are set. The usual norm is that the concessionaire is obliged to pay a *canon* to the port authority or the institution responsible for the concession. Usually, the canon is set as fixed annual fee by square meter or as a variable amount by ton or TEU handled, or as a combination of both.

Recent initiatives of the European Union are aimed at the recovery of public funds invested in ports by means of charges to infrastructure users. This would require knowledge of the level of investments and, most importantly, of the long run cost structure such that marginal costs, economies of scale and scope can be calculated and used to establish adequate fare policies for infrastructure use. In this respect, the studies by Jara-Díaz *et al.* (1997) and Jara-Díaz *et al.* (2002) analysing infrastructure costs in Spanish ports by means of a long-run multioutput cost function are particularly interesting.

Superstructure Machinery and Mobile Equipment

Above the port infrastructure there is the port superstructure, which are the buildings (warehouses, workshops and offices). Infrastructure and superstructure are complemented by the fixed and mobile equipment, and the information and automation systems.

There is no uniform scheme for the provision of cargo handling services within ports. On one extreme, the PA provides and controls exclusively the supply of cargo handling services (*comprehensive port*). On the other extreme, the PA plays no direct operational role (*landlord port*). In this latter case, control for the provision of such services is assigned to private companies allowed to operate within the port zone. The conditions under which these firms are authorised to operate vary from port to port. In some ports, private stevedores are allowed to operate even without a financial arrangement or contract. In other ports, however, the private operator can be requested to contribute to investments on port superstructure, machinery and mobile equipment, and to be part of an agreement that involves renting the basic infrastructure (owned by the PA) for a specific time period. This gives the private operators a more stable position diminishing their

[8] E.g. dam, breakwater, and navigations aids as buoys.

[9] National road and rail network plus connections with the local network of the port area.

risk, promoting a policy of finance participation. This practice has been widely applied by the principal ports in Europe.

Although the *landlord* model dominates, there is a wide range of administrative arrangements and contracts throughout the world that lay between the two extremes described (Goss (1990)).

Labour

Labour in a port can be classified grossly in two groups: those workers directly involved in cargo handling operations (stevedores or port workers) and those who are not (mostly administrative and maintenance personnel). Traditionally the former group has been strongly regulated, although changes have occurred within the last decades worldwide.

The origins of port workers protection are in the characteristics of such a job, particularly the discontinuous demand and the low degree of specialisation. Cargo handling was almost exclusively reserved for registered workers. Labour protection seems to have gone beyond reasonable limits, allowing the workers to exercise monopoly power over port operations. This is the reason why many countries have been introducing legal reforms to increase efficiency by diminishing costs through team adjustments. This is an ongoing process in Europe in general.

The most striking example of labour deregulation took place in Great Britain during 1989, where excess labour was subject to mandatory elimination. Once deregulation was complete, the entrepreneurs were able to legally reduce the average workers' age and to change labour practice dramatically. The new labour rules in British ports have translated into contracts that introduce numerical[10], temporal[11], functional[12] and financial[13] flexibility. This way, the wage rate, labour assignments and labour practices are established locally such that they are adjusted to meet the variable requirements of the customers at the ports. Of course, this has allowed labour cost reductions and a better (more intensive) utilisation of labour at the ports (Turnbull and Weston (1993a, 1993b)).

Abolishing labour regulation in Great Britain has generated increasing competition within and between ports and has pushed wages down. As a result, and as an answer to the increasing pressure by the customers in the areas of cost and quality, port operators believe that their performance has improved in most areas, including reliability and ship time at the port, beyond direct cost and quality of the service (Turnbull and Weston (1993a, 1993b)).

---

10 Numerical flexibility is obtained by means of part time contracts that can respond to demand fluctuations. Another important element is the effect of the elimination of the predetermined size of a team.

11 Temporal flexibility is attained by means of weekend work and extra hours.

12 Functional flexibility implies that there is no pre-determined type of work assigned to a particular individual. Each worker has to perform according to needs.

13 Financial flexibility is obtained through variable wage levels and a closer relation between work and salary. This has translated in many ports in reductions of salary, where the monopoly powers of workers let high wages.

13.6 Port Regulation in the European Union

As stated earlier, due to its strategic role ports have been traditionally under some form of government control, although the legal regime and autonomy varies from country to country. This variety is also present within the European Union, where the attempts at the homogenisation of the different regulations within each of the member states have been, so far, unsuccessful.

In Europe, the strategic role of ports has been explicitly recognised by all members of the Union. Their economic relevance is not only reflected by the volume of goods moved (90% of the total imports and exports to and from the EU but also by the fact that maritime transport is presently in charge of 35% of the total commerce among the members, plus some 200 million passengers per year (EU Commission (1997)). Furthermore, it is likely that congestion in roads will push part of the land transport towards the sea.

In spite of the evident economic relevance of ports for the EU, they are not mentioned explicitly in the Rome Treaty. This omission generated a debate: are European ports subject to the general provisions of the Treaty? In 1974 the European Supreme Court ended the discussion providing an answer (Case 167/73) that, in essence, states that maritime transport is not under the rules of Title IV but under the general principles of the Treaty.

Nevertheless, for a long time European ports have been operating as if the Rome Treaty was not existent. There has been various initiatives aimed at including ports within the Transport Common Policy, but they have failed mainly due to the different views and beliefs of the members regarding the economic role of ports. The European Parliament had commended a series of studies with the objective of clarifying the issues: the Kapteyn, Seifriz and Seefeld reports (EU Parliament (1961, 1967 and 1972)). These reports stimulated indeed the activities related with the potential development of a common port policy. Among these, a work group was formed, including a representative of the Economic Commission and representatives of the main European ports. The group released a Report on European Ports (EU Commission (1977 and 1985)) that identified the main organisational and economic differences, pointing out that here were no substantial differences among the ports regarding services and technical equipment.

Based upon a series of previous reports (EU Parliament (1981, 1982, 1983); EU Commission (1985)), the EU Commission released a document (EU Commission (1992)) containing the main challenges to a common transport policy identifying without ambiguity the need to consider a transport system at a European scale and to establish the basic elements for the development of so-called Trans-European Networks. The report concluded that maritime cargo transport in Europe (cabotage) should be encouraged as a way to alleviate land transport congestion and to contribute to a sustainable mobility strategy that combines fulfilling the socio-economic goals with proper environmental care.

Another report regarding the importance of a common port policy within a unified European market was released by the parliament during 1993 (EU Parliament (1993)). The purpose of this study was to provide information about

the criteria that should guide that common policy. The main recommendations were in fact very similar to those contained in previous reports and the only new topics were those related with the identification, selection and evaluation of projects that presented common interest, and safety procedures. The report also contained explicit suggestions aimed at changes in the law at a national level in order to eliminate those legal or factual conditions that led to non-competitive practices against articles 85 and 86 of the Rome Treaty, as the existence of exclusive rights and other forms of dominant situations. On this topic, it is interesting to mention the two main decisions by the European Court of Justice on port issues. In 1991 the Court declared illegal the assignment of exclusive rights to organise port labour to a national company by a State Member, as well as to make it mandatory to employ national port workers only. In 1994 the Court eliminated fares for piloting that discriminated between ships doing short sea shipping and those that carried international cargo. The Court stated that such a practice meant an abuse of a dominant position.

The interest and efforts of the European Union to establish a Trans-European Network can be clearly seen in articles 129 of the Maastricht Treaty (1992), although infrastructure planning is still the responsibility of each Member State. In the communication of the Commission (EU Commission (1992)) the need of integrating ports in a Trans-European Network was mentioned.[14]

During April 1994 the Commission approved a proposal to establish guidelines for the development of a Trans-European Transport Network. As a result, a port experts group was created within the General Office for Maritime Transport of the EU Commission, whose objective was both to provide guidelines or directions and to identify those ports that should be part of the network. During the discussions the idea that a port network was not needed gained momentum, due to the intense competition among ports. Finally, the EU Commission stated explicitly that "no port of community interest would be identified because this could distort the principle of free and fair competition among ports". In spite of this, during the discussions many Member States and the EU Parliament emphasised the need to include geographical locations of ports in order to establish an actual maritime transport network. This was resolved by means of a commitment of the EU Commission to prepare a report during 1997 identifying a set of eligible ports following the approach previously taken with the air transport component of the Trans-European Network. The European projects could include only those from the selected set, which would be checked and updated periodically. Guidelines were finally approved on July, 1996 (Aragón (1996)).

In 1997 the EU Commission released the *Green Book on Ports and Maritime Infrastructure* aimed at feeding the debate on efficiency, on the application of competitive rules, and on the integration of ports to the multimodal European network. The Green Book concludes that regulation at a European level should be developed in order to achieve a systematic liberalisation of services in the main ports with international traffic. The debate that followed the release of the Green Book was centred upon three aspects: including ports within the Trans-European

[14] If ports are integrated into the Trans-European Network, it means that they will be considered as an integrating part of the European transportation infrastructure, which implicitly means that they will be treated as a public service, as any other transportation infrastructure.

Transport Network, deregulation of port services, and public finance of ports and port infrastructure (EU Commission (2001)). This translated into a proposal of a Directive on access to the market of port services. After a series of amendments a legislative resolution was approved on March 11, 2003, by the European Parliament (EU Parliament (2003)).

This resolution on market access to port services is applicable to ports with an average annual traffic of at least 1.5 million tons or 200.000 passengers[15]. It authorises the liberalisation of port services excluding the pilotage, and reinforcing social rules for the self-assistance[16]. Besides, self-assistance is limited to ships' crew. Various amendments deal with this point. Other amendments emphasise the need for transparency in financial relations, particularly when state funds are involved, which is aimed at guaranteeing loyal competition among ports. Nevertheless, as pointed out by Farrel (2001), the approved proposal Directive does not impose requirements besides asking the PA to keep separate accounting when acting as service providers.

The position of each State Member regarding the role of ports within the EU has always depended on the economic significance of the ports in their transport systems. Traditionally, both the State Members and the most important ports individually have been opposed to the attempts at building a common port policy by the European institutions, as they perceive a loss or at least reduction of autonomy. However, the reactions to the proposal Directive regarding the access to the market of port services have been varied. Some countries are strongly opposed, as Sweden and Great Britain, and most are supportive, particularly Spain (Editorial (2002)).

For synthesis, in spite of the formation of different task forces and the existence of a multiplicity of reports, integration is a process that has just begun for the European ports. It is likely that the lack of progress reflects the wide variety of policy objectives, financial structures and property schemes prevailing in the ports of the EU. Some countries follow a policy that translates into users bearing all costs (the Anglo-Saxon approach) while others intend to encompass all benefits and costs associated to the region in which the port is located (Continental approach). In this latter case, the macroeconomic objectives as employment generation are considered very important. This has evident implications regarding the financial aspects within a port (including pricing and subsidies policies), which generates enough friction to arise to a common attitude and to reach agreements.

13.6.1 Port Regulation in Spain

Spanish ports are subject to a tight regulation of the basic conditions in which economic agents deliver their services within the port area. This regulation takes form through law 27/1992 on State Ports and Merchant Navy (Jefatura del Estado (1992)), modified by law 62/1997 (Jefatura del Estado (1997)). These meant an important change with respect to the rules before 1992. On one hand ports are given greater autonomy (de-centralisation) and, on the other hand, commercial management of ports are pushed forward.

15 The Memeber States can exclude those ports with large seasonal variation in traffic.

16 Self-assistance exists when a firm that could normally hire port services, does it by itself.

Within the Spanish port system, two large groups of ports can be distinguished: those considered as of general interest, owned by the State (article 149.1.20 of the Spanish Constitution) and those that are not, namely fishing, sport oriented and non-comercial ports, owned by the corresponding Comunidades Autónomas (article 148.1.6 of the Spanish Constitution). According to article 5 of law 27/1992, ports of general interest are those involved in international maritime commerce, those whose commercial zone of influence affect in a relevant way more than one Comunidad Autónoma, those that serve industries or entities of strategic importance from a national economy viewpoint, those whose traffic or maritime commercial activities reach a relevant level or respond to an essential need of the general activity of the state, and those considered essential for the security of the national maritime traffic because of technical or geographical reasons, particularly in insular territories.

The basic scheme set by laws 27/1992 and 62/1997 involves a single model for the organisation and management of general interest ports. These duties are assigned to a public PA, with management and legal autonomy that has its own budget, that operates under the coordination and control of the Ente Público Puertos del Estado (EPPE). This Ente Público is in charge of the governmental port policy and has general responsibility for the whole of the port system.

From a financial-economic viewpoint the EPPE gets resources from the whole port system and forms a compensation fund for investments within that system such that it is self financed as a whole, thus reducing the need for subsidies and transfers coming from the rest of the state general budget. This implies that the income perceived by the PA must respond to the general objective of reaching the global survival of the system, i.e. cover total costs, as well as the financial equilibrium of each particular port.

PA management should be based on a multi-dimensional criterion that involves "efficiency, economy, productivity and safety". They should guarantee at each port that certain services are indeed offered (article 66, law 27/1992). Such services could be offered directly or through indirect management by means of concessions or contracts.

The cargo handling services are regulated beyond law 27/1992. These services have specific laws as well (Jefatura del Estado (1986) and Ministerio de Relaciones con las Cortes y de la Secretaría de Estado (1987)). Since the Royal Law-Decree 2/86 cargo handling in Spanish ports of general interest is regarded as a public service under State responsibility[17]. The aforementioned Royal decree establishes that a state owned firm (called *Sociedad Estatal de Estiba y Desestiba,* SEED) will be formed at each port included in the decree[18]. The Royal Decree permits the access to port activities to loading/unloading firms that would like to do it through the system of administrative contracts. Each SEED started operating financially with the contribution of the private firms. Thus, all firms willing to participate in the management of the public service have to participate mandatory

[17] Also in those autonomous ports where the Port Workers Organisation existed before 1986.

[18] These are public limited companies whose objectives are to ensure that port workers are professionals and that such services are regularly provided.

in the capital of the SEEDs according to some pre-established objective criteria[19], although the participation of the State in a SEED will be larger than 50% in order to guarantee decision power. On the other hand, the corresponding PA has to set the maximum prices that the loading/unloading firms can charge for their services.

For synthesis, the regulation of the Spanish port system is based upon a scheme that allows the combination of public property of the port infrastructure (docks, land, and so on) with private property of the superstructure (warehouses, cranes, and so on). The public authority determines the conditions under which the private initiative can operate by fixing maximum prices, length and characteristics of concessions, and other conditions.

Presently, a first draft of a Pre-project of law dealing with the production of economic services at ports of general interest is being studied legally. The proposed law assigns a new role to PA, which can become entities in charge of regulation and infrastructure provision only, providing cargo handling and other services subsidiary. This way, the *landlord* PA model begins to gain power with the declared objective of promoting the private sector participation in the financing and exploitation of port facilities and in the provision of services through concessions. The proposed regulation aims at two key objectives: to extend the general rules on cargo handling service (treated as a singular case so far), and to adapt the law to the European framework designed by the EU.

13.7 Conclusions

Ports are thought and designed to transfer goods between two transport modes efficiently. To achieve this, a number of activities have to be developed within the port premises. As reviewed here they can be organised and managed in many different forms in terms of property and regulation. The relevant goal is to make the whole set work efficiently (Friedrichsen (1999)). As the private sector is usually more effective in this type of activities (Drucker (1986)), the trend towards the fruitful private-public partnership seems advisable.

Although regulation of activities within the port premises has a long tradition, there has been changes in maritime transport within the last decade that have intensified competition among ports, inducing deregulation processes and increasing private participation within the sector, leading in some extreme cases to total privatisation of a port, a trend also induced by international lending agencies. There are strong and good reasons for the presence of a central public agency that should carefully analyse which activities or aspects do need to be regulated and in which way unnecessary constraints or pressure can be avoided. In any case, planning and coordination is needed at a central level along with the necessary initiatives and controls to ensure safety and avoid negative externalities.

Economic activities within a port are multiple and heterogeneous. Among them, cargo handling has been one of the most affected by technological changes

[19] Fixed labor available, equipment investment, annual rental payment for using port land and facilities, annual volume of cargo handling, participation in port traffic of the different State ports, volume of annual port payroll.

on one hand and by competition among ports on the other. The importance of this activity is evident when one realises that it means from 70% to 90% of a vessel's bill of load (De Rus *et al.* (1994)). Besides, the role of the port terminals within the logistic systems makes them key actors of the port industry, playing a central role in the increasing competition within the sector.

A relatively large proportion of ports manage cargo handling in terminals through concessionary schemes. As a consequence, the contracts signed between the public and private entities acquire special relevance. The need to establish price and quality regulations on the services provided by these firms will be a function of the competitive pressures at every particular case[20]. The quantitative estimation of key concepts that synthesise information regarding the cost structure of those firms is indeed necessary to inform the job of the regulators and to facilitate the design of contracts. These key concepts include marginal costs by product handled, and economies of scale and scope, which are essential to determine optimal sizes, product combinations, optimal prices and so on. Perhaps the main challenge for the correct regulation within the sector is to facilitate obtaining the relevant information directly from the sources to flow and feed the technical analysis (Tovar *et al.* (2003)). This should be the most relevant duty of a central agency if an efficient set of regulations is to be set. For one of the few examples of the type of information needed and the rich analysis that can be done with it, see Jara-Diaz *et al.* (2003).

Since the creation of the European Union, European ports serve a single *hinterland* that encompasses the common market, and are under the general rules of the treaty that are of interest to port issues. These rules are mainly those that deal with free competition, monopolies and state help. In fact, from the viewpoint of the EU many aspects that are derived both from the institutional characteristics and from the port organisation have important consequences that have to be appraised within the free competition paradigm. One possible example is the potential existence of cross-subsidies. British ports claim that continental ports that receive subsidies to finance their infrastructure are larger than if they did not. Others, in turn, claim that British ports were sold to the private sector at a price that is below the market value, which is yet another form of disguised subsidy (Fleming *et al.* (1999)).

For synthesis, it seems advisable to reach a large degree of agreement and homogenisation of port policies and regulations among the members of the EU, including those legal provisions that are not directly related with ports but affect them. The variety of activities and property forms can and should be used in an adequate combination to achieve both an economically efficient use of port resources and sustainable forms of goods traffic development. In this way, fair competition will be promoted when necessary and regulation will play its role. Releasing information on port activities (costs, production) is a key requisite for a good analysis and right decisions.

[20] An example of price regulation in the port of La Luz and Las Palmas can be found in Trujillo *et al.* (1996).

Acknowledgements

This research was partially funded by Fondecyt, Chile, Grant 1010687, and the Millennium Nucleus "Complex Engineering Systems". The hospitality of the Universidad de Las Palmas de Gran Canaria is gratefully acknowledged by Prof. Jara-Díaz.

References

Aragón, F.: La red transeuropea y su componente marítima. Encuentro: el impacto económico de los puertos. Cursos de Verano de la UIMP. Santander 1996

Baird, A.: Privatization trends at the world's top-100 container ports. Maritime Policy and Management Vol. 29, 271-284 (2002)

Bennett, M.: Trade or treasury: who benefits from port privatization?. Portus Vol. 7, N1 1, 10-15 (1992)

Cullinnane, K., Song, D. W. and Gray, R.: A stochastic frontier model of the efficiency of major container terminals in Asia: assessing the influence of administrative an ownership structure. Transportation Research Part A: Policy and Practice (2001)

De Monie, G.: Mission and role of Port Authorities. Proceedings of the World Port Privatization Conference. London 1994

De Rus, G., Román, C. and Trujillo, T.: Actividad Económica y Estructura de Costes del Puerto de La Luz y de Las Palmas. Ed. Cívitas. Madrid (Spain) 1994

De Rus, G., Trujillo, L., Tovar, B., González, M. and Román, C.: Competitividad de los Puertos Españoles. Working Paper. Tribunal de Defensa de la Competencia. Madrid (Spain) 1995

Drucker, P.: The practice of management. Harpercollins Publishers. New York 1986

Editorial: EU port policy. Maritime Policy and Management. Vol. 29, 1-2 (2002)

Estache, A., González, M. and Trujillo, L.: Efficiency gains from port reform and the potencial for yardstick competition: lessons from México. World Development, Vol. 30, N° 4, 545-560 (2002)

EU Commission: Report of an inquiry into the current situation in the major community sea-ports, drawn up by the port working group (revised and enlarged in 1986), 1977

EU Commission: Progress towards a Common Transport policy. COM (85) 90 final, 1985

EU Commission: The future development of the common transport policy. COM (93) 701, 1992

EU Commission: Transport Infrastructure. COM (92) 231 final, 1992

EU Commission: Green Paper on Sea Ports and Maritime Infrastructure. COM (97) 678 final, 1997

EU Commission: Reinforcing quality service in sea ports: A Key to European Transport. COM (2001) 35 final, 2001

EU Parliament: Los problemas relativos a la política común de transportes dentro del marco de la CEE. Informe Kapteyn. Documento PE 106. Luxemburgo 1961

EU Parliament: La política común de tráfico portuario. Informe Seifriz. Documento PE 140. Luxemburgo 1967

EU Parliament: La política portuaria en el marco de la Comunidad Europea. Informe Seefeld. Documento PE 10/72. Luxemburgo 1972

EU Parliament: The common seaport policy. Documento PE 73.762. Luxemburgo 1981

EU Parliament: El papel de los puertos en la política común de transportes. Informe Carossino. Documento PE 1-844/82. Luxemburgo 1982

EU Parliament: Towards a common transportation policy. Luxemburgo 1983

EU Parliament: Progress towards a common transportation policy. Memorandum 2. Luxemburgo 1986

EU Parliament: European Sea Port Policy. Directorate General for Research. Transport Series E-1, 7-1993, 1993

EU Parliament: European Parliament legislative resolution on the Council common position for adopting a European Parliament and Council directive on market access to port services. P5_TA-PROV(2003)03-11. Provisional Edition. PE 328.825. 2003

Eyre, J.: Maritime privatization. Maritime Policy and Management, Vol. 17, N1 2, 113-121 (1990)

Farrell, S.: Comment. If it ain't bust, don't fix it: the proposed EU directive on market access to port services. Maritime Policy and Management Vol. 28, 307-313 (2001)

Fleming, D. and Baird, A.: Comment. Some reflections on port competition in the United States and Western Europe. Maritime Policy and Management Vol. 26, 383-394 (1999)

Friedrichsen, C.: Benchmarking of Ports. Possibilities for Increased Efficiency of Ports. Transport Benchmarking. Methodologies, Applications and Data Needs. European Conference of Ministers of Transport. European Commission. Paris 1999

Goss, R.: Economic policies and seaport: 2. The diversity of port policies. Maritime Policy and Management Vol. 17, 221-234 (1990)

Goss, R.: Port privatization: the public interest. Department of Maritime Studies and International Transport at the University of Wales college of Cardiff. UK 1992

Guasch, J. L.: Concessions: Bust or Boom? An Empirical Analysis of Ten Years of Experience in Concessions in Latin America and Caribbean. The World Bank Institute, Washington DC 2003

Harris, F.: Port privatization: a survey of global trends. Frederic R. Harris Research News. Fall (1989)

Heaver, T.: The implications of increased competition among ports for port policy and management. Maritime Policy and Management Vol. 22, 125-133 (1995)

Jara-Díaz, S., Cortes, C., Vargas, A. and Martínez-Budría, E.: Marginal costs and scale economies in spanish ports. 25th European Transport Forum, Proceedings Seminar L, PTRC, London, 137-147, 1997

Jara-Díaz, S., Martinez-Budría, E., Cortes, C. and Basso, L.: A multioutput cost function for the services of Spanish ports' infrastructure. Transportation Vol 29, Nº 4, 419-437 (2002)

Jara-Diaz, S., Tovar, B. and Trujillo, L.: Economies of scale and scope for cargo handling firms in Spanish ports. Proceedings of the European Transport Conference, Strasbourg, October 2003

Jefatura del Estado: Real Decreto-Ley 2/86, sobre el servicio público de estiba y desestiba de buques. Boletín Oficial del Estado Nº 23 (1986)

Jefatura del Estado: Ley 27/1992, de 24 de noviembre, de Puertos del Estado y de la Marina Mercante. Boletín Oficial del Estado Nº 283 (1992)

Jefatura del Estado: Ley 62/1997, de 26 de diciembre, de modificación de la Ley 27/1992 de 24 de noviembre, de Puertos del Estado y de la Marina Mercante. Boletín Oficial del Estado Nº 312 (1997)

Jeffrey, D.: The functions of the Public and Private sector in Ports. Mimeo. Port of London Authority. London 1994

Juhel, M.H.: Government Regulation of Port Activities: What Balance Between Public and Private Sectors?. Mimeo. World Bank Document 1997

Kent, P. E. and Hochstein, A.: Port reform and privatization in conditions of limited competition: the experience in Colombia, Costa Rica and Nicaragua. Journal of Maritime Policy Management Vol. 25, N°. 4, 313-333 (1998)

Liu, Z.: The comparative performance of public and private enterprises. Journal of Transport Economics and Policy. September, 263-274 (1995)

Ministerio de Fomento: Anteproyecto de Ley de Régimen Económico y de Prestación de servicios de los puertos de interés general. Madrid 2002

Ministerio de Relaciones con las Cortes y de la Secretaría de Estado: Orden de 15 de abril, por la que se fijan las bases para la gestión del servicio público de estiba y desestiba de buques en puertos de interés general. Boletín Oficial del Estado N° 95 (1987)

Pope, D.: The policy of seaports in the European Union. European Commission. Port Policy Unit, DG VII, 1994

Ra'anan, A.: Ports administration: should public ports be privatized?. Working Paper. World Bank. Washington 1991

Thompson, L. S. and Budin, K. J.: Global Trend to Railway Concessions Delivering Positive Results. Public Policy for the Private Sector. Note no. 134, December, World Bank, Washington DC 1997

Thurnbull, P. and Weston, S.: The british port transport industry: 1 Operational structure, investment and competititon. Maritime Policy and Management Vol. 20, 109-120 (1993a)

Thurnbull, P. and Weston, S.: The british port transport industry: 2 employment, working practice and productivity. Maritime Policy and Management Vol. 20, 181-195 (1993b)

Tovar, B.: Análisis multiproductivo de los costes de manipulación de mercancías en terminales portuarias. El Puerto de La Luz y de Las Palmas. Ph D. Departamento de Análisis Económico Aplicado. Universidad de Las Palmas de Gran Canaria. España. (Available on: http://www.fcee.ulpgc.es/~btovar/tesis.pdf.) 2002

Tovar, B., Jara-Díaz, S. and Trujillo, L.: Funciones de producción y costes y su aplicación al sector portuario. Una revisión de la literatura. Forthcoming Working Paper. Facultad de CC EE y Empresariales. Universidad de Las Palmas de Gran Canaria. España 2003

Trujillo, L. and Nombela, G.: Seaports. Privatization and Regulation of Transport Infrastructure. Guidelines for Policymakers and Regulators. Edited by Antonio Estache y Ginés de Rus. The World Bank, Washington, D.C. 113-170, 2002

Trujillo, L., Tovar, B., and González, M.: Revisión de las tarifas de los servicios portuarios. Working Paper. Autoridad Portuaria del Puerto de La Luz y de Las Palmas. Las Palmas 1996

UNCTAD: Port Pricing. United Nations Conference on Trade and Development. New York 1975

UNCTAD: Strategic planning for port authorities. United Nation Conference on Trade and Development. Geneva 1993

World Bank: World Bank Port Reform Tool Kit. World Bank, Washington DC 2001

14 Positive Theory of Regulation: an Application to Spanish Foreign Trade

G. Carrera-Gómez
University of Cantabria (Spain)

P. Coto-Millán
University of Cantabria (Spain)

J. Villaverde-Castro
University of Cantabria (Spain)

14.1 Introduction

Regulation theory has attracted a great deal of interest in economic literature during the last three decades. One aspect that has been given most attention is the question of whether regulatory measures may be influenced by the action of pressure groups. Recent empirical work on determinants of inter-industry structure of tariffs seems to support the view that some industry characteristics may influence the ability of protection-minded industries to exert pressure in order to get protection. Although it is possible to find increasing amounts of literature concerning this subject in Canada and USA, empirical work studying this issue for Europe is still scarce.

The present study aims to apply the positive theory of regulation to the case of Spanish foreign trade. Its main purpose is to analyze the potential association between the inter-sectorial changes in tariff protection and a series of variables representing industry organization, such as scale economies, market concentration, penetration of foreign direct investment, product differentiation and level of intra-industry trade. The work follows the lines pointed out in earlier papers by Caves (1976), Saunders (1980) and Greenaway and Milner (1994), sharing with them its multi-sectorial character.

P. Coto-Millán (ed.), *Essays on Microeconomics and Industrial Organisation*

The rest of the paper is organized as follows. In section 14.2, data and variables employed in the analysis are presented. In section 14.3 the main empirical results obtained in the study are shown and commented. Finally, section 14.4 puts forward the most relevant conclusions achieved.

14.2 Data and Variables

Table 14.1. Classification of industrial sectors. Clasificación Nacional de Actividades Económicas (National Classification of Economic Activities, CNAE-74)

Denomination	CNAE-74 code
Solid fuel, coke, hydrocarbons, radioactive minerals and petroleum refineries	11 to 14
Electric power, water supply and gas	15, 16
Metallic mineral products	21,22
Non-metallic mineral products	23, 24
Chemicals	25
Fabricated metal products (except machinery and transport material)	31
Agricultural and industrial machinery	32
Office and computing machinery, optical accuracy instruments and similar tools	33, 39
Electrical and electronic machinery and material (except computers)	34, 35
Automobiles, pieces and accessories	36
Other transport material	37, 38
Food, drinks and tobacco	41, 42
Textiles, leather and footwear	43 to 45
Wood and cork	46
Paper, graphic arts and publishing	47
Rubber and plastic products	48
Other manufacturing industries	49

As stated before, the main purpose of this paper is to analyze -in the Spanish case- the relationship between inter-industry variations of the degree of tariff protection and a series of variables related to industrial structure that may be indicative of the industries ability to exert pressure for protection. It is interesting to mention that the set of trade and industry data used to construct the variables employed in the analysis were recorded in various classification systems or nomenclatures. Therefore, and due to the lack of official tables of correspondence between such classifications, we have carried out a previous task of conversion between the different nomenclatures with the aim of having the data homogeneity required for carrying out the work[1]. The data were finally grouped together into the industrial

[1] Namely, the following classification systems have been used:
- Standard International Trade Classification Rev. 3 (SITC-Rev. 3)

sectors listed in table 14.1 and the test carried out is referred to a cross-section of these sectors for the year 1992.

For the purpose of explaining inter-industry variations in the level of protection in terms of a set of industry characteristics, we have estimated a linear-logarithmic regression model of the form:

$$y_j = \beta' x_j + \varepsilon_j \tag{14.1}$$

where β is a vector of unknown parameters, x_j is the vector of explanatory variables which incorporates the industry characteristics, and y_j is the degree of tariff protection.[2]

The dependent variable used in our analysis is therefore the level of tariff protection, being defined as follows:

$$T_j = \frac{\frac{1}{m}\sum_{k=1}^{m} T_k \cdot M_j^R}{M_j^R + M_j^{CE}} \tag{14.2}$$

where T_k is the average common external tariff,[3] M_j^{CE} are imports from the European Union, M_j^R are imports from the rest of the world and k denotes the tariff sectors classified under sector j.

The independent variables employed to explain inter-industry variations in tariff protection are product differentiation, market structure, scale economies, involvement of multinational firms in the sector and level of intra-industry trade. As it is well known, direct measurement of these variables presents a number of difficulties, so several proxies of them (detailed below) have been used.[4]

Regarding product differentiation, a distinction has been made between technological and horizontal differentiation. It can be expected that sectors in which the former has a particular relevance will enjoy a competitive advantage due to specific technological knowledge, being less interested in getting a high degree of protection. On the contrary, a positive association between horizontal product differentiation and pressure for protection can be anticipated.

The following alternative proxies of product technological differentiation have been used:

- Research and development expenditure (RD) to value added (VA) ratio:

- Instituto Nacional de Estadística: Encuesta Industrial 1989-1992 (National Institute of Statistics' Industrial Survey, EI-INE)
- Clasificación Nacional de Actividades Económicas (National Classification of Economic Activities, CNAE -74)
- Tariff sectors by Melo and Monés (1982)
- Ministerio de Industria y Energía: Encuesta de Coyuntura Industrial (Ministry of Industry and Energy's Industrial Conjuncture Survey, ECI-MINER)

2 The equation was estimated using ordinary least squares.

3 See Melo and Monés (1982).

4 More detailed information regarding this issue can be obtained in Carrera (1996).

$$RDR_j = \frac{\sum_{r=1}^{s} RD_r}{\sum_{r=1}^{s} VA_r} = \frac{RD_j}{VA_j} \tag{14.3}$$

where r refers to the sectors of the CNAE-74 nomenclature classified under sector j.

- Technological development expenditure (TD) to value added ratio:

$$TDR_j = \frac{\sum_{r=1}^{s} TD_r}{\sum_{r=1}^{s} VA_r} = \frac{TD_j}{VA_j} \tag{14.4}$$

Concerning horizontal product differentiation, two alternative proxies, both of them related to advertising intensity, have been considered:

- Advertising expenditures (AE) to sales (S) ratio:

$$AES_j = \frac{\sum_{l=1}^{t} AE_l}{\sum_{l=1}^{t} S_l} = \frac{AE_j}{S_j} \tag{14.5}$$

where l stands for the sectors in INE Industrial Survey classified under sector j.

- Relative advertising expenditure:

$$RAE_j = \frac{AE_j}{\sum_{j=1}^{17} AE_j} \tag{14.6}$$

As regards market structure, economic theory suggests that the benefit obtained from protection may increase when the firms are operating in highly concentrated markets. Concentration increases market power, reduces the likelihood of free-riders and makes it easier to achieve agreements. Accordingly, the anticipated sign for this variable is positive.

In order to capture the effect of market structure we have used a concentration ratio adjusted to take into account foreign competition. This internationally adjusted concentration ratio (ACR) has been constructed as follows:

$$ACR_j = \left(1 - \frac{M_j}{P_j - X_j - M_j}\right) \cdot CR_j \tag{14.7}$$

where P refers to value of production, X and M denote exports and imports, respectively, $CR_j = \sum_{l=1}^{t} \left(\frac{\sum_{n=1}^{3} P_n}{P_j} \right)$ and $\sum_{n=1}^{3} P_n$ is the production of the 3 largest establishments in the sector.

On the other hand, the existence of large economies of scale propitiates a smaller number of free-riders, therefore implying better chances to pressure for protection. The expected sign for the economies of scale variable is, consequently, positive. For the construction of this variable, the following expression has been employed:

$$SE_j = \sum_{l=1}^{t} SE_l \cdot \frac{P_l}{P_j} \tag{14.8}$$

where SE_l is the quotient between the minimum efficient size (MES)[5] and the cost disadvantage ratio.[6]

As regards the involvement of multinational corporations, two different possibilities have been considered. On the one hand, a high participation of foreign direct investment may give rise to more intense pressure for protection as the firms try to benefit from the increase in prices brought about by the tariff. On the other hand, downward protection pressures may be exerted in order to facilitate trade flows between the foreign-controlled subsidiaries and the parent company. The sign of the relationship between protection and share of foreign direct investment is therefore ambiguous. In order to capture the influence of multinational companies the following variable has been employed:

$$FDI_j = \sum_{r=1}^{s} FDI_r \,, \tag{14.9}$$

where FDI stands for foreign direct investment.

Finally, a number of authors have argued that adjustment to trade expansion may be easier in a setting of intra-industry trade and therefore resistance to trade liberalization will be lower. A negative sign is therefore expected for the variable related to the level of intra-industry trade, which has been constructed using the aggregated Grubel-Lloyd index, according to the following expression:

---

5 The MES is computed as the size (value of production) of the median establishment divided by the value of the sector's total production. By median establishment we mean the establishment which corresponds to the median of the accumulated distribution of the sector's production.

6 This relative cost disadvantage is computed as the quotient whose numerator is the value added per employee in establishments smaller than the MES and whose denominator is the value added per employee in the remaining establishments.

$$IIT_j = 1 - \frac{\sum_{i=1}^{v} \left| X_{ij} - M_{ij} \right|}{\sum_{i=1}^{v} \left(X_{ij} + M_{ij} \right)}, \tag{14.10}$$

where *i* refers to SITC-Rev.3 groups (3 digits).

14.3 Empirical Results

Table 14.2 shows the main results of estimating eq. (14.1). As can be seen from the table, some support is founded for the hypotheses presented in the previous section. All the estimated coefficients show signs that are generally in line with prior expectations, being statistically significant in most cases.

The coefficient of the explanatory variable reflecting the degree of technological differentiation, whether measured by R&D expenditure relative to value added (RDR) or by technological development expenditure relative to value added (TDR), is statistically significant in most cases, except for regression (5). This reveals a negative relationship between the level of tariff protection and the existence of this type of differentiation. The sectors in which these activities are more relevant seem to have a competitive advantage provided by their specific technological knowledge, which leads to an intense inter-industry trade, therefore lacking the need to press for tariff protection.

The sign on the regression coefficient of the variables related to product horizontal differentiation is positive, suggesting that such a type of differentiation positively influences the level of tariff protection. The estimated coefficients are significant in the case of AES in all the regressions and in regression (5) in the case of RAE.

The positive sign and strong significance of the coefficient of the scale economies variable (SE) in all the regressions where it is included gives wide support to the hypothesis that large scale economies imply a small number of free-riders and, consequently, bigger pressuring power and higher tariff protection. As regards the variable related to market structure ACR, it has been scarcely significant in the different estimations.[7]

The sign on the coefficient of the variable FDI is positive and this coefficient has been shown to be statistically significant in all the equations in which it is included. This result is consistent with the hypothesis that sectors with bigger involvement of multinational companies have a greater market power, producing higher levels of tariff protection. Nevertheless the result differs from that found by Saunders (1980) for Canada, which would also appear to be economically reasonable. Furthermore, it must be pointed out that the sign obtained to illustrate the association between these two variables may change if the time horizon of work were increased, which could probably reveal the existence of some

[7] It is interesting to mention the existence of a high correlation between ACR and the scale economies variable (SE), as well as the horizontal product differentiation variables (AES and RAE).

relationships that, for a more restricted period of time, may not be so clearly demonstrated.

Table 14.2. Determinants of the degree of tariff protection: empirical results

Variables (expected sign)	(1)	(2)	(3)	(4)	(5)
LRDR (-)	-0.1725 (- 2.154)[b]		-0.1426 (- 1.521)[c]	-0.1933 (- 1.886)[b]	-0.1312 (- 1.259)
LTDR (-)		-0.1485 (- 2.038)[b]			
LAES (+)	0.2248 (1.926)[b]	0.2273 (1.916)[b]		0.2565 (1.492)[c]	
LRAE (+)			0.0941 (1.213)		0.1192 (1.359)[c]
LSE (+)	0.2717 (2.654)[a]	0.2563 (2.503)[b]	0.1563 (1.756)[b]		
LACR (+)				0.5725 (1.638)[c]	0.2633 (1.127)
LFDI (+/-)	0.0756 (1.822)[b]	0.0593 (1.539)[c]	0.0880 (1.949)[b]	0.0946 (1.519)[c]	0.0888 (1.422)[c]
LIIT (-)	-2.8522 (- 4.928)[a]	-2.7588 (- 4.781)[a]	-3.0456 (- 4.312)[a]	-2.2275 (- 3.590)[a]	-2.7949 (- 3.815)[a]
R^2	0.707	0.697	0.650	0.607	0.594
$\overline{R}^2$	0.590	0.576	0.510	0.449	0.432
SER	0.386	0.392	0.422	0.447	0.454
SSR	1.488	1.540	1.778	2.000	2.064
F-statistic	6.039	5.754	4.646	3.853	3.657
N.O.	15	15	15	15	15

Notes:
Variables are preceded by L, which means in logarithmic
t statistic in parenthesis
[a] 1% level of significance
[b] 5% level of significance
[c] 10% level of significance

Finally, it can be observed that the intensity of intra-industry trade (IIT) seems to play an important role in the level of tariff protection. The negative sign on the

coefficient of the variable IIT and the strong statistical significance that this coefficient shows in all the equations suggest that adjustment factors intensely influence the structure of tariff protection.

14.4 Conclusions

Government regulation may respond to public interest motivations and, in fact, provide some corrections for the market failures, which may be profitable for the society as a whole. However, from the moment these public interventions take place, the risk is run of political powers responding to the demands of specific pressure groups. In fact, the results of the empirical analyses performed up to date find evidence that this may take place.

In this paper, we present some estimates for Spanish foreign trade in 1992, which offer a certain support for the previous view. We have, concretely, analyzed the potential relationship between tariff protection (taken as a particular case of regulation) and a series of variables which account for industry characteristics including technological factors, horizontal product differentiation, scale economies, market structure, involvement of multinational companies and intensity of intra-industry trade. In particular, by means of a conventional econometric model, we have tested, for the Spanish case, a set of hypotheses regarding some determinants able to affect the degree of tariff protection.

The results obtained suggest that there is a deterministic relationship between the inter-industry pattern of tariff protection and some characteristics of industrial structure. The empirical evidence found give support to the existence of a positive relationship between tariff protection and the degree of horizontal product differentiation, the presence of scale economies and the share of multinational companies in the sector. On the other hand, technological product differentiation and intra-industry trade show a negative influence on the degree of tariff protection. These results reinforce the findings of previous studies, which offer support for the role of pro-protection interest groups. Finally, it must be pointed out that these results, although interesting, must be taken with some caution since their robustness depends on their being confirmed by employing other time periods and different analytical approaches. This would be an obvious direction for further research. Another potential area of interest would be to perform the analysis at the European level.

References

Carrera-Gómez, G.: Comercio intra-industrial: análisis del caso español. Ph. D. Dissertation. Department of Economics, University of Cantabria (1996)

Caves, R.: Economic Models of Political Choice: Canada's Tariff Structure. Canadian Journal of Economics 9, 278-300 (1976)

Greenaway, D. and Milner, C.: Determinants of the inter-industry structure of protection in the UK. Oxford Bulletin of Economics and Statistics 56 (4), 399-419 (1994)

Grubel, H. G. and Lloyd, P. J.: Intra-industry trade: the theory and measurement of international trade in differentiated products. John Wiley &Sons (1975)

Melo, F. and Mones, M. A.: La integración de España en el Mercado Común. Un estudio de protección arancelaria efectiva. Economic Studies Institute (1982)

Saunders, R.: The Political Economy of Effective Tariff Protection in Canada's Manufacturing Sector, Canadian Journal of Economics 13 (2), 340-348 (1980)

15 Structure, Functioning and Regulation of the Spanish Electricity Sector. The Legal Framework and the New Proposals for Reform

F. J. Ramos-Real
University of La Laguna (Spain)

E. Martínez-Budría
University of La Laguna (Spain)

S. Jara-Díaz
University of Chile (Chile)

15.1 Introduction

An electrical system consists of a series of distinct stages: generation, transmission, distribution and supply (merchandising) of electricity services to the end-users. The traditional organisational model assumes, implicitly or explicitly, the extension of a natural monopoly condition from some of these stages to others. This is a consequence of the presumptive existence of strong, vertically-integrated economies. On the other hand, an increasing number of studies have proposed the vertical disintegration of the sector, suggesting that the common ownership of the different stages of the electric sector should be replaced by the introduction of competition wherever possible. These ideas have been developed within the context of a critique of the traditional control structure, characteristic of natural monopolies, which has been emerging in the industrialised world since the 1980's. The emphasis has now shifted towards the internal efficiency of the companies involved, and to uncovering those faults in the regulatory system which do not allow the product to be obtained at minimum cost.

P. Coto-Millán (ed.), *Essays on Microeconomics and Industrial Organisation*

In this paper, we study the structure, operation and regulation of the Spanish electricity system from 1983 to 2000. This system reflected that the general trend of reform was operating in Spain in 1983. The basic aim of the regulation was to ensure both the recovery and adequate financial return on investments made in the sector at a time of economic crisis. Furthermore, the regulatory system was particularly concerned with introducing incentives as a means for efficiency. The sector began a period of transition from a traditional system of control towards competition in generation and merchandising in 1997.

This paper is organised as follows: in section 2, the structure, functioning and regulation of the sector from 1983 to 1996 is analysed. The modification and improvement process, the basic principles of regulation, the companies financial returns systems, and their influence on the behaviour of companies, are described. In section 3, we summarise the main improvements proposed in 1997. Finally, in section 4, we present the most important conclusions which can be drawn from the study.

15.2 Structure, Functioning and Regulation of the Spanish Electricity Sector between 1983-1996

The structure and operation of the electric sector after the implementation of modifications to the system in 1984, along with the financial returns system operating in the companies until 1996, meant a great change, which had important economic consequences for the Spanish electricity industry[1]. We shall now describe the operation of the Spanish electricity board during that period.

The Spanish electricity sector[2], until 1996, operated as an integrated system. The transmission of electrical power and the short term management of the capacity for generation were in the hands of an independent entity operating under the name of Red Eléctrica de España (REE) or Spanish Electricity Network. The power generation needs for the entire network were defined by the National Power Plans (Plan Energetico Nacional, PEN). Distribution, for the most part, was the responsibility of large companies vertically integrated with generation; these companies were responsible for the supply within certain geographic regions and had the exclusive right to do so. These companies were integrated into the sector's managerial group UNESA[3].

---

1 This regulatory framework was in effect, in fact, until 1997 when the Electricity Act of 27 November 1997 came into effect, as the directions needed to apply the Ordering of the
2 National Electricity System Act passed in December 1994 were never developed.
We shall be looking only at the mainland's electricity system as the non-peninsular companies such as GESA in the Balearics and UNELCO in the Canaries operate as complete cycle
3 systems independent of the electricity network on the mainland.
Furthermore, there are some small distribution companies that acquire power generated by UNESA companies and resell it to the consumer at the end of the chain. Likewise, there also exists a series of so-called "self-producers" who produce electricity for their own industrial processes and who sell the excess to the electricity companies who, in turn, are obliged to acquire this power at prices set by the legislation. They are also obliged, under the same terms, to buy up the power from independently produced renewable energy sources.

Fig. 15.1, is a simplified flowchart showing how the system works. The UNESA companies transfer their production to the transmission network. This power plus the balance arising from international transfers, form a pool where the distributors obtain electrical power to distribute to the consumer. The operational features that are peculiar to the Spanish network are in the transmission stage, which operates and is managed independently of generation and distribution.

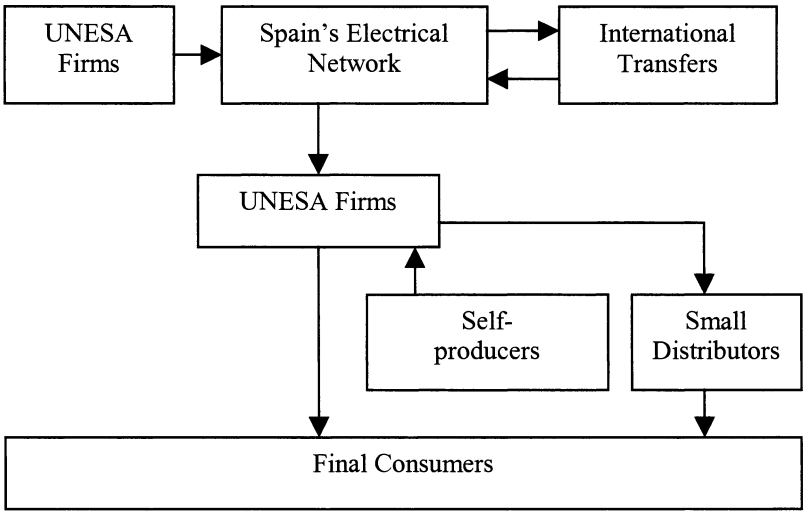


Figure 15.1. Simplified operations of the electricity system in Peninsular Spain

In 1996, the companies that constituted UNESA accounted for 88,9% of the gross production of energy and more than 90% of the distribution. UNESA was made up of ten vertically integrated companies, operating as regional distribution companies. Furthermore, a great proportion of generation was consolidated in the parent-company ENDESA , which acted only at the supply stage and which has been a public owned company for some time now. The production structure of the companies forming UNESA in 1996 is showed in Table 15.1.

Table 15.1. Production by UNESA firms in 1996

Generation.	Mill.Kwh.	%
Hydroelectric	37.694	24,1
Fossil-fueled	62.640	40
Nuclear	56.329	35,9

Source: UNESA Annual Reports

Regarding the installed capacity, the UNESA companies account for 92.65% of the total. The structure is shown in Table 15.2.

Table 15.2. Power installed in UNESA firms in 1996.

Generation	Power installed MW	% of total
Hydroelectric	16.547	36,4
Coal	10.925	24,1
Oil	8.065	17,8
Oil-Gas	2.355	5,2
Total fossils-fueled	21.345	47,1
Nuclear	7.498	16,5

Source: UNESA Annual Reports.

The generation field shows great diversification in the source of the energy. Compared with other countries, the Spanish electricity industry is characterised by a high proportion of hydroelectricity. However, there are important differences between the structure of installed capacity and the production structure. The role of coal and nuclear power in production is much higher than in capacity.

In the 1990's a rapid process of concentration took place thanks to various mergers which gave the ENDESA group (allowing for the absorption in 1996 of FECSA and SEVILLANA) 52% of the generation and 40% of the distribution market. The second group, IBERDROLA, holds a generation quota of 29% and 38% for distribution. The third and fourth producers, Union Fenosa and Hidrocantábrico, have 13% and 6% in generation and 15% and 5% in distribution, respectively.

15.2.1 The Reform Process and the Basic Principles Regulating the System's Operation (1983-1996)

The regulation until the end of 1996, which had been in effect since the early 1980's, had arisen in response to the sector's financial crisis. This crisis was the result of large investment programmes that started after the oil crisis in the 1970s. In an attempt to regulate the situation, government intervention increased during this period, setting a pattern based on negotiations between the companies and the government. In May 1983, an agreement was signed between the main companies in this sector and the Administration. The regulation and legal ordering of the Spanish electricity sector was set by law 49/84 of December 26th, which dealt with the unified administration of the sector. The Stable Legal Framework (Marco Legal y Estable, MLE) set by Royal Decree 1538/1987 regulated the economic environment in which companies should operate. The general outlines defining the Spanish regulatory framework during the period studied were three: centralised planning of the electricity systems by means of National Energy Plans, the unified

control of generation and transport, and the setting of standard rates for the entire country.

The agreement made the National Network and the company, public-owned for the most part, responsible for running the Spanish electricity system. The aim of this move was to ensure optimum efficiency, to maintain the National Network, and to promote international transfers of energy. The basic running practices were regulated by law 49/1984 of December 26th. On the 28th of January 1985, Red Electrica de España S.A. (the Spanish Electricity Board) came officially into existence, and assumed the controlling role.

The second additional clause of law 49/1984 established the need for approval of a general plan for the sector regarding the transfer of assets by the Ministry of Industry and Energy, aimed at achieving greater financial-economic equilibrium as well as power equilibrium. The previous unbalance was a consequence of dissimilar investments made by the different companies in response to the oil crisis. In 1985, the negotiations regarding the transfer of assets between the main electricity companies developed and came to an end. These negotiations lasted throughout the period of 1983-1996.

15.2.2 Unified Management and Central Planning

The National Energy Plan (PEN) 1983-1992 considered, for various reasons, paralysing the construction work on five nuclear power stations being built at that time. An order from the Ministry applied an extra charge to the price of electricity in order to finance this moratorium. Simultaneously, a plan was set up, whereby the construction of coal power plants was sped up and the work on oil-gas power plants was also paralysed. The existing Oil-Gas power plants should be used to cover the peak hours demand. REE decided on the extensions to be made to the distribution network, allowing no newcomers and keeping up the local monopolies. In accordance with the aims of the PEN this network extension was carried out by these monopolies. When growth caused overlapping between the areas of two local monopolies demand, the Government assigned it to one of the distributors.

The centralised running of the Spanish electricity sector is in the hands of REE. REE decides which power plants should operate, according to the so-called order of merit, which means the increasing order of variable costs. The system of operation aimed to reduce to an absolute minimum the supply costs while maintaining them within limits set by general criteria regarding safety and energy policy[4]. The policy of unified running is carried out within a structure whereby companies involved in the generation and distribution stages, and the one parent-company ENDESA (specialised in power generation) work hand in hand. Therefore, power transfers during each timetable block become necessary so that the production planned in each company's power plants, plus the balance of exchanged power, coincides with the demand.

The assignment of energy transfers, as well as its cost, took place through a pool formed by energy surplus from firms with excess capacity, since the REE

[4] In this respect it is worth mentioning the restrictions resulting from quotas and limitations on the use of national coal.

programme assigned them a production that exceeded their market necessities. To the surplus of each firm, a marginal cost was assigned equal to that of the energy delivered at the highest marginal cost. From these reference values, a weighted average price was calculated with the marginal costs of all the firms' surplus. This is the price that is taken into account to calculate the standard cost[5] of transfers. In the case of firms whose variable cost was larger than that of the pool, it was supposed that they deliver energy at this price and buy again at the new average price of the resultant pool.

In Table 15.3 we summarise the reform process in the period 1983-1997.

Table 15.3. Summary of the reform process

Year	Event	Development
1983	Agreement convention signed between firms	Revision of the PEN
1984	Law 49/1984 for integrated operations of the sector	Compensation system begins.
1985	_____	Development of assets exchange. January: creation of REE
1987	Legal and Stable Framework, Royal Decree 1538/1987.	
1988	_____	Development of the new compensation system of firms and MLE compensations.
1994	Approval of Law Ordering the National Electricity Sector (LOSEN)	Proposal of creation of independent system.
1997	Approval of New Electricity Law	End of traditional regulation system of the MLE.

15.2.3 The Rates and Financial Return Policy of the Legal and Stable Framework (MLE)

The guidelines regulating the economic and financial returns in the sector was finalised at the end of 1987 with the promulgation of the Stable Legal Framework, which came into effect in January 1988.

Although the mechanism of the MLE is rather complicated, it basically implied that a company involved in the generation and/or distribution stages received payments equal to its standard cost. The standard costs are a value set across all the companies involved in generation and distribution, based on both the fixed and variable costs, and including sufficient return for invested capital. Income from sales, according to the methodology of the MLE, should cover the cost of the service of the entire system. This cost is calculated by finding the aggregate of all the recognised standard costs. Furthermore, a series of extra charges is added to the rates.

[5] In this way no firm covers its demand with energy whose variable cost is more expensive than that of the pool. This mechanism allows afterwards the compensation system that needs to identify standard costs of trade for each firm.

The Ministry of Industry and Energy, by means of the General Board of Energy, determines the standard values following particular economic and energy parameters which define each concept involved in the cost. The costs that make up the total expenses to be included in the final prices are:

1. Fixed costs of generation. These cover the investments in the infrastructure and include the depreciation charges and returns on the assets.
2. Operation and maintenance costs. One part is considered fixed and another part variable according to the power installed or energy generated.
3. Variable costs arising from fuel used and transfers. This includes the costs of fuel and other fungible materials used in generation, the net cost of transfers with the pool and other transfers such as that with self-producers and international contracts.
4. Fixed and running cost in distribution. Levels of tension above or below 36KV are distinguished. In lower levels of tension this is calculated by means of the quantity of energy circulated. For higher levels, the type of investment is used for fixed cost and physical entities are used for running costs.
5. Merchandising costs. This comprises activities related to the upkeep and development of the market. This is standardised through the number of contracts and by the power turnover in tensions larger than 1 KV.
6. Cost of distribution and generation structure. This embraces costs that are not linked to productive activity and financial expenditures of clients' accounts.
7. Miscellaneous costs. This includes the quota of the Spanish Electricity Network and other surcharges in the invoice like the nuclear moratorium, the quota of the Office of Compensation (OFICO), basic stock of uranium and research funds. These surcharges together represented 14.38% of the electrical tariff in 1989 and 13% in 1996.

The Compensation System

There is just one rate for the whole country, but the different companies have both different generation equipment and different market structures, which lead to different distribution costs and different revenue per kwh sold. The acknowledged income of the companies is not the actual sum paid by clients, but the total of the acknowledged standard costs instead. Therefore a compensation system between companies becomes necessary in order to balance out the final income received by each firm with the sum of its recognised costs.

A more detailed description of the calculation of inter-firm compensation is provided by Rodríguez and Castro (1994). The compensation system aims to even out each company's unit cost regarding generation (generation compensation) and, on the market side, the idea is to even out each company's average income as compared with the average income of the system (market compensation). The algebraic sum of the compensations equals zero.

Generation compensation is calculated as follows:

$$Z_g^i = (CF_g^i + \frac{CV_g^i}{1+\pi}) - \Sigma(CF_g^i + CV_g^i)\frac{D_g^i}{\Sigma D_g^i} + \beta_i \frac{\pi}{1+\pi}\Sigma CV_g^i \qquad (15.1)$$

where:

Z^i_g: generation compensation of the firm i.

CF^i_g: standard fixed cost of generation of the firm i.

CV^i_g: standard variable cost of the generation of firm i.

D^i_g: demand of the firm i in Plant[6]

π: percentage which is taken from the variable costs to reward firms with the lowest variable costs.

β_i: coefficient of efficiency in variable costs of the firm i.

The first two terms of equation (15.1) reflect the difference between the company's average generation cost and the system average, multiplied by the company's market share. The parameter π represents the percentage of the variable costs not considered in the compensation. This creates a fund (generation margin) to be redistributed among the companies, according to the coefficient β_i[7]. The third term in equation (15.1) may be interpreted as the share of the generation margin due to each company's subsystem, based on technical efficiency. Thus those companies who contribute to the reduction of the cost of the service are rewarded.

Market compensation includes compensations for distribution costs, income from sales, and other revenue.

$$Z^i_m = \left(\frac{D^i_d}{\sum D^i_d}\sum I^i - I^i\right) - \left(\frac{D^i_d}{\sum D^i_d}\sum C^i_d - C^i_d\right) \qquad (15.2)$$

where:

Z^i_m: market compensation of the firm i.

C^i_d: fixed and variable costs of distribution and commercial management of the firm i.

D^i_d: demand in Plant, obtained in each tariff from the consumption of the subscribers for the firm i.[8]

I^i: collects the net sales turnover and other incomes from each firm.[9]

---

6 This is the sum of the energy generated in Plant from all the installations of each firm. This is standardized by a coefficient of their own consumption so that any saving in real consumption

7 means an additional profit for the firm in question.

8 This coefficient is calculated as being inversely proportional to variable costs.

The invoiced energy declared in each tariff is multiplied by a standard coefficient of losses to convert it in demand in Plant. In this way each kwh not invoiced suppose a loss for the firm. The standard coefficient is the average value of the system for which will coincide with that of generation for all the system but not for each individual firm.

The first component in equation (15.2), if positive, means that the company receives less revenue than is due them, according to the average for the sector[10]. The second component has a similar meaning to the generation compensation in that the company is compensated for the difference between its acknowledged cost and the average for the distribution sector.

We can sum up the basic principles that make up the MLE as follows (Rodríguez and Castro, 1994):

a) The administration determines for each firm a standard cost CS, according to its generation equipment and distribution structure.
b) Each firm conducts its production activity according to the directives from the managing firm from the integrated operations, incurring a cost C, and receives from sales in its market an income R thus obtaining a gross profit:

   GB=R-C.

c) Each firm receives a compensation (T) equal to the difference between the standard costs and their income (or payment if negative):

   T=CS-R.

d) The net profit received (GN) by each firm will be:

   GN = GB+T = (R-C) + (CS-R) = CS-C

The regulating method supposes that the aim of each company is to maximise the difference between standard and real costs. Regarding the productive efficiency of the system, it may be stated that the reduction of production costs is favoured, as any reduction in real costs benefits the company.

Effects of the MLE on Firms' Behaviour

On many occasions, the MLE has been classified as a case of yardstick competition, where the fixing of the price in any company is decided according to the average cost across the other companies. As Schleifer (1985) suggests, any improvement in efficiency in the sector becomes a modification of the 'yardstick'.

Rodríguez and Castro (1994), consider that calculation of the individual standard cost figures should be carried out in an *ad hoc* manner, arising from a specific price index (the Consumer Price Index, the Industrial Price Index or an average of the two). For this reason, the standard cost should be taken as a maximum price and updated periodically, independent of the average efficiency of the sector[11].

---

9 Gross income is converted into net income by detracting the charge for invoicing valid in each period and adding amounts received from the Office of Compensation (OFICO) for special tariffs such as off-peak electricity provision.

10 To give more detail, payment made by way of compensation is calculated , tariff by tariff, by comparing the average income of the company with the average income of the whole system. If the result is positive, the company keeps half, and if it is negative, it loses half. Thus the standardisation of revenue could encourage companies to increase their sale prices.

11 The classic *price cap* formula allows for price increases equal to the rate of inflation minus a factor X which reflects the average growth of productivity of the companies.

Kühn and Regibeau (1998) consider that the regulation system of the MLE has brought about incentives to reduce costs, but they point out a series of aspects which could have a negative effect on the behaviour of the companies during this period, against the intentions of the regulator. On one hand, the incentives for cost reduction were not applied equally to all types of costs. In the case of REE, such incentives did not even exist as the standard cost established was to be the same as its income. On the other hand, the aim of maximising the difference between real and standard costs can be achieved by increasing the standard costs after complicated negotiation between the government and the companies.

Crampes & Laffont (1995) studied, within the framework of the theory of incentives, how the MLE's financial return system created incentives for efficient behaviour. The standard costs CS and real costs CR for each company i are separated into fixed F and variable V:

$$CS^{i}{}_{average} \equiv \frac{CS^{i}}{q^{i}} = cvs^{i} + \frac{FS^{i}}{q^{i}}$$

$$CR^{i}{}_{average} \equiv \frac{CR^{i}}{q^{i}} = cvr^{i} + \frac{FR^{i}}{q^{i}}$$

The standard values depend, above all, on the company's decisions regarding investments, but they also depend on the regulator's assessment of the company's fixed and operative costs. A variable e_1 will be used to refer to the effort made by the management regarding equipment or technical issues ex ante, e. g. the choice of power plant size, which is beyond the control of the regulator. The variables of the real costs depend on e_1 and on the appropriate use of the equipment, thus we will call e_2 the variable associated with the appropriate use of equipment or technical effort ex post. Although the management does not decide on the price, they do have a say in the decision of supplying to each area of the market. As each company sells different products[12], we can consider e_1, e_2 and e_3 as vectors. The optimising model that explains the behaviour of the company may be expressed as follows:

$$\max_{e1,e2,e3} \sum_r \left(CS_r^i - CR_r^i\right) - \psi^i(e_1, e_2, e_3))$$

$$= \sum_r \left[(cvs_r^i(e_{1r}) - cvr_r^i(e_{1r}, e_{2r}))q_r^i(e_{3r}) + FS_r^i(e_{1r}) - FR(e_{1r})\right] - \psi^i(e_1, e_2, e_3))$$

where r represents each type of tariff and ψ the disutility or cost of the effort.

The technical decisions depend on e_1 and e_2, such that the first order conditions of the problem are:

$$q_r^i\left(\frac{dcvs_r^i}{de_{1r}} - \frac{\partial cvr_r^i}{\partial e_{1r}}\right) + \frac{d(FS_r^i - FR_r^i)}{de_{1r}} = \frac{\partial \psi^i}{\partial e_{1r}} \qquad (15.3)$$

$$-q_r^i \frac{\partial cvr_r^i}{\partial e2r} = \frac{\partial \psi^i}{\partial e2r} \tag{15.4}$$

The condition (15.4) shows that the marginal disutility of effort in variable costs coincides with the marginal profits derived from the reduction in variable costs. So, technical effort *ex post* leads to minimise costs through the compensation mechanism. From condition (15.3) it cannot be deduced that the technical effort *ex ante* is adequate, that is, the marginal disutility of effort in fixed costs does not coincide with the marginal profits derived in fixed cost reductions, for this it must be that:

$$-\frac{\partial CR_r^i}{\partial e1r} = \frac{\partial \psi^i}{\partial e1r} \tag{15.5}$$

We can deduce that the incentives derived from the regulatory framework can produce bias in investment decisions. As regards market effort the first order condition is:

$$(cvs_r^i - cvr_r^i)\frac{dq_r^i}{de3r} = \frac{\partial \psi^i}{\partial e3r} \tag{15.6}$$

This shows that the company is interested in concentrating its sales at those rates where the cost is the lowest in relation to the standard value defined by the regulator. The company does not study social welfare, measured in terms of the individual surplus of each type of consumer; nor does it find a solution to the secondary problem presented by a competitive balance whose solution would be:

$$(pr_r^i - cvr_r^i)\frac{dq_r^i}{de3r} = \frac{\partial \psi^i}{\partial e3r} \tag{15.7}$$

Crampes & Laffont highlight a further series of facts derived from the regulation and financial returns system, which we detail below:

1. Firms are remunerated on the basis of the equipment available encouraging the management to declare in total avoiding selective declarations.
2. The MLE determines two complementary mechanisms to correct inefficiencies. The first was a share of the margin on variable costs (generation margin) to redistribute it between the firms according to the coefficient β. This produces an incentive that approximates efficient behaviour *ex ante* although limited by the lack of weight that this margin has since it does not influence fixed costs.
   The second mechanism is to try to create incentives for the adequate behaviour of the market effort. For this the company only recognised half of the difference

[12] Considering the quantity sold in each tariff as a separate product.

in each tariff between the average sector income and that of each firm. The advantage of this mechanism allows consideration of the prices as a decision variable so that firms internalise the market structure.

3. The system resembles yardstick-competition but taking as a reference standard costs instead of the sector average. These standard costs take into account the sector's heterogeneity and avoid the production of huge profits or losses that would result from the pure application of a pure reference system.
4. From the dynamic point of view efficiency can be affected in different ways. The revision of standard costs is achieved in a discretional way[13] so that firms fear that real reductions of costs mean reductions in standard values and consequently a possible decrease in future income. This can discourage firms from investing appropriately.
5. A difficult element for the regulator to control is product quality. This is not easy to distinguish in the case of network investment if the aim is to expand or to improve the service. In the regulatory framework in force, if firms tend to minimise cost against quality, other firms will not be remunerated appropriately so that this is a problem of overall regulation. In the MLE this is a personalised problem in the context of the revision of the standard costs.
6. Another problem emerges because the system does not provide incentives to save energy since a co-ordination mechanism does not exist to reduce production. The compensation system in fixed costs means that firms have equipment ready to produce, and the mechanism of variable standard cost ensures a safe profit for any quantity that is produced.

The Evolution of Productivity

There may be certain reservations regarding incentives for efficiency present in the terms of the MLE, but the majority opinion is that considerable improvements have been made in the technical efficiency and profitability of the sector.

Kuhn and Regibeau (1998) point out certain indications that would support this opinion. For them, the prices of electricity in Spain are below the average of the surrounding countries in the majority of consumer categories. This fact, along with the high profit level in Spanish generation companies would suggest that the price of generation in Spain is somewhat lower than in many other industrialised countries.

The report produced by UNESA (1997) also tells us that, during the period that the SLF was in effect, –and more specifically from 1988-1995– utilities achieved increased efficiency which was transferred to the consumer in part through the drop in electricity rates in real terms (10.6% during that period). Equally, the report points out that the rates in Spain, both for domestic and industrial use, have been kept below the average of the main European countries.

Arocena and Rodríguez (1998) assess the consequences of the regulation on productivity in coal-based electricity generation during the period 1988-1995, using the Malmquist productivity index. The unit of analysis is the generating

[13] Apart from the factor depending on the Consumer Price Index or the Industrial Price Index, it is not known exactly what other adjustments are involved in the calculation of standard costs.

group and capital; work and fuel are factors considered. The main conclusions of the paper are the following:

- Productivity increases are observed for all groups during these years apart from 1989 to 1990. The annual average rate of productive growth between 1988 and 1994 is 3.2%. This productivity index can be broken down into the rate of technical efficiency and technical progress.
- The rate of technical efficiency shows its greatest increase the first year (4.7%) and the last year (2%). The first case can be explained by the immediate effects of the MLE coming into effect. In the second case, the explanation is to be found in the improvements brought into the running of the system thanks to the competitive environment created by the new Law Ordering the National Electricity System (LOSEN) in 1994.
- The rate of technical progress shows a moderate increase, except between 1991 and 1992, when it was 5.6% as a consequence of the environmental measures which required large investments and improvements in the thermal efficiency of the plants.
- This index should be modified, bearing in mind the effect of the rate of installed energy used, given that the greater use of fixed factors could explain, in part, the improvements in productivity. The new index shows improvements each year, with an average of 2.8% between 1988 and 1994.

Ramos (2000) has carried out a study of the evolution of productivity in the Spanish electricity sector, by means of the estimation of a multiproduct long run cost function, where the unit studied is the company. The results suggest that productivity has improved almost 20% during the period of 1989-1996, with an annual rate of 2.62%. The most significant improvements occurred between 1989 and 1993, during the first years of the MLE.

The improvements in productivity have, for the most part, been expressed as increased profit for the companies, as the adjustment of price rates did not take into account the possible gain in productivity.

15.3 Regulation Reform in the Spanish Electricity Sector from 1997

In this section we will deal with the most significant aspects of the renewal process undertaken in the Spanish energy sector, starting in 1997. Fundamentally, it has been a case of developing the system from a traditional control model to a market model based on the generators and on the final demand for energy. We shall detail the views and opinions of different writers about the process underway, specifically the analyses by Kühn and Regibeau (1998), Marín (1999), and Rodríguez (1999).

The reform of the regulation directed towards the market was discussed initially in 1993, and was started by the promulgation of the Law Ordering the National Electricity Sector (LOSEN) in 1994. This legislation permitted the gradual introduction of competition in the sector without totally dismantling the system

established by the MLE. The idea was to create a competitive energy market parallel to the existing system. The National Commission for the Electricity System (CNSE), which was a regulatory institution independent of the Ministry of Industry and Energy, was created although the latter retained the power of final decision. The problems arising from the system designed by the LOSEN, in combination with the change of government in 1996 accelerated the reform process. The companies reached an agreement with the government at the end of 1996, called the Electricity Protocol, which provided the basis for the new Electricity Law of November 27th, 1997.

The Electricity Law of 1997 (LSE)

The 1997 Electricity Law (LSE) extended the liberalisation brought about by the LOSEN and created the electricity wholesale market, with an initial transition period to deregulate prices and re-structure the market. This liberalising process has to meet the standards laid out in European Guidelines 96/92 EC on the community rules governing the internal energy market. The liberalising process suggested in the EU guidelines has been, however, slower than that followed in Spain.

The final aim is to completely deregulate generation and merchandising. Transmission and distribution, networks by definition, will continue to be regulated, as will be the tolls applied for their use. The law only specifies, within the areas of these regulated activities, that the tolls should be related in some way to the costs and should be uniform across the State. Third party access to the network will be guaranteed, assuming available capacity.

The agents who will take part in the electricity market are the generators who produce the electricity, the companies whose high-voltage wires carry the electricity, the distributors who serve the non-eligible customers, the customers eligible because of the volume of energy they consume, and the new marketing companies. The market will be overseen by two operating companies: the system operator who physically manages both the network and the delivery of power, and the market operator who directs the energy transfer system to determine the market price. The CNSE will inspect the system.

A company with sufficient financial and technological means will be allowed to enter the generation segment; any agent with sufficient financial resources will be able to sell energy to any type of consumer. The income of these marketing agents will depend solely on the contracts signed with their customers. The prices obtained by the generating companies will be decided by the market or bilateral contracts.

The spot market functions as a double auction where there are sales and purchase offers on the demand side. A regulatory surcharge, known as the power guarantee, is added to the spot market price, to avoid insufficient supply. The initial amount was fixed at 1.3 pesetas/kwh and corresponds to all the capacity available during the 4,500 peak demand hours of a year. At the end of 1998, the average charge was around 1.26ptas/kwh.

The merchandising segment will be liberalised gradually, so that, in year 2001, all high voltage customers will be able to choose their supplier, (which represents 50% of the power consumed). The customers who cannot choose freely will still

be within the influence area of particular distributors; the rate will be set by the government, and will be standard across the country. This rate will be based on the permanent costs of the system (operators and CNSE), the purchase price of the electricity, distribution and transmission costs, and two types of financial returns from transition costs: the nuclear moratorium and the expenses incurred by transition to competition.

The regulating regime imposes some type of vertical separation of activities in relation to property and accounting regulations. The operators will be private companies. Companies and consumers operating on the spot market will be allowed to participate, though with a maximum limit to the number of shares. The companies taking part in any of the regulated activities will not be allowed to participate directly in the non-regulated areas. Although there may be a legal separation, the presence of holding companies operating in both fields will be permitted. The accounting regulations require keeping separate accounts in the case of firms with shares in more than one regulated area; this is required from those companies that only participate in the areas subject to competition.

The Nuclear Moratorium and the Costs of Transition to Competition (CTC)

The payments to the firms affected by the nuclear moratorium have been extended indefinitely. The companies receive compensation by means of a surcharge on the price of electricity, which cannot exceed 3.54% of the income obtained. The Transition to Competition Costs (CTC) are aimed at compensating the loss of capital of the companies constituting the MLE on December 31st 1997, due to the introduction of competition. This payment will be expressed in pts/kwh and will reflect the difference between the average revenue obtained by these companies under the previous system and that obtained in the spot market. If the average price on the market exceeds 6 ptas/kwh, the difference will be deducted from the discounted value of the compensation. These payments will be made during a period of ten years and will increase the utility rates.

The discounted value of the compensation will not exceed 1,988,561 million pesetas (including the incentives related to coal). These coal-related incentives have been a load on the sector, since the national coal cost is twice as much as coal on the international market. These incentives will last throughout the transition period, as the law explicitly permits the authority's interference in the rules so that the use of national sources of primary energy may reach 15%.

Controversies

The CTC has been very controversial since the start of the liberalisation process. The calculations noted in the Protocol were widely criticised by the CNSE and consumer associations. For some existing assets, the price obtained on the market may suppose a loss in value (the remuneration through market price being lower than the costs acknowledged under the previous remuneration system); however,

the opposite may also occur[14]. The controversy has reappeared with the claim on part of the CTC (worth 1.3 billion pesetas) by the power firms, a claim backed by the government. There is a debate regarding the exact final quantity of money to be paid. Moreover, the European Commission's intervention considering the CTC as a disguised grant system to favour national companies, places the whole process in question.

The opinion that the sector is highly concentrated on a few firms seems fairly unanimous, as is the opinion that these companies operate following the vertical integration structure in the area of generation and distribution. The privatisation of ENDESA could have been carried out segregating the assets beforehand, but the opposite route was chosen. The government allowed ENDESA to acquire other companies and create a larger group. Rodríguez (1999) has pointed out that this can be analysed from two different angles. In the national context, and bearing in mind the scarce capacity for international connection, the level of concentration could be considered excessively high for the market to function efficiently. If, on the other hand, we consider the international market and adopt a mid-term and long-term perspective, any policy of de-concentration could have an influence on the future competitive ability of the Spanish companies. Ramos (2000) noted the existence of moderate economies of vertical integration between generation and distribution and, to a greater extent, the presence of economies of horizontal integration between the different types of generation and distribution. The existence of savings obtained from undertaking different activities together should be kept in mind when restructuring the sector, but it is not incompatible with the vertical disintegration of the sector as long as the markets allow effective competition in each area.

Another problem is the distortion that the CTC can bring about in the spot market. The control of the spot market price by the same companies that receive the CTCs generates unethical incentives. The established companies could try to keep the prices in this market down, in order to claim maximum compensation and, at the same time, make it difficult for other companies to enter the field of generation. Rodríguez (1999) notes that the average price of the market in 1998 has settled to around 6 pts/kwh, thus maximising the income derived from the market without affecting the maximum quantity recoverable through the CTC.

The regulatory effort is insufficient to introduce competition into the sector in Spain, according to Kühn and Regibeau (1998). We now detail the opinion offered by these authors, who compare the Spanish situation with that of the United Kingdom. Basically, they analyse three issues: the concentration of the generation field, the slow liberalisation of merchandising and the high level of vertical integration.

[14] One illustrative example is that of the Austrian regulator. In Austria, as in Spain, hydroelectric energy is remunerated according to costs, independent of when it was generated. However, liberalisation allows it to be remunerated at the price of the pool, which at peak times is much higher, providing extra revenue for the companies. The Austrian regulator considered that this improvement in the remuneration system more than compensated for the CTCs and so did not award any further compensation. Each European country has followed different criteria in their liberalisation process, regarding this type of compensation for transition costs. In England and Wales, for example, it has been incorporated into the company sale price.

1. In the United Kingdom, the high level of concentration on the supply side allowed only a few companies the control of marginal supplies on the spot market; the gains obtained in productivity were not felt by the consumer. In Spain, only two companies (ENDESA & IBERDROLA) control the majority of the assets that determine the marginal price of the market: the coal power plants and the hydroelectric plants. The problems of market structure are worsened by other characteristics of the Spanish sector. In Great Britain, the larger companies' share began to deteriorate with the introduction of combined-cycle technology. In Spain, there was a greater capacity surplus in the sector and the primary energy source, natural gas, is practically a monopoly. These companies make agreements with the existing generators to keep new firms from entering the field.
2. The extended period of transition for the liberalisation of the merchandising allows the distribution monopolies already in existence to set up barriers protecting themselves from competition in distribution. The manipulation of the final price for consumers is possible given that it is set by an implicit agreement between the government and the companies. Although the government and the firms have committed themselves to annual price reductions of 3%, the high margins in generation allow greater price reductions.
3. The high level of vertical concentration does not seem appropriate, nor does the delay in the freedom to choose the supplier by final consumers. These two circumstances could make the price hardly vulnerable to competition pressure, given that the same companies will bid on demand as distributors and/or merchandising agents.

Finally, the regulating institutions have been designed to give MIE greater control over the CNSE. Moreover, the government has certain prerogatives for the fixing of tolls and for other important decisions.

15.4 Conclusions

The organisation structure of the electricity sector between 1983 and 1996 drew together the features of a vertically integrated and a non-integrated structure. The transmission stage operated separately from the generation and distribution stages, and its management was also separated from those two stages. The regulation system assumed that the aim of the company was to maximise the difference between standard and real costs, in order to favour the reduction of production costs, given that any decrease in real costs supposed an increase in gains for the company. The studies carried out by different analysts suggest that, while the Stable Legal Framework was in effect, the electricity companies achieved increased productivity mainly thanks to management improvements, which had positive repercussions on the efficiency of the companies.

The reform of the sector that got underway in 1997 had as its goal the complete deregulation of the areas of generation and merchandising. The new scheme for operating and regulating will be developed gradually. Some experts express specific doubts about the future of the liberalisation, basing their opinion on the

point of departure of this process. The factors that encourage this opinion are largely focussed on four issues: the low capacity for international connection, the excessive concentration in the area of generation, the slow liberalisation of the marketing area, and a high degree of vertical integration.

References

Arocena, P. and Rodríguez, L.: Incentivos en la regulación del sector eléctrico español (1988-1995) (Incentives for regulation in the Spanish Electric Sector, 1988-1995). Revista de Economía Aplicada 18, 61-84 (1998)

Crampes, C. and Laffont, J. J.: Transfers and Incentives in the Spanish Electricity Sector. Revista Española de Economía. Monográfico Regulación, 117-140 (1995)

Kühn, K. and Regibeau, P.: ¿Ha llegado la competencia?. Un análisis económico de la reforma de la regulación del sector eléctrico en España (Has competition arrived? An economic analysis of the regulatory reform in the Spanish Electric Sector). Instituto de Análisis Económico 1998

Marín, P. L.: Liberalización y competencia en el sector eléctrico (Liberalisation and competition in the electric sector). Economistas 80, 62-71 (1999)

Ramos Real, F. J.: Economías de integración y productividad en el sector eléctrico español en el periodo 1983-1996. Un enfoque multiproductivo (Economies of Integration and Productivity in the Spanish Electric Sector. A Multiproduct Approach). Ph. D. Thesis, Departamento de Analisis Economico, Universidad de La Laguna, 2000

Rodríguez Romero, L.: Regulación, estructura y competencia en el sector eléctrico español (Regulation, structure and competition in the Spanish Electric Sector). Economistas 82, 121-132 (1999)

Rodríguez Romero, L. and Castro Rodríguez, F.: Aspectos económicos de la configuración del sector eléctrico en España: ¿ Una falsa competencia referencial? (Economic aspects of the electric sector in Spain. A false yardstick competition?). Cuadernos económicos de I.C.E. nº 57, 161-183 (1994)

Shleifer, A.: A theory of yardstick competition. Rand Journal of Economics 16-3, 319-327 (1985)

UNESA: Evolución económico-financiera del sector eléctrico 1988-1995 (Financial-Economic Evolution of the Electric Sector). UNESA 1997.

16 Effects of a Reduction of Standard Working Hours on Labour Market Performance*

C. Pérez-Domínguez
University of Valladolid (Spain)

In recent years many people have proposed a general reduction of the number of working hours as an effective measure to reduce unemployment rates in the European countries. This proposal has had a strong effect on public opinion, since the "working-less-for-everyone-to-be-able-to-work" assumption seems to be a self-evident truth. But there is a fallacy involved in this assumption: the labour market is rather dynamic and neither the jobs available nor the number of applicants have to remain fixed when the standard working hours are reduced by legal means.

This paper will develop a theoretical model that will enable us to ascertain how a reduction of the standard working hours affects the labour market performance.

The first section of this paper studies the expected effects of a reduction of standard working hours on employment. The second section analyses the effects of this measure on labour force participation. The third section combines the results of the previous sections in order to evaluate the effects of such a reduction on the unemployment rate. The fourth and final section summarises the main results of the present paper.

* I am grateful to José Miguel Sánchez-Molinero for helpful comments and suggestions. I acknowledge financial support for Spanish Ministry of Labour.

P. Coto-Millán (ed.), *Essays on Microeconomics and Industrial Organisation*

16.1 Effects on Employment

16.1.1 The Basic Model

We shall start by assuming that the labour demand function for an individual firm is given by the following expression

$$w = f(H);\ f(h)>0;\ f'(H)<0 \tag{16.1}$$

where H stands for the total number of hours demanded and w is the real wage per hour.

We assume that the length of the working day, h, is legally fixed. The firm has to decide how many workers, N, is going to hire, so that the total number of working hours, H, is optimal. We assume that overtime is not allowed.

The previous assumptions allow us to write

$$W/h = f(h{\cdot}N) \tag{16.2}$$

where W is the wage per person. We assume that W is fixed either by the law or by collective bargaining. Taking logarithms and reordering this equation, we have

$$\log W = \log h + \log f(h{\cdot}N) \tag{16.3}$$

Taking differentials on both sides of this equation, we obtain

$$\dot{W} = \dot{h} + \frac{\mathrm{dlog}f}{\mathrm{d}H}(h \cdot \mathrm{d}N + N \cdot \mathrm{d}h) \tag{16.4}$$

where the dotted variables indicate logarithmic derivatives (that is, growth rates). If we now multiply and divide the right hand side by H, we obtain

$$\dot{W} = \dot{h} - \frac{1}{\varepsilon}(\dot{h} + \dot{N}) \tag{16.5}$$

where ε accounts for the elasticity of the working hours with respect to the hourly wage in absolute value.

Regrouping and solving for $\dot{N}$ we have

$$\dot{N} = (\varepsilon - 1)\dot{h} - \varepsilon \dot{W} \tag{16.6}$$

This expression tells us how the number of employees varies in response to changes in the standard working hours and in the wage per person.

Next, we shall analyse the effects of a reduction of the standard working hours ($\dot{h} < 0$) on the number of workers demanded under two extreme theoretical assumptions: first, the wage per worker does not change ($\dot{W} = 0$); and second, the wage per worker falls in the same proportion as the working hours.

If the wage per worker is not altered, the effects of a measure that reduces the standard working hours on employment is given by

$$\dot{N} = (\varepsilon - 1) \cdot \dot{h} \tag{16.7}$$

This means that employment could increase, decrease or remain constant depending on whether the working-hours elasticity (in absolute value) is lower, higher or equal to one.

Some economic works on the Spanish economy have found that the most feasible values for ε are those close to one[1]. In the case $\varepsilon = 1$, a reduction of the number of working hours would not cause any effect on employment, unless it included some wage cutting measures.

In the case the wage per worker is adjusted in such a way that payment per hour remains constant, $\dot{w} = 0$, expression (16.6) implies that[2]

$$\dot{N} = -\dot{h} \tag{16.8}$$

regardless of the value of ε. In this particular case, the decrease in the number of working hours would increase employment in the same proportion.

When $\varepsilon > 1$, we already know that cutting the standard working hours causes a negative effect on employment unless a wage reduction takes place. Hence, we may ask the following: in what proportion should the wage per worker fall in order to avoid the negative effect on employment?

We want $\dot{N} = 0$; hence, expression (16.6) implies

$$\dot{W} = \left(1 - \frac{1}{\varepsilon}\right)\dot{h} \tag{16.9}$$

Given that $\varepsilon > 1$, the term within brackets must be positive and lower than one. Therefore, the wage per worker falls in order to keep the employment level constant.

To sum up, the conclusion of the previous argument can be stated as follows:

If the wage rate does not vary, a reduction of the standard working hours increases employment, provided that the elasticity of the working hours with respect to wages, ε, is less than one, and reduces employment when $\varepsilon > 1$. When $\varepsilon = 1$, the effect of the reduction of working hours on employment is null.

The negative effects of this type of measure on employment are, then, constrained to the case where $\varepsilon > 1$. But these effects can be alleviated if the reduction of the working hours is accompanied by a wage cut. In order to avoid a decrease in the employment level, the wage per worker should fall (although in a lower proportion than that of the working hours).

16.1.2 The Role of Fixed Labour Costs

To extend the basic model, we must consider the existence of fixed costs associated to the hiring of workers. These costs do not vary when the number of working hours (given a particular number of workers) diminishes. Such costs have to do with employee screening expenditures, as well as with worker training, monitoring, and welfare expenditures.[3]

---

[1] Dolado J.J. (1991, p. 675).

[2] Note that, given the wage per hour $w = W/h$, a constant w requires that $\dot{W} = \dot{h}$.

[3] For a complete classification of this costs, see Hamermech (1993, p. 47).

When such costs exist, the total hourly wage (w) can be written as the sum of two components: a variable component (w^V), and a fixed component (w^F), that is

$$w = w^V + w^F = (W^V + W^F) / h \tag{16.10}$$

where W^V and W^F are the variable and fixed hiring costs per employee.

Let us define w^F as a proportion α ($0 < \alpha < 1$)[4] of w^V. The demand function of working hours can be written as

$$w = w^V + \alpha\, w^V = (1 + \alpha)\, w^V = (1 + \alpha)\, W^V / h = f(h\, N) \tag{16.11}$$

Taking logarithms in both sides of the above equation and reordering terms we obtain

$$\log W^V = \log h - \log (1 + \alpha) + \log f(h\, N) \tag{16.12}$$

and fully differentiating

$$\dot{W} = \dot{h} - d\log(1+\alpha) + \frac{d\log f}{dH}(h \cdot dN + N \cdot dh) \tag{16.13}$$

This expression allows us to write

$$\dot{N} = (\varepsilon - 1) \cdot \dot{h} - \varepsilon \cdot \dot{W}^V - \varepsilon \cdot d\log(1+\alpha) \tag{16.14}$$

If we use the following approximation

$$\log (1 + \alpha) \approx \alpha \tag{16.15}$$

expression (16.14) becomes

$$\dot{N} = (\varepsilon - 1) \cdot \dot{h} - \varepsilon \cdot (\dot{W}^V + \alpha \cdot \dot{\alpha}) \tag{16.16}$$

We may now assume that the flexible part of the wage per worker adjusts to the variation in the working hours, so that w^V does not vary; that is

$$\dot{W}^V = \dot{h} \tag{16.17}$$

We may also assume that the fixed part of the wage per worker is not altered in spite of the fact that the working hours are being reduced; that is

$$\dot{W}^F = 0 \tag{16.18}$$

Given that $W^F = \alpha \cdot W^V$, expression (16.18) implies that

$$0 = \dot{W}^F = \dot{\alpha} + \dot{W}^V = \dot{\alpha} + \dot{h} \tag{16.19}$$

This means that, when h falls and W^F remains constant, α must increase in the same proportion as the standard working hours fall; that is

$$\dot{\alpha} = -\dot{h} \tag{16.20}$$

Substituting expressions (16.17) and (16.20) into (16.16), we obtain

[4] This assumption can be justified on the basis of experience. Experience suggests that training costs, monitoring costs, etc. are, on the whole, less than total variable wage costs over the entire period during which the worker remains attached to the firm.

$$\dot{N} = (\alpha \cdot \varepsilon - 1) \cdot \dot{h} \tag{16.21}$$

In the case there were no hiring costs, α would become zero and the above expression would boil down to

$$\dot{N} = -\dot{h}$$

This coincides with (16.8), and we already know the meaning of this expression.

If hiring costs are positive, $\alpha > 0$, provided that $\dot{h} < 0$, it must be true that

$$\dot{N} < -\dot{h}$$

This means that the existence of hiring costs reduces the potential increase of employment caused by a reduction of the working hours, when this reduction is matched by a wage cut in the same proportion.

When the reduction of the working hours is not matched by any wage cut, $\dot{W}^V = \dot{W}^F = 0$, (meaning that $\dot{\alpha} = 0$), expression (16.16) becomes

$$\dot{N} = (\varepsilon - 1) \cdot \dot{h}$$

This means that changes in N adjust to changes in h according to the value of ε. This is exactly the same situation that we met before and does not need any further comments.

In sum, we may conclude that the potential positive effect on employment of a reduction of the working hours will be lower in industries where hiring costs represent a higher proportion of the wage bill.

16.1.3 Effort Effect and Organisational Effect

So far, we have not distinguished between the number of working hours really hired by employers ($H = h \cdot N$) and the effective working hours (H^e).

This distinction arises when the "productivity" of an hour's work varies depending on the level of effort, and this effort is nor always the same.[5] Organisational factors, such as the size of the working team or the efficiency of monitoring may be affected by the reduction of the working hours, which, in turn, might have some influence on the efficiency of labour.

When the efficiency of labour varies due to a change in effort, we shall speak about an "effort effect". When the efficiency of labour changes due to the organisational factors above mentioned, we shall speak about an "organisational effect".

In principle, the effort effect would have a permanent character, whereas the organisational effect appears to be essentially temporary. The reduction of the working hours reduces work weariness, and this is a permanent effect. The organisational effect, however, can be regarded as a short-term distortion in the firm's organisational plans.

[5] Nickell, S.J. (1996, p. 634) also distinguishes between actual and effective working hours. Fallon and Verry (1988, p. 118) also analyse the role of work weariness and the length of the working day on labour productivity.

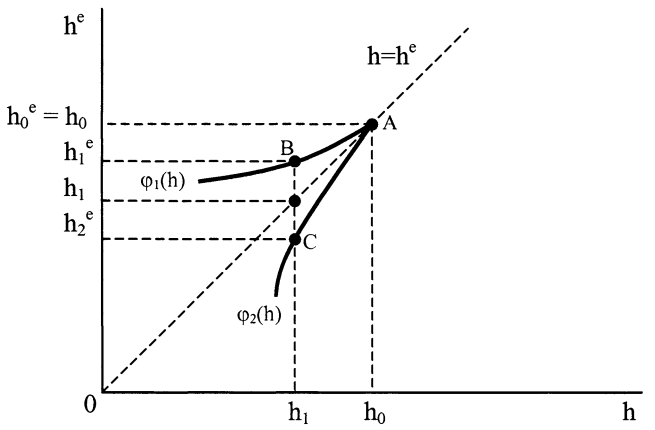


Fig. 16.1

In order to model these two effects, we are going to assume that the firms' effective working hours are defined as

$$H^e = N \cdot h^e = N \cdot \varphi\,(h) \tag{16.22}$$

where the number of effective hours, h^e, is a function φ of the actual number of hours, h. Function φ behaves as shown in figure 16.1. This figure can be explained as follows:

1. Prior to the reduction of the working hours, the working day has a length h_0. In such a situation we assume that $h^e_0 \equiv \varphi\,(h_0) = h_0$. This basically implies that production processes have been designed for a working day such as h_0.
2. The effort effect is reflected in the $\varphi_1(h)$ function. This function is defined for $h \leq h_0$, is increasing ($\varphi'_1 > 0$), and strictly convex ($\varphi_1'' > 0$). It also lies above the 45° line, meaning that, when the working hours are reduced by legal means (h becomes less than h_0) work weariness decreases and labour becomes more productive. Thus, the effective number of working hours can be expected to decrease less than the actual working hours ($h^e \equiv \varphi_1(h) > h$, $\forall\ h < h_0$), (point B of figure 16.1).
3. The organisational effect is reflected in the $\varphi_2(h)$ function. This function is defined for $h \leq h_0$, is increasing ($\varphi'_2 > 0$) , and strictly concave ($\varphi_2'' < 0$). It also lies below the 45° line, meaning that, when the working hours are reduced by legal means, the firm has to readjust its production processes to a shorter working day, and that implies a reduction in the efficiency of labour. Thus, the effective number of working hours can be expected to decrease more than the actual working hours ($h^e \equiv \varphi_2(h) < h$, $\forall\ h < h_0$), (point C of figure 16.1).

According to the above assumptions, the demand for working hours would be expressed by

$$w = f\,(H^e) \Rightarrow W / h = f\,[N \cdot \varphi(h)];\ f\,(H^e) > 0;\ f'\,(H^e) < 0 \tag{16.23}$$

Taking logarithms in both sides of this equation and fully differentiating, we obtain

$$\dot{W} = \dot{h} - \frac{1}{\varepsilon^e}\left[\dot{N} + \frac{h \cdot \varphi'(h)}{\varphi(h)} \cdot \dot{h}\right] \tag{16.24}$$

If we compare this expression with (16.5), we can observe that the only difference between them lies in the elasticity term, $h.\varphi'(h) / \varphi(h)$, that multiplies $\dot{h}$ in the right hand side of (16.24). This elasticity (from now on represented as χ) measures the sensitivity of the effective working hours with respect to changes in the standard working hours.

According to the above, expression (16.24) can be rewritten as

$$\dot{N} = (\varepsilon - \chi)\cdot\dot{h} - \varepsilon\cdot\dot{W} \tag{16.25}$$

This expression is very similar to (16.6). However, our conclusions will be different depending on the value of χ.

In those industries where the effort effect prevails, the χ term (now called χ_1) is positive but lower than one. This situation is represented in point B in figure 16.1. On the contrary, in those industries where the organisational effect prevails, the χ term (now called χ_2) is higher than one (as shown by point C in figure 16.1).

If the wage per person were to fall in the same proportion as the working hours ($\dot{W} = \dot{h}$), expression (16.25) would become

$$\dot{N} = -\chi\cdot\dot{h} \tag{16.26}$$

If the effort effect prevailed, ($0 < \chi < 1$), employment would increase, but at a lower rate than that of the fall of the working hours. This effect would be permanent. If, on the contrary, the organisational effect prevailed ($\chi > 1$), the reduction of the working hours would increase employment in a greater proportion than that of the fall of the working day. Nevertheless, in this case, the effect is likely to be temporary and may disappear as firms readjust to the new standard working hours.

As a general conclusion of this part of the paper, we may state the following:

The effect on employment of a reduction of the standard working hours depends on the relative strength of the effort and the organisational effects. When the former prevails, the potential increase of employment is proportionally less than the reduction of the working hours. This effect is likely to be permanent. On the contrary, if the latter prevails, the effects on employment are proportionally greater than the reduction of the working hours. In this case, the effects are likely to be temporary.

The organisational effect is probably stronger in large firms, with complex work structures, whereas in small and medium-size firms the effort effect is likely to be the most relevant.

16.2
Effects on Labour Market Participation

16.2.1 Labour Market Participation with Compulsory Working Hours[6]

The traditional choice model between income and leisure enables us to obtain the optimum number of working hours supplied by a utility maximising individual. However, in the present model, the individual's decision becomes a discrete variable: the individual must choose between working the standard hours or not working.[7] Figure 16.2 illustrates this situation.

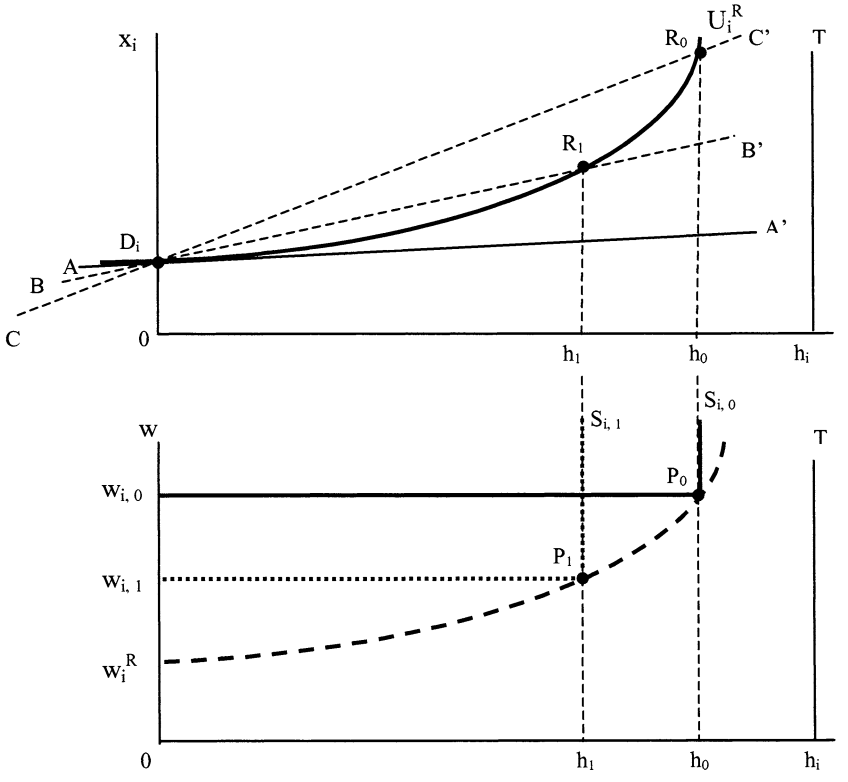


Fig. 16.2

6 I thank Angel Martin-Roman's comments and suggestions on this section.

7 Zabalza, Pissarides and Barton (1980) use a similar model but with a budget restriction consisting of three points.

Let us assume an individual, i, who derives utility from consumption, x_i, and from leisure, $T-h_i$, where T represents total available time and h_i the working hours. The individual has a non-labour income which allows him to consume D_i if he chooses not to work ($h_i=0$). The level of utility at that particular point is represented by the U_i^R indifference curve in the upper graph. The slope of the AA' line would be his reservation wage, w_i^R . Only when the wage rate is above that level, the individual would be willing to work.

Assuming that the length of the working day, h_0 ($h_0 > 0$), is fixed by the law, the reservation wage would be $w_{i,0}$ (the slope the CC' line). For any wage below that level the individual would be better off, if he chose not to work. This individual would only improve his situation by working the standard hours at wages higher than $w_{i,0}$. The individual's labour supply would now be represented by the broken solid line $S_{i,0}$ in the lower graph. This supply is described by the following function

$$h_i(h_0) \begin{cases} 0 & \forall w \le w_{i,0}(h_0) \\ h_0 & \forall w > w_{i,0}(h_0) \end{cases} \qquad (16.27)$$

If the number of working hours were reduced to h_1 ($h_1 < h_0$), the new reservation wage would be $w_{i,1}$ (the slope of the BB' line), which is lower than $w_{i,0}$. Therefore, the reduction of the working hours would imply a new labour supply for the individual such as that represented by the broken dotted line in the lower graph of figure 16.2, $S_{i,1}$.[8]

If we assume that there are several workers and that each worker has a different reservation wage, we could build an aggregate labour supply "curve" (for any given length of the working day) such as the solid broken line $S_0(h_0)$ or the dotted broken line $S_1(h_1)$ in figure 16.3. If the number of workers is large enough, these broken lines could be approximated by continuous curves.

The labour supply function in this model can be expressed in terms of hours (hours of labour supplied at each particular wage, given the length of the working day) or in terms of persons (number of individuals available for work at each

[8] It can be easily shown that there is a positive relationship between the individual's reservation wage and the standard working hours. This reservation wage is a value of w such that:

$$U^R_i(T, D_i) = U_i(T - h_i, D_i + w \cdot h_i)$$

Let us call $w_{i,0}$ the w value that satisfies the above equation for a length of the working day such as h_0; and let us assume that the working hours are exogenously reduced ($dh_0 < 0$). Then, it follows that:

$$0 = \partial U^R_i / \partial h_0 = - \partial U_i / \partial (T - h_0) + \partial U_i / \partial x_i \cdot [w_{i,0} + h_0 \cdot \partial w_{i,0} / \partial h_{0i}]$$

which implies that:

$$\pi_{i,0} = (RMS_{i,0} / w_{i,0}) - 1 > 0$$

where $\pi_{i,0}$ is the elasticity of the reservation wage with respect to the working hours, and $RMS_{i,0}$ is the Marginal Rate of Substitution between income and leisure.
Given any positive value of the working hours, and given the strict convexity of the indifference curves, it must be true that $RMS_{i,0} > w_{i,0}$. Hence, $\pi_{i,0}$ must be strictly positive, which means that changes in the working hours give rise to changes of the reservation wage in the same direction.

particular wage, given the length of the working day). The graphical representation would be identical in both cases. We would only need to take into account that each block of h hours represents an additional worker.

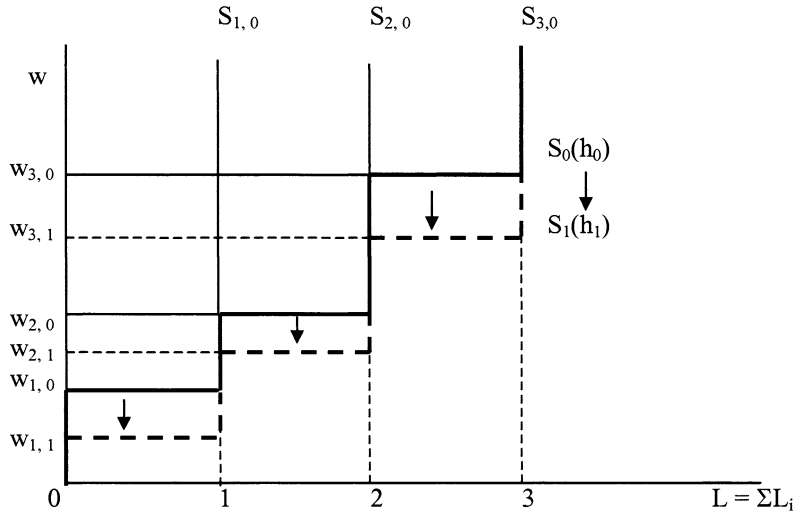


Fig. 16.3

Aggregate labour supply, measured in terms of people available for work, can be written as L=L(w, h). If the number of workers is large enough and there is enough diversity among them, such function could be regarded as continuous. This would allow us to express the effects of changes in w or in h on L in terms of derivatives. In sum, the aggregate labour supply function can be written as

$$L = L(w, h); \qquad L_w > 0; \qquad L_h < 0; \tag{16.28}$$

16.2.2 Effects of a Reduction of the Working Hours on Aggregate Labour Supply

The full differential of (16.28) is

$$dL = L_w\, dw + L_h\, dh$$

which allows us to write

$$\dot{L} = \sigma_w \cdot \dot{w} - \sigma_h \cdot \dot{h} \tag{16.29}$$

where σ_w is the elasticity of labour supply with respect to the hourly wage and σ_h is the elasticity of labour supply with respect to the length of the working day (taken in absolute value). If we now take into account that $\dot{w} = \dot{W} - \dot{h}$ and substitute this expression into the previous one, we obtain

$$\dot{L} = \sigma_w \cdot \dot{W} - (\sigma_w + \sigma_h) \cdot \dot{h} \tag{16.30}$$

If the wage per worker is not altered after reducing the number of working hours, the resulting effect on participation is given by

$$\dot{L} = -(\sigma_w + \sigma_h) \cdot \dot{h} \tag{16.31}$$

When the working day becomes shorter ($\dot{h} < 0$), labour supply increases due to two reasons. One reason is that a shorter working day with a constant wage per man implies a higher hourly wage and that tends to increase participation, as measured by σ_w. Another reason is that a shorter working day has a direct effect on participation as measured by elasticity σ_h.

When the reduction of the working hours is accompanied by a fall in the wage per person such that the hourly wage remains constant, ($\dot{w} = 0$), the effect of such a measure on participation can be calculated from expression (16.29) and is given by

$$\dot{L} = -\sigma_h \cdot \dot{h} \tag{16.32}$$

The basic conclusion of the last two sections is that a reduction of the working hours increases participation both directly and indirectly (through changes in the hourly wage).

16.3 Effects on the Unemployment Rate

16.3.1 The Basic Case

This section analyses the effects of a reduction of the working hours on the economy's unemployment rate (u). Our scenario will be an economy where firms do not have to face any fixed costs when hiring their workers. We shall also ignore, for the moment, the possible influence of the length of the working day on work efficiency.

The unemployment rate (u) is defined as

$$u \equiv U/L \equiv 1 - N/L; \text{ hence: } N/L = 1 - u \tag{16.33}$$

and given that

$$\log (1 - u) \approx -u \tag{16.34}$$

we may write

$$\log (N/L) \approx 1 - u$$

which implies

$$du \approx \dot{L} - \dot{N} \tag{16.35}$$

If we replace $\dot{L}$ with expression (16.30) and $\dot{N}$ with expression (16.6), we have

$$du \approx \sigma_w \dot{W} - (\sigma_w + \sigma_h)\dot{h} - (\varepsilon - 1)\dot{h} + \varepsilon \dot{W} \tag{16.36}$$

Reordering terms, the above expression becomes

$$du \approx (\varepsilon+\sigma_w)\dot{W}+(1-\varepsilon-\sigma_w-\sigma_h)\dot{h} \tag{16.37}$$

This expression allows us to make some predictions about the effect of a shorter working day on the unemployment rate.

If the wage per worker does not vary when the working day becomes shorter, expression (16.37) boils down to

$$du \approx \left[1-(\varepsilon+\sigma_w+\sigma_h)\right]\cdot\dot{h} \tag{16.38}$$

which means that, when the working hours are reduced, the unemployment rate increases, provided that the sum of the hour elasticity of labour demand and the two labour supply elasticities (all of them in absolute value) is higher than one; that is, provided that $(\varepsilon + \sigma_w + \sigma_h) >1$. This condition will be true as long as $\varepsilon >1$.

If we assume that the wage per person falls in the same proportion as the working hours ($\dot{W}=\dot{h}$), the hourly wage remains unchanged, and expression (16.37) becomes

$$du \approx (1-\sigma_h)\cdot\dot{h} \tag{16.39}$$

which means that the reduction of the working hours will reduce the unemployment rate, provided that σ_h is less than one.

Finally, we may calculate the wage reduction required to prevent the unemployment rate from being affected when the working hours are reduced, under the assumption that $(\varepsilon + \sigma_w + \sigma_h)$ is higher than one. Assuming du=0, expression (16.37) allows us to write

$$\dot{W}=\frac{(\varepsilon+\sigma_w)+(\sigma_h-1)}{(\varepsilon+\sigma_w)}\cdot\dot{h} \tag{16.40}$$

This means that the effects of a shorter working day on the unemployment rate could be alleviated by decreasing the wage per person. If σ_h is lower than one, the required reduction in the wage per person will be proportionately smaller than that of the working hours. If $\sigma_h >1$, the required change in W will be proportionately greater.

As a general conclusion of this section we may state the following:

The effect of a reduction of the working hours on the unemployment rate depends on the changes of the wage per person that accompany that measure. If the wage per person does not change at all, the reduction of working hours is likely to increase the unemployment rate, given that the sum of the three elasticities (the elasticity of labour demand -ε-, plus the two elasticities of labour supply $-\sigma_w$ and σ_h-) is likely to be greater than one. Only in the case of a rigid labour demand ($\varepsilon <1$) and a quite rigid labour supply ($\sigma_w + \sigma_h <1$), it could occur that $\varepsilon +\sigma_w + \sigma_h <1$, which is the condition required for the unemployment rate to go down when the working day becomes shorter and the wage per man does not change. In any case, the positive effect of the reduction of the working hours on the unemployment rate could be mitigated if such reduction were accompanied by a fall in the wage per person.

16.3.2 Effects on the Unemployment Rate under a More General Model

This section analyses the effects caused by a reduction of the working hours on the unemployment rate within a context different from that considered in the previous section. Here, we shall take into account the existence of fixed as well as variable wage costs, as described in section 16.1.2. The influence of the length of the working day on work efficiency will also be taken into account, following the lines established in section 16.1.3.

First of all, we must take into account that the fixed component of the wage (either per hour or per person) does not enter the labour supply function. Hence, this function must be rewritten as

$$L = L\,(w^v,\, h)$$

which means that

$$\dot{L} = \sigma_w \cdot \dot{W}^v - (\sigma_w + \sigma_h) \cdot \dot{h} \tag{16.41}$$

It must be observed that the elasticity σ_w is now referred to the variable component, w^v, of the hourly wage. For simplicity, we have kept the same notation as before.

As regards the labour demand function, some changes must be made in order to take into account (a) the existence of hiring costs; and (b) the fact that the length of the working day may influence the efficiency of labour.

When both (a) and (b) are considered, the labour demand function can be rewritten as

$$(1+\alpha)\, w^v = f(H^e) \Rightarrow [(1+\alpha)\, W^v] / h = f[N \cdot \varphi(h)];\; f(H^e) > 0;\; f'(H^e) < 0;$$

which means that

$$\dot{N} = (\varepsilon - \chi) \cdot \dot{h} - \varepsilon \cdot (\dot{W}^v + \alpha \cdot \dot{\alpha}) \tag{16.42}$$

If we substitute (16.41) and (16.42) into (16.35), we obtain

$$du \approx (\varepsilon + \sigma_w) \cdot \dot{W}^v + (\chi - \varepsilon - \sigma_w - \sigma_h) \cdot \dot{h} + \varepsilon \cdot \alpha \cdot \dot{\alpha} \tag{16.43}$$

If the working hours go down and the wage per man remains constant ($\dot{W}^v = 0$, which implies $\dot{\alpha} = 0$), the variation of the unemployment rate is given by

$$du \approx [\chi - (\varepsilon + \sigma_w + \sigma_h)] \cdot \dot{h} \tag{16.44}$$

Comparing this expression with (16.38), we observe that the only difference between them lies in elasticity χ, which measures the sensitivity of the effective working hours with respect to changes in the standard working hours. We would like to compare the effect of a reduction of the working hours according to (16.44) and according to (16.38). In order to do so, we rewrite (16.44) as

$$du^* \approx [\chi - (\varepsilon + \sigma_w + \sigma_h)] \cdot \dot{h} \tag{16.45}$$

By subtracting (16.38) from (16.45), we obtain

$$[du^* - du] \approx (\chi - 1) \cdot \dot{h} \tag{16.46}$$

Hence, when the working day becomes shorter ($\dot{h} < 0$), if the effort effect prevails (χ <1), du* > du. This means that the effect of a shorter working day on the unemployment rate is stronger now than in the basic case: the unemployment rate increases now *more* than before. The previous statement assumes that a shorter working day increases the unemployment rate; but we already know that this may not be so. In some presumably rare cases, a shorter working day can reduce unemployment. In such cases, the above expression tells us that du* is *less negative* than du, meaning that the unemployment rate falls *less* (as a result of a shorter working day) than in the basic case.

The opposite will happen when the organisational effect prevails (χ >1). Nevertheless, this effect will only be temporary, following the interpretation given in section 16.1.3.

Now, let us suppose that the flexible part of the wage per person falls in the same proportion as the working hours ($\dot{w}^{v} = \dot{h}$), which means that the flexible part of the hourly wage remains unchanged. On the other hand, since the fixed part of the wage per worker (hiring costs) does not change, α must increase in the same proportion as the standard hours fall; that is, $\dot{\alpha} = -\dot{h}$. In this case, expression (16.43) becomes

$$du \approx [\chi - (\sigma_h + \varepsilon \cdot \alpha)] \cdot \dot{h} \tag{16.47}$$

If we rewrite du in expression (16.47) as du**, and subtract (16.39) from (16.47), we obtain

$$[du^{**} - du] \approx [(\chi - \varepsilon \cdot \alpha) - 1] \cdot \dot{h} \tag{16.48}$$

We want to know whether this expression is positive or negative, when the working hours fall ($\dot{h} < 0$). If the effort effect prevails (χ < 1), given that $\varepsilon \cdot \alpha$ is always positive, [(χ – $\varepsilon \cdot \alpha$) – 1] is necessarily negative. Hence, du** > du. This means that the effect of a shorter working day on the unemployment rate is stronger than in the basic case. We must also notice that that effect is also stronger the higher the proportion of hiring costs over total labour costs (the higher α).

As a general conclusion of this section we may state the following:

When there are hiring costs and when the reduction of the working day reduces work weariness, it becomes more difficult to reduce the unemployment rate through a reduction of the working hours.

16.4 Conclusions

In this paper, we have analysed the effects of a legal measure reducing the standard working hours on employment, participation, and unemployment rates. Our basic conclusions are the following:

1. If the wage rate does not vary, a reduction of the standard working hours increases employment, provided that the elasticity of the working hours with respect to wages (ε) is less than one and reduces employment when ε >1. When ε =1, the effect of the reduction of working hours on employment is null.

The negative effects of this type of measure on employment are, then, constrained to the case where $\varepsilon > 1$. But these effects can be alleviated if the reduction of the working hours is accompanied by a wage cut. In order to avoid a decrease in the employment level, the wage per worker should fall (although in a lower proportion than that of the working hours).

2. The potential positive effect on employment of a reduction of the working hours will be lower in industries where hiring costs represent a higher proportion of the wage bill. Such costs have to do with employee screening expenditures, as well as with worker training, monitoring, and welfare expenditures.
3. The effect on employment of a reduction of the standard working hours depends on the relative strength of the effort and the organisational effects. The former takes place when the reduction of the working hours reduces work weariness. The latter occurs when factors such as the size of the working team or the efficiency of monitoring are affected by the reduction of the working day.
   When the effort effect prevails, the potential increase of employment (provided the wage rate falls in the same proportion as the working day) is proportionally less than the reduction of the working hours. This effect is likely to be permanent. On the contrary, if the organisational effect prevails, the effects on employment are proportionally greater than the reduction of the working hours. In this case, the effects are likely to be temporary.
4. When the working day becomes shorter, labour supply increases due to two reasons. One reason is that a shorter working day with a constant wage per man implies a higher hourly wage and that tends to increase participation. Another reason is that a shorter working day has a direct effect on participation (it induces people to participate more).
5. The effect of a reduction of the working hours on the unemployment rate depends on the changes of the wage per person that accompany that measure. If the wage per person does not change at all, the reduction of working hours is likely to increase the unemployment rate. Only in the case of a rigid labour demand and a quite rigid labour supply, the unemployment rate is likely to decrease when the working day becomes shorter and the wage per man does not change. In any case, the positive effect of the reduction of the working hours on the unemployment rate could be mitigated if such reduction were accompanied by a fall in the wage per person.
6. When there are hiring costs and when the reduction of the working day reduces work weariness, it becomes more difficult to reduce the unemployment rate through a reduction of the working hours.

References

Dolado, J.J.: Valoración Crítica de las Estimaciones Econométricas Disponibles de la Relación entre los Precios Relativos y el Empleo en la Economía Española. In: Bentolila, S., Toharia, L., Estudios de Economía del Trabajo en España, III El Problema del Paro, C. 20. MTSS 1991

Fallon, P., Verruy, D.: The Economics of Labour Markets. Phillip Allan 1988

Hamermesh, D.: Labour Demand. New Jersey, Princeton University Press 1993
Nickell, S.: Dynamic Models of Labour Demand. In: Ashenfelter,O., Layard, R., Handbook of Labour Economics, Chapter 9. North-Holland 1987
Zabalza, A., Pissarides, C., Barton, M.: Social Security and the Choice Between Full-Time Work, Part-Time Work and Retirement. Journal of Public Economics 14 (1980)

17 Transitional Dynamics and Endogenous Growth Revisited: the Case of Public Capital

B. Sánchez-Robles
University of Cantabria (Spain)

17.1 Introduction

The last two decades have seen a remarkable flourishing of dynamic models in the field of Macroeconomics. Within this broad framework, a particular area of research that has been especially active- after decades of being almost dormant– is the one that focuses in economic growth. Following the pathbreaking contribution of Romer (1986), economists have been working intensively in this range of issues in order to ascertain which are the crucial factors that promote economic development or, on the contrary, condemn a country to poverty and stagnation for long periods of time. In parallel to theoretical contributions, the empirical tests of these models have also been abundantly carried out following the seminal contribution of Barro (1991). However, perhaps one of the particular aspects of economic growth models that has been somehow neglected, in part due to its intrinsic difficulty, is the analysis of the transitional dynamics of the models towards the Balanced Growth Path (BGP) [1]. In effect, it is frequent for economist to solve their models under the assumption of a constant rate of growth of relevant variables, since this procedure makes the analysis tractable[2] and, on the other hand, the ultimate goal of these models is to predict the long run behaviour of the economy. Notwithstanding this practice, the study of the transition of a particular model to the steady state may prove useful for some reasons: first,

[1] Two exceptions are the papers by Caballe and Santos(1993) and Xie (1994) that do pursue the study of the transitional dynamics of the Lucas (1988) model analytically.

[2] By means of providing a close form solution to a system of non linear differential equations that otherwise would have a cumbersome solution (if any).

P. Coto-Millán (ed.), *Essays on Microeconomics and Industrial Organisation*

because it helps ascertain the short run implications of the model and secondly, because it can provide an integrated framework for the joint analysis of economic growth and fluctuations. Finally, it performs as a good test of the oft-debated hypothesis of convergence.

This paper intends to fill in this gap by means of analysing the transitional dynamics of two alternative models that coincide in one crucial factor: in both of them public capital plays a prominent role as an input in production. However, the implications of the models as regards the behaviour during the transition are different. In order to carry out this research and establish comparisons between these models a very powerful technique, the time-elimination model designed by Mulligan and Sala-i-Martin (1991, 1993) will be employed. The particular details of the implementation of this method will be explained as the paper proceeds.

The structure is as follows: in section 17.2 the basic features of the models employed are described. Section 17.3 describes the BGP solution. Section 17.4 focuses on the time elimination technique. Section 17.5 makes some comments on the simulations implemented, and section 17.6 concludes with some final remarks.

17.2 Setup of the Models

The basic models that will be analysed in the paper may be characterised by the following assumptions:

Preferences

1. The economy is composed by infinitely lived agents, in the same fashion as in Ramsey (1928)[3]. The agents of this economy want to maximise their intertemporal utility function given by eq. (17.1)

$$U(0)=\int_0^{\infty} e^{-\rho t} u(c(t))L(t)dt \tag{17.1}$$

in which U(0) means utility of the family discounted to the present, ρ is the discount rate, L(t) is the size of the family and c(t) means consumption per capita. More specifically, the utility function is of the variety of constant relative risk aversion (eq. (17.2)), in which σ, the coefficient of relative risk aversion, also expresses the inverse of the intertemporal elasticity of substitution.

$$u(c(t)) = \frac{c(t)^{1-\sigma}-1}{1-\sigma} \tag{17.2}$$

$$\sigma=\frac{-u''(c)c}{u'(c)} \tag{17.3}$$

[3] As it is well known, it is more intuitive to grasp the rationale under this assumption if we think of families or dynasties, linked by altruistic bonds.

2. There is not technological progress or population growth, and thus the analyses in per capita terms and in aggregate terms are the same. Population at the initial moment is normalised to 1 for simplicity. The subscript t will be suppressed from now on, whenever it is possible, in order to alleviate notation.

Technology

Two different specifications will be used in order to describe the production side of the economy. In a sense, they are slightly more realistic that the one of the AK type, employed in the seminal paper of Barro (1990) and in Rebelo (1991), in that they allow for some degree of convergence. Furthermore, they permit a direct impact of public capital on growth, as has been reported by some empirical evidence (Sánchez-Robles, 1998a), whereas the AK class of models only predicts an indirect impact of public capital on growth, via increases in private investment. As Barro and Sala-i-Martin (1995) correctly point out, sometimes more sophisticated models than the AK are necessary in order to match the empirical evidence better.

The specifications of the production function that will be considered below are basically the following:

b.1. CES production function: (Arrow *et al.*, 1961)

$$Y=A\left\{(bK)^{\psi}+[(1-b)G]^{\psi}\right\}^{1/\psi}$$

$$0<\psi<1 \tag{17.4}$$

b.2. Jones-Manuelli (JM) production function (Jones and Manuelli, 1990):

$$Y=A(K+G)+bK^{\beta}G^{1-\beta} \tag{17.5}$$

$$0<\beta<1$$

As it is apparent from eq. (17.5), the JM production function may be considered as a combination of an AK function (first part) together with a Cobb Douglas technology (second part). The AK term warrants the existence of endogenous growth whereas the other component allows the model to have transitional dynamics.

Labor does not appear as an input in production in neither of these cases for simplicity (similarly to Barro, 1990). K is private capital. G should be interpreted here as the stock of public capital. In these two cases the production function is homogeneous of degree one in both inputs, and marginal productivity of each factor is decreasing. The crucial feature that allows for endogenous growth in both cases, however, is the violation of the Inada condition (Inada, 1963), that provides a lower bound to marginal productivity of the inputs, thus ensuring sustainable growth in the steady state.

More specifically, and focusing in the CES function, A may be regarded as the efficiency parameter, in the terminology of Arrow *et al.* (1961). b is the

distribution parameter, and captures the relative role of each input in the production process. The parameter ψ in the CES function is positively related to the elasticity of substitution ε by the expression $\varepsilon=1/1-\psi$ (see Arrow *et al.*, 1961). The employ of this production function in growth models has been, to our notice, very sparse (an exception is Easterly, 1993). As far the Jones-Manuelli specification is concerned, A and b are technological parameters.

c) The dynamics of private and public capital are given by:

$$\dot{K}=(1-\tau)Y-c-\delta K \tag{17.6}$$

$$\dot{G}=\tau Y-\delta G \tag{17.7}$$

in which τ represents the constant (and unique) tax rate and δ is the depreciation rate, common for the two types of capital for simplicity[4]. A dot over a variable represents its derivative with respect to time. Intuitively, the resources that can be devoted to new private capital formation (savings) come from the output net of taxes, once the depreciation of the existing capital stock has been provided for. Public capital is financed out of taxes, and a balanced budget is run in every period. Since the economy is assumed to be closed, international capital flows are explicitly ruled out. This assumption may be relaxed in further research in order to allow for an open economy scenario, but it is not particularly relevant for the analysis carried out here.

17.3 Solution for the BGP in the CES Case

The paper shall present the analysis for the CES case in a more detailed way. The extension to the Jones-Manuelli case is straightforward, and the main results of this case are also presented along the paper.

As it was said above, the BGP may be regarded as the long run equilibrium of the economy, in which all relevant variables grow at constant rates.

More formally, define x* as the long run equilibrium solution, where

$$x^* \in X$$

X being the space of states, a subset of R^n.

If we denote by *f* the map implied by the system of differential equations that governs the dynamical behaviour of the model, then the BGP solution is a fixed point of the map *f*, such that it verifies

[4] This is, perhaps, a debatable assumption since the depreciation rates of public and private capital do not need to be the same. However, and since the main goal of the paper is to compare the nature and speed of the transitional dynamics implied by the two types of production function considered, and since the assumption is made for the two cases, final results probably would not vary very much if the assumption was to be changed. Finally, to assume different rates for each types of capita would introduce a large degree of complexity in the analysis. It is beyond the scope of this paper, although it can be a promising avenue for future research.

$$x^* = f(x^*)$$

Intuitively, once the economy reaches the BGP, it shall remain there unless an external shock takes place. We should be careful at this point, however, since the long run equilibrium may be either a node (stable equilibrium) or a saddle point (unstable equilibrium), this last case being the most frequent in endogenous growth models[5]. From an economical viewpoint, and since externalities arise in this model, two types of solutions can be distinguished in this setting, the social planner's solution and the competitive equilibrium outcome. The social planner's solution will be presented first.

a) Social planner solution

The social planner must choose the optimal path of consumption such that it maximises the utility of the agents (eq. 17.1), subject to the aforementioned constraints (eq. (17.6) and eq. (17.7)). He can also choose the optimal tax rate. Hence, in this setup c and τ are control variables, whereas K and G are the state variables. He takes the initial level of private capital and the public expenditure as exogenous.

$$K(0) > 0$$

$$G(0) > 0$$

The Maximum Principle of Pontryagin (1962) may be applied. The current value Hamiltonian is:

$$H = \frac{c_{(t)}^{1-\sigma} - 1}{1-\sigma} + \lambda\left[(1-\tau)Y - c - \delta K\right] + \mu(\tau Y - \delta G) \tag{17.8}$$

where λ y μ represent the shadow prices of K and G, respectively. The first order conditions are as follows (where MPK, MPG and APK, APG represent marginal and average products of the inputs K and G, respectively).

$$H_c = 0 \rightarrow c^{-\sigma} = \lambda \tag{17.9}$$

$$H_\tau = 0 \rightarrow \lambda = \mu \tag{17.10}$$

$$\rho\lambda - H_K = \dot{\lambda} \rightarrow \dot{\lambda} = \rho\lambda - \lambda(1-\tau)MPK + \lambda\delta - \mu\tau MPK \tag{17.11}$$

$$\rho\lambda - H_G = \dot{\mu} \rightarrow \dot{\mu} = \rho\lambda - \mu\tau MPG + \delta\mu - \lambda(1-\tau)MPG \tag{17.12}$$

and the transversality conditions

$$\lim_{t\to\infty} e^{-\rho t}\lambda_t K_t = 0 \tag{17.13}$$

[5] Depending on which class of equilibrium takes place, should a shock occur, the model would converge again to the equilibrium or diverge from it.

$$\lim_{t\to\infty} e^{-\rho t}\mu_t G_t = 0 \tag{17.14}$$

Equations (17.6), (17.7) and (17.9) to (17.12) together with transversality conditions (17.13) and (17.14) form a non linear system of differential equations with boundary conditions, solvable only by numerical methods.

Nonetheless, the solution for BGP rate of growth can be obtained. In order for the utility to be bounded, another condition regarding the parameters should be made explicit. The condition is the following: $\rho > 0, \sigma > 1$, which is totally consistent with the values of these parameters that are commonly used in the calibration of this models.

To see that this condition is enough, notice that, if consumption grows at a rate γ, consumption in time t will be $c_t = c_o\, e^{\gamma t}$. This means that the expression inside the integral in equation 1 will be $e^{-(\rho-\gamma(1-\sigma))t}$. This expression will tend to 0 as t tends to infinite whenever $\rho > \gamma(1-\sigma)$, which is true if $\rho > 0, \sigma > 1$.

Taking logarithms and differentiating with respect to time in (17.9)-(17.12) yields

$$\gamma^* = \frac{\dot{c}}{c}^* = \frac{1}{\sigma}\left\{A\left[b^{\psi} + (1-b)^{\psi}\left(\frac{1-b}{b}\right)^{\frac{\psi}{1-\psi}}\right]^{\frac{1-\psi}{\psi}} b^{\psi} - \rho - \delta\right\} \tag{17.15}$$

As it is common in this class of models, in the BGP all relevant variables - c, K, G, Y - grow at the same rate, γ^*.

The interpretation is the usual one in this kind of models: i.e. consumption grows in this economy insofar as the marginal productivity of capital, net of depreciation, exceeds the rate of temporal preference. The higher the elasticity of substitution in consumption, the greater the response of the agents to a difference between the net rate of return and the rate of time preference. It is worthwhile to notice that growth arises endogenously in this model due to the homogeneity of the production function in both inputs (eq. 17.4). Saving of the individuals is devoted to private investment, which in turns increases output and therefore revenues associated to taxes, that are translated into public capital due to the balanced budget assumption. Hence K and G grow at the same pace, and the same happens to output due to the homogeneity of degree one of both production functions. This feature is also present in other types of models, such as Barro (1990) and Rebelo (1991).

It is also worth noticing that, from equations (17.10), (17.11) and (17.12) the ratio of private to public capital in the BGP can be obtained. It turns out to be

$$\frac{K}{G} = \left(\frac{b}{1-b}\right)^{\frac{1}{1-\psi}} \tag{17.16}$$

Intuitively, the fact that the tax rate can be chosen optimally ensures that the marginal productivity of both types of capital in the long run equilibrium is equal and this, in turn, entails a fixed ratio K/G. This ratio is positively related to the distribution parameter b, as it should be expected.

b) Competitive equilibrium

In this setup there is an effect similar to an externality. As it is apparent from the simple inspection of the model, public capital enhances the marginal productivity of private capital. Individual agents, though, will neglect in their optimising decisions the impact that their investment choices have in the public capital stock, and hence in the rate of return of his own capital by means of this feed back, since this last effect is very small[6]. The externality does not prevent the existence of a competitive solution (for a proof in a framework of increasing returns, see Romer, 1983), but it does introduce a wedge between the individual and the social rate of return. Therefore the amount of private investment allocated by individual agents and the competitive equilibrium rate of growth will be smaller than the one achieved in the social planner's solution.

Thus, the set up of their problem is to maximise (17.1) subject to (17.6). Individuals only have consumption as a control variable, since the tax rate is exogenous for them.

The optimisation conditions are, in this case, the following (obviously the first one is the same as in the social planner case).

$$Hc=0 \quad \rightarrow c^{-\sigma}=\lambda \tag{17.17}$$

$$\rho\lambda - HK = \dot{\lambda} \rightarrow \dot{\lambda} = \rho\lambda - \lambda(1-\tau)MPK + \lambda\delta \tag{17.18}$$

Dividing through by λ we get

$$\frac{\dot{\lambda}}{\lambda}=\rho-(1-\tau)MPK+\delta$$

Now we proceed in the traditional way. We take logs in (17.17), differentiate with respect to time and plug in the equation for the variation in the shadow price we just got. The final expression is:

$$\frac{\dot{c}}{c} \text{ c.e.} = \left[\frac{1}{\sigma}(1-\tau)MPK-\rho-\delta\right] \tag{17.19}$$

Since $(1-\tau)$ is less than one, the competitive equilibrium rate of growth is smaller than the one attained in the social planner's solution, as it should be expected.

The procedure of obtaining the BGP rate of growth is analogous for the Jones-Manuelli case. It is detailed in Sánchez-Robles (1998b) and briefly described in the Appendix.

[6] There is, of course, a negative externality in terms of the marginal contribution to congestion by the individual, but we shall omit this consideration in the model for simplicity.

17.4 The Time Elimination Method and the Analysis of the Transitional Dynamics

In order to analyse the transitional dynamics, the time elimination method of Mulligan and Sala-i-Martín (1991, 1993) has been used. It is a specially suitable algorithm for dynamic systems consisting of differential equations with boundary conditions. The basic strategy is to suppress time explicitly of the equations, working instead with the so called *policy functions*, that relate a continuum of optimal values of the control variable to optimal values of the state variable. Next, and since one of the main difficulties of these models is dealing with terminal conditions, the system is rewritten backwards, in order to get an initial value problem in which the differential equations may be solved by numerical methods very easily. This procedure is clearly superior to others as the shooting, that proceeds by trial and error and therefore may be very time-consuming. For a detailed explanation of the time elimination method, see Mulligan and Sala-i-Martín (1991). Nonetheless we shall make a brief sketch of this technique along the rest of the paper. Simulations for the competitive market outcomes will be presented here. Simulations for the social planner solution are left for future research.

The models presented here, however, requires a further transformation. Since the rate of growth along the BGP is greater than zero for both production functions, the system has to be transformed in another one in which the relevant variables do not grow. This is accomplished by means of defining the so called *control like* and *state like* variables, whose main feature is that they do not grow in the long run, thus making the analysis tractable.

The steps are the following:

Following Mulligan and Sala-i-Martín (1993), *control like* variable a and *state like* variable z are defined. Their expressions are

$$a = c/K \tag{17.20}$$

$$z = K/G \tag{17.21}$$

Since c, K y G grow at the same rate in the BGP, a and z will be constant in the long run equilibrium. Therefore

$$\frac{\dot{a}}{a} = \frac{\dot{z}}{z} = 0 \tag{17.22}$$

The expressions of the rate of growth of c, K and G are redefined in terms of a and z.

The rate of growth of c is displayed in equation (17.19). Growth rates for K and G are obtained dividing expressions (17.6) and (17.7) by K and G, respectively. Therefore the expressions for a and z are

$$\dot{a}/a = \dot{c}/c - \dot{K}/K = (1/\sigma)[(1-\tau)(MPK(z)) - \rho - \delta] - (1-\tau)APK(z) + a + \delta \tag{17.23}$$

$$\dot{z}/z=\dot{K}/K-\dot{G}/G=(1-\tau)APK(z)-a-\tau APG(z) \tag{17.24}$$

where

$$MPK(z)=A\left[b^{\psi}+(1-b)^{\psi}z^{-\psi}\right]^{\frac{1-\psi}{\psi}}b^{\psi}$$

$$MPG(z)=A\left[b^{\psi}z^{\psi}+(1-b)^{\psi}\right]^{\frac{1-\psi}{\psi}}(1-b)^{\psi}$$

$$APK(z)=A\left[b^{\psi}+(1-b)^{\psi}z^{-\psi}\right]^{\frac{1}{\psi}}$$

$$APG(z)=A\left[b^{\psi}z^{\psi}+(1-b)^{\psi}\right]^{\frac{1}{\psi}}$$

We can obtain a numerical approximation to the policy function a(z) using expressions (17.23) and (17.24) and the fact that

$$a'(z)=\dot{a}/\dot{z}$$

Now we have a differential equation in z, which represents the slope of the policy function, and is as follows:

$$a'(z)=\dot{a}/\dot{z}=(a/\sigma)\left[(1-\tau)MPK(z)-\rho-\delta\right]-$$

$$a(1-\tau)APK(z)+a^2+\delta a/z(1-\tau)APK(z)-za-\tau APG(z) \tag{17.25}$$

We can solve the equation if we take the values of a and z in the BGP as the initial conditions and proceed backwards in time. If values are assigned to the parameters, many software packages may solve the equation numerically[7].

The simulation implemented in the previous step has provided pairs of values (z,a) that belong in the policy function. We can interpolate a polynomial, in order to get an algebraic expression of the policy function.

If we replace a in expression (17.24) by this polynomial, we have a new differential equation, this time of z in t. It can also be solved by means of adding an initial condition. From there on it is straightforward to graph the temporal behaviour of the relevant variables in the model and their rates of growth. A similar procedure in order to get the correspondent modified differential equations (not displayed for lack of space, but available under request), has been done for the Jones-Manuelli case.

[7] More specifically, the prompt ode 23 of Matlab has been used.

17.5
Discussion of the Main Results of the Simulations

Simulations for alternative set of parameters have been carried out. They shall be described below, stressing their similarities and differences.

CES Function. Baseline

Table 17.1. Sensitivity analysis. CES and JM Production functions (absolute values in the BPG)

			a*	z*	gc*	t*
Ces case 1	$\Psi = 0.75$		0.3	0.475	0.0211	15.52
Ces case 2	$\Psi = 0.6$		0.36	0.49	0.0542	13.01
JM case 1	$\beta = 0.3$		0.84	0.21	0.0259	9.61
JM case 2	$\beta = 0.7$		0.38	0.42	0.0265	15.18
Baseline parameters CES cases 1 and 2				**Baseline parameters JM cases 3 and 4**		
	δ	0.05		δ	0.05	
	τ	0.3		τ	0.3	
	ρ	0.1		ρ	0.1	
	σ	3		σ	3	
	A	0.3		A	0.1	
	b	0.6		b	0.25	
			a*	**z***	**gc***	**t***
Ces case 3	$\Psi = 0.75$		0.1	2.9	0.015	15.03
Ces case 4	$\Psi = 0.6$		0.11	3.2	0.0452	9.15
JM case 3	$\beta = 0.3$		0.48	1.3	0.0427	5.11
JM case 4	$\beta = 0.7$		0.43	1.57	0.0683	6.11
Baseline parameters CES cases 3 and 4				**Baseline parameters JM cases 3 and 4**		
	δ	0.1		δ	0.1	
	τ	0.15		τ	0.15	
	ρ	0.08		ρ	0.08	
	σ	3		σ	3	
	A	0.3		A	0.3	
	b	0.6		b	0.25	

As it was stated before, the production function is:

$$Y=A\left\{(bK)^{\psi}+[(1-b)G]^{\psi}\right\}^{1/\psi}$$

$$0<\psi<1$$

The first simulation has employed the CES specification, with the following values of the parameters: A= 0.3, b=0.6, ρ =0.1, δ = 0.05, σ = 3, τ = 0.3, Ψ = 0.75. These values are consistent with those employed in similar calibrations, and in turn yield the numbers a* = 0.3, z* = 0.475 for a, z in the BGP.

Fig. 17.1 to 17.3 display the behaviour of the model under these assumptions and Table 17.1 gives some quantitative results. Fig. 17.1 shows the policy function (optimal value of a as a one-to-one mapping of the starting condition z), that is remarkably similar to the policy function of the Ramsey model. Fig. 17.2 displays the behaviour over time of the ratio of private to public capital. The assumption here is that the ratio increases over time until the BGP value. Analogous results would be reached under opposite conditions. It is interesting to notice that the graph shows the convergence behaviour very neatly. The ratio increases very quickly while the economy is farther off the long run equilibrium, whereas it reaches a stable value corresponding to the BGP in around 15.5 years.

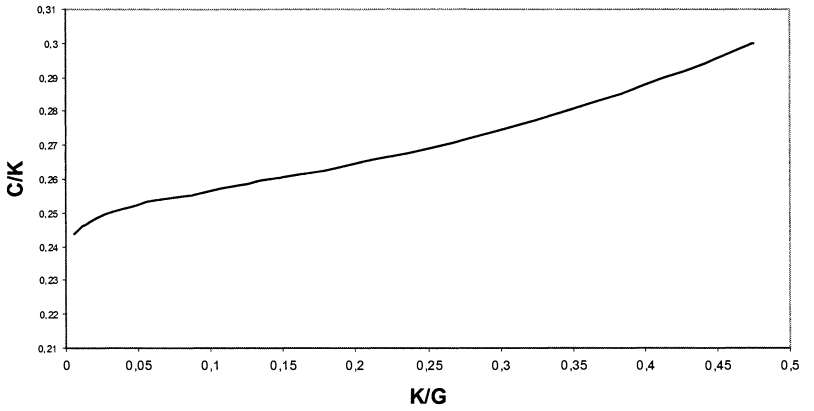


Fig. 17.1. Policy function, CES case 1

Finally, Fig. 17.3 shows the time trajectory of the growth rate of consumption. The convergence property is again apparent. The rate of growth is large while the economy is far from the steady state (and hence the marginal productivity of private capital is large) while it diminishes progressively until it is stabilised around the steady state value (in this particular case, 0.0211). A first interesting result can be stressed: this particular model, while generating endogenous growth by the interplay of private and public capital, does have a transitional dynamic towards the steady state that mimics fairly well the one displayed by the

neoclassical models. This is not a surprise, however, since the CES production function may still be regarded as neoclassical in some sense, thanks to the assumption of decreasing marginal productivity of capital, which is of course the force underlying convergence.

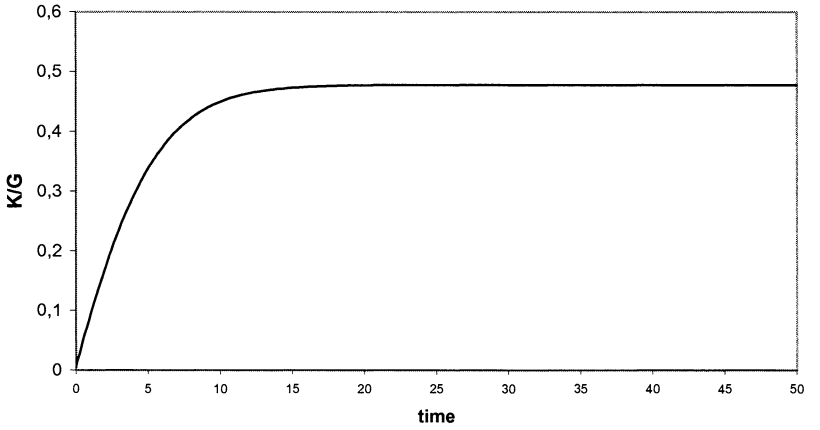


Fig. 17.2. K/G versus time, CES case 1

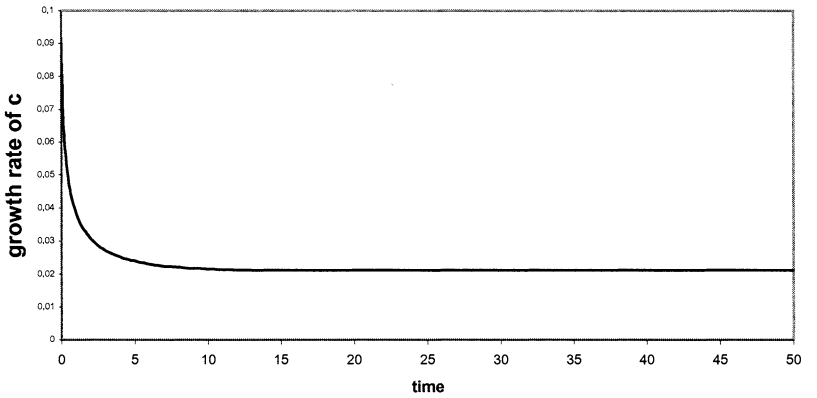


Fig. 17.3. Growth rate of c, CES case 1

CES Function, Case 2

The same procedure has been implemented with a different value of the parameter Ψ, $\Psi = 0.6$. The rest of parameters employed in this case remain the same. Now

the level of both a and z in the long rum equilibrium are higher, 0.36 and 0.49, respectively.

The evolution of the model under these assumptions regarding the parameters are displayed in Fig. 17.4 and Fig. 17.5, that also show the evolution of the model with Ψ=0.75 to ease the comparisons between them. The steady state rate of growth in this second case is higher, of 0.0542. The performance of the relevant variables is very similar to the previous case, but it should be noticed that the transition to the steady state takes place in the second scenario within a shorter period of time (13 years).

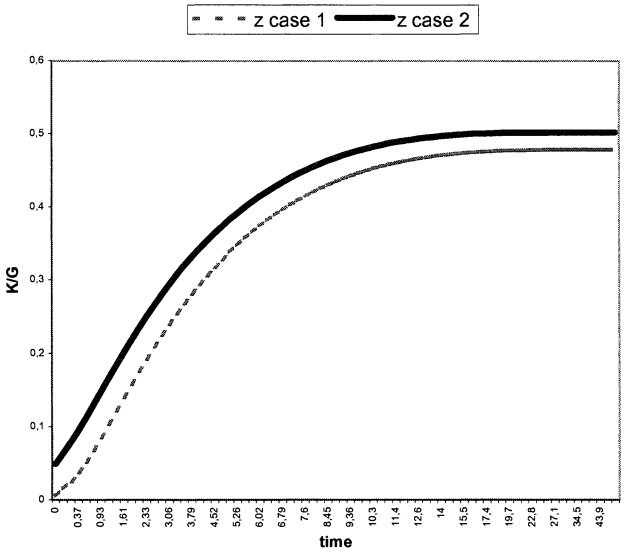


Fig. 17.4. K/G versus time, CES cases 1 and 2

The intuition for this fact is very simple: with a lower value of Ψ, which is the only parameter that has been changed in comparison with the first case, the marginal productivity of private capital is higher. Since its relevance in production is large (as shown by a distribution parameter b of 0.6), it means that the growth rate of consumption will also be larger and reach the steady state level before than in the previous case. It can also be shown (see Sánchez-Robles, 1995), that the transition is not as smoother as in the previous case. This is an interesting implication since it can be argued that smaller values of Ψ will be associated to larger fluctuations in economic activity. Therefore, smaller values of Ψ and larger elasticities of substitution have growth effects (increase the rate of growth), level effects (are linked to higher values of K/G and C/K in the BGP), and reduce the time necessary to reach the long run equilibrium.

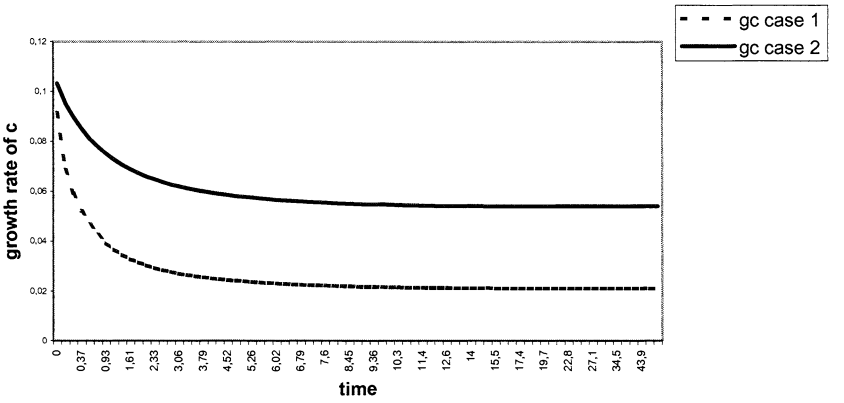


Fig. 17.5. Growth rate of consumption, CES cases 1 and 2

Jones-Manuelli Production Function. Baseline

This sub-section takes as the benchmark production function the Jones–Manuelli specification, as presented in eq. (17.5) above and rewritten below. The specification takes the form in this case

$$Y=A(K+G)+bK^{\beta}G^{1-\beta}$$

$$0<\beta<1$$

The set of parameters employed in the baseline simulation are: A= 0.1, b=0.25, σ=3, ρ=0.1, τ=0.3, β = 0.3, δ=0.05.The BGP values for a and z are 0.84 and 0.21[8], respectively. The BGP growth of consumption is, in this case, 0.0259.

Fig. 17.6 and Fig. 17.7 show the basic features of the models under this alternative specification of the production side of the economy. As we can see in Fig. 17.6, at first sight the time path of K/G is very similar that the one obtained for the CES cases. It should be noticed, however, that the model converges to the steady state faster now that in the two previous CES cases. The steady state is approached in around 9.6 years, which is well below the time required for the CES specification in both cases considered above. The changes in the growth rate of consumption (Fig. 17.7, dotted line) over time are also more pronounced. This already points out that transition under these class of models with β = 0.3, A = 0.1,

[8] It has been attempted to choose the parameters in such a way that the BPG values of a and z are the same in the CES and JM case. This has proved to be very difficult, since due to the differences between the functions, BGP figures for a and z coincided only if very odd values for the parameters were assumed. We have preferred to keep the value of the parameters more similar, despite the fact that a and z in the BGP will be different under CES and JM assumptions.

b = 0.25, are more rapid and less smooth that under the CES specifications. In turn, fluctuations will also be more pronounced upon the assumption of a Jones-Manuelli production function with the values of β and the other technological parameters as specified above. The convergence property still holds, however, as it should be expected. This is apparent in Fig. 17.7, which displays the inverse relationship between the rate of growth and distance to the steady state that is typical of the convergence property.

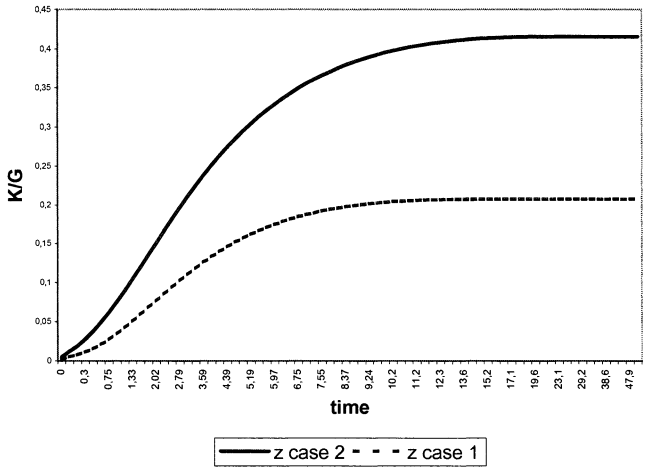


Fig. 17.6. K/G Versus time, Jones-Manuelli cases 1 and 2

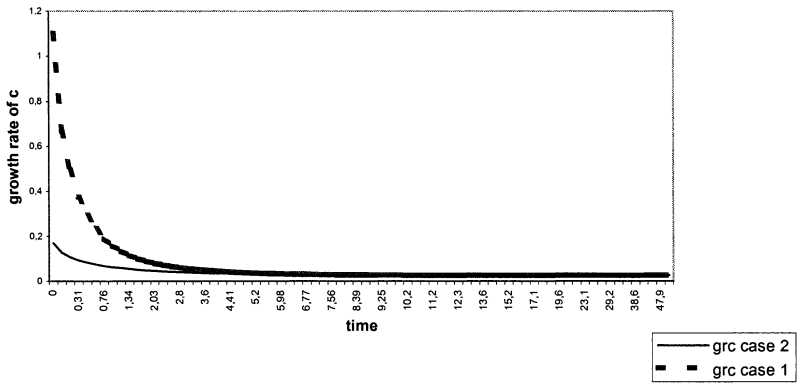


Fig. 17.7. Growth rate of c, Jones-Manuelli cases 1 and 2

Jones-Manuelli Production Function. Case 2

Another simulation has been implemented for the JM case, using a value of 0.7 for β. The BGP values of a and z are 0.38 and 0.42, and the equilibrium rate of growth is 0.0265. The rate of growth is slightly higher this time due to an increase in the productivity parameter β. The dynamic behaviour of the system is displayed also in Fig. 17.6 and Fig. 17.7. The time pattern is again similar, but the more remarkable fact is that the speed of convergence is rather different: now the model displays a slower convergence. The number of years necessary to reach the BGP is 15.18. If we compare cases JM 1 and 2 (Table 17.1), we see that a change in β brings about basically a level effect in the BGP values of a and z. In particular, z is two times higher in the second example. The speed of transition becomes smaller, whereas the change in the growth rate is very small. It is interesting to notice that these results are similar to those obtained for the Ramsey case. (Barro and Sala-i-Martin, 1998). Therefore we could conclude tentatively that the JM with β=0.7 resembles somehow the Ramsey model.

A Change in the Tax Rate

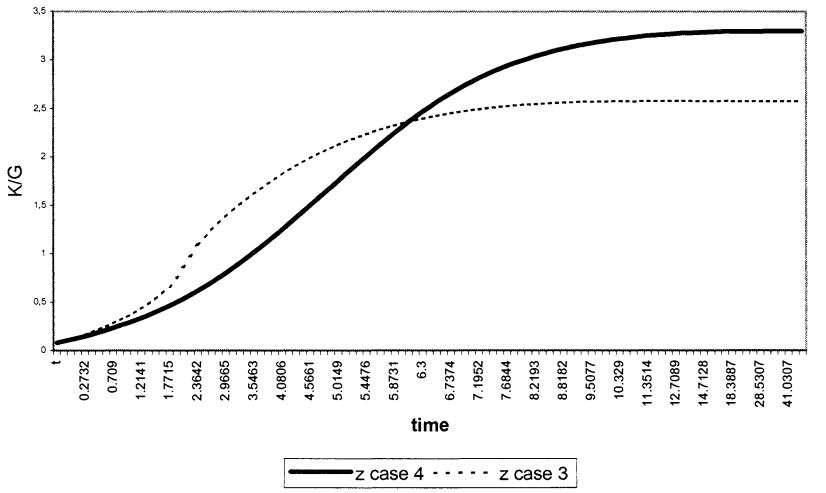


Fig. 17.8. K/G versus time, CES cases 3 and 4

We have performed a further experiment by means of reducing the tax rate to 0.15. In order to get realistic values of the rate of growth, small adjustment in two other parameters have been required. In particular, we assumed ρ = 0.08, δ=0.1, for both production functions, and A = 0.3 in the JM case. Basic results for both production functions under the same two scenarios than above are reported in the second part of Table 17.1. The specifications under these new assumptions are called CES 3 and 4, JM 3 and 4. The patterns over time of the relevant variables

are displayed in Fig. 17.8 to 17.11. Visual inspection of the graphs shows that the main conclusions of the first scenarios carry over to this one. There are, however, some interesting differences.

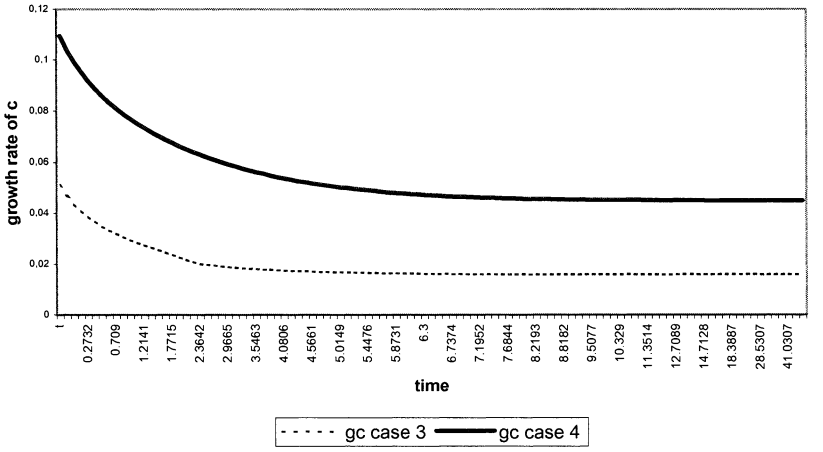


Fig. 17.9. Growth rate of c, CES cases 3 and 4

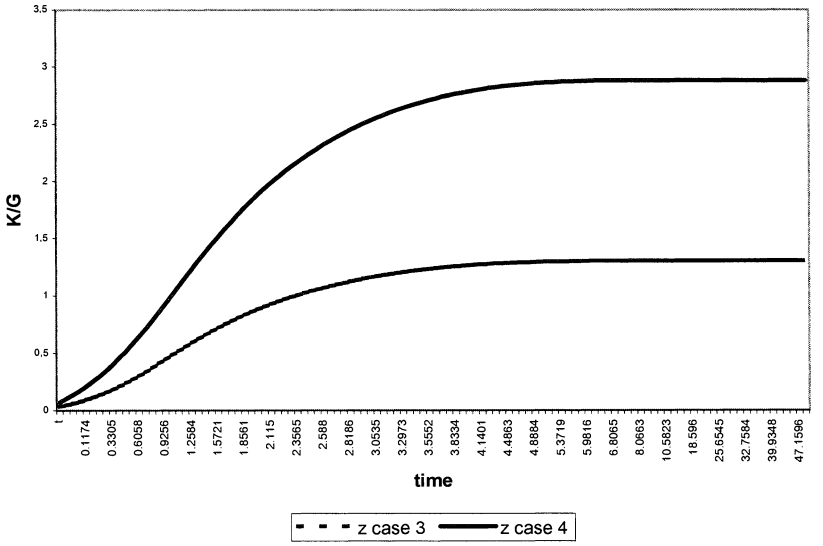


Fig. 17.10. K/G versus time, JM cases 3 and 4

The main effects of a reduction of the tax rate are as follows: First, as it should be expected, the BGP value of z rises in all cases if more resources are devoted to the private sector. The number for a* is smaller in all 3 cases except JM 4, due to an

increase in the numerator of the expression (remember that a=C/K). Rates of growth are smaller and transitions slightly quicker for the CES cases. The intuition for the first claim is the following: the growth rate is positively related to the MPK (see eq. 17.19), but has no connection with the MPG in the competitive equilibrium solution. Under concavity conditions in the production function, larger values of z in the BGP means that the MPK and hence the rate of growth will be smaller.

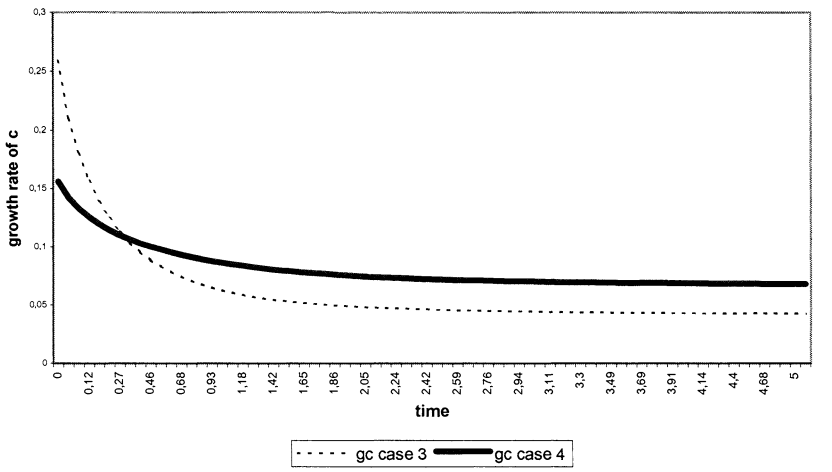


Fig. 17.11. Growth rate of c, Jones Manuelli cases 3 and 4

The JM specification deserves further attention. The value of A=0.3 that was chosen this time means that more importance is attributed now to the AK part of the JM function, in detriment of the neoclassical-Cobb Douglas one. Thus, it is not a surprise that the transitions will be extremely fast under these assumptions. It is also interesting to notice that the growth effect is also higher this time: the growth rate rises from 0.0427 (JM 3) to 0.0683 (JM4). Notice that implications for the growth rate are exactly the opposite now: higher z* are associated with larger growth rates. This fact is due, again, to the ambivalent nature of the JM. As it can be seen from simple inspection of the production function, there is a part of the MPK that is constant and independent of the level of K (the term A). Therefore, larger levels of A are consistent with bigger z*, higher MPK and larger rates of growth.

Sensitivity Analysis: Elasticities

Table 17.2 tries to capture the sensitivity of these models to the changes in parameters by means of analysing the correspondent elasticities (computed as per cent change in the variable x over per cent change in the i^{th} parameter).

The CES function with τ=0.3 (CES 1 and 2) displays a large sensitivity of the growth rate to changes in Ψ. The change in the number of years necessary to reach the BGP (computed just as the percentage change) is, instead, quite small.

With a smaller tax rate (CES 3 and 4), the elasticity of the growth rate with respect to Ψ is still higher. Notice that the change in the BGP value of z is larger in CES 3 and 4, and therefore the same applies to the MPK, that is closely linked to the growth rate as argued above. The reduction in the years needed to get the equilibrium under either value of Ψ (0.75 or 0.6) is now more pronounced.

Let's consider the JM function. Now a variation in the parameter β brings about only a small relative change in the growth rate, although indeed larger when the tax rate is smaller. In effect, the relevance of β in the growth rate is limited by the fact that, again, the JM may be regarded as a combination of an AK and Cobb-Douglas production functions. The impact upon the transition of a change in β is high for the JM in cases 1 and 2, when the AK part of the function is less important, whereas when the JM resembles the AK model the change is comparatively smaller. The opposite holds true for the elasticity of the rate of growth.

We can summarise the messages contained in Table 17.2 by saying the following:

CES cases 1 and 2: the elasticity of the growth rate to changes in Ψ is relatively large. A decrease in Ψ is associated to higher growth, a large increase in a*, a moderate increase in z* and a small reduction in the duration of the transition.

With a smaller tax rate in the CES case the changes in a* are smaller and the variation in z* are larger than in the previous case. The direction of the change is the same. The growth rate is now still more sensitive to changes in Ψ, and the same holds true about the duration of the transition.

JM cases 1 and 2. An increase in β reduces a* and increases z* in a remarkable amount. The variation in the growth rate is, instead, quite small. The speed of transition decreases quite a lot (the number of years necessary to reach the BGP is higher) if β rises.

Table 17.2. Sensitivity Analysis. Elasticities CES and JM Production Functions

	CES 1 and 2	CES 3 and 4	JM 1 and 2	JM 3 and 4
	Ψ	Ψ	β	β
a*	1%	0.50%	-0.41%	-0.08%
z*	0.16%	0.52%	0.75%	0.16%
gc*	7.84%	10.07%	0.02%	0.45%
Δ t	-0.16%	-0.39%	0.58%	0.20%

Notes:

Δ t: measured as per cent change in years necessary to reach BGP

Baseline values for rest of parameters: see Table 17.1

For interpretation of the Table, see text

Table 17.3. Sensitivity Anlysis. JM Production Function. The Role of A

Response of relevant variables to changes in A							
A	**a***	**z***	**gc***	**A**	**a***	**z***	**gc***
0.1	0.62	0.64	0.0093	0.5	0.76	1.2	0.1015
0.2	1.1	0.33	0.015	0.7	1.07	1.1	0.1582
0.3	2.46	0.135	0.1053				
Rest of parameters				**Rest of parameters**			
δ	0.05	σ	3	δ	0.1	σ	3
τ	0.3	b	0.2	τ	0.15	b	0.25
ρ	0.1	β	0.3	ρ	0.08	β	0.3

JM cases 3 and 4. Now this specification keeps the main features of the AK production function: a larger elasticity of the growth rate and a smaller sensitivity of the duration of transition to changes in β. The changes in a* and z* are now more sparse.

If a change in the parameter β does not have a crucial influence in the BGP rate of growth in the JM, as it has been discussed above, which is the main factor that determines it? Table 17.3 goes further into this issue by means of looking at the impact of changes in the technological parameter A for the JM case. It is apparent how the variation in A is positively and greatly associated with changes in the growth rate. This experiment has been carried out under alternative value of the rest of the parameters, and the results are basically the same. For example, a change of A from 0.2 to 0.3 increases the growth rate from 0.015 to 0.1053.

17.6 Conclusions

This paper has analysed the transitional dynamics of two models of endogenous growth with public capital. The main conclusions we can draw from this paper are the following:

The two production function that have been considered in this paper - the CES and the Jones-Manuelli specifications - share two important features: they generate both endogenous growth and convergence, and this prediction seems to match the empirical evidence rather well. They differ in one main point, though, namely, the speed of transition to the BGP. Calibrations employing fairly similar range of parameters show that the duration of convergence to the equilibrium is smaller in the Jones-Manuelli case (except in one case, with a large β), whereas the transition is slower and smoother, instead, in the CES specification.

The duration of convergence is closely related to the parameter Ψ in the CES production function: intuitively, if Ψ is lower, the marginal productivity of capital is higher and thus the economy will grow faster since the relative share of K in the production function is large.

The rate of growth and the speed of transition in the JM are crucially dependent on A and β. The rate of growth in the BGP is positively related to the value of the parameter A. As regards the duration of the transition, it is considerably higher if the parameter β increases, under the assumption of A relatively low (i.e. if the function is closer to a Cobb-Douglas). If A is large and therefore the function is more similar to the AK model, then the change of β has a smaller impact on the numbers of years necessary to reach the BGP although, again, a larger β is associated with a slower transition. Growth sensitivity to changes in β are also more pronounced for large values of A.

Finally, it should be added that these results are tentative and many other calibrations should be performed in order to have a clear idea of the performance of the two production functions that have been employed. In particular, a promising avenue for future research will be to compare these results with those obtained for the social planner solution.

Appendix

Analytical Solutions of the Model in the Jones-Manuelli Case

a) Social planner's solution

Similarly as in the CES case, two types of solution can be distinguished in this setting, the social planner's solution and the competitive equilibrium outcome. The social planner's solution will be presented first. We shall denote the shadow prices of K and G by λ and ν, respectively.

In this case the current value Hamiltonian is:

$$H=\frac{c(t)^{1-\sigma}-1}{1-\sigma}+\lambda\left[(1-\tau)Y-c-\delta K\right]+\nu\left[\tau Y-\delta G\right] \tag{17.26}$$

The first order conditions are

$$H_c=0\rightarrow c^{-\sigma}=\lambda \tag{17.27}$$

$$H_\tau=0\rightarrow\lambda=\nu \tag{17.28}$$

$$\rho\lambda-H_K=\dot{\lambda}\ \rightarrow$$

$$\dot{\lambda}=\rho\nu-\lambda\left\{(1-\tau)\left[A+\beta BK^{\beta-1}G^{1-\beta}\right]-\delta\right\}-\nu\tau\left(A+\beta BK^{\beta-1}G^{1-\beta}\right) \tag{17.29}$$

$$\rho\nu-H_G=\dot{\nu}\rightarrow$$

$$\dot{v}=\rho v-v\left\{\tau\left[A+(1-\beta)BK^{\beta}G^{-\beta}\right]-\delta\right\}-\lambda\left\{(1-\tau)\left[A+(1-\beta)BK^{\beta}G^{-\beta}\right]\right\} \tag{17.30}$$

and the transversality conditions

$$\lim_{t\to\infty} e^{-\rho t}\lambda_t K_t=0 \tag{17.31}$$

$$\lim_{t\to\infty} e^{-\rho t}v_t G_t=0 \tag{17.32}$$

As it was done before, the steady state solution is found by means of taking logarithms and differentiating with respect to time in eq. (17.27) to (17.30):

$$\gamma c^{*}=\frac{1}{\sigma}\left\{A+\beta^{\beta}(1-\beta)^{1-\beta}-\rho-\delta\right\} \tag{17.33}$$

where $\gamma c^{*}=\dfrac{\dot{c}}{c}$

The ratio of private to public capital in the steady state is in this case,

$$\frac{K}{G}=\frac{\beta}{1-\beta} \tag{17.34}$$

b) Competitive equilibrium.

We shall follow the same strategy to get the competitive equilibrium rate of growth as in the CES case. Therefore, and denoting by MPK and MPG the marginal productivity of K and G, the final expression is:

$$\frac{\dot{c}}{c}\text{ c.e.}=\frac{1}{\sigma}\left[(1-\tau)\text{MPK}-\rho-\delta\right] \tag{17.35}$$

References

Arrow, K. J., Chenery, H. B., Minhas, B. S. and Solow, R.: Capital-Labor substitution and economic efficiency. Review of Economics and Statistics 43, 225-250 (1961)

Barro, R.: Government spending in a simple model of endogenous growth. Journal of Political Economy 98 part 2, S103-S125 (1990)

Barro, R.: Economic growth in a cross section of countries. Quarterly Journal of Economics 106, 407-443 (1991)

Barro, R and Sala i Martin, X.: Economic Growth. Cambridge Mass: The MIT Press (1998)

Caballé, J. and Santos, M.: On endogenous growth with physical and human capital. Journal of Political Economy 101 (6), 1042-1067 (1993)

Easterly, W.:How much do distortions affect growth. Journal of Monetary Economics 32, 187-212 (1993)

Inada, K.: Some structural characteristics of Turnpike Theorems. Review of Economic Studies 31, 43-58 (1963)

Jones, L. and Manuelli, R.: A convex model of equilibrium growth: Theory and policy implications. Journal of Political Economy 98, 1008-1038 (1990)

Lucas, R.: On the mechanics of economic development. Journal of Monetary Economics 22, 3-42 (1988)

Mulligan, C. and Sala i Martin, X.: A note on the Time-Elimination method for solving recursive economic models. NBER Technical Paper no. 116 (1991)

Mulligan, C. and Sala i Martin, X.: Transitional dynamics in two-sector models of endogenous growth. Quarterly Journal of Economics 108, 737-773 (1993)

Pontryagin, L. et al.: The mathematical theory of optimal processes. New York: Interscience Publishers (1962)

Ramsey, F.: A mathematical theory of saving. Economic Journal 38, 543-559 (1928)

Rebelo, S.: Long run policy analysis and long run growth. Journal of Political Economics 99, 500-521 (1991)

Romer, P.: Dynamic competitive equilibria with externalities, increasing returns and unbounded growth. Ph.D. Dissertation, University of Chicago (1983)

Romer, P.: Increasing returns and long run growth. Journal of Political Economics 94, 1002-1037 (1986)

Sánchez-Robles, B.: Capital público y crecimiento económico: un modelo alternativo. Cuadernos de Economía 23, 66, 349-371 (1995)

Sánchez-Robles, B.: Infrastructure investment and growth: Some empirical evidence. Contemporary Economic Policy 16, 1, 98-108 (1998a)

Sánchez-Robles, B.: The role of infrastructure in development: some macroeconomic considerations. International Journal of Transport Economics 25, 2, 113-136 (1998b)

Xie, D.: Divergence in economic performance. Journal of Economic Theory 63, 1, 97-112, (1994)

18 Unions, Wages and Productivity. The Spanish Case, 1981-2000

N. Sánchez-Sánchez
University of Cantabria (Spain)

B. Sánchez-Robles
University of Cantabria (Spain)

18.1 Introduction

The labor market is a crucial institution in any economy. It supplies firms with one of the inputs in the production process, labor. Moreover, it allows potential employees to find a job in accord with their preferences and skills. The smooth functioning of the labor market is thus a key piece in order for economic resources to be allocated efficiently. The performance of the labor market has also implications for relevant macroeconomic variables such as productivity, the unemployment rate or inflation.

The labor market, however, differs from conventional markets. One of the main divergences between these two categories is related to the fact that each potential employee does not negotiate the price and quantity of his services alone. Rather, he delegates an important part of this negotiation in the unions. Hence, to understand the mechanisms of a particular labor market we need to ascertain in depth the behavior of the unions operating in it and the consequences of such behavior.

It has been considered traditionally that one of the main role of unions is to ensure an adequate wage for employees. Accordingly, they strive for higher wages and, if they are successful, employees will earn a wage that exceeds the equilibrium wage (the wage that would prevail under perfect competition). The

P. Coto-Millán (ed.), *Essays on Microeconomics and Industrial Organisation*

outcome of the negotiation, therefore, will be a wage that is typically higher than the one clearing the market.

More recently, unions have also been considered as promoters of *social capital* in the economy. According to this last view, they represent workers' petitions, exert a pressure aimed to improve conditions at the workplace and, more generally, act as a vehicle of transmission among employers and employees (the so called *exit voice* mechanism). The basic intuition underlying this idea is that unions may act as the voice of employees, thus easing communication with the employer and helping reduce the degree of job turnover and the training cost of new workers. The enhancement of working conditions, in turn, may increase inputs' productivity. (For a full explanation of this voice mechanism see Freeman (1980), Freeman and Medoff (1984) and Booth (1995)). Checchi and Lucifora (2002) argue that unions can provide an insurance towards unemployment to workers.

The *exit voice* mechanism may contribute to improve the atmosphere at the work place and increase labor productivity, but it could also be the case that unions introduce distortions in the organisation of the firm and spur antagonism among different categories of employees. These distortions, in turn, might damage efficiency and induce lower levels of output per worker. It could also be the case that unions reduced productivity if they imposed *make work*[1] practices, entailing that the number of employees exceed the optimal. The net impact of unions on productivity is thus ambiguous.

Empirical evidence has shown that that unions exert and upward pressure on wages (see, for example,). However, there is not such consensus on the sign of the effect of unions on productivity according to the available evidence.

This lack of consensus can be partly attributed to the fact that unions affect productivity through two different channels. First, they have a direct impact on the degree of efficiency of the firm, which in turn will be positive if the exit voice effect is large enough or negative if the outcome of the unions' activity is a disruption in the social climate in the firm. Second, unions exert an indirect effect on productivity through changes in wages.

The pressure induced by unions on the relative price of labor changes the quantity demanded by the firm and thus the magnitude of output per worker. Thus, even in the cases in which the exit voice mechanism induced substantial rises in productivity, the sign of the total effect would still be unclear. It is not surprising that, while there is a certain accord on the sign of the impact of unions on wages[2], such a consensus is lacking when addressing the effect of unions on productivity.

A set of contributions that flourished following the seminal contribution of Brown and Medoff (1978) have studied empirically the changes in productivity induced by unions. Examples of papers that report a positive impact of unions on productivity are Brown and Medoff (1978), Allen (1983, 1984, 1986, 1988), Clark (1980, 1984), Freeman and Medoff (1984), Benson (2000) and Green *et al.* (1996). More recently, Machin and Stewart (1996) have argued that financial performance (an indirect proxy for productivity) is lower in unionised establishments. García Serrano and Malo (2002) analyse Spanish data and find

[1] This term refers to some procedures imposed by unions within labor deals.

[2] See, for example, Blanchflower and Bryson (2002).

that unionisation reduce gross worker flows (although an impact on job flows is not detected). Instead, other papers such as Delery *et al.* (2000) or Pencavel (2003) finds little evidence in support of the voice mechanism. Paradoxically, studies of this sort for the case of the Spanish economy – where the power of unions has been historically larger - are sparser.

The Spanish labor market has some specific and interesting features. One of the most significant is perhaps the principle of *general efficiency* of agreements or mandatory extension of collective contracts that entails that collective bargaining has a large impact on the whole economy. Moreover, and since in practical grounds this principle makes the services of unions tantamount to public goods, this principles renders irrelevant, on practical grounds, the distinction between members and non members of the unions when trying to measure the true influence of unions. This kind of arrangement is common in Europe, whereas in the US workers can chose between unionised and non unionised workplaces[3]. Another characteristic of the Spanish labor market is the large degree of wage inertia present in it, despite the high level of unemployment (the unemployment rate exceed 20% in the 80s and the early 90s). This inertia has been especially acute in the 80s although has decreased in the last decade.

Furthermore, the behavior of Spanish unions over time has not been uniform. This fact is consistent with some reforms that affected the Spanish labor market in the 90s and tried to reduce some of its rigidities. In the early 90s the government implemented a package of measures intended to increase flexibility in the labor market. In particular, temporary agreements were encouraged. This policy, however, entailed a heavy segmentation of the labor market. The main goal of the social agents was not so much get increases in wages but improve the degree of stability in the job. As a result of this performance, the 90s have envisaged a considerable effort in the reduction of labor costs, tacitly accepted by unions[4]. Whereas in the second half of the 80s the average increase in labor costs was 6%, in the 90s this figure did not reach 3%. Increases in wages agreed by collective bargaining have been even negative in 1994, 1995 and 2000.

This chapter pursues an empirical analysis of the connection between unionisation, wage increases and productivity changes for the Spanish economy over the period 1981-2000.

The structure of the chapter is as follows: Section 2 designs a model that distinguish two channels through which unions may affect productivity and provides a theoretical background for the empirical analysis. Section 3 describes the data and main results of the estimations pursued. Section 4 offers some concluding remarks.

---

3 For an interesting description of the features of European unionization, see Checchi and Lucifora (2002).

4 Blanchflower and Bryson (2002) also document a reduction in the wage premium induced by unions in the US and UK economies from 1994 onwards.

18.2 Theoretical Background

As it was said above, the influence of unions on the firm may be decomposed in two main sorts of effects:
A direct impact on productivity, particularly through mechanisms of the exit-voice sort.

A pressure in wages, that indirectly alters the quantity of labor hired by the firm, and ultimately brings about changes in productivity.

To pin down the impact of unions on productivity is a complex task since the total effect will be the result of several partial influences. A natural way to ascertain whether unions increase or decrease productivity is to estimate both the direct and indirect effects mentioned above (direct impact on productivity and indirect effect through changes in wages) separately and add up the results[5].

This section presents a model, inspired in Clark (1984), that tries to disentangle the two effects mentioned above by means of using elasticities.

We define the elasticity of labor productivity to unions as the changes in productivity induced by the presence of unions (measured in percentage points). This elasticity, in turn, can be decomposed in two terms: the impact of unionisation on labor productivity through changes in wages (elasticity-wages), that we denote by $\varepsilon^{w}_{Q,U}$ and the impact of unionisation on labor productivity through modifications in efficiency, $\varepsilon^{A}_{Q,U}$ (elasticity-efficiency). These ideas are summed up graphically in Fig. 18.1.

Next we proceed to derive and obtain an analytical expression for these elasticities, under the following assumptions:
The production function is Cobb Douglas of the form

$$Q = A K^{\beta} L^{\alpha}$$

$$\alpha + \beta = 1$$

where Q is output, L is labor, K is capital and A is Total Factor Productivity (TFP). α and β are technological parameters.

Firms are price-takers in the products markets, and the industry is in a long run equilibrium position.

The demand function of the good supplied by the firm exhibits constant elasticity η and has the form $P = Q^{-1/\eta}$.

Trade unions press through collective bargaining in order to induce wage increases for workers. Following Lewis (1963), we assume that the wage that comes up from negotiation W_u is related to the competitive wage W_n (the one that would prevail in absence of unions) through eq. (18.1):

$$w_u = (1 + mU) w_n \qquad m > 0 \tag{18.1}$$

[5] Most papers that deal with the connection unions-productivity focus in the first of these two effects. They do not consider, however, the indirect effect that unions may exert on productivity by their impact on wages.

where U captures the degree of unionisation, as measured, for example, by the number of workers affiliated to the union, or, alternatively, by the number of workers affected by collective bargaining.

Fig. 18.1. Impact of unions on productivity

18.2.1 Elasticity of Output with Respect to Unions through Wages

The program of minimisation of the cost for the firm allows to compute the ratio K/L, (eq. (18.2)):

$$\frac{K}{L} = \left(\frac{(1 + mU)w_n}{r} \right) \frac{\beta}{\alpha} \tag{18.2}$$

where r is the rental price of capital.

Total costs are of the form

$$CT = \left(\frac{Q}{A} \right) \left[w_n^{\alpha} r^{\beta} \left(\left(\frac{\alpha}{\beta} \right)^{\beta} + \left(\frac{\beta}{\alpha} \right)^{\alpha} \right) \right] \tag{18.3}$$

Equating average costs to the price of output[6] yields the equilibrium level of output in this industry (eq. (18.4)):

[6] Notice that the equilibrium quantity produced by the firm in the long run when the industry is in perfect competition is obtained from equating prices to the minimum of the long run average cost. When there are constant returns to scale, the curve of average costs in the long run is horizontal. The number of firms is undefined and the equilibrium in the industry can be analyzed as if the production was offered by a single firm.

$$Q = \left(\frac{1}{A}\right)^{-\eta}\left[w_n^{\alpha} r^{\beta}\left(\left(\frac{\alpha}{\beta}\right)^{\beta} + \left(\frac{\beta}{\alpha}\right)^{\alpha}\right)\right]^{-\eta} \tag{18.4}$$

and labor productivity can be written as

$$\frac{Q}{L} = \frac{A}{\left(\frac{\alpha}{\beta}\right)^{\beta} w_n^{-\beta} r^{\beta}} \tag{18.5}$$

From eq. (18.5) we can compute the elasticity of productivity with respect to wages (eq. (18.6))

$$\varepsilon_{Q/L,w} = \left(\frac{\delta Q/L}{\delta w}\right)\left(\frac{w}{Q/L}\right) = \beta \tag{18.6}$$

Using assumption 4 above and substituting w_n by w_u yields:

$$\frac{Q}{L} = \frac{A}{\left(\frac{\alpha}{\beta}\right)^{\beta} [w_n(1+mU)]^{-\beta} r^{\beta}} \tag{18.7}$$

and the elasticity of productivity with respect to unions via wages $\varepsilon_{Q/L,u}^{w}$ can be computed as:

$$\varepsilon_{Q/L,U}^{w} = \left(\frac{\delta Q/L}{\delta U}\right)\left(\frac{U}{Q/L}\right) = \beta\left[\frac{mU}{1+mU}\right] \tag{18.8}$$

Eq. (18.8) is increasing in U: higher presence of unions will induce larger values of this elasticity. $\varepsilon_{Q/L,U}^{w}$ will be positive as long as the margin m is positive.

18.2.2 Elasticity of Productivity to Unions through Changes in Efficiency

So far we have considered the changes that can be induced in output through variations in wages. Unions, however, not only can affect wages but also may conceivably exert some impact on the organisation of the firm. In effect, as argued above, unions may lead to gains in productivity through improvements in the climate of social relations within the firm and in the motivation of workers.

The influence of unions on production through this last channel can be also computed from a Cobb-Douglas production function such as eq. (18.9):

$$Q = A\,(U)\,K^{\beta}L^{\alpha} \tag{18.9}$$

where A(U) is a function of the degree of unionisation of the form:

$$A(U) = A\,(1+dU) \tag{18.10}$$

To grasp the intuition behind eq. (18.10), we can think of A as the level of total factor productivity (TFP) that would be achieved in absence of unions whereas

A(U) is TFP considering the presence of unions. U captures the degree of unionisation in that particular sector. d measures the magnitude and sign of the impact of unions on the organisation of the firm. If this impact is positive (negative) d will also be larger than (smaller than) zero. The rest of the assumptions are the same as in the first case.

The elasticity of productivity to unions via improvements in efficiency can be computed following a similar procedure as above. For an industry under perfect competition and constant returns to scale, labor productivity and elasticity are given by eq. (18.11) and (18.12) respectively:

$$\frac{Q}{L} = \frac{A_n(1+dU)}{\left(\frac{\alpha}{\beta}\right)^{\beta} w^{-\beta} r^{\beta}} \tag{18.11}$$

$$\varepsilon_{Q,U}{}^{A} = \left(\frac{\delta(Q/L)}{\delta U}\right)\left(\frac{U}{(Q/L)}\right) = \left[\frac{dU}{1+dU}\right] \tag{18.12}$$

18.2.3 Total Elasticity of Productivity to Unions

Now it is straightforward to compute the total sensibility of the ratio Q/L to unionisation, that will be given by the sum of eq. (18.8) and (18.12):

$$\varepsilon_{Q/L,U} = \left(\frac{\delta(Q/L)}{\delta U}\right)\left(\frac{U}{(Q/L)}\right) = \left[\frac{\beta mU}{1+mU} + \frac{dU}{1+dU}\right] \tag{18.13}$$

The first term in the right hand side of eq. (18.13) captures the sensibility of productivity to unions as far as wage negotiations are concerned. If m is positive, then the presence of unions is positively correlated with productivity. The second term captures the sensibility of productivity to unions via changes in the organisation of the firm. The sign of this term depends crucially on d. If $d<0$, then unionisation ends up in lower levels of productivity. If $d>0$, the sign of the total effect of unions on productivity depends on the relative magnitude of the individual effects. If both effects are equal in magnitude, then unions will not affect productivity.

18.3 Empirical Analysis

This section will pursue an empirical exercise that estimates the impact of unions on productivity. We proceed in two steps. First, we estimate the wedge m that unions impose on wages. Second, we estimate the influence of unions on efficiency by means of estimating d.

As a first step we need to define three categories of wages. W_{ni} is the equilibrium wage in the i^{th} industry in absence of unions. It captures idiosyncratic features of the sector such as productivity or unemployment. W_{ei} and W_{oi} are

defined as the wages negotiated at the firm level and at a higher level (usually the sector's level), respectively. They are affected by unionisation and by the nature of collective bargaining. The links between the three categories of wages can be described as:

$$W_{ei} = W_{ni} (1 + m_1) \tag{18.14}$$

$$W_{oi} = W_{ni} (1 + m_2) \tag{18.15}$$

where m_1 and m_2 are the wage premiums achieved by bargaining at the firm level and at a higher level, respectively.

Eq. (18.14) and (18.15) establish that wages established under collective bargaining, either at the firm level or at a higher level, can be computed as the wages that would be fixed in absence of negotiation plus a margin due to unionisation and bargaining.

Dividing (18.14) and (18.15) through L (the number of employees) and taking logs we get:

$$\ln (W_{ei}/L) = \ln (W_{ni}/L) + \ln (1 + m_1) \tag{18.16}$$

$$\ln (W_{oi}/L) = \ln (W_{ni}/L) + \ln (1 + m_2) \tag{18.17}$$

The next step consists in assuming that the wage in each sector can be computed as a geometric average of the following terms: wages fixed under bargaining at the firm level, wages fixed under bargaining at a higher level and wages fixed in absence of bargaining. Thus wages in sector i^{th} can be computed as (18.18):

$$W_{ui} = W_{ei}^{Pe} W_{oi}^{Po} W_{ni}^{Pn} \tag{18.18}$$

where P_e, P_o, and P_n are the percentage of employees covered by bargaining at the firm level, bargaining at a higher level and covered by no bargaining, respectively. Dividing through by L and taking logs we have:

$$\ln (W_{ui}/L) = P_e \ln (W_{ei}/L) + P_o \ln (W_{oi}/L) + P_n \ln (W_{ni}/L) \tag{18.19}$$

Plugging in (18.16) and (18.17) in (18.19) yields:

$$\ln (W_{ui}/L) = P_e \ln (W_{ni}/L) + P_e \ln (1 + m_1) + P_o \ln (W_{ni}/L) + P_o \ln (1 + m_2) +$$

$$P_n \ln (W_{ni}/L) \tag{18.20}$$

Employing the approximation $\ln(1+m_1) = m_1$, and rearranging terms we get:

$$\ln (W_{ui}/L) = P_e m_1 + P_o m_2 + (P_e + P_o + P_n) \ln (W_{ni}/L) \tag{18.21}$$

$$\ln (W_{ui}/L) = P_e m_1 + P_o m_2 + \ln (W_{ni}/L) \tag{18.22}$$

According to (18.22) the average salary in sector i^{th} is a function of the proportion of employees covered by each sort of agreement together with the margins imposed in the various extents of negotiation.

Generally speaking, the exact value of W_{ni} will be unknown. However, it can be estimated assuming that wages depend on a set of variables according to a function of the form:

$$\ln(W_{ni}/L) = f(X_i) + \varepsilon_{i1}$$ [7]

The variables that encompass the X vector will be detailed below. Plugging this expression in (18.22) yields:

$$\ln (W_{ui}/L) = P_e\, m_1 + P_o\, m_2 + f(X_i) + \varepsilon_{i1} \qquad (18.23)$$

From the estimation of (18.23) we can recover the coefficients m_1 and m_2.

18.4 Data and Variables

The empirical analysis pursued has been divided in two subperiods: 1981-1992 and 1993-2000. The reason for this division is related to the data. In 1992 the methodology employed by the National Institute of Statistics changed and the series constructed before and after this year are not homogeneous.

For the subperiod 1981-92 we have used a data panel of 88 activities obtained from the Industrial Survey of Firms. The dependent variable is wage per worker, measured in real terms. The deflator employed is the index of industrial products. The regressors, following Fernández and Montuenga (1997), include two sets of variables. The first one considers the aspects that determine the wage internally, while the second set refers to conditions in the labor market. Among the first we have included output per worker, the average size of the firm, a proxy of human capital and hours worked per employee. In the second group we have included the first lag of the wage in that particular sector and unemployment (both at the aggregate and the sector's level)[8].

The degree of unionisation has been captured by the number of employees affected by each agreement. The statistics compiled by the Ministry of Social Affairs discriminate among employees affected by agreements at the firm level (denoted above by Pe) and employees covered by agreements of higher scope (Po). We have also considered the total number of employees affected by bargaining (Pt). This measure is probably too crude a proxy of unionisation; unfortunately, and to our knowledge, other alternatives to capture the degree of unionisation were not available for the Spanish economy.

In the subperiod 1993-2000 we have used data from 100 branches. The source is the *Encuesta Industrial de Empresas* (Industrial Survey of Firms). The variables included in the analysis are the same as in the first subperiod. The only exception is T (establishments over employees, a proxy of the size of firms) not available in this second subperiod.

---

[7] Alternatively, we could assume that Wni is fixed within a insider-outsider model, along the lines of Layard *et al.* (1991).

[8] The variable average hours worked per employee has been included to control for the fact that collective bargaining may affect not only wages but also the number of hours that encompass the working day.

18.5 Main Empirical Results

a) Wages

We have pursued the estimation of the wage equation by means of the Generalised Method of Moments (GMM). The estimation includes a lag of the wage cost in order to capture wage inertia.

We have also considered the possibility of treating productivity (Va) and sectoral unemployment (Us) as endogenous variables. Accordingly, we have used as instruments the first lag of both variables. Other regressors are: hours worked (Hr), the number of establishments per worker (T), some proxies for the degree of unionisation –as above, the percentage of employees under firms agreements (Pe), under general agreements (Po), and under agreements of total coverage (PT)- and the stock of human capital, (H) measured by the percentage of employees that enjoy a certain level of studies (primary studies).

Results are displayed in table 18.1. The main messages of the estimation can be summarised as follows:

1. The point estimate of the first lag of the wage exhibits a rather high value. It is close to 0.9 in the subperiod 1981-92 and reaches the value 1 when two lags are introduced in the equation. In the second subperiod, however, the coefficient is not higher than 0.5. A preliminary interpretation of this result is that wage inertia has been acute in the 80s but has decreased over time. This is consistent with the larger degree of flexibility that the Spanish labor market has acquired in the 90s when compared to the 80s.

Results suggest that unions have indeed imposed a premium to wages in collective bargaining over the subperiod 1981-92. The coefficients associated to Pe and Po are positive and significant at conventional levels. This is not the case, though, in 1993-2000, since the signs of both coefficients are negative. This result is in accord with the idea that unions pursued a strategy of moderation in bargaining from 1994 onwards (Fina *et al.* (2001a, 2001b)).

The link between sectoral unemployment and labor costs over the subperiod 1993-2000 is found to be positive. In other words, the decrease in unemployment has been accompanied by a reduction in wages. This can be attributed, in turn, to the greater social commitment of agents, and in particular of unions, that arose in the 90s. This attitude entailed a moderation in wage demands when negotiating, despite the fact that unemployment was also decreasing.

Finally, the coefficients associated to productivity and hours per employee are positive and significant in both subperiods[9].

The second variable employed to capture the degree of unionisation is the total coverage of all the agreements that have been signed (Pt). Results obtained when using this variable are displayed in table 18.2. The estimations pursued are very similar to those carried out in the previous analysis. The main conclusions are also similar. Pt has a positive sign and is significant in all the estimations pursued for

[9] Fernández and Montuenga (1997) suggest that wages are influenced by productivity only in laggard sectors, whereas in more dynamic sectors this link can not be detected.

the period 1981-93, while exhibits a negative sign (but is not significant) in the following decade.

Table 18.1. Wage cost estimation (unionisation captured by Po and Pe)

Again, time dummies are important in order for instruments to be considered as appropriate. If dummies are not included in the estimation, the Sargan statistic rejects the null hypothesis of the validity of the instruments[10].

[10] Arellano (2000) documents a similar result, in the sense that dummies are necessary in order for the instruments to be valid when working with cross sectional data with a common trend. He attributes this result to the high variability of the dependent variable over the period considered. This variability implies that the ortogonality conditions in this particular case are not fulfilled.

The theoretical model presented above, together with the results got from the estimation, allows to compute the sensibility of productivity to the presence of unions $\varepsilon_{Q/L,u}^{w}$, by using eq. (18.24):

$$\varepsilon_{Q/L,u}^{w} = \beta \times [mU/(1+mU)] \quad (18.24)$$

where U captures the degree of unionisation and has been considered to be the point estimates of Po, Pe and Pt in the estimations pursued above.

The value for the parameter (β) has been taken as 0.33, in line with most of the literature. The elasticity may be computed for each sector and year. Since we are interested in the differences across sector, however, we shall not compute the elasticity on a yearly basis. Rather, we take the average of the measure of unionisation over the period considered.

The values of the elasticity for each sector over the subperiod 1981-92 are shown in table 18.3.

Columns 1 and 2 of table 18.3 display the values of the elasticity $\varepsilon_{Q/L,u}{}^{w}$ when it is computed using employees under general coverage agreements Po and employees under firms agreements Pe as proxies of the degree of unionisation, respectively. The elasticity measured by means of general coverage agreements varies between zero and 0.02. This last value correspond to those sectors where the coverage due to general agreements exceeds the 80% of the employees. The elasticity measured through firms agreements varies between 0 and 0.037.

In some particular branches (precision tools and leather) both elasticities are rather low. Anyhow, most sectors display a high elasticity when unionisation is captured by agreements of higher scope, whereas elasticity is lower when referred to firms agreements.

Column number 3 displays the values of elasticities obtained when unionisation is proxied by the proportion of employees subject to agreements of general coverage. The values vary within the range of 0.008 (shoes) and 0.27 (oil extraction, chemistry, textile, metallic products and food).

The computation of the elasticity of productivity to unions through wages for the 90s is not pursued here, since the variables that capture unionisation were, in general, not significant in the estimations. It is possible that the proxies employed do not capture accurately the role of unions in the 90s. The search of adequate proxies of unionisation and the computation of this elasticity is left for future research.

b) Elasticity of productivity to unions through gains in efficiency

So far we have computed the response of productivity to unions through changes in wages. Next we shall analysis that answer via changes in efficiency. The impact of unionisation can be measured by two different ways: considering that unions increase labor efficiency, along the lines of the model of Brown and Medoff (1978), or assuming that unions increase the efficiency of all inputs. We have chosen this second option since it seems more adequate for the Spanish situation: as it was stated above, the distinction between members and non members of the unions does not show the true influence of unions due to the principle of general efficiency of agreements.

In accord with the analysis stated above, now we shall capture the presence of unions through the parameter A in the production function[11]. In particular:

$$Q_i = A_n (1+dU_i) K_i^{\beta} L_i^{\alpha} \qquad (18.25)$$

[11] Following Serrano (1996) we could introduce as an additional input in the production function the stock of human capital. Thus the production function would be:
$Q = A K\alpha H\beta L(1-\alpha-\beta)$
Under constant returns to scale, dividing by the number of employees yields:
$Ln (Q/L) = \ln A + \alpha \ln (K/L) + \beta \ln (H/L)$

dividing through by L_{it} and taking logs yields:

$$\ln (Q_{it}/L_{it}) = \ln A_n + dU_{it} + \beta \ln (K_{it}/L_{it}) + (\alpha + \beta - 1) \ln L_{it} \quad (18.26)$$

where $\alpha + \beta$ represent the (constant) returns to scale, i indexes sectors and t time.

The dependent variable is measured as added value per worker, deflated with the index of industrial prices. The regressors are: capital stock (K), the number of employees (L), the number of establishments per worker (T), some proxies for the degree of unionisation –as above, the percentage of employees under firms agreements (Pe), under general agreements (Po), and under agreements of total coverage (PT)- and the stock of human capital, (H) measured by the percentage of employees that enjoy a certain level of studies (primary studies).

It could be the case that capital and labor were not strictly exogenous variables and thus the prerequisites for a valid within estimator would not be fulfilled. To overcome this possibility we have employed an instrumental variables estimator. The method of estimation chosen is Generalised Method of Moments (GMM) (Arellano and Bover (1990)). Table 18.4 displays the main results obtained from the estimations pursued when capital and labor are instrumented by three of their own lags. We show the results got from One Step GMM (GMM1) and Two Steps GMM (GMM2). Results from GMM1 seem more plausible. The comparison of the results obtained from GMM1 and GMM2 suggests that GMM2 estimates may be less precise due to a downward finite sample bias[12].

The point estimates of K and L display values in accord with the literature and are significant at conventional levels. Human capital is also positively and significantly correlated with productivity. The size of firms shows a negative sign, which can be attributed to some kind of agglomeration effect whereby in larger establishments it is easier to achieve a higher level of productivity.

Proxies of unionisation display negative signs but are not significant in the GMM1 specifications.

Analogous estimations have been made for the subperiod 1993-2000 (Table 18.5). The main conclusions obtained for the first subperiod carry over to the second. Again, the coefficients associated to proxies of unionisation are not significant.The coefficients of Pt and Po are now positive but not significant.

Once the parameter d is estimated, we could compute the elasticity of productivity to unionisation using eq. (18.27):

$$\varepsilon^{A}_{Q/L,u} = [dU/(1+dU)] \quad (18.27)$$

where U is the measure of unionisation. Since this variable has not been significant in the estimations for the 90s, we assume that d is equal to zero and the elasticity is also zero.

[12] A similar result is documented in Arellano and Bond (1991)

This result should be taken with caution. It is perhaps too risky to say that unionisation has not had any influence in the productivity of firms over the period considered. Nonetheless, the tentative conclusion we may derive from this exercise is that the measures of unionisation employed here do not suggest a relevant impact of unions on productivity via gains in efficiency of the firm. Remember that these proxies, although imperfect, did capture the impact of unions on wages, at least to a certain extent.

In order to combine both effects, through wages and through the productive process, we use eq. (18.13) above. The resulting expression is:

$$\varepsilon_{Q/L,\,u} = \beta\, mU/(1+mU) + dU/(1+dU) \qquad (18.28)$$

Since the elasticity to unionisation via productivity is zero, total elasticity is tantamount to the response of productivity to unions via wages (already displayed in table 18.3).

These results suggest that in the 80s Spanish unions have had a positive influence in the productivity of labor, especially via increases in wages. This effect is not apparent in the subperiod 93-2000. The impact of unions on productivity via productive process is negative in 1981-92 and positive but not significant in 1993-2000.

18.6 Concluding Remarks

Unions may affect output through two channels: first, unions exert pressures on wages, and this alters the demand of labor by the employer and hence the productivity of labor. Second, unions may affect the level of efficiency in the firm and thus have impact on labor productivity.

This paper has described a procedure that allows to disentangle and measure these two effects. The main originality of the models rests on the decomposition of elasticity in two components that can, in turn, be added up: the elasticity of productivity to unions through changes in wages and the elasticity of productivity to unions via changes in the productive process.

Next we have pursued an empirical exercise using data from the Spanish economy. The data cover two subperiods: 1981-92 and 1993-2000.

Since Spanish employees do not need to belong to a union in order to profit from the outcomes of collective bargaining, data of union affiliation are not very representative. We have used as a proxy of unionisation the percentage of employees covered by firms agreements, Pe and total coverage agreements, Po.

The main messages of the estimations are the following:

a) Unions seems to have brought about positive margins on wages in the 80s. This margin is not observed, however, in the 90s. This result is in accord with the recent economic history of Spain: the 90s have envisaged a considerable effort of reduction of labor costs.
b) Wage inertia, understood as sluggish adjustment of labor costs, is rather high in the 80s but smaller in the 90s.
c) Unionisation, as captured by Po and Pt is positively correlated with wage increases in the 80s but not in the 90s.
d) The sensibility of labor costs to changes in employment is small. This is consistent with the high degree of rigidity in the Spanish labor market.

The influence of unions in the productive process does not appear as relevant. The data available for the Spanish economy do not confirm the conclusions of the exit voice models, according to which unions may improve communication between employers and employees, thus improving motivation and the atmosphere at the work place.

Finally, these results should be considered with caution since the scope of conclusions are limited by the availability of data, in particular those that capture the degree of unionisation.

References

Allen S. G.: Unionization and productivity in office building and school construction. NBER Working Papers, June (1983)

Allen S. G.: Unionized construction workers are more productive. Quarterly Journal of Economics 99, may, 51-74 (1984)

Allen S. G.: Unionization and productivity in office building and school construction. Industrial and Labor Relations Review 39, Jan., 187-201 (1986)

Allen S. G.: Productivity levels and productivity change under unionism. Industrial Relations vol. 27-1, 94-113 (1988)

Arellano, M.: Panel data econometrics, forthcoming (2000)

Arellano, M. and Bond, S.: Some Tests of specification for panel data: Monte Carlo evidence and an application to employment equations. Review of Economic Studies 58, 277-297 (1991)

Arellano, M. and Bover, O.: La econometría de datos de panel. Investigaciones Económicas (Segunda época) vol. XIV-1, 3-45 (1990)

Benson, J.: Employee voice in union and non-union Australian workplaces. British Journal of Industrial Relations 38-3, 453-459 (2000)

Blanchflower, D. and Bryson, A.: Changes over time in union relative wage effects in the UK and the US revisited. NBER Working Paper 9395, December (2002)

Booth, A. L.: The economics of trade union. Cambridge University Press 1995

Brown, C. and Medoff, J.: Trade unions in the production process. Journal of Political Economy vol. 86-3, 355-378 (1978)

Checchi, D. and Lucifora, C.: Unions and labour market institutions in Europe. Economic Policy 35, 363-498 (2002)

Clark K. B.: The impact of unionization on productivity: a case study. Industrial and Labor Relations Review 33, July, 451-469 (1980)

Clark K. B.: Unionization and firm performance: The impact of profits, Growth and Productivity. American Economic Review 74, December, 893-919 (1984)

Delery, J. E., Gupta, N., Shaw, J. D., Jendins, G. D. and Ganster, M. L.: Unionization, compensation and voice effects on quits and retention. Industrial Relations 39-4, 625-645 (2000)

Fernandez, M. and Montuega, V.: Salario y productividad sectorial: ¿existe evidencia de un comportamiento dual?. Cuadernos Económicos de ICE 63, 79-103 (1997)

Fina, L, Gonzalez, G. and Perez, J. I.: Negociación colectiva y salarios en España. Colección de Estudios del Consejo Económico y Social 110. Madrid 2001a

Fina, L, Perez, J. I. and Toharia, L.: El reto del empleo, Macgraw Hill. Madrid 2001b

Freeman, R. B.: The exit-voice tradeoff in the labor market: unionism, job tenure, quits, and separations. Quarterly Journal of Economics vol. 94-4, 643-673 (1980)

Freeman, R. B. and Medoff, J. L.: What do unions do? Basic Books, Inc. New York 1984

Garcia-Serrano, C. and Malo, M. A.: Worker turnover, job turnover and collective bargaining in Spain. British Jounal of Industrial Relations vol. 40-1, 69-85 (2002)

Green, F., Machin, S. and Wilkinson, D.: Trade unions and training practices in British workplaces. LSE Centre for Economic Performance Discussion Paper 278, February (1996)

Layard, R., Nickell, S. and Jackman, R.: Unemployment. Oxford University Press 1991

Lewis H. G.: Unionism and relative wages in the United States: an empirical enquiry. University of Chicago Press. Chicago 1963

Machin, S, and Stewart, M.: Trade unions and financial performance. Oxford Economic Papers 48, 213-241 (1996)

Pencavel, J. H.: The surprising retreat of Union Britain. NBER Working Paper 9564, March (2003)

Serrano, L.: Indicadores de capital humano y productividad. Revista de Economía Aplicada 4-16, 177-190 (1996)

19 Comparative Analysis of Port Economic Impact Studies in the Spanish Port System (1992-2000)

J. I. Castillo-Manzano
University of Seville (Spain)

P. Coto-Millán
University of Cantabria (Spain)

M. A. Pesquera
President of Gijón Port Authority (Spain)

L. López-Valpuesta
University of Seville (Spain)

19.1 Introduction

This work is a comparative approach of the results obtained from a representative sample of economic impact studies made during the nineties on the ports of the various sectors of the Spanish Port System (Southern, Northern and Eastern Port Subsystems). Before starting this comparison, we will set forth the adequacy of this mathematic tool to evaluate port investment projects. We have homogenised over time and space the results obtained from a representative sample of these studies and have included technical recommendations to standardise the results of the studies in order to facilitate the comparison.

P. Coto-Millán (ed.), *Essays on Microeconomics and Industrial Organisation*

19.2 The Economic Impact Studies as a Tool Employed by the Port System

In spite of being one of the most widespread tools used in port analysis, as shown by the studies made in the Spanish Port System, it has been always criticised for its methodological shortcomings. We will remark below the main criticisms received by these studies.

Verbeke and Debisschop (1996) pose certain issues for discussion which are often criticisms of the use of port economic impact studies:

1. Those who are against the use of economic impact studies state that these studies cannot deal with the effects of the marginal changes produced in the prices of the *inputs* and *outputs* as a consequence of carrying out the investment project, however, these effects are accounted for in the cost-benefit studies. Nevertheless, for Verbeke and Debisschop, an impact study compares the economic situation given by the execution of the project with a situation in which this project had not been implemented, so the marginal analysis would be implicit.
2. A second criticism is based on the fact that, the economic impact studies deal with some elements such as wages, mortgages and financial costs as increases of impact, while they should have been taken as project costs as in the conventional financial analyses or in the cost-benefit analyses. Therefore, if economic impact studies are used as a tool for selection among the diversity of investment projects, we may run the risk that, when the project with the highest impact is selected, it may also be the project with the highest cost. This statement cannot be so easily held if we take into account that in the economic impact studies the above-mentioned issues are only included when they come from a *"sustainable"* activity, in other words, an activity maintained over time.
3. The economic impact studies are often criticised because they are considered to be used as a public relationships instrument rather than as a serious tool for the evaluation of projects. But, according to the authors, the same could be said of the cost-benefit analyses, which can be employed to give false credibility to the decisions made for political or income distribution reasons. In order to increase the confidence in the use of impact studies, it is proposed that the impacts calculated on added value and employment should be based upon firm market analyses including studies which examine in detail the different port users as well as the predictions on port traffic and the development of commercial and industrial activities. With this aim, it is necessary to clearly establish the relationships between the economic and port activity in the area studied.
4. Those who are against the application of economic impact studies to port analyses usually state that these studies implicitly believe in some way that the businessmen and firms' managers change their strategic behavior depending on whether a port project is carried out or not.

   However, it is certain that this belief is necessary for example, according to Verbeke and Debisschop, when a cost-benefit analysis predicts the use of one

port rather than another by a particular traffic, basing upon a simple comparison of costs.

5. Economic impact studies have a bias towards the overestimation of the effects produced by the non-existence of the port on the industrial and commercial activities in connection with it. In order to avoid this situation, authors consider that a proper estimation requires the availability of recent and reliable *input-output* tables about the economic impact area, as well as indicators which show whether the productive structures of the companies affected by the project are different from those suggested on the *input-output* tables.
6. One of the most important criticisms of the use of economic impact studies is that these do not take into account any possible negative externality caused by the development of the port project, and they do not allow the introduction of *shadow prices* in order to assess the particular *inputs* or *outputs*. Therefore, the economic impact studies should be combined with other evaluation tools in order to make decisions, so that, the importance that the economic impact study would have, would depend on the importance given by decisions-makers to the criteria of the added value generated, the jobs created and the taxes collected. However, according to Verbeke and Debisschop, it cannot be overlooked that the rest of the tools which can be employed have similar shortcomings.
7. Another criticism arises from the idea that the cost-benefit analysis is a more efficient instrument than the economic impact studies since the former may include in only one decision criterion all the economic welfare effects produced by the project, while an economic impact study, whose results are reflected in different areas (wages, jobs, added value, taxes...), may provoke confusion in those who must make a definite decision. According to Verbeke and Debisschop, we would be in the midst of a false controversy since both analytical instruments are based upon definitions which do not have anything to do with what are considered to be the grounds of economic welfare. Cost-benefit analyses especially focus on the surplus increase to be obtained by producers and consumers as a consequence of the application of the project, while economic impact studies focus on determining the sustainable contribution over time of the port project to the regional product. Given that economic welfare can be defined in different ways, it is not so evident that one decision tool may prevail over the rest.

With respet to Verbeke and Debisschop's criticism, mentioned in the fifth place, concerning the reliability of the input-output tables of the productive structures of port companies, Consultrans and the *Centro de Estudios Económicos Fundación Tomillo* (1998) criticise the economic impact studies, which they call "traditional", for quantifying the indirect and induced impact by means of input-output tables. According to them, this cannot be done thoroughly since the interrelations within the port concerning both supply and demand, are unknown. The only port references with which these effects operate are those obtained in the sample: invoicing, employment and added value. Any other economic sector with similar figures to these variables would yield exactly the same indirect and induced impacts, no matter what its cost and demand structure may be. For this reason, in order to include the ports in the input-output table, we must know the port agents and the complete structure of its purchases and sales by economic sector, rather

than only the global figure of invoicing, employment and added value. The questionnaire addressed to the companies must record in detail by sectors, the structure of the supply or production and that of the demand. According to these authors, the traditional methodology does not employ the input-output model for their analysis, but rather, it simply uses the input-output tables in order to approach the indirect and induced impacts caused by the port activity.

Although this would be the most exhaustive way of carrying out the impact study, it is very difficult, not to say impossible, to be implemented since the breakdown of the information required must be done so thoroughly (classified into as many sectors as considered in the corresponding input-output table) by the port companies that, in most cases it does not exist. In the Port Industry there is a high presence of family companies whose accounting does not have that degree of disaggregation or, on the contrary, they are big companies with branches within the area subject to study but whose accounting is carried out in their main office, so they cannot provide that breakdown level either. Moreover, even if this breakdown were possible, the increase of the costs for the study would be such - mainly when carrying out the surveys campaign - that it would be necessary to estimate the profitability of making it prior to carrying it out.

An intermediate solution is usually working with less sectors than those included in the input-output table, between 15 and 20, and trying to encourage some of the representative companies of the sector to disaggregate their purchases and investments. Obviously, this will be managed by motivating the companies as to the beneficial effects which the economic impact study will have on them. Once we have the sectorial disaggregations of the most representative companies, we will be able to use them as a base to establish specific behavior patterns for the Port Industry of the port analysed.

19.3 Spanish Impact Studies Representative Sample

In this sample we have focused our analysis on the economic impact studies done in the Spanish ports since 1992. During this year, the 27/1992 Act which regulates the current Spanish Port System, under which the old Port Boards were transformed into Port Authorities, was passed. Therefore, we have not included in our study the first works about the Spanish ports impacts, such as those carried out by Fraga and Seijas (1992), which use data from 1990 for the Ports of El Ferrol and by De Rus, Román and Trujillo (1994), for the Port of Las Palmas, with data from 1992. The exception has been the study of the Puertos Gallegos (*Gallego* Ports), made in 1994 but referring to 1992, in which, the input-output methodology proposed by the TEMA consulting company to be applied in the Spanish Port System impact studies, started to be developed[1].

In order to facilitate the analysis, we have taken into account only one study per port – among the wide variety that existed -, generally the latest one, and the results referred to the most realistic hinterland. We will mention below two

[1] This methodology has been the most widely used by the Spanish ports in the development of impact studies.

methodological aspects in order to delimit these two questions (the time period and the geographical area object of the impact studies):

- The time period must be a year which may be representative of the port activity, without big changes in its traffic statistics or in investments. Otherwise, we would have to use average data. For example, if the data about investment in infrastructures or superstructures are considered to be abnormal, we would have to work with a figure which may be representative of the average port activity. In addition, these data must be from sufficiently long ago in order to be included in the finalised accountancy of the consulted companies but at the same time sufficiently recent to give a current perspective to the study.
- As far as the geographical aspect is concerned, impact studies usually quantify the influence of the port in an area which is clearly delimited from the jurisdictional point of view (province, region, country). However, in the majority of the cases, this area does not coincide with the true hinterland of the Port, a coincidence which would provide the highest accuracy from the methodological point of view. An adequate instrument to determine the hinterland of a transport infrastructure are the statistical regressions between the port inlet traffic and the GDP of the area submitted to study. This idea is verified by the foundations of the International Trade Theory, under which imports depend on the importing region's income. For this reason, a linear regression between the port inlet traffic and the GDP/GAV of the area subject to study can be estimated. Thus, the imports concept is assimilated to that of Port inlet traffic. Another work hypothesis which determines the hinterland is the cointegration analysis study of any port traffic with variables representative of the hinterland.

It is widely known that economic impact studies usually divide port activity into two blocks: the Port Industry or set of activities directly required for the shipping transport of goods and passengers (Port Authority, ship's agents, stevedores, freight forwarders, coasters, towing, mooring); and the Port-dependent Industry or related to the Port, which includes the set of economic activities of the regional or provincial economy which presents a certain dependency relationship with the port rather than being part of the Port Industry.

In this work, we have only considered the variables obtained after calculating the economic impact of the Port Industry in each of the ports analysed. We have chosen the Port Industry due, initially, to the high homogeneity in its definition, rather than the Port-dependent Industry, whose delimitation may be very different from one to other studies, depending on the industrial or commercial character of the port. However, currently, the majority of the Spanish ports have this double function, the difference between them being marked by the specific weight of each function.

Nevertheless, we know that, the more efficiently a port operates, the easier the traffic of goods is at lower costs, the less the impact caused by the Port Industry is (as regards Jobs, Wages and Salaries, Taxes) and the higher the positive impact will be on the Port-dependent Industry as regards Added Value. Since the Port-dependent Industry is not dealt with in this work, we have not been able to verify any possible transfer of the economic impact of the Port Industry to the Port-

dependent Industry, thus leaving the extension of this analysis pending a new line of research.

With respect to the definition of both sectors; the Port Industry and the Port-dependent Industry, we must say that, logically, the more or less extensive delimitation of each of them, shall depend on the higher or lower port impact. Anyway, whatever the definition may be, it must be clearly specified at the beginning of a research in order to facilitate comparisons and avoid definition bias.

We will set out below the impact studies analysed in this comparative research and highlight the methodological differences which may exist between them:

A) South Port Subsystem.

1. Bahía de Algeciras Port Authority. This study was carried out by a group of professors of the University of Seville monitored by professors Castillo Manzano and Lebón Fernández and coordinated by professor López Valpuesta. The year chosen for the study was 1996 and they obtained results about the impact of the Autonomous Community of Andalusia and the shire of Gibraltar. A quotation of the research managed by the *Dirección General de Carreteras y Puertos del Estado* (General Management of State Roads and Ports) and developed by the Consulting Company ETT. S.A. states "the scope of influence of Algeciras Port presents little scope of influence, mainly focusing on the province of Cádiz which captures 62% of the journeys and maintains specially strong relationships only with the bordering provinces and Madrid." Therefore, - and although it seems more accurate to consider the Gibraltar area as hinterland of Bahía de Algeciras Port -, due to the few statistical data available (the shire's GDP is unknown), this research has only studied the impact of the Bahía de Algeciras Port on the Andalusian economy. The methodology employed was the Input-Output model adjusted to the Spanish Port System by the TEMA Consulting Company for State Ports.
   The Port Industry analysed in this research also includes the fishing activity of Algeciras and Tarifa Ports with an impact of 1,364,28 direct jobs.
2. Seville Port Authority. This research, like the one above, was monitored by professors Castillo Manzano and Lebón Fernández and coordinated by professor López Valpuesta. The results were obtained for the Autonomous Community of Andalusia and the province of Seville and referred to the year 1995. Another research monitored by the same professors was carried out for 2000 for the province of Seville. With respect to Seville Port's hinterland and according to ETT. S.A. "the scope of influence of Seville Port is very little; 78% of the journeys are within the province of Seville, and only the relationships established with the bordering provinces of Córdoba, Badajoz and Cádiz, as well as Madrid connections, are important". Therefore, for this comparative study, they have chosen the results of the impact on the most realistic hinterland; the province, and selected the most updated of the two which exist; that of the year 2000. The methodology used was the Input-Output adjusted to the Spanish Port System by the TEMA Consulting Company.
3. Bahía de Cádiz Port Authority. The impact study of this Port was monitored by professor Rey Juliá, from the University of Cádiz. The results obtained referred to 1998 and were calculated for the province of Cádiz. The methodology

employed was the Input-Output adjusted to the Spanish Port System by the TEMA Consulting Company.

This study deals with the fishing activity in a separate section, while, in this paper, it has been included within the Port Industry. The fishing activity accounts for about 60% of the Port Industry impact, with the exception of the Tax variable, which represents 10%, a logical consequence of the presence of the Customs in the Port Industry.

4. Huelva Port Authority. The Huelva Port impact study was carried out by a team of professors from the University of Huelva monitored by professor García del Hoyo. The period covered in the study is 1996 in the geographical areas of the province of Huelva and the Autonomous Community of Andalusia. This comparative study took the Port economic effects on the province given that, according to ETT's study, "the scope of influence of Huelva Port has little scope of influence; 64.5% of the journeys take place within the province of Huelva, and only the relationships established with Seville are important".

   This study is based upon the Input-Output methodology and the Andalusia economy's multipliers[2] and it makes a distinction between three types of impacts which, according to the authors, contradict the traditional definition given in the Input-Output analysis. According to them, the traditional model distinguishes between direct and indirect impacts depending on the companies' location (those companies located within the port's premises produce direct effects, while the ones outside these premises but dependent upon their operation, produce indirect effects). According to the authors, this definition (which is not the one employed in the methodology proposed by the TEMA consulting company) is not adequate since, many companies located within the port area are landowners. For example, the study mentions Huelva's chemical plant, which occupies the land of Huelva Port and is not considered to be a direct effect.

   The three impacts defined in this study are:

   - Direct impacts: those generated within the port facilities as a consequence of the development of port operations by public and private entities under the Port Authority's coordination, in compliance with the 27/1992 Act.
   - Indirect impacts: are derived from the economic activities developed by firms whose operations are, to a large extent, linked to the port, since it is the main channel for the firm's raw materials and/or products. According to the authors, indirect impacts include all those activities which would disappear if the Port did not exist, which include: fishing and a percentage of the chemical and basic companies, depending on the volume of goods they transport through the Port, estimated in monetary units. The rest of the companies located in the Port, not considered in the direct effect, as well as the part of the chemical and basic industry not included in the indirect effect, would be accounted for by an indicator of the strategic importance of the Port.

[2] Proposed by Otero J. M. (1995): "Multiplicadores de la economía andaluza: conceptos, medida y guía de aplicación" Contabilidad Regional y Tablas Input.-Output de Andalucía 1990. IEA, 143-269 and García Lizana, A. et al (1996): El impacto de los aeropuertos sobre el desarrollo económico. Civitas, 29-57.

- Induced impacts: are the effects generated by the multiplying effect of the direct and indirect impacts on the regional economy.

Therefore, since in this paper we only compare the results of the Port Industry, we have taken as such the direct effect plus the direct induced and the part of the indirect effect plus the indirect induced relative to the fishing sector.

5. Ceuta Port Authority. This study was monitored, like Bahía de Algeciras' and Seville's studies, by professors Castillo Manzano and Lebón Fernández and coordinated by professor López Valpuesta. The study was carried out for the year 1996 and the results obtained referred to the impact on the economy of the Autonomous Community of Ceuta. In this study, the Input-Output methodology adjusted to the Spanish Port System was used by the TEMA Consulting Company.

   The definition of Port Industry includes the fishing activity, although in the year considered in the study, this sector accounted for little significant volume.
6. Santa Cruz de Tenerife Port Authority. This study was carried out by a group of professors from he University of La Laguna monitored by professor Martínez Budría. The geographical area subject to study was the province of Santa Cruz de Tenerife, which is composed by four islands (Tenerife, La Palma, La Gomera and El Hierro), and five Ports of general interest (Santa Cruz de Tenerife, Los Cristianos, San Sebastián de La Gomera, La Estaca and Santa Cruz de la Palma). The methodology employed presents the port as a sector receiving all the expenses which are a direct consequence of the docking of vessels at the ports. Once the direct effect was obtained, the multiplying effects were quantified by using the Input-Output methodology. However, this research only calculates the direct effects (produced by the expenses which are directly caused by the docking of vessels) and the indirect effects (produced by the inter-industrial expenses of direct activities)[3], rather than the induced effects. The study referred to 1992 and 1993, although for this test we have only taken the data referring to 1993 for being the most recent.

B) North Port Subsystem

1. El Ferrol - San Ciprián, A Coruña, Marín - Pontevedra, Vigo and Vilagarcía Port Authorities. This study, carried out by the TEMA Consulting Company and monitored by professor De la Lastra, measured the impact of the Puertos Gallegos on the economy of the Autonomous Community of Galicia and on the national economy for 1992. It consisted of the application of the methodology proposed for State Ports by the consulting company, this study being the first one which employs it.

   The study includes within the Port Industry those firms linked with the fishing sector, which account for 18.6% of the Industry's Gross Added Value.
2. Santander Port Authority. Two studies about the impact of Santander Port have been carried out. One was monitored by professors Coto Millán and Villaverde Castro, and the other by Coto Millán, Gallego Gómez and Villaverde Castro. Both studies apply the Input-Output methodology proposed by the TEMA Consulting Company and calculate the impact on the Autonomous Community

[3] Both effects are defined under the criterion established by Gardner Pinfold Consulting Economists (1991): Port of Halifax Economic Impact Study, Report.

of Cantabria. The former study refers to 1993 and the latter to 1998. We have chosen the latter because it provides us with the most recent data.
This study includes fishing as a Port Industry activity.

3. Bilbao Port Authority. There are two studies about the economic impact of Bilbao Port. One was carried out by Bilbao Plaza Marítima S.L. about 1993 and calculated this port impact on the Basque Country. This study applied the Input-Output methodology and defined the direct, indirect and induced impacts according to the methodology proposed by TEMA, classifying the firms subject to study into Companies-Port (the equivalent of Port Industry) and Port's Customers located in the Basque Country (the equivalent of Port-dependent Industry).
   The second study was carried out by KPMG Consulting and also calculated the impact on the economy of the Basque Country. The period studied was the year 1999. In order to calculate the economic impact, this study analysed the activity implied by each goods unit which arrives in or leaves the Port, rather than studying the total income of each agent set up in the Port. The study only takes into account the activities concerning the transport of goods, rather than other activities. It developed the multipliers of three sectors (activities related to transport, road transport and railway transport) by using the Input-Output methodology through the Basque Autonomous Community's Input-Output Tables for 1995.

C) East Coast Port Subsystem

1. Barcelona Port Authority. This study was carried out by CONSULTRANS and the *Centro de Estudios Fundación Tomillo* under professor Clavera's advice, from the Universitat Autónoma of Barcelona. The geographical area studied was the Community of Catalonia for the year 1995.
2. Tarragona Port Authority. Like the study about Barcelona Port, this research was carried out by CONSULTRANS and the *Centro de Estudios Fundación Tomillo*, under professor Clavera's advice, about the same geographical area and period.
   Both studies employed the Input-Output methodology but they included the Port sector as a separate activity in the Input-Output Table (ports are normally included in the sector of activities related to transport). The Port sector included in Catalonia's Input-Output Table (the research updates this table from 1987 to 1995) is the addition of the input-output vectors of Barcelona and Tarragona Ports. Finally, this study used the inter-sector macroeconomic model of the economy in order to introduce the dynamics of all agents. The MIDE method (Spanish Inter-sector Macroeconomic) allows us to estimate the evolution over time of the different economic sectors, including Ports, and simulate the effect of different future sceneries.
   According to its authors, this study differs from others made on port economic impact in the definition of the effects. This study defines Port Activity (Initial Effect) as the activity indispensable to bridge the gap between road and shipping transport, it providing the effects on the rest of the economy (Direct, Indirect, Induced Effects). These effects are defined in detail as follows:
   - Initial Effect: includes the Gross Added Value composed mainly by the Wage-earners Pay and the Gross Earned Surplus of the port sector.

- Inter-sectorial Effect: includes the following three effects:
- Direct Effect: effect generated by the expenses derived from the port activity (daily functioning and investment expenses). This effect shall be mainly linked with the activities which supply the port.
- Indirect effect: effect generated by the port intermediate consumptions.
- Induced effect: includes the consumption which the employment created by the economic activity of the above-mentioned effects generates.

3. Castellón Port Authority. This study was carried out by a team of professors from the University Jaume I of Castellón and the University of Valencia monitored by professors Fuertes Eugenio and García Menéndez. The geographical area studied was the Community of Valencia for the 1997 period.
4. Valencia Port Authority. Valencia Port Authority is a state-owned company responsible for the management and administration of the Ports of Valencia, Sagunto and Gandía. A research was carried out which estimated the impact of Valencia and Sagunto Ports on the economy of the Community of Valencia. This study was monitored by professors Fernández Guerrero and García Menéndez for the year 1997.
   The methodology employed for the impact studies of the Ports of the Autonomous Community of Valencia was the Input-Output model through its production, income and employment multipliers. Like in the studies carried out for Barcelona and Tarragona, the Input-Output Table of the Community of Valencia from 1990 to 1997 has been updated by using the RAS (array-type convergence iterative method) technique.

19.4 Comparative Methodology of the Impact Studies

19.4.1 Variables Chosen

The aim of this work is to reach the maximum standardisation in the results in order to facilitate their testing. With this end in view, of all the impact variables studied by the different works, we have only chosen the ones which appear in all of them: Gross Added Value (GAV) and Employment, which, are also very significant and define the port activity. The GAV is estimated at factor cost (GAV fc) in the majority of the studies considered so this is the magnitude taken into account for testing. From the studies which did not included this value, we have taken the following variables:

- GAV at market prices (GAV mp) for the Barcelona and Tarragona Ports.
- GDP (Gross Domestic Product) for the Bilbao Port.
- Income, for the Huelva Port since, according to the authors of the study, this is the equivalent of the GAV fc.
- For the case of the Santander Port, we have taken the GAVfc variable, while in the study that variable is calculated only for the direct effect. In order to calculate the GAVfc of the indirect and induced effects we apply the relationship established between the GAVfc and GAVmp in the study.

Finally, we should highlight that the economic data units are thousands of euros for the GAVfc and jobs in the Employment variable.

19.4.2 Testing by Homogenising the Variables over Time

Given that the economic impact results covered different years, these have been unified by turning the current euros into constant euros of the year 2000 (year of the latest study carried out). With this aim, we have used a price index established from the annual mean of the national consumer price index (CPI) (base 1992).

Once the studies have been homogenised over time, the GAV fc and Employment variables obtained have been divided by the traffic volume moved in the port during the year subject to study (thousands of metric tons) in order to calculate an impact ratio generated (in euros) per ton moved in the port studied[4]. We have also calculated the GAV/Employment ratio in order to favor testing.

Table 19.1 sets forth the total results of GAV fc and Employment, as well as of the port traffic, once the economic impact variables have been homogenised over time. Table 19.2 shows the ratios calculated and table 19.3 presents the statistics calculated for the ratios series.

Table 19.1. Testing under a time period basis. Traffic, GAV fv and Employment

PORT AUTHORITY	TRAFFIC (thousands of tons)	GAV fc (thousands of constant euros)	EMPLOYMENT (jobs)
BAHÍA DE ALGECIRAS*	36,836	356,034.18	10,09.04
SEVILLE	4,492	79,274.55	2,074.53
CEUTA*	3,094	62,407.79	1,558.29
SANTANDER*	4,949	129,248.76	3,114.30
BILBAO	27,056	405,819.29	9,762.00
BAHÍA DE CÁDIZ*	4,007	160,570.20	4,490.00
PUERTOS GALLEGOS*	22,992	232,471.66	6,514.00
BARCELONA	23,293	917,104.84	16,104.00
TARRAGONA	28,705	237,343.37	3,259.00
HUELVA*	15,154	184,896.63	6,583.00
CASTELLÓN	8,382	101,199.94	2,458.00
VALENCIA	18,247	582,743.29	15,651.00
S.C. DE TENERIFE	12,269	167,203.01	4,147.00

Source: own work based on the data of economic impact studies

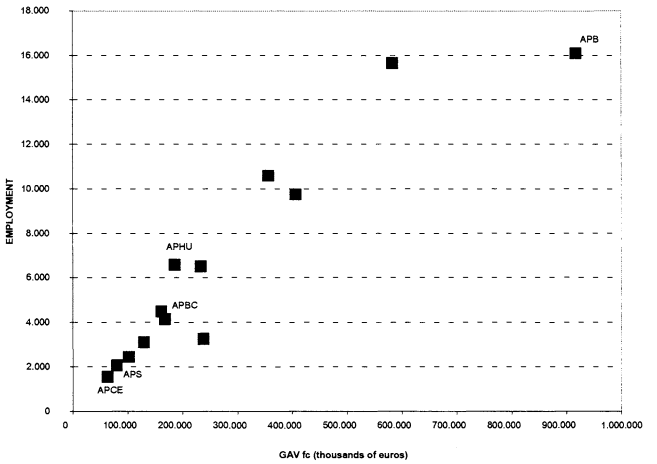
* Fishing sector included

[4] As for the testing of the studies we have only chosen the Port Industry data, it is natural to assume that the results are proportional to the traffics.

Table 19.2. Testing under a time period basis. Impact Ratios

PORT AUTHORITY	GAV/TRAFFIC	EMPLOYMENT/TRAFFIC	GAV/EMPLOYMENT
BAHÍA DE ALGECIRAS*	9.6654	0.2880	33.5595
SEVILLE	17.6479	0.4618	38.2132
CEUTA*	20.1706	0.5036	40.0489
SANTANDER*	26.1161	0.6293	41.5017
BILBAO	14.9992	0.3608	41.5713
BAHÍA DE CÁDIZ*	40.0724	1.1205	35.7617
PUERTOS GALLEGOS*	10.1110	0.2833	35.6880
BARCELONA	39.3726	0.6914	56.9489
TARRAGONA	8.2684	0.1135	72.8271
HUELVA*	12.2012	0.4344	28.0870
CASTELLÓN	12.0735	0.2932	41.1717
VALENCIA	31.9364	0.8577	37.2336
S.C. DE TENERIFE	13.6281	0.3380	40.3190

Source: own work based on the data in table 19.1



* Fishing sector included.

Fig. 19.1[5]. Dispersion chart: testing under a time period basis

[5] See Appendix in order to identify the acronyms in the charts.

Table 19.3. Testing under a time period basis. Statistics describing the impact ratios

	EMPLOYMENT/TRAFFIC	GAV/TRAFFIC	GAV/EMPLOYMENT
Mean	0.490440	19.71252	41.76397
Median	0.434407	14.99923	40.04890
Maximum	1.120539	40.07242	72.82705
Minimum	0.113534	8.268363	28.08699
Std.Dev.	0.273948	11.15817	11.40699
Skewness	0.954277	0.844190	1.740094
Kurtosis	3.267910	2.268660	5.476083
Jarque-Bera	2.011941	1.833804	9.881457
Probability	0.365690	0.399755	0.007149
Sum	6.375716	256.2627	542.9316
Sum Sq. Dev.	0.900570	1494.058	1561.432
Observations	13	13	13

Source: own work based on the data in table 19.2.

By testing over time we obtain the following results:

The highest GAV/traffic ratios are those of the Bahía de Cádiz and Barcelona Ports, with a ratio of 40, and the lowest values are for the Tarragona Port with a value of 8.2 and the Puertos Gallegos and the Port of Bahía de Algeciras with a ratio of approximately 10. The ratio for the rest of the Ports range from 12 for Huelva and Castellón and 32 for Valencia.

As regards the Employment/traffic ratio, no Port reaches the unit, with the exception of the Bahía de Cádiz Port, with a value of 1,12. The lowest value corresponds to the Tarragona Port with a value of 0.11. The Puertos Gallegos, Bahía de Algeciras, Castellón and Santa Cruz de Tenerife have also a low value around 0.3.

Finally, as regards the GAV/Employment ratio, all the ratios of the Ports studied range from 28 to 42, with the exception of the Barcelona and Tarragona Ports, whose ratios are 56.9 and 72.8 respectively. It is difficult to justify the latter ratio if we compare it with the rest of the ports, and especially - by similitude - in comparison with the Barcelona Port, which is very high. Within the interval of values between 28 and 42 the Huelva port stands out, with the lowest value, 28, and the Ceuta, Tenerife, Castellón, Santander and Bilbao Ports with a ratio ranging from 40 to 41.6.

We can assume a normal distribution for the ratios Employment/traffic and GAV/traffic with p-values associated to the Jarque Bera's tests of 0.365690 and 0.399755 respectively and mean and typical deviation parameters as recorded in table 19.3.

19.4.3 Testing by the Homogenisation of the Variables over Time and in the Geographical Area

In this case, in addition to the homogenisation over time, we have made an attempt to unify the geographical area. The economic impact of each port has been studied

on a hinterland of different size and economic importance. In order to unify these results referred to different geographical areas we have calculated indexes drawn from the regional or provincial GDP at market prices (current prices) of the year studied. The data of GDP at current prices has also been estimated at constant prices of the year 2000. Depending on the GDP deviations of each port hinterland with respect to the mean, we have calculated an index which modifies the results of the impact study either up or down. Naturally, this modification only affects the indirect and induced effects since these are calculated through the Input-Output Tables relative to each geographical area subject to study.

The spatial homogenisation index appears on table 19.4:

Table 19.4. Spatial homogenisation index

HINTERLAND	YEAR	GDP mp (current prices) (thousands of euros)	CIP	GDP mp (constant prices 2000) (thousands of euros)	INDEX
ANDALUSIA	1996	62,389,878.00	1.0989	68,559,155.27	0.43
SEVILLE	2000	19,620,943.00	1.0000	19,620,943.00	1.52
CEUTA	1996	668,097.08	1.0989	734,160.29	40.61
CANTABRIA	1998	6,531,353.00	1.0582	6,911,707.98	4.31
BASQUE COUNTRY	1999	35,958,368.00	1.0343	37,193,122.90	0.80
CADIZ	1998	10,400,278.00	1.0582	11,005,940.80	2.71
GALICIA	1992	19,546,596.47	1.3044	25,496,152.95	1.17
CATALONIA	1995	82,752,572.00	1.1380	94,171,801.52	0.32
HUELVA	1996	4,053,667.00	1.0989	4,454,504.39	6.69
VALENCIA	1997	47,233,520.00	1.0776	50,901,120.59	0.59
TENERIFE	1993	7,135,402.02	1.2474	8,900,652.87	3.35

Source: own work based on the data supplied by the INE

In order to apply this index it was necessary that the studies calculated the Direct, Indirect and Induced impact results separately. This has been done in all cases, with the exception of the Bilbao Port, for which we did not have partial results, and the Huelva Port, given that the effects chosen in this study as Port Industry (Direct Effect plus Direct Induced and Indirect Effect, plus the Indirect Induced of the fishing sector) did not allow a similar breakdown to those of the remaining studies.

A mean calculated from the other studies has been applied to these two ports[6]. As regards the Barcelona and Tarragona Ports, the definitions did not coincide

[6] Since the percentages of the Santa Cruz de Tenerife Port are not similar to those of the rest of the ports studied (a logical situation regarding that the study of the island port impact does not calculate the induced effect), they have not been taken into account when calculating the average percentages necessary to separate the direct effects from the other effects.

with those of the remaining studies, thus, the Direct Effect is known here as Initial Effect while the Indirect and Induced Effects are known as Inter-sectorial Effects.

When we analysed the total impact distribution between the Direct Effect and the Indirect Effect plus the Induced, we reached the following conclusions:

- As regards the GAV fc, the majority of the Ports register Direct Effects of 55% (the Bahía de Algeciras, Seville, Ceuta, Barcelona, Tarragona and Castellón ports). The rest of the ports register a higher percentage, around 60%. This is the case of the Santander Port, the Puertos Gallegos, the Bahía de Cádiz and Valencia Ports, with the only exception of the Santa Cruz de Tenerife Port, whose direct effect for the GAV fc is 84%, although this value, which seems very high, may be due to the fat that, in this study, we only quantify direct and indirect effects.
- As far as the Employment variable is concerned, there are more controversies. We have the same distribution as above: 55% as Direct Effect, as shown in Ports such as Ceuta, Valencia or the Puertos Gallegos; between 46% and 48% for the Ports of Seville, Santander, Castellón and Barcelona; and the Bahía de Algeciras and Tarragona are below this value, with a ratio between 35% and 40%. The exception is again the Santa Cruz de Tenerife Port, with 71% of direct effect for the employment variable due to the reasons exposed above.

In view of these data we can point out that the ports with a high direct effect – which are the majority of those studied, mainly for the GAV fc variable – show a lower integration of the port activity in the economy of their hinterland.

Table 19.5. Testing under a time and geographical area basis. Traffic, GAV fc and employment

PORT AUTHORITY	TRAFFIC (thousands of tons)	GAV fc (thousands of constant euros)	EMPLOYMENT (jobs)
BAHÍA DE ALGECIRAS*	36,836	263,710.42	6,983.02
SEVILLE	4,492	97,712.78	2,647.27
CEUTA*	3,094	1,182,947.58	28,747.54
SANTANDER*	4,949	300,311.23	8,559.35
BILBAO	27,056	371,135.68	8,771.75
BAHÍA DE CÁDIZ*	4,007	265,217.89	7,706.08
PUERTOS GALLEGOS	22,992	248,140.19	7,009.65
BARCELONA	23,293	623,145.40	10,348.30
TARRAGONA	28,705	163,495.82	1,817.68
HUELVA*	15,154	638,302.47	25,743.01
CASTELLÓN	8,382	82,570.23	1,923.16
VALENCIA	18,247	486,145.18	12,770.10
S.C. DE TENERIFE	12,269	231,588.87	6,907.77

Source: own work based on the data of economic impact studies

* Fishing included

Table 19.5 and table 19.6 set forth the total variables homogenised over time and in the geographical area, as well as the respective ratios calculated for the traffic. Finally, table 19.7 shows the statistics calculated on the ratios.

Table 19.6. Testing under a time and geographical area basis. Impact ratios

PORT AUTHORITY	GAV/TRAFFIC	EMPLOYMENT/TRAFFIC	GAV/EMPLOYMENT
BAHÍA DE ALGECIRAS*	7.1590	0.1896	37.7645
SEVILLE	21.7526	0.5893	36.9107
CEUTA*	382.3360	9.2914	41.1495
SANTANDER*	60.6812	1.7295	35.0857
BILBAO	13.7173	0.3242	42.3103
BAHÍA DE CÁDIZ*	66.1886	1.9232	34.4167
PUERTOS GALLEGOS*	10.7925	0.3049	35.3998
BARCELONA	26.7525	0.4443	60.2172
TARRAGONA	5.6957	0.0633	89.9475
HUELVA*	42.1211	1.6988	24.7952
CASTELLÓN	9.8509	0.2294	42.9347
VALENCIA	26.6425	0.6998	38.0690
S.C. DE TENERIFE	18.8759	0.5630	33.5258

Source: own work based on the data in table 19.5

* Fishing sector included

Table 19.7. Testing under a time and geographical area basis. Statistics describing impact ratios

	EMPLOYMENT/TRAFFIC	GAV/TRAFFIC	GAV/EMPLOYMENT
Mean	0.729942	25.85249	42.50206
Median	0.503646	20.31428	37.76454
Maximum	1.923154	66.18864	89.94754
Minimum	0.063323	5.695726	24.79517
Std.Dev.	0.662217	20.35477	16.34328
Skewness	0.932870	0.988011	2.074796
Kurtosis	2.227507	2.667525	6.684190
Jarque-Bera	2.038864	2.007600	16.67920
Probability	0.360800	0.366484	0.000239
Sum	8.759309	310.2298	552.5268
Sum Sq. Dev.	4.823842	4557.483	3205.235
Observations	12	12	13

Source: own work based on the data in table 19.6

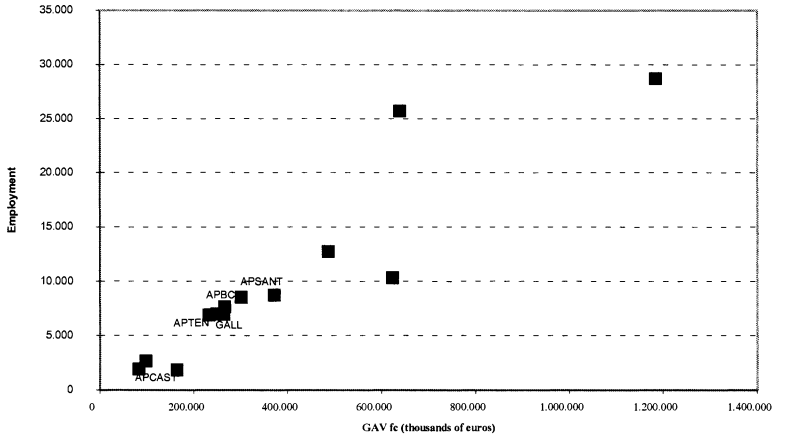


Fig. 19.2. Dispersion chart: testing under a time and geographical area basis

From the testing, by homogenising the time and taking into account the disagreement about the data shown on table 19.1, given by the economic activity of the hinterland, we obtain the following results:

- Once the results in relation with the hinterland have been standardised, we observe that the data relative to the Ceuta Port do not have any relation to the rest of the Ports, thus becoming outliers with respect to the statistics calculated in table 19.7. The economic justification is quite intuitive if we consider the absolute dependency of the city of Ceuta on its port, it being the only point of entry for goods into the city since no airport or commercial frontier with Morocco allows the city to be supplied from the south. Moreover, this high economic activity may be an indicator of the importance of the black market and smuggling to the southern area of some of the goods which come through the Ceuta port. Therefore, when calculating the statistics with respect to the GAV/traffic and Employment/traffic ratios, we have decreased population from 13 to 12 elements (see fig. 19.3).
- 69% of the Ports have a GAV/traffic ratio ranging from 6 to 27. This mean is exceeded by the Huelva Port, with a value of 42.1 and the Santander and Bahía de Cádiz Ports (with 60.7 and 66.2 respectively). Finally, as already stated, the Ceuta Port very much exceeds these with a value of 382.3.
- The same percentage of Ports has an Employment/traffic ratio which does not exceed the unit. Thus, the Huelva, Santander and Bahía de Cádiz ports have ratios between 1.7 and 2 and the Ceuta Port has a ratio of 9.3, a value which very much exceeds the mean.
- With respect to the GAV/Employment ratio we observe intervals which range from 25 to 43, with the exception of the Catalonian Ports of Barcelona and Tarragona, whose ratios are 60.2 and 89.9 respectively.
- Again, we can assume a normal distribution for the Employment/traffic and GAV/traffic ratios having p-values associated with Jarque Bera's tests of

0.360800 and 0.366484 respectively, and mean and typical deviation parameters as shown in table 19.7.

- An interesting data which is again repeated is how the variation (Std. Dev.) with respect to the mean in the GAV/Employment ratio is considerably lower than the rest of the variables. This fact facilitates the formulation of general conclusions with respect to this ratio.

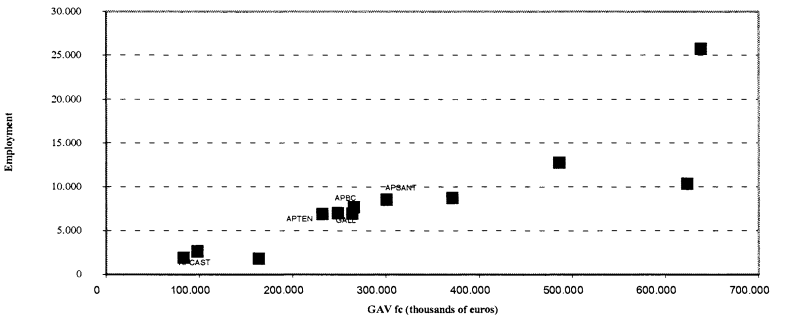


Fig. 19.3. Dispersion chart: testing under a time and geographical area basis (without outliers)

19.5 Conclusions

1. As far as the testing of the variables homogenised over time is concerned:
   - The Tarragona Port obtains lower GAV/traffic and Employment/traffic ratios with respect to the Bahía de Cádiz Port, whose ratios are higher with respect to traffic.
   - The Bahía de Algeciras Port, Puertos Gallegos and Castellón Port, have - after the Tarragona Port - the lowest GAV/traffic and Employment/traffic ratios of all the Ports considered.
   - The highest GAV/Employment ratios correspond to the Barcelona and Tarragona Ports, while the lowest ones are those of Huelva and Bahía de Algeciras.
2. With respect to the testing of the variables homogenised over time and the hinterland considered:
   - The Tarragona Port has again the lowest GAV/traffic and Employment/traffic ratios, while the Ceuta Port has the highest ratios with respect to traffic.
   - As with the testing based upon the time period, the Bahía de Algeciras, Castellón ports and the Puertos Gallegos, in addition to the Bilbao Port - in this order - have, after the Tarragona Port, the lowest GAV and Employment ratios with respect to traffic.
   - On the other hand, the highest GAV/traffic and Employment/traffic ratios are, after Ceuta, those of the Bahía de Cádiz, Santander and Huelva Ports.

- We have also that the highest GAV/Employment ratios correspond to the Ports of Barcelona and Tarragona, while the lowest ones are those of Huelva and Tenerife.

Appendix

Table 19.8. Acronyms used in the charts

APBA	BAHÍA DE ALGECIRAS PORT AUTHORITY
APS	SEVILLE PORT AUTHORITY
APCE	CEUTA PORT AUTHORITY
APSANT	SANTANDER PORT AUTHORITY
APBI	BILBAO PORT AUTHORITY
APBC	BAHÍA DE CÁDIZ PORT AUTHORITY
GALL	PUERTOS GALLEGOS
APB	BARCELONA PORT AUTHORITY
APT	TARRAGONA PORT AUTHORITY
APHU	HUELVA PORT AUTHORITY
APCAST	CASTELLÓN PORT AUTHORITY
APV	VALENCIA PORT AUTHORITY
APTEN	SC DE TENERIFE PORT AUTHORITY

References

Economic Impact Studies

Bilbao Plaza Marítima, S. L.: Estudio del impacto económico del Puerto de Bilbao en el País Vasco (1993). Department of Transports and Public Works of the Basque Government 1995

Castillo Manzano, J. I. (coord.): El Puerto Bahía de Algeciras, el motor económico del Sur. Ministry of Public Works and Bahía de Algeciras' Port Authority 2001

Castillo Manzano, J. I., López Valpuesta, L. and Castro Nuño, M.: El Puerto de Ceuta. Una pieza clave en la economía de la Ciudad Autónoma. Editorial Civitas 2000

Consultrans; Fundación Tomillo Economic Studies Centre: Análisis de impacto económico de los Puertos de Barcelona y Tarragona. Puertos del Estado 1998

Coto Millán, P.; Gallego Gómez, J. L. and Villaverde Castro, J.: Crecimiento portuario y desarrollo regional. Una aplicación al Puerto de Santander. Santander's Port Authority 2001

De Rus, G., Roman, C. and Trujillo, L.: Actividad económica y estructura de costes del Puerto de La Luz y de las Palmas. Civitas Publishing Company 1994

ETT S.A.: Accesibilidad Terrestre en las Instalaciones Portuaria de interés general de las fachadas Marítimas Mediterránea y Suratlántica. General Management of State Roads and Ports. Summary included in the journal Puertos 77, 78, 79 and 80

Fernández, J. I.; García, L.; Huet, F., Goerlich, F. and Pallardó, V.: El impacto económico del Puerto de Valencia. (data provided by the authors)

Fraga Sardiña, J. and Seijas Macías, J. A.: El Puerto de Ferrol y su influencia en la economía de la comarca. Port Board and Ría de Ferrol 1992

Fuertes, A., García, L., Fernández, J. I.., Cuadros, A. and Huet, F.: Estudio del Impacto económico del Puerto de Castellón. Bulletin of State Ports monthly Information 80, 3-8 (2000) (this information has been complemented with data of the research provided by the authors)

García del Hoyo, J. J., González Galán, D., García Ordaz, F. and De Paz Báñez, M.: Estimación de los Efectos Económicos derivados de la Actividad del Puerto de Huelva. Huelva's Port Authority, El Monte Foundation, University of Huelva 1999

KPMG Consulting: El Impacto Económico del Puerto de Bilbao. Puertos 82, 4-8 (2000)

Lebón Fernández, C., Castillo Manzano, J. I. and López Valpuesta, L.: El impacto económico del Puerto de Sevilla sobre la economía andaluza. Civitas Publishing Company 1998

López Valpuesta, L. and Castillo Manzano, J. I.: Análisis de la actividad económica del Puerto de Sevilla y su influencia provincial. Publishing Department of the University of Seville 2001

Martínez Budría, E., Gutiérrez Hernández, P., López Martín, L. J. and Martín Álvarez, F.: El impacto económico de los puertos de Santa Cruz de Tenerife sobre la provincia. Hacienda Pública Española 148, I, 175-185 (1999)

Rey Juliá, J. M.: Evaluación del impacto económico del Puerto de la Bahía de Cádiz. Puertos 100, 19-22 (2002)

TEMA consultora: Elaboración de una metodología para la evaluación de los impactos de la actividad portuaria sobre la economía. State Ports 1994

TEMA consultora: Evaluación de los Impactos de la Actividad de los Puertos de Galicia sobre la Economía de la Región. State Ports 1994

TEMA consultora: Evaluación de los Impactos de la Actividad de los Puertos de Galicia sobre la Economía Nacional. State Ports 1995

Villaverde Castro, J. and Coto Millán, P.: Análisis de impacto económico portuario: una aplicación al Puerto de Santander. Santander's Port Authority 1996

Additional References

De Salvo, J. S.: Measuring the Direct Impacts of a Port. Transportation Journal, Summer, 33-42 (1994)

State Ports Public Entity: Anuarios Estadísticos (1992-2000)

INE: Contabilidad Regional de España (series 1991-1996 and 1995-2001)

Verbeeke, A. and Debisschop, K.: A note on the use of port economic impact studies for the evaluation of large scale port projects. Internacional Journal of Transport Economics, Vol. XXIII 3, 247-266 (1996)

20 Economic Impact of Santander Airport

G. Carrera-Gómez
University of Cantabria (Spain)

P. Coto-Millán
University of Cantabria (Spain)

R. Sainz-González
University of Cantabria (Spain)

V. Inglada-López de Sabando
University Carlos III (Spain)

20.1 Introduction

In three previous works by Coto-Millán and Villaverde-Castro (1995, 1996) and Coto-Millán *et al.* (2001) research on the port of Santander was carried out in a similar way to the present study. Here we have followed the methodology used in the last work, considering it to be the most suitable for the case of Santander, and the most appropriate methodology today. The category of "agreed methodology" is mainly based upon the wide range of studies (we have reference of over two hundred studies) which employ it, and upon its use by the majority of the Spanish airports, a fact which, saving the logical differences, will allow us to make comparisons between them. The data employed in this research refers to 1998 because the last information obtained from the surveys of different partnerships and the Register of Company Reports, belongs to this period.

P. Coto-Millán (ed.), *Essays on Microeconomics and Industrial Organisation*

Airport agents are presumably grouped into two main sections: Airport Industry and Airport-dependent Industry. Airport Industry includes: Santander Airport (AENA), Customs and the Rest of the Airport Industry, in other words, those economic agents whose operations are necessary to carry out the loading and unloading of goods and passengers.

The Airport-dependent Industry includes the tourist expenses of the users of Santander Airport, the induced wages and salaries at the Travel Agencies from the operations carried out by Santander Airport, the savings generated by companies for the use by their employees of Santander Airport rather than Bilbao Airport, and the induced wages and salaries from the costs of the journeys by taxi from Airport to the city, and vice versa. These are the companies of the regional economy which are linked with the airport activity by a dependency relationship. The variables to be considered in order to estimate the economic impact of Santander Airport are: number of employees, sales, wages, salaries, gross operating surplus (GOS), paid taxes and gross added value (GAV). With the aim of estimating the economic impact for 1998, we have followed the steps below in our research:

Firstly, we have estimated the magnitude of each relevant variable for every agent of the Airport Industry: jobs, sales, wages and salaries, GOS, taxes and GAV.

Secondly, we have regionalised and updated the last input-output table available.

Thirdly, we have calculated the indirect impact vector by areas of activity from the sectorisation, by areas of activity, of the purchases and investments of Santander Airport (AENA) and of the Rest of the Airport Industry.

Fourthly, from the 1994 indirect impact vector in pesetas, we have calculated the indirect impacts of the Airport Industry by multiplying this by the product of the Technical Coefficients Matrix of the GAV of 1994 and the Regional Inverse Matrix of 1994, and the result, by a deflator vector which enables us to pass to pesetas of 1998. Once we have obtained this peseta vector for the year 1998, we have multiplied it by the Index matrix in relation with the GAV of jobs, wages and salaries, GOS, taxes and sales. Thus, we have finally obtained the indirect effect impact in terms of the magnitudes jobs, sales, wages and salaries, GOS, taxes and GAV.

Fifthly, we have calculated the induced impact vector from the sectorisation by areas of activity, of direct and indirect wages and salaries.

Sixthly, from the induced impact vector in pesetas of 1994, we have calculated the induced impacts of the Airport Industry by multiplying this by the product of the Technical Coefficients Matrix of the GAV of 1998 and the Regional Inverse Matrix of 1994, and the result obtained has been multiplied by a deflator vector which allows us to pass from pesetas of 1994 to pesetas of 1998. Once this peseta vector for 1998 has been obtained, we have multiplied it by the Index matrix in relation with the GAV of jobs, wages and salaries, GOS, taxes and sales. Then, we have finally obtained the induced effect impact in terms of the magnitudes jobs, sales, wages and salaries, GOS, taxes and GAV.

Seventhly, we have added the direct, indirect and induced effects of Airport Industry and presented a summary of the total effects of this Industry.

Eighthly, we have estimated, for the Airport-dependent Industry, the magnitude of each relevant variable from the conjectures explained.

Ninthly, we have calculated the induced impact vector of the Airport-dependent Industry by the sectorisation, by areas of activity, of the direct and indirect salaries of the Airport-dependent Industry.

Tenthly, we have calculated the induced impacts of the Airport-dependent Industry from the induced impact vector, by multiplying this by the product of the Technical Coefficients Matrix of the GAV of 1994 and the Regional Inverse Matrix of 1994, and the result obtained has been multiplied by a deflator vector which enables us to pass from pesetas for 1998. Once this peseta vector for 1998 has been obtained, we have multiplied it by the Index matrix in relation with the GAV of jobs, wages and salaries, GOS, taxes and GAV.

Eleventh, we have added the direct, indirect and induced effects of the Industry Dependent on de Airport and presented a summary of the total effects of this industry.

Twelfth, we have added the direct, indirect and induced effects of the Airport Industry and the direct, indirect and induced effects of the Airport-dependent Industry, to obtain the total effects of the Airport on the economy of Cantabria.

To conclude with, these effects have been discussed in terms relative to the economy of Cantabria.

20.2 Direct Effects of the Airport Industry

20.2.1 Santander Airport. AENA

The direct effects of Santander Airport (AENA) have been obtained from the Annual Report of 1998. Table 20.1 shows these effects for the magnitudes covered by this impact research.

Table 20.1. Santander Airport. AENA. Direct effects 1998

20.2.2 Customs and Administration Services

Data on table 20.2 have been obtained from the Customs and Aministration Services, considering for the jobs and income an estimation of the work developed by the different officers who do not exclusively carry out activities in connection with the Airport.

20.2.3 Rest of Airport Industry

This section deals with the following agents:

- Air Companies
- Handling Agents
- Commercial Concessions
- Security and Cleaning

The data have been summarised in table 20.3 and have been mainly obtained from surveys as well as from the Register of Companies' account deposits and from other indirect information from the Chamber of Commerce.

Table 20.3. Rest of airport industry. Direct effects 1998

20.2.4 Total Airport Industry

Table 20.4. Total airport industry. Direct effects 1998

By adding tables 20.1, 20.2 and 20.3 we have obtained table 20.4, which specifies the direct effects of airport industry.

20.3
Indirect and Induced Effects of Airport Industry

These effects have been calculated by using the information provided by the national input-output table, once regionalised and updated for the year 1998. The results are set forth on tables 20.5, 20.6 and 20.7.

Table 20.5. National matrix of intermediate inputs R16

20.3.1 Regionalisation of the National Input-Output Table

Since, as it is known, there is not an Input-Output Table for Cantabria, we have had to regionalise the last Input-Output Table of the Spanish Economy for 1994 in Cantabria and, then, update it for 1998. In order to regionalise this table, it is necessary to calculate the data of the Spanish Regional Accounting relative to Cantabria. The Spanish Regional Accounting classifies the data according to 17

areas of activity, while the Input-Output Table of the Spanish Economy for 1994 provides information for 57 areas of activity. For this reason, it is necessary to reduce the National Table to the 17 areas of activity and, in order to homogenise the data, join area 17 (production attributed to banking services) and area 14 (credit and insurance services), to have only area 14 (credit and insurance services), so that we have a table with the 16 areas of activity below:

1. Agriculture, forestry and fishing
2. Energy
3. Mineral and metal goods and their by-products
4. Chemical products
5. Metal goods, machinery and electric material
6. Transport material
7. Food, drinks and tobacco
8. Textiles, leather, shoes and clothes
9. Paper, printed goods
10. Miscellaneous industries
11. Building and civil engineering works
12. Restoration and repairing services. Retailing, hotels and catering
13. Transport and Communication Services
14. Credit and Insurance Services
15. Other sales
16. General administration services, teaching and researching, health services, domestic services and other services not-implying sales.

Once the National Table has been reduced to these 16 areas of activity we have added the GAV and Total Resources row in order to complete table 20.5. Table 20.6 presents the technical coefficients matrix of this reduced table elaborated by dividing each value of the transaction matrix by the Total Resources of each area of activity. Once we have made the calculations, we have proceeded to regionalise the national Table by following the location coefficients method. We have calculated the location coefficients of the GAV at market prices for each area of activity. These coefficients are provided on table 20.7. With them, we have made a matrix whose main diagonal is formed by the coefficients and the remaining elements are zero. This matrix has been multiplied by the national technical coefficients matrix on table 20.6 in order to obtain the regionalised technical coefficients matrix for Cantabria appearing on table 20.8. From this matrix we have obtained the Leontief regional inverse matrix set forth on table 20.9.

20.3.2 Indirect and Induced Impact Vectors

The indirect effects of Airport Industry have been obtained by considering the generation of jobs and the added value of this industry's main suppliers. In order to get this information we must have the Airport Industry's purchases and investments in connection with the rest of the areas of activity in the economy. However, since this information is not available by areas of activity, we have therefore proceeded to the sectorial disaggregation of the purchases of those

sectors concerning air transport, as set out on table 20.10. Something similar to purchases happens with the investment disaggregation, such that, we have carried out the operation in the same way in order to obtain table 20.11. Moreover, with the aim of obtaining the induced effects, we must have the private consumption expenses of the agents referred to by the direct and indirect effects. In order to obtain such expenses, we have proceeded to the internal private consumption sectorial disaggregation on table 20.12.

Table 20.10. Sectorial disaggregation of the purchases of sectors connected with air transport

Table 20.11. Continued

Table 20.14. Purchases and investments. Airport industry by sectors

20.3.3 Total Airport Industry

Once we have obtained the indirect and induced impact vectors, it is possible to calculate the indirect and induced effects. With this aim, we have followed the steps below:

Firstly, we have established a matrix whose main diagonal's elements are the added value coefficients obtained from the national technical coefficients matrix and the rest of cases have zero value.

Secondly, we have multiplied the above-obtained matrix by the Leontief regional inverse matrix and the indirect impact vector (or induced, as applicable), to obtain an added value vector in pesetas for 1994, to which we have applied the corresponding deflator to finally obtain pesetas for 1998.

Thirdly, taking into account the information of the Spanish Regional Accounting for Cantabria, we have calculated the wages and salaries, jobs, taxes and GOS ratios for each area of activity.

Fourthly, by multiplying the added value vector at input costs - deflated into pesetas for 1998 - by each ratio vector obtained in the previous step, we have the indirect impact values (or induced, as applicable) for the various economic magnitudes applicable in this study. The indirect and induced effects appear on tables 20.16 and 20.17.

Table 20.16. Indirect effects of airport industry

Table 20.17. Induced effects of airport industry

20.5 Effects of the Airport-Dependent Industry

The Airport-dependent Industry is that industry which would disappear if Santander Airport did not exist. It is known that the tourists who use the airport run into some expenses when they come to the region which would not be incurred if these tourists did not use the airport. We have estimated through surveys that 25% of the passengers in July and August of 1998 (42,885

passengers) used the airport for their holidays; 12% of them stayed about 7 days, 11% about 5 days and 2% about 2 days, having spent an average of 10,000 pesetas per day. With these data, the total expenses of these tourists who use the region's airport (which imply returns for the region) obtained is 2,095.6866 million pesetas.

In addition, some of the jobs in the Travel Agencies depend to some extent on Santander Airport. These effects can be estimated by considering the part of wages and salaries of these jobs aimed at consumption in the region. Only 20% of these salaries depending on Santander Airport have been considered, yielding 112. 5 million pesetas.

Another effect are the savings generated by the companies due to the existence of Santander Airport. With regard to this, the results of the surveys reveal that 60% of the journeys are for business, producing average savings of 25,000 pesetas, a result obtained by adding 15,000 pesetas for one additional hour of work of the executive and 12,000 pesetas of the return ticket by taxi or on private car, less 2,000 pesetas given that the fare is 10% less than the equivalent in Bilbao for a 20,000 peseta journey, all of which implies a total saving of 3,419.46 million pesetas for 1998.

To conclude with, the Airport generates wages and salaries in the taxis sector which would not be given otherwise. We have estimated through the surveys that 45% of the passengers use taxis and the average journey is 2,000 pesetas, thus having a total of 205.167 million pesetas of wages and salaries induced by Santander Airport.

Therefore, the total of savings and income induced by the airport-dependent industry are estimated to be some 5,832.8136 million pesetas, an amount which can be attributed to wages and salaries which generate induced effects in the region. These induced effects are calculated in a similar way to those induced effects calculated for the Airport Industry.

20.6
Induced Effects of the Airport-Dependent Industry

In addition, on table 20.19 we can observe the above-defined induced impact effects of the Airport-dependent industry

Table 20.19. Induced effects of the airport-dependent industry (1998)

Table 20.19. Continued

20.8 Summary and Conclusions

To summarise, Santander Airport generated 1,179 jobs in 1998, which constitutes 0.69% of regional employment with respect to the 171,900 jobs generated in Cantabria for that year. Moreover, Santander Airport generated 5,746 millions of GAV in 1998, which constitutes 0.56% of the regional GAV of Cantabria with respect to the 1,132,786 millions of GAV generated in Cantabria for that year.

References

Coto-Millán, P. and Villaverde-Castro, J.: El impacto económico del Puerto de Santander en la economía cántabra. Autoridad Portuaria de Santander 1995

Coto-Millán, P. and Villaverde-Castro, J.: Impacto económico portuario: Metodologías para su análisis y aplicación al Puerto de Santander. Ed. APS 1996

Coto-Millán, P., Gallego-Gómez, J. L. and Villaverde Castro, J.: Crecimiento y Desarrollo Portuario. Aplicación al Puerto de Santander. Autoridad Portuaria de Santander 2001

Other References

Calvo García-Tornel, F. and Morales Gil, A.: Potencial de Captación y Generación de Tráfico del Aeropuerto de Alicante. Editorial Civitas-AENA 1998

De Rus Mendoza, G., Trujillo Castellano, L., Román García, C. and Alonso Sosa, P.: Impacto Económico del Aeropuerto de Gran Canaria. Editorial Civitas-AENA 1996

García Linaza, A., Martín Reyes, G. and Otero Moreno, J. M.: El Impacto de los Aeropuertos sobre el Desarrollo Económico. Métodos de Análisis y Aplicación al Caso del Aeropuerto de Málaga. Editorial Civitas-AENA 1996

García Montalvo, J. and Pérez García, F.: Metodología y Medición del Impacto Económico de los Aeropuertos: el Caso del Aeropuerto de Valencia. Editorial Civitas-AENA 1996

Gutiérrez Hernández, P., López Martín, L. J. and Navarro Ibáñez, M.: Impacto Económico de los Aeropuertos de Tenerife en su Entorno. Análisis de los Pasajeros de los Aeropuertos de la Isla de Tenerife. Editorial Civitas-AENA 1999

Ministerio de Obras Públicas, Transportes y M. A.: Plan Director de Infraestructuras. 1993-2007, 2ª ed. MOPTMA 1994

Potryokowski, M. and Taylor, Z.: Geografía del Transporte. Ariel 1984

Robusté, F. and Clavera, J.: Impacto Económico del Aeropuerto de Barcelona. Editorial Civitas-AENA 1997

Valdés, L. and Ruiz, A. V. (coords.): Turismo y Promoción de Destinos Turísticos: Implicaciones Empresariales. University of Oviedo, Servicio de Publicaciones 1996

Villaverde Castro, J. and Coto Millán, P.: Port Economic Impact: methodologies and appliation to the Port of Santander. International Journal of Transport Economics. Special Issue: Infrastructure Investment and Development Vol. XXV-N. 2, 159-179 (1998)

Villaverde Castro, J. and Coto Millán, P.: Guest Editor's Introduction. International Journal of Transport Economics. Special Issue: Infrastructure Investment and Development Vol. XXV-N. 2, 109-112 (1998)

21 Dynamic Adjustments in a Two-Sector Model

F. Galera
University of Navarra (Spain)

P. Coto-Millán
University of Cantabria (Spain)

This paper presents a simple dynamic model in which only one resource is used entirely in order to produce two goods. Technology displays constant returns in both cases. The resource is owned by a few individuals who must choose between the two activities in which to use it. Here we analyse the decisions of change of activity in which the resource is to be used taking into account the different profitability per time unit of the activities directed to produce both goods. We show that, under certain conditions, it is possible to reach cyclic, or chaotic, dynamics so that the returns per factor may never be equal in both sectors.

The basic instrument for the analysis is a difference equation. It is common knowledge that the behaviour of this type of equation essentially depends on the values taken by certain parameters, and that, using simple assumptions, the solutions may behave in a cyclic or chaotic manner. This fact has led to applications in economic models and justifies apparently random behaviours in an analytical manner. See, for example, the mentioned papers by Kelsey, Lichtenberg A.H. & Ujihara, A. and Dwyer, G. P. Jr.

This paper is framed within this type of literature but its main contribution is not only to illustrate a possible example of chaotic behaviour but also, and more importantly, to state that this behaviour is more likely in certain conditions than in others.

21.1 The Formal Model

Let us assume an economy constituted by N equal individuals in which each one of these individuals owns 1/N amount of a productive resource – either labour or

P. Coto-Millán (ed.), *Essays on Microeconomics and Industrial Organisation*

any other type – at each time period. The total amount of the resource is 1. This resource may be used to produce two types of goods called C and D. The activity dedicated to produce C is named the first sector while the remaining activity is the second sector. At each time unit, the individuals decide whether they use their resource in goods C or D. We let x_t represent the amount of the resource used in the first sector during the period t, which coincides hypothetically with the amount of individuals who dedicate the resource to that sector. $1-x_t$ will be the amount of resource dedicated to the production of D.

The returns in the production of C and D are constant so that the amount produced of C will be $d(1-x_t)$. From here we will assume that d=1. There will be one price for each good, taking the price of goods D as numerary, the expression used will be

$$p_t = p_t^C$$

$$p_t^D = 1$$

In order to establish the demand functions for each goods, we will assume that all the individuals have the same tastes and these are expressed in a utility function such as the Cobb-Douglas function:

$$U = a \ln C_t^n + (1 - a) \ln D_t^n \tag{21.1}$$

where C_t^n is the amount demanded for goods C during the period t by the n individual and D_t^n has an analogous meaning. With this assumption, the demand for each one of the economic agents are given by the following equations:

$$C_t^n = a/p_t \cdot I_t^n$$

$$D_t^n = (1 - a) \cdot I_t^n \tag{21.2}$$

where I_t^n is the individuals' incomeduring the period t.

In this economy, the individuals' income available comes entirely from their resource income. With this assumption, we have the following equation:

$$\sum_{n=1}^{N} I_t^n = p_t c x_t + 1 - x_t \tag{21.3}$$

Therefore, the aggregated demand of goods C in the period t will be:

$$C_t = \sum_{n=1}^{N} C_t^n = \sum_{n=1}^{N} \frac{a}{p_t} I_t^n = \frac{a}{p_t}(p_t c x_t + 1 - x_t) \tag{21.4}$$

The demand for goods D will be similarly obtained. By equating the aggregated demand and supply for each goods, we will have the following equation:

$$c x_t = \frac{a}{p_t}(p_t c x_t + 1 - x_t), \quad 1 - x_t = (1-a)(p_t c x_t + 1 - x_t) \tag{21.5}$$

From any of these equations we can reach the equilibrium price:

$$p_t = \frac{a(1-x_t)}{(1-a)c x_t} \tag{21.6}$$

In fact, by substituting p_t in the above equation, we can observe that this is the equilibrium price.
Taking into account the fact that workers assign the amount of factor according to their own individual interest, to avoid a transfer of resources from one sector to another, it must hold that the returns per time unit are the same in both activities.

$$p_t c = 1 \tag{21.7}$$

In other words, if an individual employs his resources in the first sector, he will get the same income – per time unit – as in the second sector.

When this condition is not given, the factor will be re-assigned with the corresponding changes in prices. We will assume that the number of agents who want to change the assignment of the factor from the second to the first sector is proportional to the difference – or quotient, which is the same in this case since the returns value of the second factor is 1 - between the returns of both activities:

$$\Delta x_t = r\,(p_t c - 1) \tag{21.8}$$

From this, and from the equilibrium prices in equation (21.6), we obtain the dynamic equations of the model we are studying:

$$x_{t+1} = x_t + r\frac{a - x_t}{(1-a)x_t} \tag{21.9}$$

We are basically interested in the stability conditions as well as in the possible evolutions of the variable x_t, which represents, as already said, both the amount of resource and the proportion of individuals who assign their resource to the first sector. Now, we must solve the recurrence defined in equation (21.9). To see in a simple manner how this type of recurrences are studied, see chapter 11 in Peitgen et al.

21.2 Basic Results

To make the reading of this exposure less complex, we will summarise the essential results and technical data and will include them in the annex in a simplified form.

We would like to answer some questions about the equilibrium stability of the equilibrium, the possibility of chaotic dynamics and the stability of the system itself. In particular:

I) What conditions must be imposed to parameters "r" and "a" in order to reach a stable equilibrium? How do those restrictions depend on parameter "a", in other words, on the relative size of the first sector?
II) Under what conditions may a chaotic dynamics appear? Do those conditions depend on parameter "a"?
III) What values of "r" and "a" guarantee that a negative number of producers does not appear in the second sector? What about the first sector? What conditions provoke the disappearance of a sector in a particular period?

These questions are answered as follows:

I) The system is stable if $r < U(a) = 2a(1-a)$.
II) The system enters in the so-called chaotic zone when $r \geq Y(a) \approx 2.63626a(1-a)$.
III) A negative number of producers does not appear in the first sector when

$$V(a) = \begin{cases} 4a\,(1-a) & \text{if } a \leq \frac{1}{2} \\ 1 & \text{if } a > \frac{1}{2} \end{cases}$$

A negative number of producers does not appear in the second sector if $r \leq W(a)$, as defined in the annex.

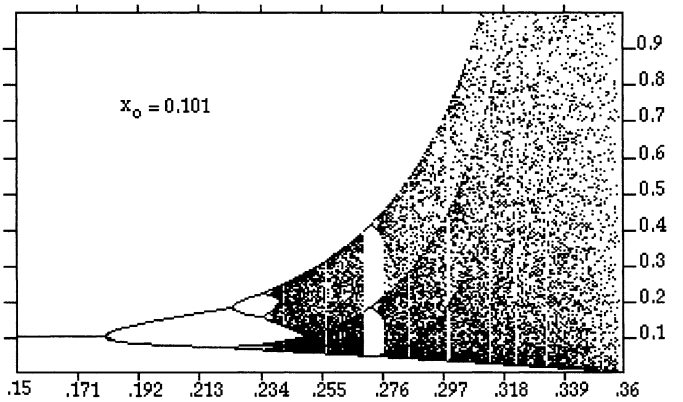


Fig. 21.1. Bifurcation diagram of x with respect to r with a=0.1

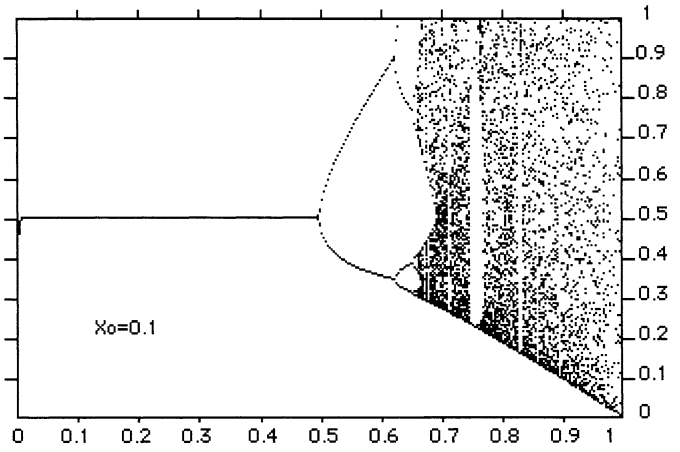


Fig. 21.2. Bifurcation diagram of x with respect to r with a=0.5

Figures 21.1 and 21.2 present the bifurcation diagrams of the variable x with respect to the parameter r with a=0.1 and a=0.5 values. For those who are not familiarised with these diagrams, an interpretation of them can be seen in the works by Peitgen et al, Collet & Eckmann, Devaney or in any other work dealing with chaos from an elementary point of view.

21.3 Conclusions

The dynamic possibilities of the model essentially depend on parameters "r" and "a". As indicated, parameter r measures the response of the agents to the differences in the profits. The higher "r", the greater the amount of individuals and resources moving form one sector to another and, initially, that mobility should make the system more stable since any desequilibrium would be more rapidly corrected. However, we can observe that, if the value of r exceeds a certain limit, the only equilibrium presented by the dynamic system becomes unstable, which generates a cyclic, or even chaotic, behaviour in some cases. If the value of r becomes even higher, any of the sectors may disappear from the economy.

Moreover, parameter "a" basically measures the individuals' tastes. The higher "a", the greater the relative weight of the first sector within the economy. This relative weight affects the dynamic features of the system. To be precise, if the sectors of economy have a similar relative weight (that is to say, if $a \cong 0.5$), the system proves more stable since, as seen in figure 1, the value of r must be higher to produce both instability and a chaotic situation. In fact, when a=0.5 a chaotic situation is not possible because, before entering the chaotic zone, one of the sectors disappears and the dynamics become pointless. On the other hand, if the volume of a sector is much lower than that of another sector, it is more likely that unstable dynamic and even chaotic behaviour appears.

In my opinion, this fact is significant since it suggests –taking into account all the restrictions of a model with so simple assumptions and only two sectors – that, when the productive resources are assigned and the size of the production sectors are similar, we may expect a more stable and calm dynamics of factor transfer from one activity to the other. However, when the sectors have different sizes, it would be more likely that unstable dynamics appear permanently without ever equalling the factor returns in the different activities and also that some productive sectors disappear.

Annex

In order to justify the answers to questions I), II) and III) formulated above, we will study the interaction in the function:

$$y = f(x) = x + r\frac{a - x}{(1-a)x}$$

where we have that the only fixed point $f(x^*) = x^*$ is found in $x^* = a$.

Let us see the basis for these answers:

I) The derivative of f is:

$$f'(x) = 1 - \frac{ra}{(1-a)x^2}$$

Therefore,

$$f'(a) = 1 - \frac{r}{a(1-a)}$$

And, since $r > 0$ and $0 < a < 1$, the equilibrium point is locally stable if and only if

$$R < 2a\,(1 - a) = U(a)$$

II) From the above result, we have that the first bifurcation appears in $r=2a(1-a)$. In order to establish condition II), firstly, we will find where the second bifurcation appears. Here, we are making the same study as in I) but this time applied to function $g(x) = f(f(x))$. This means that we have to get the values for which this function is satisfied, x^* being a fixed point of g different from a that verifies:

$$-1 < g'(x^*) < 1.$$

In this case, solving the equation:

$$g(x) = f(f(x)) = x$$

we obtain the following three values:

$$x_1 = a$$

$$x_2 = a\,(\,5+\sqrt{5}\,)/4$$

$$x_3 = a\,(\,5-\sqrt{5}\,)/4$$

and, solving the equation:

$$g'(x^*) = f'(f(x^*))\,f'(x^*) = -1,$$

taking for x^* the x_2 or x_3 values – the same result is obtained for both of them – and omitting the estimations, we obtain the value of the parameter in which the second bifurcation is located: $r = 2.5a(1-a)$.

Secondly, in order to find the entry into the chaotic zone, we will use the Feigenbaum constant [see Peitgen (1992)]: assuming that r_1, r_2, r_3, ... are parameter r values for which the successive bifurcations appear, then:

$$\lim \frac{r_n - r_{n-1}}{r_{n+1} - r_n} = d = 4.669201$$

Moreover, assuming that not only the limit but also all these quotients are equal to the Feigenbaum constant (d) and, taking into account that we already know the value at which the two first bifurcations appear, we have that:

$$\lim r_n \cong 2.63626a(1-a),$$

which is the chaotic zone. This bifurcation has been previously represented as Y(a)=2.63626a(1–a) and, even though it is approximated, it very much coincides which the bifurcation diagrams experimentally obtained.

III) The f function always has a minimum value.

Let x_{min} be the minimum of f in the [0,1] interval, it i obtained that:

$$x_{min} = \sqrt{\frac{ra}{1-a}} \qquad \text{when } r \le (1\text{-}a)/a$$

or we have $x_{min} = 1$ in the opposite case.

Therefore,

$$f(x_{min}) = \begin{cases} 2\sqrt{ra/(1-a)} - [r/(1-a)] & \text{if } r = (1\text{-}a)/a \\ 1\text{-}r & \text{if } r > (1\text{-}a)/a \end{cases}$$

The $f(x_{min}) \ge 0$ condition is equivalent to: $r \le 4a\ (1\text{-}a)$ if $r \le (1\text{-}a)/a$. However, when this condition is not satisfied we have that $1–r \ge 0$. In this case, if we define the following function:

$$V(a) = \begin{cases} 4a\ (1-a) & \text{if } a \le ½ \\ 1 & \text{if } a > ½ \end{cases}$$

we can express the results obtained as follows: with the dynamics generated by this system, the first sector shall always have a positive resource number if and only if $r \le V(a)$.

There is a much more complicated result when a negative number of resources does not appear in the second sector. Here we can also distinguish two cases. When the minimum of the function is located in the significant zone, the equation to be solved is:

$f^2(x_{min}) = 1$

which is equivalent to

$$2\sqrt{\frac{ra}{1-a}} - \frac{2r}{1-a} + \frac{ra}{(1-a)\cdot[2\sqrt{ra/(1-a)} - r/(1-a)]} = 1$$

which is also equivalent to the following cubic equation:

$$4r^3 + 4\ (1\text{-}a)\ (1\text{-}4a)\ r^2 + (1\text{-}a)^2\ (25a^2 - 14a + 1)\ r - 4\ (1\text{-}a)^3\ a = 0$$

This, after undergoing the corresponding transformations, has a solution equivalent to the following expression:

$$(1-a)\cdot\left\{\frac{4a-1}{3} + \frac{1}{6}\sqrt[3]{1-a}\left[\sqrt[3]{388a^2 + 43a + 1 + (45a+3)\sqrt{a(75a+6)}} + \right.\right.$$

$$\left.\left.\sqrt[3]{388a^2 + 43a + 1 - (45a+3)\sqrt{a(75a+6)}}\ \right]\right\}$$

which, to abbreviate, we shall call f(a).

When the minimum is located outside the significant zone (as said before, this happens when $r > (1-a)/a$), the condition we want to get is that $r < (2-2a)/(2-a) = y(a)$. Therefore, this condition is satisfied when:

1) if $r < (1-a)/a$ then, $r < f(a)$;
2) if $r > (1-a)/a$ then, $r < y(a)$.

However, if it is verified that 2), then $(1-a)/a < r < y(a) = (2-2a)/(2-a)$, which is only true if $2/3 < a < 1$. Therefore, the function which expresses a behaviour in which no negative number of resources appears in the second sector, is the following:

$$W(a) = \begin{cases} f(a) & \text{if } a \leq 2/3 \\ y(a) & \text{if } a > 2/3 \end{cases}$$

In order to check this results, we can see that if $a = 2/3$, it must be satisfied that: $f(a) = y(a) = (1-a)/a$; which, in fact, is so.

Basic References

Chiarella, C.: The Elements of a Nonlinear Theory of Economics Dynamics. Lecture Notes in Economics and Mathematical Systems. Springer Verlag, Berlin 1990

Collet, P., Eckmann, J.P.: Iterated Maps on the Interval as Dynamical Systems. Progress in Physics, Vol. I. Birkhäuser, Boston 1980

Devaney, R.L.: An Introduction to Chaotic Dynamical Systems. 2nd Ed. Addison-Wesley, California 1989

Dwyer, G. P. Jr.: Stabilization Policy Can Lead to Chaos. Economic Inquiry, vol. XXX, 40-46 (1992)

Kelsey, D.: The Economics of Chaos or the Chaos of Economics. Oxford Economic Papers 40, 1-31 (1988)

Lichtenberg A.H., Ujihara, A.: Application of Nonlinear Mapping Theory to Commodity Price Fluctuations. Journal of Economic Dynamics and Control 13,225-246 (1989)

Peitgen, H.O., Jürgens, H., Saupe, D.: Chaos and Fractals. Springer Verlag, New York; 1992

Other References

Baumol, W.J., Benhabib, J.: Chaos: Significance, Mechanism, and Economic Applications. The Journal of Economic Perspectives, Vol. 3-1, Winter, 77-105 (1989)

Chiarella, C.: The Elements of a Nonlinear Theory of Economics Dynamics, Lecture Notes in Economics and Mathematical Systems, Springer Verlag, Berlin 1990

Collet, P., Eckmann, J.P.: Iterated Maps on the Interval as Dynamical Systems. Progress in Physics, Vol. I.. Birkhäuser, Boston 1980

Deneckere, R., Pelikan, S. : Competitive Chaos. Journal of Economic Theory 40, 13-25 (1986)

Devaney, R.L.: An Introduction to Chaotic Dynamical Systems. 2nd Ed. Addison-Wesley, California 1989

Gumowski, I., Mira, C.: Recurrences and Discrete Dynamical Systems. Lecture Notes in Mathematics 809. Springer 1980

Kelsey, D.: The Economics of Chaos or the Chaos of Economics. Oxford Economic Papers 40, 1-31 (1988)

Lakshmikantham, V., Trigiante, D.: Theory of Difference Equations, Numerical Methods and Applications. Academic Press, London (1988)

Lorenz, H.W.: Nonlinear Dynamical Economics and Chaotic Motion. Lecture Notes in Economics and Mathematical Systems, 334. Springer Verlag, Berlin 1989

Mirowski, P.: From Mandelbrot to Chaos in Economic Theory. Southern Economic Journal, October, vol. 57-2 (1990)

Preston, C.: Iterates of Piecewise Monotone Mappings on an Interval. Lecture Notes in Mathematics, 1347. Springer Verlag, Berlin-Heidelberg 1988

Rasband, S.N.: Chaotic Dynamics of Nonlinear Systems. John Wiley & Sons, New York 1989

Wiggins, S.: Introduction to Applied Nonlinear Dynamical Systems and Chaos. Texts in Applied Mathematics 2. Springer Verlag, New York 1990

22 Market Failures: the Case for Road Congestion Externalities

V. Inglada-López de Sabando
University Carlos III (Spain)

P. Coto-Millán
University of Cantabria (Spain)

22.1 Introduction

The study of the external effects on the different economic sectors and branches has experienced a spectacular boom during the past decades within the framework of the Economic Analysis. This interest has become especially strong in the field of transport, in which we only have to go over the economic literature of the past years in order to find the large number of studies, research works and papers dealing with this topic.

The present paper covers the study of the different aspects in connection with the external effects on transport. The theoretical and methodological concepts on externalities introduced are applied to road congestion in order to determine the price of congestion and the net social benefit gain given by the introduction of that price.

With this aim, we have defined the economic concept of externality and described the wide typology existing around this concept, especially focusing on the specific field of transport, in which the different internal and external components of social cost are delimited and the theoretical framework which enables the internalisation of the external components of social cost from a Microeconomic Analysis approach.

Thus, we have studied in detail the external component of transport social cost known as congestion, which derives from the fact that introducing a new vehicle

P. Coto-Millán (ed.), *Essays on Microeconomics and Industrial Organisation*

into the transport infrastructure leads to an increase in the generalised cost to the remaining users since the speed decreases due to the higher volume of vehicles and, at the same time, the duration of the trip increases, this being a component of that cost. After describing in detail the theoretical concept of congestion optimal pricing and the way in which this component of the social cost is estimated, we have determined its magnitude for the case of the Spanish road network.

22.2 Externality

22.2.1 Concept

It has been stated in the economic literature that an externality is produced in consumption when the consumer is directly affected by either production or other agents' consumption[1]. Analogously, there is an externality in the production when a company' or consumer's decisions affect the production processes of other companies. Although the externality concept is commonly associated to that of diseconomy, it is possible to capture the existence of positive externalities[2].

The essential characteristic of externalities is the impossibility of transaction in the market. When external effects exist, the market does not enable an efficient allocation of resources in a Pareto sense. However, other social institutions such as the legal system or the State involvement, may, to some extent, reproduce the market mechanism and therefore, attain economic efficiency.

In connection with this, we must extend the definition of externalities[3] if they satisfy the following conditions: they are effects caused by activities outside the market; they occur when different agents use a resource jointly without property rights being clearly defined and at least one of the parts would prefer a contract agreement.

It has been often stated that the main, perhaps even the only, cause of these effects is the inadequate definition of property rights[4] since, otherwise, they would be eliminated by the market activity itself by means of the negotiation of these rights, no matter who owns these rights.

In addition, the externalities frequently have the characteristic of being public goods; in other words, there exists no rivalry in their use, so that their use by one agent does not prevent others from using them; and they are non excludible, that is to

---

1 Varian (1992) adopts this definition. In line with this, Rothengater (1993) states that an externality is either a relevant cost or benefit which is not considered by the economic agents when making their decisions. Button (1986) states that the externalities are generated when the activities of a group (consumers or producers) affect other group's welfare without any compensation for it.

2 Classic examples of positive and negative externalities are the cases of an apiculturist, an orchard of apple trees, or a fishery which is affected by the water upflow wastage. In the field of transport, positive external effects are in general less important. Roads constitute an example of this, since, apart from their basic function of facilitating transport, they also act as firebreaks in the event of a fire.

3 INFRAS/IIVW (1994 and 2000).

4 Coase (1960).

say, they cannot be prevented from being used by others. Therefore, even if the property rights were perfectly defined, market failures could take place as a result of the free rider problem[5].

An externality is considered to be "relevant" if it significantly affects the adaptive efficiency of market economy[6]. Definitely, there are many activities outside the market but only some of them are relevant enough and therefore require the State involvement to reduce their impact[7].

We must also distinguish between technological and pecuniary externalities. Although it is somewhat difficult to differ between both types of external effects because they usually take place simultaneously, we can state that technological externalities usually appear in the production or utility function, while this does not happen in the case of pecuniary externalities. Technological externalities are concerned with optimal efficiency aims, while pecuniary ones are connected with distributive aims and therefore, do not affect the total net benefit. In connection with this, in the INFRAS/IWW study (1994), two types of externalities are exposed, reaching the conclusion that only technological externalities must be considered because in the case of pecuniary ones, the effects on other markets are transferred through the relative price mechanism and therefore, when aggregating on all markets, the final effect produced is nil. In this sense, we cannot miss that pecuniary externalities are closely related with the consumer's surplus and that, in the event they may reach a significant magnitude, they would be even revealing that the relative price mechanism is operating efficiently so there would be no reason for state involvement.

This rule would have only two exceptions: a) the case that the change in the relative prices should in turn alter income distribution and b) the case that there may be any shortcoming in the functioning of the market (lack of information or monopoly competitions).

In these two cases, the pecuniary externalities in a market may induce technological externalities in a different market. Our analysis deals with the former externalities, which affect income distribution, since the direct contribution of the transport sector to the domestic product is comparatively small. In addition, we can base our conclusions upon the assumptions that the public measures which internalise such external effects come along with the compensation mechanisms when they may significantly affect income distribution.

In view of the impossibility to estimate the degree of the second order effects, we do not consider either the technological externalities which become pecuniary externalities, whenever failures in the functioning of the market take place[8].

In connection with the external benefits, as reported by ECMT, we may consider they behave as pecuniary externalities since most of them are captured

---

5 In addition, according to Nash (1997), the negotiation may bring about significant transaction costs.

6 A simple example of a non relevant externality is the usefulness implied by the spotting of cars and trains for some people, but it is obvious that, in this case, an internalisation policy is not required unless the passengers or operators may be disturbed by any of these individuals.

7 Verhoef (1994).

8 According to Rothengatter (1994) and IWW/INFRAS (1994 and 2000). Likewise, Greenwald and Stiglitz (1988) show, by using the example of the environmental externalities, that the wrong prices in a market may cause additional effects on others (secondary multiplying effect).

through the operation processes of the markets either in a direct (time savings) or an indirect (globalisation and regional development) manner. In line with this, Nash (1997) and Rothengatter (1993) state that these external benefits do not satisfy the condition which characterises the technological externality: having been generated outside the market[9].

22.3 Transport Externalities

22.3.1 Methodological Framework

The theoretical framework described in Jansson (1993) relies on a basic purpose which consists of the maximisation of the consumer's and producer's surplus addition once the costs of the negative externalities supported by the rest of society have been subtracted. Therefore, the net social benefit must be maximised.

The gross benefit for the users is determined by the integral of the marginal utility function $MU = U(Q)$, where Q is the flow of traffic.

In order to reach the consumer's surplus we must substract from the gross benefit the paid price P and the user's average cost AC^{user}.

$$\text{Consumer's surplus} = \int_0^Q U(Q)dQ - (P + AC^{user})Q$$

Likewise, the producer's surplus is $PQ - TC^{prod}$ and the overall external cost TC^{ext}. Finally, we obtain the net social benefit.

$$NSB = \int_0^Q U(Q)dQ - Q \cdot AC^{user} - TC^{prod} - TC^{ext}$$

By specifying the cost functions[10] and taking into account that the infrastructure design variables are fixed and can be therefore overlooked, we have the following expression, which must be maximised:

$$NSB = \int_0^Q U(Q)dQ - g(Q)\,Q - f(Q) - h(Q)$$

The addition of the price P and the user's mean cost AC^{user} is usually known as "generalised cost": $GC = P + g(Q)$

An equilibrium condition which must be satisfied is that the generalised cost is equal to the marginal utility:

---

9 Rothengatter (1994) and Verhoef (1994) also state that the external benefits of transport services are small and therefore, we may overlook them.

10 The total cost is the addition of the infrastructure, user's and external costs. The first component is a design function of the infrastructure, the physical characteristics (orography, etc.) and the volume of traffic. The user's cost component is a design function of the infrastructure and the capacity use rate. Finally, the third component accounts for the overall cost of the rest of society and is a design function of the infrastructure and the volume of traffic.

$$P + g(Q) = U(Q)$$

We impose below the first-order condition for a maximum which consists on the fact that the derivative of NSB with respect to the amount Q is equal to zero:

$$\frac{\partial NSB}{\partial Q} = U(Q) - Q\frac{\partial g}{\partial Q} - g(Q) - \frac{\partial f}{\partial Q} - \frac{\partial h}{\partial Q} = 0$$

By replacing $P + g(Q)$ with its equivalent $U(Q)$ we obtain for the price *P* the following final result:

$$P = \frac{\partial f}{\partial Q} + Q\frac{\partial g}{\partial Q} + \frac{\partial h}{\partial Q} = MAC^{prod} + Q\frac{\partial AC^{user}}{\partial Q} + MAC^{ext}$$

Therefore, we have achieved an important result which is that, in order to produce the maximum net social benefit, we must have the condition that the price is equal to the addition of the short-run marginal cost of the producer of transport infrastructure services plus the cost imposed on the remaining users of the infrastructure and the cost imposed on the rest of society by an extra user of the transport infrastructure.

The first addend is related with the infrastructure maintenance while the second one includes the congestion cost and part of the cost of accidents. The second and third addend, which contains various types of different negative externalities (accidents and environmental, for example), represent the main price components in the urban areas.

22.3.2 Typology

Table 22.1 shows the different types of social costs which exist in transport, including those of infrastructure (building and maintenance) and those associated with the use of infrastructure for transport activities. They are the addition of internal or private costs plus the external costs. Part of these costs are private, in other words, the user receives them as soon as he decides when and how to use the transport infrastructure. These may be either private since the beginning (users' own time employed in the trip, petrol without taxes and amortisation of the vehicle) or become private through the government's involvement[11]. The rest of the costs are external (accidents, environmental, etc.).

In order to study the externalities in transport we should use the base – following the common criterion used among the analysts – of a classification in four clearly differentiated groups: infrastructure deterioration through the use and wear of the pavement, congestion costs, costs of accidents and the costs closely related with the damage caused to the environment such as noise and pollution[12].

[11] By means, for example, of taxation, installation of a catalyst or the compulsory insurance. In public transport, the private cost would correspond to the fare paid for the ticket. The environmental externalities act on the rest of society while other externalities such as accidents and congestion, would affect the rest of the users of the infrastructure.

[12] Among other analysts, Carbajo (1991) and Newbery (1990 and 1994) use this classification of externalities in transport, as the ECMT (1998) and De Borger and Swysen (1998) do.

Table 22.1. Typology of social costs in transport

22.3.3 Internalisation Instruments

The four huge types[14] of internalisation instruments used to correct the effects of externalities are: economic incentives, regulatory instruments, actions on the infrastructure and service supply, and finally, information and persuasion measures.

Among the economic instruments we find the measures on prices in the form of either negative (taxation) or positive (subsidies) incentives. Other economic instruments are those aimed at regulating quantities, as it is the case of property rights. Regulatory instruments establish technical standards (speed limits and emission standards) or limit the demand (prohibitions on traffic in some particular areas). The third group of instruments includes the actions aimed at extending and improving the infrastructures, as well as the quantitative and qualitative improvements in public transport services. Finally, the last group contains those measures directed to the improvement of the degree of information to users about the transport impact on the environment.

The main aim of internalisation is to provide incentives with the purpose of reducing external costs[15] to the optimal level. Obviously, these incentives must be clearly identified by the users in order to be considered in the decision process which

[13] For example, landscaping effects, land occupation, energy consumption, vibrations, effects on the fauna and flora, etc.

[14] According to Rothengatter (1993) and ECMT (1998).

[15] In ECMT (1998), it is pointed out that the revenues generated by the internalisation may provide the Government with resources either to reduce taxes or to invest in environmental protection.

determines transport demand. In line with this - as the Economic Theory teaches us - from the viewpoint of the economic efficiency, the economic instruments based on actions on the prices, are the most adequate[16].

As regards the behavior of the different economic internalisation instruments of the external effects of transport (petrol tax, generalised toll, vignette, vehicle insurance premium, etc.), these may be classified, firstly, with respect to their degree of efficiency to reduce the social cost connected with the different transport externalities and, secondly, taking into account their high or low degree of practical application (short-run availability, implementation, legal barriers, etc.).

With respect to the degree of efficiency of these instruments, for the case of roads, the electronic toll has proved to be the best internalisation method of all the external effects, except for climatic change, for which this method is overcome by petrol tax[17]. However, as far as its possible practical application is concerned, this instrument presents difficulties for the short-run generalised implementation and its cost is high[18]. Nevertheless, the petrol taxes applicable have the advantage of being easy to be implemented and show an acceptable degree of efficiency in the internalisation of all the externalities, with the exception of congestion.

22.4 Congestion

22.4.1 Concept

The external component of the social cost of transport known as congestion has its origin in the fact that the increase in the volume of traffic on a stretch of road due to a higher number of vehicles incorporated to the circulation flow, generates a reduction on the average speed of the vehicles and therefore, a higher mean trip time. This effect leads to an increase in the generalised cost of the trip when a monetary value is introduced for time. It is an external effect since individuals do not bear the increase in the generalised cost of the other users' trip, but only their own increase. The magnitude of this increase in the cost depends on the volume of traffic existing, and is higher as we reach the infrastructure capacity.

In short, it can be considered[19] that the overall time of the trip, which constitutes a part of the generalised cost of the trip, can be divided into two

16 Baumol and Oates (1988) suggest that subsidies may create an undesirable incentive and are difficult to remove. Rothengatter (1993) mentions, among the disadvantages of regulatory instruments, the fact that no incentives are given for the reduction of the pollution levels under the established limit, as well as the difficulty found in the acquisition by the regulator of the sufficient information in order to efficiently allocate the reductions of the level of external effects (for example pollution) among the diversity of externality producers.

17 Johanson and Sterner (1998) highlight that the global pollution is connected with the emissions produced by fossil petrol combustion and therefore, the petrol tax is a very efficient internalisation instrument.

18 In ECMT(1998), the compatibility and technical harmonisation problems which make its implementation difficult are also mentioned.

19 Levinson and Gillen (1998) obtain both components of the trip time. In UNITE (2000) these two components of the trip time are also distinguished.

components. The first one is the ideal situation of traffic given when there is not congestion. This component is a user internal cost and essentially a function of the distance and speed of the trip. On the other hand, the second component or congestion cost - which is external and not borne by each individual traveler - basically depends on the number of vehicles circulating by road. In order to assess this external effect we have to obtain a mean of the increase in the time cost produced by a new vehicle over the rest of the vehicles on the road.

Given its unique nature, the marginal cost of congestion used to penalise the rest of the users depends on the volume of traffic that exists and therefore, in order to evaluate it, it is necessary to adopt a dynamic approach, different from other external effects of transport. In connection with this, we must base upon an expression which relates the speed, and more specifically, the time of the trip to the volume of traffic, as well as upon the knowledge of the demand behavior with respect to the price.

Other components of the vehicle operating cost - such as petrol or lubricant consumption - can be also affected by the reduction of speed produced by the increase in the volume of traffic[20].

Traditionally, economists and engineers have based upon different methodological approaches for the study of this transport externality, while the latter define the optimal volume of traffic mainly as a function of road features (slope, width, number of lanes, etc.), the economists consider that this is also a function of demand behavior.

When analysing the extensive literature about transport pricing and externalities, we may consider congestion to be the leading research topic in the field of Economic Analysis[21]. Since the first works were carried out, it has been the most common topic of analysis in the field of transport economy.

An important characteristic of congestion consists of being an externality internal to the transport market as it is solely borne by the infrastructure users.
The consideration of territorial and time dimensions gains a particularly relevant character in the study of congestion. In line with this, magnitudes are considerably higher in the urban field and at some particular rush hours when the traffic reaches its maximum value.

22.4.2 Congestion Pricing

The optimal congestion pricing is determined from the assumption that road users obtain an average net private benefit for the difference between the amount they

---

20 Kraus et al. (1976) consider that the operating costs of each user do not significantly depend on the volume of traffic. Hau (1998) reaches the same conclusion since, with a low volume of traffic, the speed is high and so is the petrol consumption. But to a high volume of traffic also corresponds a low speed with continuous speeding up and stops which bring about a high petrol consumption.

21 According to Johansson-Stenman and Sterner (1998), in the leading works about congestion by Walters (1961) and Vickrey (1955 and 1985), the concept of road transport optimal pricing associated to the introduction of the applicable rate paid by users for the internalisation of the external effects, is only used for congestion. Later, it was extended to other types of externalities: environmental, noise, infrastructure deterioration, accidents, climatic changes, etc.

are willing to pay for the use of the road and the private costs they incur in[22]. However, due to other social costs generated by the use of transport services, the volume of traffic should be limited in such a way that the net marginal private benefit of an additional vehicle-kilometer may equal its external marginal cost. Under these circumstances, the level of the economic damage is considered to be optimal since the total benefits and social costs reach their maximum value.

The theoretical framework used for the case of congestion relies on the assumption that the capacity of a transport infrastructure has a limit and the traveler's generalised cost increases mainly due to the increase in the time of the trip when the volume of traffic approaches the level delimited by such capacity. In order to simplify this, we use the hypothesis that there is only one type of vehicles. The individual decision of traveling by road depends on the generalised cost of trip (G) which includes the variable operating costs (for the case of the private car), fares (for the case of public transport) and the value of the time used[23].

When there is congestion, the number of vehicles (Q) increases, which reduces the speed of traffic and leads to an increase in trip time and the operating costs ($dC/dQ > 0$). When users decide whether to travel or not, and which transport mode to use, they only have to take into account the additional cost derived from their decision and overlook the effect on the level of congestion and the other users' trip cost. There is therefore, a difference between the marginal private cost (C) and the marginal social cost (MSC):

$$MSC = dC/dQ = C + Q(dC/dQ) = C(1 + e_a) > C$$

where Q is the number of vehicles traveling and e_a is the elasticity of average social cost with respect to the number of vehicles. This difference is the marginal external cost, which can be obtained as the product of the volume of traffic (Q) by the derivative of the average operating cost (private) with respect to the number of vehicles (dC/dQ).

In order to bridge the difference between the average social cost (C) and the marginal cost (MSC) an optimal price per trip must be fixed T in such a way that: $T = S - C = C_0 \cdot e_a$, where C_0 is the average optimal social cost of the number of vehicles Q_0 .

Figure 22.1 shows the curves of the private marginal cost and the social marginal cost together with the demand curve. The congestion price HB is equal to the difference between the social and private marginal costs for the optimal volume of traffic, which is represented in the intersection between the demand curve and the curve of the marginal social cost.

---

22 The development of this section is based upon Carbajo (1991).

23 The variable operatiing costs (for example, petrol) may include a tax, and therefore, it could be admitted that the individual private cost is different from the average social cost (G), $C = G - t$ where t accounts for the applicable tax or subsidy. For our case, we will assume that this tax is deducted from the marginal cost of infrastructure.

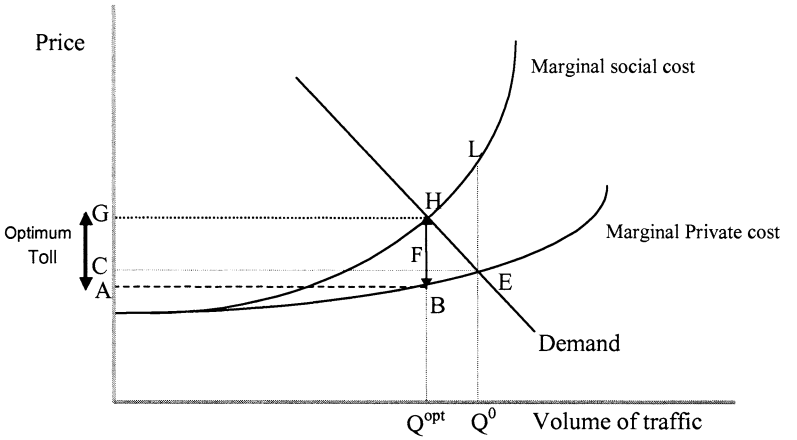


Fig. 22.1. Congestion pricing

In general, the volume of traffic which corresponds to this optimal price is lower than the existing one[24] and therefore, there is a surplus of trips which should not be made from the point of view of economic efficiency. Moreover, the optimal volume of traffic is a demand function for the shift either in the curve or in the slope, in other words, due to the price elasticity value. If no charge or lien existed the situation of equilibrium would be given at the intersection point between the demand curve and the one of the variable average costs (private costs). Those users whose value (amount they are willing to pay) of the trip is higher than the cost entailed, would make the trip. However, at this intersection point, the additional cost for the society, i.e.; the social cost which includes the other users' cost, exceeds the benefit produced by the last traveller. This premise continues to be satisfied until the intersection point between the demand curve and the curve of the short-run marginal costs is reached. The amount, for which the additional costs of the trips contained between these two points exceed the additional benefits, would represent the welfare losses caused by a non optimal price. In order to reach the situation of equilibrium with the optimal demand, a congestion charge must therefore be, imposed. This charge is equal to the negative external effect produced by congestion, in other words, the difference between the cost imposed by travellers to the society and the cost they necessarily bear when travelling.

[24] Prud'homme (1998) states, when analysing the consequences of optimal pricing, that, except when the demand curve cuts the private cost curve at its flat part, the level of use of the road is always higher than the optimum. Mohring (1976) shows that the intersection point with the average cost curve determines the stable equilibrium point since the curve of the travelers' willingness to pay – the inverse of the demand curve – equals the curve of average variable costs, the function upon which travelers base their decisions to travel. The difference between both curves is the congestion marginal cost (external): the additional delay imposed by a driver on the others.

Another method often used for the calculation of the marginal cost consists of having the difference between the average costs corresponding to the volumes of traffic of $n+1$ and n vehicles, in other words, the increase produced in the average cost when a new vehicle is added[25].

Hau (1991) explains in detail the diverse methodological choices to assess the impact produced by pricing based upon the marginal social cost, on social welfare, measured as net benefit. One of these methodologies consists of measuring the areas of gain or loss of welfare. This method is known as the methodology of quantities and is mainly used in the USA. The loss caused by the decrease in the trips from Q^O to Q^{opt}, as a consequence of the increase in the generalised cost of the trip from P^O to P^{opt} , is represented in the vertical trapezoidal area Q^{opt} H E Q^O. On the other hand, the savings in the cost of resources for the travelers caused by the traffic decrease, associated with a lower level of congestion and external cost is represented in the vertical area Q^{opt} H L Q^O. Therefore, the net benefit to the society derived from optimal pricing would be represented in the triangular area H L E.

A variant of this methodology is the net benefit methodology. The net benefit, in the case of the optimal level of traffic Q^{opt}, is the triangle formed by the demand function and the marginal cost curve. Analogously, the net benefit for the present level (non optimal) of Q^0 is given by the difference between the former area and the triangular area H L E. Therefore, this area represents the gain in welfare caused by savings in the social cost produced by the decrease in the volume of traffic from Q^0 to Q^{opt}. This way of calculating the benefit helps us to graphically understand how the net benefit is maximised when the pricing is based upon the marginal social cost. In fact, any movement of point Q^{opt} would generate a decrease in the area which measures the maximum net benefit. To the left (or right of Q^{opt}) the marginal value of travelers would be higher (or lower) than the marginal social cost.

Moreover, we cannot overlook the fact that there exists a transfer payment to the Government of a value equal to the area AGHB which is precisely the toll paid by road users. This payment must be excluded from the calculations carried out under the cost-benefit analysis for being a mere transfer. Paradoxically, however, it is actually the imposition of this toll that makes it possible for those road users to benefit from trip time savings equivalent to the area ACFB since, without this charge, the decrease in the volume of traffic which causes these time savings would not take place. However, road users suffer a loss in the consumer's surplus equal to the area CGHF. This process is similar to that of a monopolist who owns part of the consumer's surplus. In addition, the savings of time equivalent to the area ACFB also become revenue for the Government in the form of a transfer.

In short, the triangle HLE[26] accounts for the net social benefit gain produced by pricing, equivalent to the marginal social cost of congestion.

25

26 For example, Shah (1990) employs this methodology.

Lee (1997) highlights that the loss of efficiency shown by this triangular area may be classified into two components: the loss of time caused by a vehicle surplus above the optimal number and the loss of consumer's surplus represented by the bottom triangular area.

22.5 The Case of Spain

The above-mentioned methodology has been applied to the case of Spain with the aim of determining the social optimal congestion charge in the Spanish road network. It has been considered that the components of the vehicle functioning cost: petrol and lubricant consumption, maintenance and depreciation are independent from the level of traffic[27] and therefore, we only need to obtain the increase in the time component of the generalised cost.

22.5.1 Assessment Process

The following elements have been used:

a) Volume-Speed Functions.

Firstly, we need a function which connects the trip speed with the volume of traffic. The inverse of this function enables us to know how the average trip time varies depending on the number of vehicles. Apart from the algebraic expression taken for this function, there are other important factors such as the capacity, geometric and design characteristics of the road, percentage of heavy vehicles and equivalent number of cars allocated to each type of heavy vehicles in connection with their consumption and capacity.

Therefore, it is indispensable to know how the speed varies depending on the volume of traffic at different times. In connection with this, the typology of the functions used in the studies carried out present a wide spectrum. In order to achieve the purpose of this work, we have considered those types of functions which better adapt the interurban environment, and have overlooked the urban specifications.

The following relationships have been considered in this work:

$V_1 = 48 + 72(1 - I/C)^{0,5}$; for a road with various lanes[28].

$V_1 = 100 - 42(I/C)$; for a conventional road with only one lane.

Where V_1 accounts for the speed of light vehicles expressed in Km./h, I represents the hourly flow of vehicles and C is the road hourly capacity defined as the maximum time volume of vehicles which enables a stable traffic without significant interruptions.

The magnitudes of the capacity of each stretch of the road networks have been obtained by applying the methodology of TRB (1975) giving as a result the values of diverse factors which determine the capacity. Thus, for example in the case of roads with different lanes, this capacity depends on the number of lanes, width of the lane, type of traffic, percentage of heavy vehicles, type of terrain, etc.

---

27 The same hypothesis is used in Mohring (1976).

28 The relationship used comes from the approximation of a second-degree function to the curve proposed by the USA Road Capacity Guidebook (TRB, 1985).

Moreover, the speed of heavy vehicle is obtained by means of the following formula[29]: $V_p = 0{,}59 \cdot V_l + 28{,}85$

b) Demand functions.

Another fundamental element is the function which connects the demand with the price (or its inverse function, which is the willingness-to-pay)[30]. The exponential-like demand functions have been often used: $q = \alpha \cdot e^{(-p/\mu)}$, where α accounts for the number of vehicles-kilometer per time unit when the price p is nil, in other words, the potential demand. Moreover, μ is the average consumer's surplus per traffic unit. In this case p/μ is the demand elasticity with respect to the generalised price or cost. Some works have also used the linear demand function[31].

Once the demand price elasticity is known, we can predict the users' reaction when introducing the congestion load so that, it would be feasible to determine the equilibrium point which corresponds to the optimal congestion price. In this sense, we must point out – although it may seem obvious – that the magnitude of the external marginal costs of congestion of the current level of traffic is different from the one of the external marginal costs of congestion associated to the optimal level of traffic.

In order to determine the price demand variations, we have used the values obtained in Coto-Millán *et al.* (1997) for the petrol and diesel consumption elasticities relative to their respective prices: 0.25 and 0.16. From the values of the petrol consumption costs and the above-mentioned elasticities we have determined the successive points of the demand curve until an optimal equilibrium is achieved.

c) Value of time.

Likewise, it is necessary to use a trip time value estimation which enables us to turn time into the user's monetary cost. Very few studies have been carried out in Spain for the estimation of the time value in the different modes of transport. In this work, we have considered for road transport, the values supplied by the administrative Department responsible for transport in Spain, as set forth in MOPT(1991). These values expressed in prices of 2001 are 11.86 and 20.34 euros per hour for car and truck respectively.

---

29 Where V_l is the speed of the light vehicle expressed in kilometers/hour.

30 Gomez Ibañez and Okeeffe (1985) distinguish between the demand growth due to the demography and income and that generated by the price, also including time in the price. They consider a wide range of elasticities with respect to the price within the interval (-0,3,-3), depending on diverse factors: existence of an alternative via, road features, etc.. Evans (1992) uses the exponential function.

31 For example, the linear demand function is used by Decorla-Souza and Kane (1992).

22.5.2 Results

As shown in table 22.2, the Spanish road network - once the urban roads under the town councils' jurisdiction have been excluded - is composed by three subnetworks whose management is under the authority of three different public administrations. The one with the highest volume of traffic is under the authority of the State Administration, with a daily average volume of traffic of 12,144 vehicles. This subnetwork includes 8,082 kilometers of turnpikes and free highways. In the other two networks, the kilometers of highway and levels of traffic are low, with a daily average volume of traffic of 3,616 and 530 vehicles respectively.

Table 22.2. Non urban Spanish road network: characteristics of the different subnetworks. Year 2001

	Length	Vehicles-kilometer (Millions)	Daily Average Volume of Traffic
State Administration Network	24,458	110,826	12,414
Autonomous Community Network	70,844	93,499	3,616
Rest (Local Governments)	68,487	13,242	530

Source: Our own elaboration

In order to estimate the congestion price we have chosen an interurban network which is considered to be sufficiently representative given that it includes the road stretches with the highest volume of traffic. It is 20,932 kilometers long, of which 8,383 kilometers are free highways or turnpikes. Its daily average volume of traffic is 13,482 vehicles, of which 2,540 are heavy vehicles (19% of the total).

Given the importance of the territorial component in the evaluation of congestion costs, we have divided the road network chosen into 432 stretches so that this important territorial dissagregation provides us with more accurate results. In order to choose these stretches, we have mainly taken into account that they present certain homogeneity in their levels of traffic and that the municipalities with a certain level of population form part of these stretches' nodes. Each stretch is associated with a traffic counting station which help us to know the volume and time distribution of the corresponding traffic, as well as its distribution rate per type of vehicle (cars, trucks and buses). From the inventory elaborated we obtain the geometric features of each stretch (type of terrain, slope, etc...) necessary to determine the capacities, speeds and operating costs.

Analogously, we have observed a high degree of time dissagregation for traffic data since all the time volumes of traffic are considered in the 81 traffic counting stations chosen. In order to obtain a high efficiency we have formed three groups depending on the value of the daily average flow of traffic (IMD) at the station[32]. In addition, time intervals groups have also been formed according to the IMD

[32] These groups correspond respectively to IMD values of: less than 5,000 vehicles, from 5,000 to 12,000 vehicles and more than 12,000 vehicles.

rate they represent. Given the similitude between these time distributions, in the end, we have used the distribution frequencies of the average hourly flows of traffic as shown in table 22.3.

Table 22.3. Hourly traffic distribution in the Spanish road network

	Between 0% and 2% of the IMD	Between 2% and 4% of the IMD	Between 2% and 4% of the IMD	Between 2% and 4% of the IMD	More than 8% of the IMD
Number of hours (rates with respect to the total of annual hours)	26%	20%	33%	16%	5%

By applying the above-described process we have obtained for each stretch the congestion costs in road interurban transport with respect to the levels of traffic existing in 2001[33], and finally, the congestion prices of each type of vehicle associated to the optimal equilibrium situation. In this optimal equilibrium situation, the magnitudes of the decrease in the levels of traffic of light and heavy vehicles with respect to the initial situation are 6,9% and 4,2% respectively.

The average values of the congestion prices obtained for each type of vehicle in 2001, as show in table 22.4, are relatively low (2,61 cents of euro per vehicle-kilometer for cars, 5,74 cents of euro for buses and 7,83 cents of euro for trucks) only at rush hours, when the time volumes of traffic are over 8% of the IMD, the congestion costs reach significant magnitudes (15,88 cents of euro per vehicle-kilometer for cars, 34,93 cents of euro for buses and 47,63 cents of euro for trucks).

Table 22.4. Optimal congestion charges in the Spanish interurban road network. (Cents of euro per vehicle-kilometer. Year 2001[34]

	Mean value	Rush hour (with volume of traffic > than 6% of the IMD)	Rush hour (with volume of traffic > than 8% of the IMD)
Car	2.61	6.19	15.88
Bus	5.74	13.61	34.93
Truck	7.83	18.56	47.63

Source: Our work

[33] The external congestion costs obtained for the existing levels of traffic expressed in cents of euro of 2001 are 2.19; 4.81 and 6.56 for car, bus and truck respectively. The equivalence in terms of contribution to congestion, of the bus and truck with respect to the car is 2.2 and 3 respectively.

[34] The number of rush hours for which the congestion costs have been calculated account respectively for 21% and 5% of the total.

Finally, we have determined the increase in the social benefit produced, which is the triangular area HLE of figure 22.1. The introduction of the charge equivalent to the marginal social cost of congestion brings about a net social benefit gain of 89.37 millions of euros.

22.6 Conclusions

As regards the study carried out about the congestion price chosen in the Spanish road network - which characterises by covering all road stretches with a high volume of traffic - we must highlight that the results obtained about the levels of congestion in the Spanish road network are in general medium. The mean values obtained for each type of vehicle are relatively low (2.61 cents of euro per vehicle-kilometer for cars, 5.74 cents of euro for buses and 7.83 cents of euro for trucks).

However, at rush hours, when the volume of traffic is higher, this congestion price may reach relevant magnitudes. Therefore, when the volume of traffic at rush hours is higher than 8% of the IMD, congestion costs reach values of 15.88 cents of euro for vehicle-kilometer for cars, 34.93 cents of euro for buses and 47.63 cents of euro for trucks.

In addition, the study shows that the net social benefit gain obtained when the congestion price is introduced in the road network chosen, is of 89.37 millions of euros for 2001.

However, we must point out the high values of congestion in certain stretches of the road network, as is the case in the approach roads to big cities and in the roads alternative to turnpikes, which suggest recommendations to transport policies. For the case of approach roads to the big metropolitan areas (for example Madrid) we may propose congestion pricing by means of new technologies of electronic toll.

Finally, in order to avoid the inefficiencies of high levels of congestion in the conventional roads alternative to tolls, we must suggest the revision of the mechanisms and magnitudes of existing tolls, aiming these to the maximum efficiency and optimisation of social welfare in global transport system. In this way, they would become an efficient method of interurban traffic management in the Spanish road network.

References

Baumol, W. J. and Oates, W. E.: Economics, Environmental Policy and the Quality of Life, Englewood Cliffs. NJ: Prentice-Hall 1988

Button, K. J.: Transport Economics. Gower, Aldershot 1986

Carbajo, C. J.: El coste social de los accidentes de carretera y la contaminación del aire. Investigaciones Económicas, Vol. XV, no 2, 269-283 (1991)

Coase, R. H.: The problem of social cost. Journal of Law and Economics 3, 1-44 (1960)

Coto-Millán, P., Baños-Pino, J. and Inglada, V.: Marshallian demands of intercity passenger transport in Spain: 1980-1992: an economic analysis. Transportation Research E (Logistics and Transportation Review),Vol. 33, 2, 79-96 (1997)

De Borger, B. and Swysen, D.: Optimal pricing and regulation of transport externalities: a welfare comparison of some policy alternatives. In: Environment and Transport Modelling, 10-39 chapter 2, Roberto Roson and Kenneth Small publishers. Kluwer Academic Publishers. London, Great Britain 1998

Decorla-Souza, P. and Kane, A. R.: Peak period tolls: precepts and prospects. Transportation, Vol. 19, n. 4, 293-311 (1992)

ECMT: Efficient transport for Europe, policies for internalisation of external costs. OECD, Paris 1998

Evans, A. W.: Road congestion pricing: when is it a good policy?. Journal of Transport Economics and Policy 26, 213-243 (1992)

Gastaldi, M. et al.:Valuation of environmental externalities: from theory to decision making. Transportation Planning and Technology, Vol19, 207-219 (1996)

Greenwald, B. C. and Stiglitz, J. E.: Externalities in economic with Imperfect Information and incomplete markets. Quarterly Journal of Economics, 228-264 (1988)

Gomez Ibañez, J. A. and O'keeffe, M. M.: The benefits from improved investment rules: a case study of the Interestate Highway System. Report for the U.S. Department of Transportation, Washington 1985

Hau, T.D.: Economic fundamentals of road pricing: a diagrammatic analysis. The World Bank, Washington 1991

Hau, T.D.: Congestion pricing and road investment, 39-79. In: Button, K. J. and Verhoef, E. (ed.): Road pricing, traffic congestion and the environment: issues of efficiency and social feasibility. Cheltenham: Edward Elgar 1998

INFRAS/IWW: External Effects of Transport. Study carried out for the UIC, Paris 1994

INFRAS/IWW: External Costs of Transport. Study carried out for the UIC (International Railway Union), Paris 2000

Janson, J. O.: Government and transport infrastructure. In: Polak, J. and Heertje, A. (editors): European Transport Economics. CEMT/Blackwell 1993

Johansson-Stenman, O. and Sterner, T.: What is the scope for environmental road pricing?, 150-171. In: Button, K. J. and Verhoef, E. (ed.): Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility. Cheltenham: Edward Elgar 1998

Kraus, M., Mohring, H., and Pinfold, T.: The welfare costs of non optimum pricing and investment policies for freeway transportation. The American Economic Review, Vol. 66, no 4, September, 532-547 (1976)

Lee, D.: Uses and meanings of full social costs estimates. In: Greene, D. L., Jones, D. W. and Delucchi, M. A. (eds.): The Full Costs and Benefits of Transportation: Contributions to Theory, Method and Measurement, 113-148. Springer Verlag, Berlin Heidelberg 1997

Levinson, D. M. and Gillen, D.: The full cost of intercity highway transportation. Transportation Research D Vol. 3, no 4, 207-223 (1998)

Mohring, H.: Transportation Economics. Cambridge, Massachusetts: Ballinger 1976

MOPT: Manual de Evaluación de Inversiones en Ferrocarriles. Ministry of Transport and Public Works, Madrid 1991

Nash, C.: Transport externalities: does monetary valuation make sense? In: De Rus, G. and Nash, C. (eds.): Recent Development in Transport Economics, chapter 7, 232,255. Ashgate, Aldershot 1997

Newbery, D. M.: Pricing and congestion: economic principles relevant to pricing roads. Oxford Review of Economic Policy, vol. 6, no 2 (1990)

Newbery, D. M.: The case for a Public Road Authority. Journal of Transport Economics and Policy 28, September, 235-250 (1994)

Prud'homme, R.,: Road Congestion Costs in the Paris Area. 8TH World Conference on Transport Research, Antwerp 1998

Rothengatter, W.: Externalities of Transport In: Polak, J. and Heertje, A. (editors): European Transport Economics. CEMT/Blackwell 1993

Rothengatter, W.: Do external benefits compensate for external costs of transport ?. Transportation Research-A, Vol. 28A, no 4, 321-328 (1994)

Shah, A. M.,: Optimal pricing of traffic externalities: theory and measurement. International Journal of Transport Economics, Vol. XVII, no 1, February, 3-19 (1990)

TRB: Highway Capacity Manual. Special Report 209. Transportation Research Board, Washington D.C. 1985

UNITE:Unification of Accounts and Marginal Costs for Transport Efficiency. V European Framework Program 2000

Varian, H.: Microeconomic Analysis, 3rd edition. Norton & Company 1992

Verhoef, E.: External effects and social costs of road transport. Transportation Research-A, Vol. 28A, no 4, 273-287 (1994)

Vicrey, W.: Some Implications of marginal cost pricing for public utilities. American Economic Review 45, 605-620 (1955)

Vicrey, W.: The fallacy of using long-run costs for peak load pricing. Quaterly Journal Economics, no 100, 1331-1334 (1985)

Walters, A. A.: The theory and measurement of private and social cost of highway congestion. Econometrica, October, vol. 29, no. 4, 676-699 (1961)

23 Social Benefits of Investment Projects: the Case for High-Speed Rail

V. Inglada-López de Sabando
University Carlos III (Spain)

P. Coto-Millán
University of Cantabria (Spain)

23.1 Introduction

In line with the trend towards greater liberalisation of diverse markets of products and factors observed in modern economies, the role of the State has become increasingly relevant, not only in regulating the way the market operates, but also in orchestrating certain policies inserted into the welfare state.

One of the policies where the public sector continues to play a leading role is in transport infrastructure investments, particularly since Aschauer's seminal work (1989) in which he defined a new role for public investment as the engine of productivity and, basically, of the competitiveness of any economy, complementing the traditional Keynesian role based on demand. It is not therefore surprising that infrastructure investment is one of the mainstays of the economic programs of many governments.

However, these undeniably beneficial economic and social effects should not cause us to forget the important costs society has to bear for financing these infrastructure projects. We need to compare the costs and benefits of any investment proposal from the point of view of society as a whole, not merely from that of the agents or people directly involved. In this regard, the methodological tools of the welfare state economy, such as cost-benefit analysis, are particularly useful for evaluating the variation in the net social benefit of the investment project.

P. Coto-Millán (ed.), *Essays on Microeconomics and Industrial Organisation*

The aim of this paper is the social evaluation of investment projects. We begin with a description of the methodological lines for evaluating projects in the framework of the welfare economy. These methodological concepts are then used to assess the social rate of return on the introduction of a new product that has revolutionised the transport sector: high-speed rail (hereinafter, the AVE). This evaluation discusses the case of the Spanish Madrid-Barcelona-French border rail corridor.

23.2 Methodological Framework

In project evaluation, the ACB (Cost-benefit Analysis) is the most widely used methodology. This technique measures the costs and benefits on a project by including goods and services which are not negotiated in the markets. The benefits to be estimated are the utility increases measured through the variations in the consumer's surplus. The theoretical foundations of the ACB are found in the welfare economy, of which it constitutes the main empirical instrument.

Figure 23.1 allows us to check that the variation in the consumer's surplus provide us with a measurement of welfare variation. This figure shows the demand curve of the good q. Subindexes a and h represent the situations before and after the project respectively.

Fig. 23.1

The gross benefit for consumers in both situations is given by their respective willingness-to-pay:

$DAP_1 = Oabq_a$; $DAP_2 = Oadq_h$

The cost of the good to the consumer in the situation before and after the project is expressed with the product of the price by the quantity. Here, the prices are p_a y p_h, however, when we refer to the modes of transport, it is more appropriate to use the concept of generalised cost before the project g_a and generalised cost after the project g_h, and the difference between the DAP and the cost of the good is known as consumer's surplus (EC).

$$EC1 = DAP_1 - g_a\, q_a \; ; \; EC_2 = DAP_2 - g_h\, q_h + p_h\, q_h - p_a\, q_a$$

where $p_h\ q_h$ are the returns obtained after the project and $p_a\ q_a$ those obtained before the project. Finally, the variation in the consumer's surplus (VEC) between both situations is obtained as follows:

$$VEC = EC_2 - EC_1$$

The VEC measures the difference between the welfare in both situations and is given by the area $g_a bdg_h$. However, the use of the consumer's surplus as a welfare measure poses different problems. Firstly, there is a variation in the real income and the income marginal utility along the demand curve observed. For this reason, the measurement of the changes in welfare by means of a monetary instrument which varies when the prices vary, will yield wrong results. The solution is using the compensated demand curve introduced because it poses the practical problem of its determination. In any case, the errors committed by the use of the demand observed are not significant. A second problem arises with the question of price variations of a goods within the general equilibrium context. Therefore, in view of the price modification of a goods we must expect changes in the demand curves of other goods, that these are substitutive or complementary of the former and, consequently, generalised variations in prices. The monetary measurement of the change in utility in a general equilibrium environment is expressed by the following equation:

$$dU = \sum_{i=1}^{n} \int_{g_{i2}}^{g_{i1}} q_i(p_i)$$

where:

dU = utility change

q_i = goods, $i = (1,...,n)$

g_i = prices (generalised costs of the modes) of goods

The problem derived from the equation above is that the value of the second member varies depending on the way of integration. The widely accepted way is that the change in prices takes place at a constant rate.

The costs to be estimated in the ACB methodology are the choice costs, in other words, those costs which must measure the benefit on the best alternative project. However, the cost of a project C_i may not measure the value of the goods which would have been given in the event that the best choice had been developed, that is to say, they may not be choice costs. On the other hand, the C_i is used in the evaluation of individual costs for two reasons. On the one hand, the task of finding choice costs is sometimes virtually impossible and, on the other hand, many costs are not appropriate and they must be corrected by trying to find the costs from shadow prices. When costs from shadow prices are used, these

show the social value of the outputs and inputs implied in the projects. Thus, the costs are social although they are not real operating costs but costs which allow us to evaluate the project.

The net benefits on a project are measured as the difference between the gross benefits and those costs understood as welfare gains and losses, and they are measured as indicated above.

The decision rules used in the acceptance or rejection of public investment projects have been basically two: the net present value rule (VAN) and the return internal rate or return rule (TIR). The former evaluates the following expression

$$VAN = \sum_{t=1}^{n} \frac{B_t}{(1+r)^{t-1}} - \sum_{t=1}^{n} \frac{C_t}{(1+r)^{t-1}} \tag{23.1}$$

where the first addend accounts for the current net value of the benefits (VAN(B)) and the latter that of the costs (VAN(C)).

As observed, we have previously decided to used r as a discount rate. If VAN > 0, the project is accepted and the opposite case is rejected.

The domestic return rate rule - TIR - consists of finding the value of r, which is the solution of the above equation (23.1) when the VAN is zero, and comparing it with the minimum type of return rate admitted. If the value is higher than this return rate, the project is accepted and the opposite case is rejected.

Both rules may give rise to two different ordering of projects, so there may be doubts about which projects to accept and which ones to reject when there is a budget restriction. On the other hand, the TIR rule presents the following additional shortcomings: it must be modified with regard to mutually exclusive projects since there is a discrimination against a higher cost of capital. Likewise, it discriminates against long-life projects and those projects with longer period of construction. Moreover, multiple solutions may be given for the same project. These shortcomings, among others, have led most ACB authors to be in favor of the VAN rule.

23.3 High-Speed Rail

23.3.1 Main Features

The emergence of high-speed rail as a mode of transport essentially for passengers was a particularly important event for the transport sector. Since it was introduced in Europe in the 1970s to the Paris-Lyon rail corridor, the high-speed train has revolutionised the transport market and allowed railrways to reverse the downward trend that had converted them into a marginal mode of transport for travelers for over 40 years.

Since its inauguration in Spain, in April 1992, the Madrid-Seville high-speed train has had considerable commercial success and has been very well accepted by travelers. The qualities of this new railroad technology (speed, comfort, punctuality and safety) are rated highly by users (Inglada and Coto, 2002). This leads to high occupancy rates and positive results for the operating company.

However, we must remember that these results do not take into account the costs of maintaining and amortising the infrastructure.

The large number of trips generated by this new product and its high degree of absorption of demand from other modes of transport, particularly air transport, has made the AVE the predominant mode of transport along this route, constituting an indisputable historical milestone for railroad transport. The satisfaction of the users of this mode of transport and its commercial success are undeniable.

Basing their arguments on environmental considerations and other components of the social cost of transport, numerous influential voices in the European Commission are pushing to ensure that, in the future, the railroad plays an important role in the European market for passenger and freight transport.

It is not therefore surprising that, in this context, high-speed rail is the most important heading of transport infrastructure investments planned this decade in Spain, surpassing those planned for the road system.

However, without wishing to ignore its undeniable advantages, we must not forget that a project of this nature involves mobilising a considerable sum of public resources that could otherwise be put to a different use. From the point of view of economic efficiency, this social opportunity cost must be considered when undertaking a project of this nature by comparing it to the undeniable benefits of the AVE.

Besides its positive aspects, the AVE has other issues concerning the considerable cost of introducing the high-speed train (infrastructure and rolling stock), which need to be considered in order to evaluate the variation in net social benefit produced by this railroad investment policy from an economic standpoint.

We will use the methodological tool of cost-benefit analysis to determine the social loss or benefit of the introduction of the AVE, and we will obtain the mode distribution before and after the AVE from data provided by operators and the experience of the Madrid-Seville rail corridor.

The work is structured as follows: we will begin with a description of the main features of the rail corridor and the different types of offer of the AVE: shuttles, variable- gauge and high speed. We will then move on to compare the values of the different parameters characterising the other modes of transport using this corridor. Using surveys and data from the transport operating companies for this route, it is feasible to estimate the two components of demand for the AVE – generation and substitution – required for an appraisal of this project. As a basic reference, we will consider the impact of the AVE on mode distribution along the Madrid-Seville route.

Finally, using cost-benefit analysis methodology and the results obtained in earlier headings, we will estimate the social costs and benefits of introducing the AVE. The current value of the stream of these net benefits defines the profit (or loss) of net social benefit as a result of adopting this policy of investing in railroad transport.

23.3.2 Corridor Features

Traditionally, the most important features of the Madrid-Barcelona corridor are:

- High income and population.

The metropolitan areas of Madrid and Barcelona have over 4 million inhabitants and Zaragoza has over 0.5 million inhabitants.
Catalonia and Madrid are two of Spain's regions with the highest renta per capita, which exceeds the EU average.
This leads to considerable movement between these regions. One fact worth noting is that over 4 million people used the "shuttle service" by air between Madrid and Barcelona in 2001.
However, in the case of the Madrid-Seville AVE, economic activity is poor and the geographical position is peripheral. There is also little movement and work is of a marked seasonal nature.

- Connection with France
  The Madrid-Barcelona-French Border corridor is one of the two most important land links with Europe. It channels a significant flow of tourists to the south of the Spain and the Levante. Catalonia and Aragón also have important commercial and cultural relations with the south of France. It has some very important network economies.
  On the other hand, network economies in the Madrid-Seville AVE are almost non-existent.

23.3.3 Features of the AVE

This new mode of transport is marked by its high speed, more than double that of conventional railways, but also by its high infrastructure cost, a fixed cost that is almost independent of the number of travellers and therefore needs high levels of demand if it is to obtain an acceptable return rate. The introduction of the high-speed train significantly reduces the generalised cost of railroad transport. This reduction does not lie in the money component of this cost, but rather in the rest of the components (time, comfort, etc.).

A more detailed analysis of the Madrid-Barcelona-French Border corridor – the basis of our study – will enable us to distinguish between three very different subproducts within the "high speed" product. These are: Shuttles, Long-distance and variable-gauge[1].

"Shuttles" offer lower-priced trips due to their reduction at peak times, discounts and the use of units with more tourist class seats. They are a high-quality form of regional railroads that will help make the infrastructure more profitable.

Lastly, the "variable gauge" segment will allow use of the same rolling stock on the new UIC width lines as on the traditional Spanish width lines.

Table 23.1 reveals the most significant features of each type of offer, which differentiate the high-speed products. The basic discriminating factors between Shuttles and Long-distance trains are the type of demand catered for and the price, while the variable-gauge segment is basically distinguished from the rest by its use of a different kind of rolling stock, because of the need to use infrastructure with different track gauges. This, combined with the need to change gauge at the

[1] The new high-speed line has been constructed using UIC gauge, which is different from the Spanish gauge. Therefore, systems are required that permit the use of both types of track with a minimum loss of time during the changeover between the two infrastructures.

interchange, will mean that the average speed, and hence, the reduction in the generalised cost, will be lower than in other segments of the high-speed offer.

Table 23.1. Discriminating factors of the submarkets of the Madrid-Barcelona AVE

	-SHUTTLES-	-LONG-DISTANCE-	-VARIABLE-GAUGE-
Journeys	Barcelona-Tarragona Barcelona-Gerona Barcelona-Lérida, etc.	Madrid-Barcelona Madrid-Zaragoza Others	Madrid-Soria Madrid-Logroño Others
Stock	Alsthom	Talgo & Siemens	CAF
Infrastructure	New high-speed lines	New high-speed lines	New and conventional lines
Prices (Average revenue per traveler-km. in euro cents 2002	8,11	10,82	9,32
Occupancy Rate	0.60	0.65	0.60
Type of demand	Regional rail with a high percentage of commuter trips	Long-distance	Long-distance

A more detailed analysis reveals that the shuttle segment, with low prices and predominantly discounted tickets, generates a demand consisting mainly of commuter trips, where both legs of the journey are made on the same day. This is due to a reduction in the generalised cost of transport – mainly in time and price – which will also yield a reduction in the price of a complementary good, such as housing.

23.3.4 Mode Parameters and Features

In this section, we will consider the following mode parameters and features: a) monetary cost; b) travel-times; c) other generalised cost components; and d) value of time and distances.

a) Monetary cost.

Table 23.2 illustrates the monetary components of the generalised cost for each mode of passenger transport competing on the Madrid-Barcelona corridor. It shows how the AVE fare is higher than all modes of transport except for the air transport, but even here, if we consider the top quality category of the AVE, its price comes very close to that of the air transport.

c) Other components of the generalised cost.

The introduction of high-speed rail does not only yield a significant reduction in travel-time, but also in other components of the generalised cost, such as safety and comfort. This is illustrated in table 23.5 which is based on a survey of users of the Madrid-Seville AVE. It shows that these components figure highly in passengers preferences. In line with surveys carried out on the Madrid-Seville

[2] Includes penalty for delay, safety margin, departure time, and access times from the station to the departure point and destination of the journey. For the airplane, we have used 1hr for travel, 25' safety margin, including departure time, and a 10' penalty for delays. For the bus and the AVE, we have used 10' safety and departure, and 40' for traveling from and to the station. Lastly, for conventional trains, we have estimated the same time as for the AVE plus a 10' penalty for delays.

AVE, of the main reasons for choosing the AVE, comfort (29%) is rated at almost the same level as time (30%) and substantially above price (11%). This is particularly significant with the group of airplane users, where the comfort component (31%) is much higher than price (19%) and equal to the sum of the "traditional" components of the generalised cost: price and time.

If we analyse each mode of transport separately in greater detail, we can see that the main reasons for choosing the AVE over the airplane are comfort (31%), price (19%), speed (13%), novelty (11%) and safety (6%). With regard to the car, the most important reason is time (42%), followed by comfort (35%), and safety (10%). For conventional trains, the reasons for the substitution are: speed (57%), comfort (19%) and novelty (9%). Lastly, for the bus, the main reasons are speed (67%) and comfort (13%).

d) Value of time and distances.

Lastly, to carry out the cost-benefit analysis of this project, we also need to know the monetary values of time for each mode of transport. For this article, we have used the updated values provided by the government department in charge of transport in Spain, detailed in the Ministry of Public Works and Transport journal – MOPT (1991). Table 23.6 illustrates these values:

Table 23.6. Values of time (euro cents 2002 per passenger)

-Car-	-Airplane-	-Bus-	-Train-
6.04	25.56	3.25	13.01

Source: MOPT (1991)

23.3.5 Demand

Through the drastic reduction in the values of non-monetary components of the generalised cost, the introduction of high-speed rail has two clearly distinguishable effects on transport demand: inducement and substitution. The first of these effects refers to trips that would not have been made if this new service did not exist and the second, to those that would have been made on another mode of transport.

The AVE demand component, usually termed 'induced', includes all new trips. Besides passengers who have never made such a trip, this "generating effect" should also include another component of people who made this trip before the AVE existed and who now make more trips. As the table 23.7 shows, the average number of annual trips made by users of the Madrid-Seville route increases very significantly from 11.1 to 15.2.

Table 23.7. Evolution of the frequency of trips on the Madrid-Seville route

	BEFORE THE AVE	AFTER THE AVE
Twice a week or more	3%	6%
Once a week	6%	8%
Every two weeks	7%	9%
Once a month	16%	16%
Once every three months	18%	16%
Less	30%	17%
Did not travel	21%	---
First time	---	28%
Average trips per year	11,1	15,.2

Source: Based on surveys

The table 23.8 illustrates that, due to the magnitude of the effect of substitution, the introduction of high-speed rail will have has a very significant effect on the demand for the other modes of transport competing with the latter on the Madrid-Barcelona corridor. A more detailed analysis allows us to confirm the sheer magnitude of this effect in the case of the conventional train, which loses around 75% of the traffic it had before the AVE, and causes it to all but disappear from the corridor. In the case of air transport, the introduction of high-speed rail will cause demand on the Madrid-Barcelona route to drop significantly, by around 40%, as table 23.8 illustrates. Losses for the car are lower than in the above cases, around 20% on the Madrid-Barcelona route. Lastly, there does not appear to be a strong impact on the bus sector (10% losses), because the two products do not replace each other.

Therefore, we could conclude that the introduction of high-speed rail will yield a dramatic change in the respective shares of transport modes, and we can now speak of a transport market before and after the AVE. The railroad, with a

substantial 43.5%, would be the main mode of transport for the Madrid-Barcelona route, easily surpassing air transport.

Table 23.8. Change in modal split on the Madrid-Barcelona route due to the introduction of the AVE. (Thousands of passengers in both directions)

Transport mode	Before the AVE	(%)	After the AVE	(%)
Car	1312.1	19.9	1049.7	14.6
Air transport	4582.8	69.4	2749.7	38.2
Bus	300.8	4.6	270.7	3.7
Conventional train	408.0	6.1	100.8	1.4
AVE	------	------	3034.9	42.1

23.4 Cost-Benefit Analysis

23.4.1 Social Costs

As with any product, the costs of the AVE can be classified as fixed, semi-fixed and variable, depending on the length of time we take into account. Fixed costs correspond to the construction of the infrastructure (broadly-speaking) and its maintenance (although these costs probably evolve in line with demand in the long-term). Semi-fixed costs correspond to the purchase of rolling stock and, lastly, variable costs are commonly called operating costs and are very sensitive to the evolution of demand. All taxes have been excluded from the costs in each heading. So it is considered that prices (net of tax) of the infrastructure, trains and operating costs measure opportunity costs except in the case of labour.

The different costs are quantified below.

- Infrastructure Construction Costs

The infrastructure of the AVE includes the track, earthworks, signposting, stations, catenary, etc. Mainly the Railroad Infrastructure Management Agency (GIF) has constructed it. Work began in 1996 and will end in 2005. The schedule of expenses has been taken from information from the GIF and RENFE. The total cost of 855 kilometers is 7928.27 million euros of 2002. Each kilometer of construction of the infrastructure cost 9.27 million euros of 2002.

- Infrastructure Maintenance Costs

An infrastructure maintenance cost[3] of 10.22 euro cents per km has been estimated.

This costs heading will probably be slightly sensitive to demand in the long-term. We have not considered this effect as we have assumed that it will be counter-balanced by a likely downward trend in unit maintenance costs.

[3] This includes other concepts, such as station maintenance.

- Rolling Stock Costs

The costs of the three types of rolling stock are: 20.7 million euros for long-distance trains; 9.6 million euros for shuttle trains and 11.4 million euros for variable-gauge trains.

In order to calculate these costs, we need to obtain the units of rolling stock required. To find out the units required, we have based our calculations on the following:

The average capacity of Siemens and Talgo trains is 361 passengers. An occupancy rate of 0.65[4] has been assumed for Siemens and TALGO stock, which yields 235 passengers transported per unit. If we multiply this value by the annual route made by this unit under normal operating conditions (450,000 kilometers), we will need to add a new set of wagons every 105.75 million passengers-km per year.

We have assumed an occupancy rate of 0.60 for shuttles. As their capacity is 238 passengers and we suppose that a train covers 400,000 kilometers a year, we will need to add a new train every 57.12 million passengers-km.

For variable-gauge rolling stock, we have assumed an occupancy factor of 0.60. As their capacity is 228 passengers per train and the average standard route of each unit is 400,000 kilometers per year, we will need a new train every 54.72 million passengers-km.

We should point out that we have assumed the ideal operating conditions or maximum efficiency for incorporating new rolling stock. Therefore, there would be no discontinuities and new units would be added when necessary, evolving in line with costs.

- Operating Costs

This heading includes all costs involved in operating the AVE[5], and has been obtained separately for both types of stock. For long-distance stock, we have calculated the cost at 7.90 euro cents per passenger-km; for other segments of the offer the cost is 6.73 euro cents, in 2002.[6]

- Residual Value

The useful life of rolling stock is considered to be 20 years, and we have assumed a linear depreciation over this period.

For the infrastructure heading, MOPT (1991) estimates technical lifetime values separately, according to the different headings of the infrastructure, such as the installation of electrification and safety. We have estimated a value of 45 years as the average of the different headings and areas considered, for the entire infrastructure. Therefore, if we assume a linear depreciation, the residual value after 40 years' operation would be minimal, approximately 10% of the value of the investments made. For a technical lifetime of 30 years, the value of residual stock will be approximately 33% of the initial investment.

---

4 This occupancy corresponds to the Madrid-Seville corridor.

5 Therefore, we have included headings on energy, personnel, maintenance of rolling stock and services, restoration, video, etc. We have also included the "major maintenance operation" carried out every seven years, in this cost.

6 In the cost-benefit analysis, we should subtract the VAT payable on stock and services (around 85%) from the total.

23.4.2 Social Benefits

The methodology used for social benefits is that used in Dodgson (1984) and described fully in De Rus and Inglada (1994 and 1997). The basic points are summarised as follows:

The social benefits of introducing high-speed rail are generated by travel-time savings and the generation of trips.

Firstly, we will refer to the social benefits generated by travel-time savings. As the figure below shows, for users of conventional trains and the bus, the generalised cost (g_t), made up of the rate (p_t) and the value of the total time invested in the journey ($g_t - p_t$), declines to the value of the generalised cost of travelling on the AVE (g_h). We can express the benefits of this reduction for each mode of transport as:

$$(g_t - g_h)q_t + 1/2(g_t - g_h)(q_h - q_t) + p_h q_h - p_t q_t - C_t + C_h .$$

This is equivalent to the areas of the rectangles $g_t b g_h e$ and $p_h f p_t j$, less the net cost of obtaining these benefits. This cost is that of introducing the AVE less the savings generated by eliminating conventional train and bus services.

However, if we look at this figure and disregard who reaps the surplus, the social benefit can be obtained for trips diverted from other modes due to the travel time savings yielded by a faster transport mode. Therefore, we simply need to calculate the reduction in access and travel times, and multiply by the value of time.

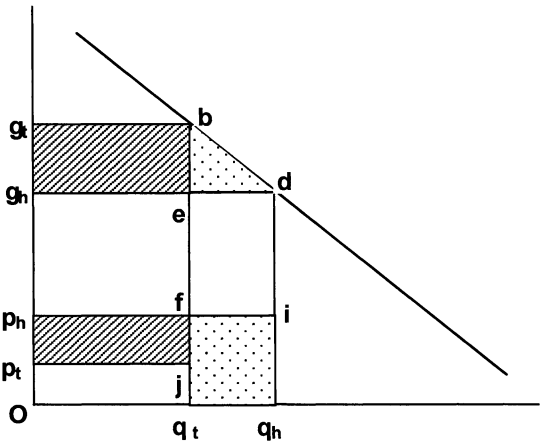


Figure 23.2. Users diverted from rail and bus

We will now deal with the social benefits from the trips generated by this mode of transport. The total benefit for generated trips in figure 23.2 is the area below the function of demand (the trapezium $bdq_h q_t$) less the area of the rectangle (edif) representing consumed travel-time. It can be broken down into two components:

representing consumed travel-time. It can be broken down into two components: the first component, rectangle (fiq_tq_h) in this figure, is obtained from the revenue from these trips $(q_h-q_t)\bullet(p_h)$ and the second, triangle (bde), is half the difference of the generalised costs for generated trips $1/2\ (q_h-q_t)\bullet(g_t-g_h)$.

As we have survey data, it is feasible to distinguish between trips generated by an increase in the frequency of trips (of users diverted from other modes of transport) and newly-generated trips. For those who traveled previously, we have taken as a reference the generalised cost of the primitive mode of transport. For new trips, we have used the weighted average of the generalised costs of the different modes, whose weightings have been determined on the basis of data from the survey carried out on the substitution effect.

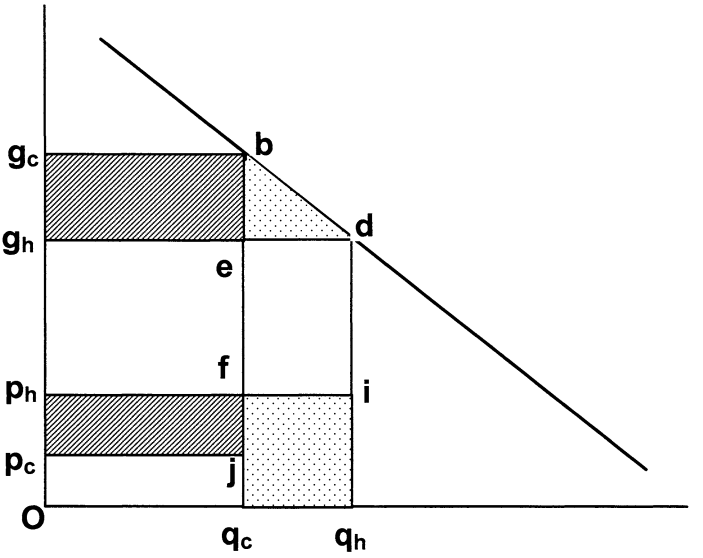


Fig. 23.3. Benefits of the AVE (users diverted from cars)

This methodology is valid for the conventional railroad and bus, but we will need to modify it for the airplane and the car. In the case of the car, we must add the savings in operating costs through not traveling in the car (p_cjoq_c) to the savings in resources in trips diverted from other modes ($g_cbg_he + p_hfp_cj$), as figure 23.3 illustrates. Given that there are a number of alternative modes, we only need to add the costs of the AVE in one mode to avoid a double accounting.

Finally, in the case of the air transport, illustrated in figure 23.4, although the generalised cost is lower for the AVE, this reduction in costs is due exclusively to the lower price, because the time component is still lower in the airplane. Therefore, we obtain a negative benefit value if we apply the above methodology directly. However, there are other important components of the generalised cost: comfort, safety, etc., which reduce the disutility of the AVE in line with price. Therefore, we have opted to assume a decline in the generalised cost of the AVE

compared to other modes of transport, in line with the results of the survey on reasons for choosing the AVE over the airplane[7].

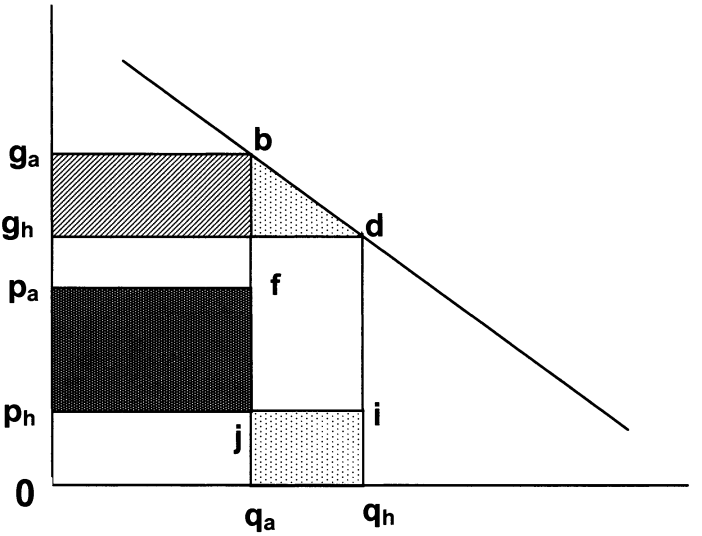


Fig. 23.4. Benefits of the AVE (users diverted from the airplane)

23.4.3 Other Social Benefits

There are other social benefits besides those in the above heading; these are the social benefits of time savings and trip generation. Here, we will consider:

- Costs reduction in other transport operators:
  - Reduction in conventional train costs
  - Reduction in airplane costs
  - Reduction in bus costs
  - Reduction in car costs
- Reduction in external costs:
  - Reduction in accident costs
  - Reduction in congestion costs
  - Reduction in environmental impact (local, global and noise pollution).
  - Reduction in the costs of maintaining the road and conventional railroad infrastructures.

Therefore, we have not directly considered the macroeconomic, sector and regional effects analysed in detail by Alvarez and Herce (1993) and Inglada

[7] In line with the evaluations of users in the surveys, we have increased the generalized cost of alternative modes of transport to introduce other components of utility, such as: comfort, safety, etc. Therefore, these costs have been multiplied by 1.8 for the car, 1.2 for the train, 1.15 for the bus and 1.5 for the airplane.

(1993). An initial reading of these works reveals that the macroeconomic effects produced during the construction of the infrastructure, common to all types of public investment, are diverse. Positive effects include the heavy impact on the generation of employment, while other macro magnitudes, such as inflation and salaries in the public and foreign sectors, decline. We are also well aware of the heavy pull on other public investment sectors as well as the transport sector.

However, in our case it could be lower than in other alternative projects because part of the inputs is from other economies. Finally, from the point of view of regional balance, introducing the AVE offers a clear benefit in the relevant territories, bearing in mind Spain's peripheral situation as regards decision-making centers and the production of the European economy. However, this effect is difficult to quantify, particularly if we try to separate the inherent benefit from reductions in time and transport generated by the AVE, which is what we are considering here. Bonnafous (1987) describes the regional effects of the TGV, pointing out that the greatest impact was in the Paris region and that certain sectors, such as the hotel industry, had declined.

The methodology used to evaluate the headings of benefits is detailed below.

An initial heading of the inherent benefits of operating the new product consists of the reduction in costs in alternative modes of transport: conventional railroad, airplane, buses and car.

- Reduction in other transport operator costs:

The reduction will be noted in other transport operators: conventional trains, airplane, bus and car.

- Reduction in conventional train costs.

The intermodal substitution caused by the AVE is particularly significant in the case of conventional trains, which become an all but marginal mode of transport on this corridor in terms of passenger transport.

To determine the corresponding reduction in costs, we have used the production cost structure[8] of an average daily train, which is most clearly affected by the reduction in offer. Of all of its components, we have only excluded half of the cost of the stations heading. However, the entire heading on amortization has been included as the train could be used on an alternative route.

As the offer in this corridor all but disappears, we may use a unit cost ratio per passenger-km to determine the total cost; we obtained this by applying a coefficient (representing the difference in occupancy rate between this corridor and the national average) to the average national value. We have estimated 3.85 euro cents of 2002 per passenger-kilometer from conventional trains.

- Reduction in airplane costs.

The diversion of passengers from air transport to the AVE generates a reduction in costs for air transport operating companies on the Madrid-Barcelona corridor. For this mode of transport, unlike conventional trains, the reduction of the offer is generated, besides a reduction in the number of flights, by a reduction in average load factor.

[8] We have not therefore considered the costs of central organization, also called structural.

Therefore, instead of using a unit cost per passenger-km, we have based the methodology on calculating the reduction in the number of flights to obtain the cost saved on the Madrid-Barcelona route, by applying the average cost of a flight, which is finally extrapolated from the rest of the routes.
We have obtained the unit cost per flight by eliminating fixed entries from the total cost, such as those for structure and part of marketing, and all taxes. The amortization heading has been maintained because the aircraft could be used on another line, for example, by the operating company hiring out the airplane. The cost is 12.61 euro cents per passenger-km from the airplane.

- Reduction in bus costs.
The effect of intermodal substitution produced by the AVE in the case of the bus generates a saving in costs for operating companies, essentially by reducing the number of journeys. To calculate this saving, we have based our calculations on the cost structure of a representative bus in MOPT (1991).
This structure also includes the saving in road maintenance costs due to a reduction in the number of buses circulating on roads. The saved cost includes all components, including amortization, because the vehicle may be used on another route in the short term, in the form of opportunity cost. Therefore, we have only eliminated the costs of the tax heading, because we assume that structural or organisation costs are included in the data on the bus itself. The saving is 3.19 cents per passenger-km diverted from the bus.

- Reduction in car costs.
In the case of private vehicles, unless they are hired, the planned journey cannot be substituted by another journey to another destination in the same period of time. Therefore, the "fixed" headings, such as part of amortization and insurance should not be included in the cost saving calculation. The part of the amortization heading that we have included is based on MOPT (1991), which estimates that half of this heading corresponds to the lapse of a length of time, while the other half is closely linked to vehicle use.
On the basis of this premise, we obtain the structure of saved costs by removing the fixed and taxes headings for each journey on the corridor. For more accurate results, we have broken these routes down into homogenous stretches related to traffic, in accordance with the above work. The costs for each stretch, for which we have applied specific methodologies, are: vehicle amortization, maintenance, fuel consumption, lubricant consumption and tire consumption.
The unit saving is 6.67 cents per passenger-km.

- Reduction in external costs:

Table 23.9. Marginal external social costs by transport mode (Euro cents.of 2002)

	Environmental	Infrastructure maintenance	Accidents	Congestion
Car	1.64	0.73	2.65	1.93
Train	0.52	2.28	0.15	–
Bus	0.58	0.14	0.68	0.33

Table 23.9. Continued

	Environmental	Infrastructure maintenance	Accidents	Congestion
Airplane	2.14	–	–	–
AVE	0.43	–	–	–

Source: Inglada (2003)

To evaluate external costs, we have used the methodology and results of Inglada (2003). The final values are illustrated in table 23.9.

We set out below a brief summary of the way in which these costs are calculated:

- Reduction in accident costs.

The introduction of the high speed Rail in the Madrid-Barcelona-French border rail corridor leads to a decrease in passengers (by car and bus) in these roads, with the subsequent decrease in the number of accidents. According to Inglada (2003), we have assumed a unit value for the elasticity of the number of accidents with respect to the traffic and therefore, in order to calculate the external social cost of the accidents we only need to take into account the non internalised costs by means of the applicable policies. A base value of 0.93 million euros of 2002 has been calculated for human life according to the European Commission's recommendations conveniently updated and an alternative value of 1.39 million euros of 2002 (50% higher) in the sensitivity analysis. Analogously, we have determined the external costs of the accidents in conventional railroads. The addition of both costs is the benefit of the decrease in accidents due to the lower number of travellers on the road and railroad.

- Reduction in Congestion Costs.

The introduction of the AVE has led to a reduction in the traffic of roads in the Madrid-Barcelona rail corridor. This reduction brings about an increase in the speed and therefore, a decrease in travel-time for the vehicles (cars, buses and lorries) which are on the road. In order to quantify this externality, it is necessary to determine the relationship that exists between the volume of traffic and the speed.

In TRB (1985) two figures simulate the desired relationship for the cases of motorway and conventional railroad. In order to obtain a higher efficiency, two curves in the form of a parabola and line have been adjusted to these figures respectively, which yield the following algebraic expressions:

$V = 48 + 72\ (1 - i/c)^{1/2}$ For Motorway and Dual Carriageway

$V = 100 - 22\ (i/c)$ For Conventional Railroad

Where:

V = Speed of light vehicle.

i = Flow of hourly traffic.

C = Road hourly capacity.

The speed of heavy vehicles is determined by means of the following relationship:

$V_p = 0{,}52\ V + 28.85$

The speed for each time of the year should be calculated by means of the above-mentioned formulae in order to determine the average speed. However, having verified that the error margin is very small and obviously, this being more operative, the times of the year have been classified into frequency intervals according to the values of the volumes of hourly traffic.

From the speeds obtained for each interval and by means of the application of the corresponding weighted mean, we have the average annual speed for each stretch of the corridor chosen by using the volume of traffic provided by the respective traffic measuring station.

Using this methodology, we can determine the speeds and therefore, the travel-times at each given[9] stretch for the options "with or without AVE". Through the aggregation of the stretches, we obtain the time reductions both for light and heavy vehicles in the routes which compose the corridor analysed. Finally, by means of the monetary values of time in table 23.6 we obtain the cost savings due to the reduction in road congestion. This benefit is only considered for road transport.

Finally, in order to obtain a higher efficiency, we have assumed that, once the full road capacity has been reached for certain times, the users change the time of their movements and use the nearest possible time periods in which the volume of traffic is lower than the capacity[10].

- Reduction in environmental costs (local, global and noise pollution.)

In this heading we must differentiate between the effects produced during the building of the infrastructure and those produced during its use. The former are mainly visual intrusion and land occupation, whose cost, especially in the case of the latter, should be included in the infrastructure implementation cost. On the other hand, according to Nash (1991), while railroads require less surface area than roads, we must also take into account the volume of traffic borne by each mode of transport. The railroad has even less advantages if we compare the new railroad line with the alternative extension - by means of additional lanes - of an existing road.

Among the effects produced by the use of the infrastructure, noise and pollution stand out. As regards noise, even though the results of some surveys evaluate the AVE impact as lower than that of the road, we cannot say that this result is unanimously accepted and it is difficult to reach an easily quantifiable conclusion in connection with this point, and more especially for the case of interurban commuting using the AVE. We have used the results exposed by Inglada (2003) as shown in table 23.9.

The product obtained from the difference between the unit environmental cost in each mode of transport and that of the AVE by the number of travelers-km diverted to the AVE yields the reduction in the environmental cost in each mode

9 The choice has been made from the assumption that the stretches have very similar intensities of traffic.

10 To be perfectly fair, we should include assumption of extension of the road at the time the capacity is used up, assuming a supply policy of immediate response to demand requirements. However, the error margin incurred for not taking into account this aspect would be contained within the limits of the work.

of transport. Finally, these four headings together determine the environmental benefit caused by the introduction of the AVE. We must also highlight that the environmental costs in the AVE produced by the induced demand segment have also been considered.

- Reduction in the road and conventional railroad infrastructure maintenance costs.

Due to the substitution effect, the introduction of the AVE in this corridor leads to a reduction in the number of vehicles (cars and buses) in the road network affected, which in turn leads to a reduction in the maintenance costs borne by the respective state administrations rather than by the user. The same occurs in conventional railroad, in which a reduction in the maintenance costs of the railroad infrastructure of this corridor (not applied on the user's ticket) takes place due to the decrease in travellers given by the introduction of the AVE.

In order to calculate the unit costs per vehicle-kilometer for the car and bus, we have based on Inglada (2003). Essentially, the methodology used consists of classifying the different sections, which compose the maintenance costs of road infrastructure and are borne by the respective state administrations, into the different types of vehicles (cars, buses and lorries) according to the criteria chosen (weight per axle, equivalent vehicles, etc.). The addition of these components for each type of vehicle determines its unit cost of maintaining the infrastructure. The final values of the unit costs for the car and bus are shown in table 23.9. Analogously, as regards the railroad, the costs of maintaining the railroad infrastructure are distributed between travellers and goods by using a similar methodology. The unit costs of maintaining the railroad infrastructure per traveller-km - the ones applicable in this work - are shown in table 23.9.

- Miscellaneous

Another heading not included in the analysis is the one related to the benefits associated with the delay in making public investments, caused by the decrease in the traffic borne by the infrastructures of the mode of transports alternative to the AVE. However, mainly due to the wide supply expansion, which took place during the years of implementation of the new project[11], there will be an excess of capacity in the alternative modes, especially in air transport, which makes the magnitude of this effect be very small. Therefore, we have decided not to include it in the quantification made, assuming that it is part of the error margin of the work.

For example, the decrease – in relative terms – in traffic at Barajas (Madrid) airport caused by the new project would be very small[12] and would not constitute a decisive fact when deciding a possible extension of this airport. Another benefit of this project would consist of delaying of the extension of the Madrid-Zaragoza infrastructure with respect to the alternative choice of maintaining the "status quo", estimated approximately for two years. In any case, congestion would not

11 The extensions in progress of Barajas (Madrid) and El Prat (Barcelona) airports would lead to an important increase in the supply capacity of both airports, at practically the same time as the launch of the AVE.

12 According to the estimations carried out, the reduction of passengers in Barajas airport due to the introduction of the Madrid-Barcelona AVE would not exceed 7%.

exist in this road until after the year 2015, and therefore, its influence on the results of the evaluation made would be of little significance.

23.4.4 Other Evaluation Hypotheses

a) Demand.

The initial demand of the corridor before the introduction of the AVE has been determined by using the data of the operating transport companies and the surveys on mobility carried out. From the results obtained about the Madrid-Sevilla AVE (Inglada, 1993) we have estimated the substitution and generation factors of the new demand for trips.

At the same time as the process of introduction of a new product in the market, the full potential national demand for the AVE is reached after a maturing period. For the evolution during this period, estimated in four years, we have based on the commonly used hypothesis that the evolution of the product demand is similar to the logistic curve. During the first year (2005), it will reach 50%, at the second, 70%, at the third, 90% and finally, in 2008 the full potential demand estimated at 5,055.8 million passenger-km will be reached.

For the international demand, we have based on the assumption that it will take place in 2007 once the extension of this line is concluded (Figueras-Perpignan stretch with a tunnel crossing the Pyrenees), which would connect with the French high speed train network. We have assumed that the evolution of this international demand throughout he 4-year maturing period will follow the same guidelines as national demand since 2007.

Given the uncertainties existing about the date of completion of the French high-speed rail network[13] we have considered two scenarios in the demand evaluation which correspond to the national demand and the addition of the national plus international demand[14].

In order to apply this methodology it is necessary to know, not only the initial demand for the AVE but also its evolution during the operating period considered disaggregated into its two components: generated and diverted traffic, as well as the revenues that are necessary to measure the benefits produced by the induced travelers.

In order to solve this question, we have considered that the demand for the AVE develops with an elasticity with respect to the GDP of 1.4. This value is similar to that obtained in air transport in our country (Coto et al. 1997) and is in accordance with the values obtained in other countries (Owen and Philips, 1987).

In the base scenario we have assumed that the GDP annual growth rate is 2.5% over the whole project. This value would correspond to the growth rate of the Spanish potential GDP. Alternatively we have considered a trend GDP growth scenario of 3.5%.

---

13 For example, for the Nimes-Montpellier stretch, which connects both lines the French government has not still decided on a date for initiating its construction.

14 International demand is very important since it would represent - in the event that France would complete its high speed network - about 40% of the national demand and 29% of the total (6,680 thousand million travelers-km).

b) Lifetime of the project.

Likewise, we have considered two lifetime values for the project of 30 and 40 years. The latter period seems to be appropriate given the importance of the project. The evaluation is made at constant prices of 2002. We have assumed the maintenance of the relative prices of the different goods and services during the life of the project.

c) Social Discount Rate.

All costs and benefits are updated at a social discount rate of 6% in real terms which is the value normally used in the Spanish official evaluation guidebooks[15]. The social profitability of this project is obtained by comparing the above-described costs and benefits updated to the base year through the social discount rate chosen. With this aim, the current net value of this flow of net social benefits is determined.

23.4.5 Social Profitability

Table 23.10. Social benefit of the AVE in the Madrid-Barcelona-French border rail corridor. Hypothesis: only national traffic. (Million euros of 2002)

	Basic result of the AVE[16]	Result in the most favorable scenario[17]
Total infrastructure, maintenance and operating costs		
Infrastructure	–4658,0	4285,3
Residual value	252,9	71,9
Rolling stock	–919,7	1026,5
Infrastructure maintenance	–624,3	683,4
Operation	–3593,6	4402,2
Time savings of users diverted from:		
Other modes of transport	1333,1	1636,6
Generated trips	1168,8	1434,9

[15] This social discount rate has been used in other infrastructure projects (especially roads) carried out in Spain during the nineties. In Riera (1993) even higher rates are used (8 and 10%) for the economic evaluation of the Barcelona ring road.

[16] This corresponds to a basic value of accidents (0.93 millions of euros), a GDP growth of 2.5% and 30 years of duration of the project.

[17] It corresponds to a maximum value of accidents of 1.39 euros, a GDP growth of 3.5%, 40 years of duration of the project and estimation of the labor factor shadow price. In order to calculate this value, it has been assumed that labor costs represent 25% of the total of infrastructure costs and that 25% of the workers were in a situation of structural unemployment.

Table 23.10. Continued

	Basic result of the AVE[18]	Result in the most favorable scenario[19]
Reduction in costs of:		
Conventional railroad	496.1	609.0
Air transport	1763.7	2165.3
Bus	50.0	61.4
Operating costs of cars	1097.8	1347.7
Congestion	187.1	229.7
Accidents	458.4	844.1
Environment	372.8	457.7
Infrastructure maintenance	416.8	511.7
Net present value of the AVE	-2198.0	-1027.5

Given the great doubts that exist about the date of extension of the French high speed train until the Spanish border to connect with the Spanish rail network, we have preferred to carry out a first evaluation of the project only considering potential Spanish traffic.

Table 23.10 shows the results of the social evaluation of the Madrid-Barcelona-French border high-speed train when only the national traffic is taken into account. We observe that social costs exceed social benefits in 2,198 million euros of 2002. Even with the most favorable hypotheses to the return rate on project (an estimated growth of 3.5%, maximum value of the human life, etc.) the net social benefit is negative (-1,027.5 million euros of 2002). The basic reason of this negative result relies on the existence of a very low level of demand which makes the willingness-to-pay for the capacity be lower than the capacity costs themselves. In this sense, when comparing with other modes of transport, high-speed train return rate depends much more on the density of traffic in the corridor, since the supply of additional units of railroad service incorporates a much lower additional cost due to the intense effect of scale economies[20].

On the contrary, the estimation of the international potential traffic to be captured by the AVE if the French high-speed rail is extended to the Spanish border, leads to a considerable change in the results as regards the return rate on the project.

[18] This corresponds to a basic value of accidents (0.93 millions of euros), a GDP growth of 2.5% and 30 years of duration of the project.

[19] It corresponds to a maximum value of accidents of 1.39 euros, a GDP growth of 3.5%, 40 years of duration of the project and estimation of the labor factor shadow price. In order to calculate this value, it has been assumed that labor costs represent 25% of the total of infrastructure costs and that 25% of the workers were in a situation of structural unemployment.

[20] The difference between fixed and variable cost is higher in the high-speed train than in road transport and even more in air transport. The fixed costs of this new technology may become twice or three times higher than those of the road.

As observed in table 23.11, if we take into account the international demand segment, the net social benefit generated by the Madrid-Barcelona-French border AVE is of 617.1 million euros for the base scenario with 40 years of project lifetime.

Table 23.11. Social benefit of the AVE in Madrid-Barcelona-French border rail corridor. (Hypothesis: with optimistic international traffic).(Million euros of 2002)

	Benefits[21] of AVE
Total infrastructure, maintenance and operating costs	
Infrastructure	–4568.0
Residual value	97.3
Rolling stock	–1338.3
Maintenance	–683.4
Operation	–5793.3
Time savings of users diverted from:	
Other mode of transports	3252.4
Generated trips	1978.8
Cost reduction in:	
Conventional railroad	740.7
Airplane	2796.1
Bus	115.2
Car operating costs	1808.0
Congestion	310.2
Accidents	753.4
Environment	595.3
Maintenance	642.7
Net present value of AVE	617.1

23.5 Conclusions

The introduction of the high-speed train brings about a significant generalised cost reduction of the railroad mode. This reduction, which is due to non monetary components of this cost (time, comfort, etc.) gives rise to two clearly distinguishable effects on the demand, which are known as induction and substitution effects. Due to the high magnitude of the substitution effect, the introduction of the high- speed train gives rise to very significant effects on the

[21] It corresponds to a basic value of accidents (0.93 euros, GDP growth of 2.5% and 40 years of duration of the project. In view of the scope of this project, it seems logical to consider this duration period.

demand for the rest of the alternative means of transport competing with it in the Madrid-Barcelona-French border rail corridor. In line with this, it can be stated that railroad is the mode which will prevail in the Madrid-Barcelona route, with a higher market share than that of the air transport.

Therefore, this type of improvement policy for the railroad infrastructure as regards supply with the subsequent slight reduction in the time component of the overall cost of railroad transport proves to be very efficient to alter modal distribution although it leads to a high social cost.

By comparing the results obtained in the Madrid-Sevilla rail corridor (See De Rus and Inglada (1997)) we may observe that in the Madrid-Barcelona-French border corridor the social profitability of AVE is really higher.

But when only the national traffic is taken into account, we have obtained that social costs exceed social benefits in 2,198 million euros of 2002. Even with the most favorable hypotheses to the profitability of this project (an estimated growth of 3.5%, maximum value of the human life, etc.) the net social benefit is negative (-1,027.5 million euros of 2002). The basic reason of this negative result relies on the existence of a very low level of demand which makes the willingness-to-pay for the capacity be lower than the capacity costs themselves.

On the contrary, the estimation of the international potential traffic to be captured by the AVE if the French high-speed rail network is extended to the Spanish border, leads to a considerable change in the results. If we take into account the international demand segment, the net social benefit generated by the Madrid-Barcelona-French border AVE is of 617.1 million euros for the base scenario with 40 years of project lifetime.

It is important to highlight that the sensitivity analysis carried out shows us how heavily the return rate on the project depends on the extension of the French high speed train to the border. In the event that this extension did not take place, the net social benefit would be estimated negative, even if we consider the most favorable hypothesis for the project. In short, this research is an interesting proof that the "network externalities" are very important for the social rate of return on a project.

References

Alvarez, O. and Herce, J. A.: Líneas ferroviarias de alta velocidad en España. Economía Aplicada 1, no 1, 5-32 (1993)

Bonnafous, A.: The regional impact of the TGV. Transportation 14, 127-137 (1987)

Coto-Millán, P., Baños-Pino, J. and Inglada, V.: Marshallian demands of intercity passenger transport in Spain: 1980-1992. An economic analysis. Transportation Research- E 33(2), 79-96 (1997)

De Rus, G. and Inglada, V.: Análisis Coste-Beneficio del Tren de Alta Velocidad en España. Revista de Economía Aplicada 3 (vol. I), 27-48 (1993)

De Rus, G. and Inglada, V.: Cost-Benefit Analysis of the High-Speed Train in Spain. The Annals of Regional Science 31, 175-188 (1997)

Dogson, J.: Railroads Costs and Closures. Journal of Transport Economics and Policy, vol. XVIII, no 3, 219-235 (1984)

Inglada, V.: El papel de las infraestructuras en la competitividad y el desarrollo económico. Estudios Territoriales 97, 397-409 (1993)

Inglada, V.: Análisis empírico del impacto del AVE sobre la demanda de transporte en el corredor Madrid-Sevilla. Revista de Estudios de Transportes y Comunicaciones 62, 35-51 (1994)

Inglada, V.: Competencia intermodal, externalidades y determinación del equilibrio social en el transporte. Mimmeo 2003

Inglada, V and Coto, P.: Introduction of an Innovative Product: The High Speed Train. In: Coto P. (ed.): Essays in Microeconomis and Industrial Organisation, Chapter 3. Springer Verlag, Heidelberg, Germany 2002

MOPT: Manual de Evaluación de Inversiones en Ferrocarril de Vía Ancha. Ministery of Public Works and Transport, Madrid 1991

Nash, C. A.: The case for High Speed Rail. Investigaciones Económicas, vol. XV, no 2, 337-354 (1991)

Owen, A. D. and Phillips, G. D. A.: The characteristics of railroads passenger demand. Journal of Transport Economics and Policy 21, no 3, 231-253 (1987)

RENFE: Encuestas realizadas a los viajeros del AVE. RENFE 1993

TRB: Highway Capacity Manual. Special Report 209, Washington D.C. 1985