INSTITUTO UNIVERSITARIO
de Análisis Económico y Social

Universidad
de Alcalá

www.iaes.es

# Urban income and city size: Ecological Inference with Entropy Econometrics for the Spanish municipalities.

*Esteban Fernández-Vázquez*
*Fernando Rubiera-Morollón*
*Elizabeth Aponte-Jaramillo*

**SERIE** DOCUMENTOS DE TRABAJO
07/2009

## INSTITUTO UNIVERSITARIO DE ANÁLISIS ECONÓMICO Y SOCIAL

DIRECTOR

***Dr. D. Tomás Mancha Navarro***  
Catedrático de Economía Aplicada, Universidad de Alcalá

DIRECTOR FUNDADOR

***Dr. D. Juan R. Cuadrado Roura***  
Catedrático de Economía Aplicada, Universidad de Alcalá

AREAS DE INVESTIGACIÓN

ANÁLISIS TERRITORIAL Y URBANO  
***Dr. D. Rubén Garrido Yserte***  
Profesor Titular de Universidad  
Universidad de Alcalá

ECONOMÍA LABORAL  
***Dr. D. Carlos Iglesias Fernández***  
Profesor Contratado Doctor  
Universidad de Alcalá

ESTUDIOS SECTORIALES, FINANCIEROS Y PYME  
***Dr. D. Antonio García Tabuenca***  
Profesor Titular de Universidad  
Universidad de Alcalá

SERVICIOS E INNOVACIÓN  
***Dr. D. Luis Rubalcaba Bermejo***  
Profesor Titular de Universidad  
Universidad de Alcalá

## WORKING PAPERS

The Institute of Social and Economic Analysis - IAES (before Servilab) edits Working Papers, where are included advances and results of some research projects done as part of the research done by the Institute's staff and other researchers in colaboration with the Institute.

Those Working papers are available in:

http://www.iaes.es/iuaes_sp/publicaciones.htm

**ISSN:1139-6148**

## LATEST WORKING PAPERS

**WP-12/08 CONVERGENCIA REGIONAL EN PRODUCTIVIDAD Y CAMBIOS EN LA ESTRUCTURA PRODUCTIVA**
Juan Ramón Cuadrado Roura y Andrés Maroto Sánchez

**WP-13/08 EVOLUCIÓN RECIENTE DE LA SEGREGACIÓN LABORAL POR GÉNERO EN ESPAÑA**
Carlos Iglesias y Raquel Llorente

**WP-01/09 EL IRPF ANTE UNA ENCRUCIJADA: OPCIONES DE REFORMA**
José M. Domínguez Martínez

**WP-02/09 LA PROYECCIÓN DE LA CRISIS EN ESPAÑA**
Antonio Torrero Mañas

**WP-03/09 LA ACTIVIDAD EMPRESARIAL EN LA COMUNIDAD DE MADRID Y EN SUS PROVINCIAS LIMÍTROFES: CONCENTRACIÓN Y DIFUSIÓN ESPACIAL**
Maria Teresa Gallo Rivera, Rubén Garrido Yserte y Tomás Mancha Navarro

**WP-04/09 LACRISIS DE 2008 Y LA NATURALEZA DE LA POLÍTICA ECONÓMICA**
Xosé Carlos Arias

**WP-05/09 LA CALIDAD DEL EMPLEO EN UN CONTEXTO REGIONAL, CON ESPECIAL REFERENCIA A LA COMUNIDAD DE MADRID**
Diego Dueñas Fernández, Carlos Iglesias y Raquel Llorente

**WP-06/09 LA INTERVENCIÓN DEL SECTOR PÚBLICO EN LA EDUCACIÓN**
José Dominguéz Martínez

INSTITUTO UNIVERSITARIO
de Análisis Económico y Social

Universidad
de Alcalá

Plaza de la Victoria, 2. 28802. Alcalá de Henares. Madrid - Telf. (34)918855225
Fax (34)918855211 Email: iaes@iaes.es. WEB: www.iaes.es

# URBAN INCOME AND CITY SIZE: ECOLOGICAL INFERENCE WITH ENTROPY ECONOMETRICS FOR THE SPANISH MUNICIPALITIES

**ABSTRACT:**

Most of the regional empirical analyses are limited by the lack of data. Researches have to use information which is structured in administrative or political regions not always economically meaningful. Any aggregation of a territory in regions with economic sense requires data of the main economic variables at basic spatial units, like GDP at a local level. In this paper a methodology to approximate local GDP values is proposed using entropy econometrics which can be defined as an exercise of ecological inference. In addition to the analysis of the main characteristics of the techniques proposed, the paper illustrates how the procedure works taking as empirical application the estimation of income for the Spanish municipalities according to their size. As an example of the possibilities open by this methodology a regional classification based on the relevance of the cities size, which allows us to measure the relevance of agglomeration economics, is empirically applied to the Spanish case obtaining some interesting first results.

KEY WORDS: urban size and income relationship; entropy econometrics, ecological inference and Spain

JEL: C15, C21, R11 and R12

# RENTA MUNICIPAL Y TAMAÑO URBANO: INFERENCIA ECOLÓGICA CON MÁXIMA ENTROPÍA PARA LOS MUNICIPIOS ESPAÑOLES.

**RESUMEN:**

La mayoría de los análisis regionales están limitados por la ausencia de datos. Los investigadores tienen que usar la información estructurada en unidades político-administrativas que carecen de sentido económico. Cualquier agregación del territorio se hace sobre la base de estas unidades espaciales administrativas que limitan mucho las posibilidades del análisis. Para poder superar estas limitaciones es preciso enfrentarse a la obtención de datos a escala local. En este trabajo se propone una metodología que puede ser de gran utilidad para superar esta limitación y que nos permite estimar variables como el PIB local mediante máxima entropía. Se propone la técnica que podría ser más apropiada para un ejercicio de inferencia ecológica de este tipo y se ilustran sus posibilidades mediante una aplicación al caso de la economía española estimando los niveles de renta de los municipios españoles. Ello nos permite hacer una clasificación de los mismos atendiendo a su posición y tamaño que nos posibilita hacer un estudio del efecto de las economías de aglomeración y la influencia de la localización en los municipios españoles. Algunos de los resultados permiten valorar la importancia de las grandes metrópolis españolas así como la distancia respecto a las mismas.

PALABRAS CLAVE: Relaciones entre tamaño urbano y renta, máxima entropía, inferencia ecológica y España.

JEL: C15, C21, R11 y R12.

## AUTORES:

ESTEBAN FERNANDEZ-VAZQUEZ. Faculty of Economics, Campus del Cristo, 33006 Oviedo (Spain); Phone: (+34)985105056; e-mail: evazquez@uniovi.es

FERNANDO RUBIERA-MOROLLON. University of Oviedo (Spain)

ELIZABETH APONTE-JARAMILLO. University Autonoma of Occidente (Colombia)

## INDEX

## 1. Introduction: Why is Relevant to Obtain Local Data?

With some relevant exceptions, like USA, data a local level are not normally available. The most important economic data, as the .GDP, are usually presented by administrative or political divisions of the national territories with non-extensive spatial desegregations. If we use this information the empirical possibilities of our studies are clearly limited. But if we are able to use a proper delimitation of the Regions, with an economic meaning taking in account a particular theoretical framework, a specific objective or hypothesis contrast, the analysis will be clear and we will be able to obtain more relevant conclusions (Behrens and Thisse, 2007).

For example, one of the most relevant concepts from the urban economic point of view are the *agglomeration economies* and *diseconomies* and their effect of that over the location decisions, economic structure and growth between larger and smaller cities (Henderson, J.V. and Thisse, J.F., 2004). *Agglomeration economies* could be approximated by the *size* of the main urban area of a particular region. As Melo et al. (2009) show there exist an international regularity that connects the *size* of the main city with the GDP *per capita* average in the area. But these regularities could be tested only for some countries and for some specific cities in which local information are available (see Melo et al., 2009). In most of the cases there isn't enough desegregated information to conduct this type of empirical researches.

Other relevant and more classical concept in the regional studies is the importance of the *distance* and how the transportation costs can affect to the business localization and throw these microeconomic decisions the macroeconomic structure could be transformed and, consequently, the levels of  GDP *per capita* could change. In the literature authors use to test all this ideas with the employment data, that usually is available a local level. But the final contrast referred to the income changes connected with the position of each area is not possible in most of the cases due to this general lack of information at a local level.

Similarly to these two questions we can make others like: how we can evaluate the impact of a regional polity at a local level?, how we can test the relevance of a new infrastructure?, how we can compare the different cities and economical efficiency of the different models of urban growth?.. The list of relevant questions could be extended widely and always we must face to the basic problem of lack of data, especially GDP, at a local level in most of the cases.

The objective of this paper is developing a useful approach to obtain this information based on the entropy econometrics. The technique could give us information at a local level organize by the size of the main urban area which is specially interesting from the point of view of most of the models and analysis of regional and urban economies. Our intention is to propose a procedure that could be applied in different scenarios with minor adaptations. Nevertheless in this first step we test the possibilities of the approach applying it to the Spanish case.

The paper is divided in three more sections. Next section discusses about the entropy econometrics solution to an ecological inference problem and presents our methodological proposal. Section 3 presents an empirical application to Spain for 2001 and discusses the results obtained applying a typical and mean fully set of *Regions*. Main conclusions and further researches complete the paper. Additionally, an appendix reports the outcomes of a Monte Carlo experiment that tests the liability of the empirical results.

## 2. The Methodology: Ecological Inference with Entropy Econometrics.

### 2.1. The Maximum Entropy (ME) and Cross Entropy (CE) solutions to pure inverse problems.

In this section, the basics of Entropy Econometrics will be introduced for estimate unknown probabilities in the context of *pure inverse problems*. More extensive introductions can be found in Kapur and Kesavan (1992), Golan *et al*. (1996) or, much more recently, Golan (2006).

Traditionally, probability has been used as a measure of the uncertainty about an event. Let us assume that this event that can take *K* possible outcomes $E_1,E_2,...,E_K$ with the respective distribution of probabilities $\boldsymbol{p} = [p_1, p_2, .., p_K]$ such that $\sum_{i=1}^{K} p_i = 1.$ Following the formulation of Shannon (1948), the entropy of this distribution $\boldsymbol{p_x}$ will be:

$$H(\boldsymbol{p}) = -\sum_{i=1}^{K} p_i ln p_i \qquad (1)$$

that takes its maximum when $\boldsymbol{p}$ is a uniform distribution ( $p_i = \frac{1}{K}; \ \forall i = 1,..,K$). This entropy measure gives the uncertainty of the outcomes of the event, but this univariate framework can be extended to situations where we are interested in the study of bidimensional distributions given by the pair of variables (*x,y*), where variable *x* can take K different values $\{x_1, x_2, ..., x_K\}$ and variable *y* can take T values $\{y_1, y_2, \dots, y_T\}$. In this situation, the joint probability of a pair of random observations $(x_i, y_j)$ will be denoted as $p_{ij}$ and the Shannon's entropy measure for the $K \times T$ possible outcomes will be:

$$H(\boldsymbol{P}) = -\sum_{i=1}^{K}\sum_{j=1}^{T} p_{ij} ln p_{ij} \qquad (2)$$

Again, the entropy measure reaches its maximum when $\boldsymbol{P}$ is uniform. Apart from measuring the uncertainty associated to a random process, Shannon's entropy can be used for recovering an unknown probability distribution form partial or incomplete data.

We will base our explanations on the matrix-balancing problem depicted in Golan (2006, page 105), where the goal is to fill the (unknown) cells of a matrix using the information that is contained in the aggregate data of the row and column sums. Graphically, the point of departure of our problem is a matrix like Table 1.

Table 1.
**Known and unknown data in a matrix balancing problem.**

| | $z_{\cdot 1}$ | … | $z_{\cdot j}$ | … | $z_{\cdot T}$ |
|---|---|---|---|---|---|
| $z_{1\cdot}$ | $z_{11}$ | … | $z_{1j}$ | … | $z_{1T}$ |
| … | … | | … | | … |
| $z_{i\cdot}$ | $z_{i1}$ | … | $z_{ij}$ | … | $z_{iT}$ |
| … | … | | … | | … |
| $z_{K\cdot}$ | $z_{K1}$ | … | $z_{Kj}$ | … | $z_{KT}$ |

The $z_{ij}$ elements of the matrix are the unknown quantities we would like to estimate, where $\sum_{j=1}^{T} z_{ij} = z_{i\cdot}$, $\sum_{i=1}^{K} z_{ij} = z_{\cdot j}$, and $\sum_{i=1}^{K}\sum_{j=1}^{T} z_{ij} = z$. These elements can be expressed as sets of (column) probability distributions, simply dividing the quantities of the matrix by the corresponding column sums $z_{\cdot j}$. Note that. In such a case, the previous matrix can be rewritten in terms of a new matrix $\boldsymbol{P}$ that is composed by a set of *T* probability distributions (Table 2).

Table 2

**The matrix balancing problem in terms of probabilities.**

|  | $y_1$ | … | $y_j$ | … | $y_T$ |
|---|---|---|---|---|---|
| $x_1$ | $p_{11}$ | … | $p_{1j}$ | … | $p_{1T}$ |
| … | … |  | … |  | … |
| $x_i$ | $p_{i1}$ | … | $p_{ij}$ | … | $p_{iT}$ |
| … | … |  | … |  | … |
| $x_K$ | $p_{K1}$ | … | $p_{Kj}$ | … | $p_{KT}$ |

Where the $p_{ij}$'s are defined as the proportions $\frac{z_{ij}}{z_{.j}}$, and the new row and column margins as $x_i = \frac{z_{i.}}{z}$ and $y_j = \frac{z_{.j}}{z}$ respectively. Consequently, the followings equalities are fulfilled by the $p_{ij}$ elements[1]:

$$\sum_{i=1}^{K} p_{ij} = 1\,;\ \forall j = 1, \dots, T \tag{3}$$

$$\sum_{j=1}^{T} p_{ij} y_j = x_i\,;\ \forall i = 1, \dots, K \tag{4}$$

These two sets of equations reflect all we know about the elements of matrix $\boldsymbol{P}$ . Equation (3) shows the cross-relationship between the (unknown) $p_{ij}'s$ in the matrix and the (known) sums of each row and column. Additionally, equation (4) indicates that the $p_{ij}'s$ can be viewed as (column) probability distributions. Note that we have only $K+T$ pieces of information to estimate the $K \times T$ elements of matrix $\boldsymbol{P}$ , which makes the problem ill-posed. In such a situation, usually called a *pure linear inverse problem,* the Maximum Entropy (ME) principle can be applied to recover the unknown $p_{ij}$ probabilities. This principle is based on the selection of the probability distribution that maximizes (5) among

[1] Note that in such a case, these $p_{ij}$ elements can be seen as conditional probabilities to each column.

all the feasible probability distributions that fulfil (6) and (7). So, the following constrained maximization problem is posed:

$$\underset{\boldsymbol{P}}{\mathrm{Max}}\, H(\boldsymbol{P}) = -\sum_{i=1}^{K}\sum_{j=1}^{T} p_{ij} ln p_{ij} \tag{5}$$

Subject to:

$$\sum_{j=1}^{T} p_{ij}y_j = x_i\,;\ \forall i = 1, \dots, K \tag{6}$$

$$\sum_{i=1}^{K} p_{ij} = 1\,;\ \forall j = 1, \dots, T \tag{7}$$

In this problem the equations (7) are just normalization constraints that guarantee that the estimated probabilities sum to one, and equations (6) ensure that the recovered distributions of probabilities are compatible with the aggregate data of $\boldsymbol{x}$ at all *K* observations. The Lagrangian function for such a problem will be:

$$L = -\sum_{i=1}^{K}\sum_{j=1}^{T} ln p_{ij} + \sum_{i=1}^{K}\lambda_i\left[x_i - \sum_{j=1}^{T} p_{ij}y_j\right] + \sum_{j=1}^{T}\mu_j\left[1 - \sum_{i=1}^{K} p_{ij}\right] \tag{8}$$

And the solutions (taking into account the first-order conditions) are:

$$\hat{p}_{ij} = \frac{exp\left[\hat{\lambda}_i y_j\right]}{\sum_{i=1}^{K} exp\left[\hat{\lambda}_i y_j\right]}\,;\ \forall i = 1, \dots K; j = 1, \dots, T \tag{9}$$

where $\hat{\lambda}_i$ are the Lagrangian multipliers associated with restrictions (6).

Alternatively to this case, it might be also possible a situation where, in addition to the information contained in the aggregate data, we have available a set of prior probabilities $q_{ij}$. In other words, we want to transform an *a priori* probability matrix $\boldsymbol{Q}$ into a posterior matrix $\boldsymbol{P}$ that is consistent with the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. This type of problem is frequent in some fields of economic research: for example in input-output analysis the researchers often must update an input-output matrix of coefficients to make it match with actual known row and column sums, using as *a priori* information the data collected in a previous table.

The solution to this type of problems is obtained by minimizing a divergence measure with the prior probability matrix $\boldsymbol{Q}$ subject to the set of constraints (6) and (7). The ME problem is therefore transformed into a so-called Cross-Entropy (CE) problem, which can be written in the following terms:

$$\underset{\boldsymbol{P}}{\text{Min}}\, D(\boldsymbol{P}\|\boldsymbol{Q}) = \sum_{i=1}^{K}\sum_{j=1}^{T} p_{ij}\, ln\left(\frac{p_{ij}}{q_{ij}}\right) \qquad (10)$$

Subject to the same restrictions given by the set of equations (6) and (7). The divergence measure $D(\boldsymbol{P}\|\boldsymbol{Q})$ is the Kullback-Liebler entropy divergence between the posterior and prior distributions. The Lagrangian function for the CE problem is:

$$L = D(\boldsymbol{P}\|\boldsymbol{Q}) + \sum_{i=1}^{K}\lambda_i\left[x_i - \sum_{j=1}^{T} p_{ij}y_j\right] + \sum_{j=1}^{T}\mu_j\left[1 - \sum_{i=1}^{K} p_{ij}\right] \qquad (11)$$

And the solutions are:

$$\tilde{p}_{ij} = \frac{q_{ij} exp[\tilde{\lambda}_i y_j]}{\sum_{i=1}^{K} q_{ij} exp[\tilde{\lambda}_i y_j]};\ \forall i = 1, \ldots K; j = 1, \ldots, T \qquad (12)$$

The CE estimation procedure can be seen as an extension of the ME principle (or alternatively the ME can be considered as a particular case of the CE procedure), given that the solutions of both approaches are the same ($\hat{p}_{ij} = \tilde{p}_{ij}$) when the $T$ *a priori* probability distribution contained in $\boldsymbol{Q}$ are all uniform. In other words, the ME solutions are obtained by minimizing the Kullback-Liebler divergence $D(\boldsymbol{P}\|\boldsymbol{Q})$ between the unknown $p_{ij}$ and the probabilities $q_{ij} = \frac{1}{K}\ \forall i = 1,..,K$.

## 2.2. The ME-CE approach in the presence of noisy data.

The entropy solutions depicted above to recover unknown probability distributions can be applied also to situations different from the pure inverse problems. Consider a case where, for example, the observations of vector *x* are "contaminated" by some measurement error; or, alternatively, a situation where the *x* values are affected by some uncontrolled factor different from the pure linear relationship with *y*. In both cases, the equation (13) that relates $\boldsymbol{x}$ and $\boldsymbol{y}$ will be affected by the presence of a random disturbance $\boldsymbol{\epsilon}$ in the following terms:

$$x_i = \sum_{j=1}^{T} p_{ij} y_j + \varepsilon_i\,;\ \forall i = 1, \ldots, K \qquad (13)$$

Or, more generally:

$$\boldsymbol{x} = \boldsymbol{P}\boldsymbol{y} + \boldsymbol{\epsilon} \qquad (14)$$

Entropy econometrics can also deal with the estimations of the unknown $p_{ij}$ elements in such situations, which is the typical specification of a linear econometric model[2]. A first step to estimate the $p_{ij}$ probabilities is the reparametrization of the $\varepsilon_i$ terms, given that the CE formulation is designed for dealing with elements that behave as proper probability distributions (condition fulfilled by the $p_{ij}^{\prime}s$ but not for the $\varepsilon_i^{\prime}s$ ). This reparametrization allows us to generalize the use of the CE technique (Generalized Cross Entropy or GCE hereafter) to these familiar linear models.

Oppositely to other estimation techniques, GCE does not require rigid assumptions about a specific probability distribution function of the stochastic component, but it still is necessary to make some assumptions. Basically, we represent our uncertainty about the realizations of vector $\boldsymbol{\varepsilon}$ treating each element $\varepsilon_i$ as a discrete random

[2] This section will focus only on the application of the CE techniques given that, as commented before, the ME solution can be seen as a particular case of the CE approach when $q_{ij} = \frac{\mathbf{1}}{K}\ \forall i = 1,..,K$ .

variable with $J \geq 2$ possible outcomes contained in a convex set $\boldsymbol{v}' = \{v_1, \dots, v_J\}$, which for the sake of simplicity is assumed as common for all the $\varepsilon_i$. We also assume that these possible realizations are symmetric around zero ($-v_1 = v_J$). The traditional way of fixing the upper and lower limits of this set is to apply the three-sigma rule (see Pukelsheim, 1994). Under these conditions, each element $\varepsilon_i$ can be defined as:

$$\varepsilon_i = \sum_{h=1}^{J} w_{ih} v_h \, ; \ \forall i = 1, \dots, K \tag{15}$$

Where $w_{ih}$ is the unknown probability of the outcome $v_h$ for the observation *i*, which implies that $\epsilon$ is assumed to have mean $E[\boldsymbol{\epsilon}] = \mathbf{0}$ and a finite covariance matrix. From this reparametrization, equation (15) can be written as:

$$x_i = \sum_{j=1}^{T} p_{ij} y_j + \sum_{h=1}^{J} w_{ih} v_h \, ; \ \forall i = 1, \dots, K \tag{16}$$

Or, more generally:

$$\boldsymbol{x} = \boldsymbol{P}\boldsymbol{y} + \boldsymbol{W}\boldsymbol{v} \tag{17}$$

Now we need also to estimate a $(K \times J)$ matrix $\boldsymbol{W}$ for the $(1 \times J)$ support vector $\boldsymbol{v}'$. From a matrix $\boldsymbol{W}^0$ of *a priori* probabilities, the CE program depicted before can be rewritten as a GCE in the following terms:

$$\underset{\boldsymbol{P},\boldsymbol{W}}{\text{Min}}\, D(\boldsymbol{P},\boldsymbol{W} \| \boldsymbol{Q},\boldsymbol{W}^0) = \sum_{i=1}^{K} \sum_{j=1}^{T} p_{ij} \, ln\left(\frac{p_{ij}}{q_{ij}}\right) + \sum_{i=1}^{K} \sum_{h=1}^{J} w_{ih} \, ln\left(\frac{w_{ih}}{w_{ih}^0}\right) \tag{18}$$

Subject to:

$$x_i = \sum_{j=1}^{T} p_{ij} y_j + \sum_{h=1}^{J} w_{ih} v_h \, ; \;\; \forall i = 1, \dots, K \qquad (19)$$

$$\sum_{i=1}^{K} p_{ij} = 1 \, ; \; \forall j = 1, \dots, T \qquad (20)$$

$$\sum_{h=1}^{J} w_{ih} = 1 ; \; \forall i = 1, \dots, K \qquad (21)$$

Note that this GCE program comes from introducing in the pure inverse problem the estimation of the unknown probabilities $\boldsymbol{W}$ corresponding to the stochastic term $\epsilon$ . The solutions of the GCE program are:

$$\tilde{p}_{ij} = \frac{q_{ij} exp\left[\tilde{\lambda}_i y_j\right]}{\sum_{i=1}^{K} q_{ij} exp\left[\tilde{\lambda}_i y_j\right]} ; \; \forall i = 1, \dots K ; j = 1, \dots, T \qquad (22)$$

$$\tilde{w}_{ih} = \frac{exp\left[\tilde{\lambda}_i v_h\right]}{\sum_{h=1}^{J} exp\left[\tilde{\lambda}_i v_h\right]} ; \; \forall i = 1, \dots K ; h = 1, \dots, J \qquad (23)$$

Equation (22) presents an identical structure to (12) for the estimated $p_{ij}$ probabilities. Equation (23) shows the CE solution for the estimation of $w_{ih}$ when the *a priori* probabilities are fixed as uniform ( $w_{ij}^{0} = \frac{1}{J} \; \forall h = 1, \dots, J$ ), which is the natural (and most frequently applied) point of departure to reflect the high degree of uncertainty about $\epsilon$ .

### 2.3. Recovering individual characteristics from aggregate data: Ecological Inference based on CE-GCE techniques.

The entropy-based estimation techniques sketched before can be directly applied to the field of Ecological Inference (EI), which can be roughly defined as the attempt to infer individual characteristics from aggregate information. The research in this area has experienced an enormous development in the last years, given its usefulness in many academic disciplines of social science as well as in policy analysis. The foundations of EI were introduced in the seminal works by Duncan and Davis (1953) and by Goodman (1953), whose techniques were the most prominent in the field for more than forty years, although the work of King (1997) implied a substantial development by proposing a methodology that conciliated and extended the approaches taken previously. An extensive survey of recent contributions to the field can be found in King, Rosen and Tanner (2004).

Actually, in one of the chapters of that work, Judge et al. propose the use of information-based estimation techniques in the field of EI, although their proposal is made in a different context (the estimation individual voters' behavior from aggregate election data Peeters and Chasco (2006) also combined entropy econometrics in the context with EI but in a different way to the one proposed in this paper. Roughly speaking, they used GCE for estimating a weighted regression model that allows for recovering characteristics at a regional scale from information at a national level.

To explain how the GCE technique can be applied in the context of EI, consider a geographical area (a country, for example) that can be divided in *T* smaller spatial units (regions). Besides to this first geographical partition, suppose that another division according other characteristic is also possible. Consider that the second criterion applied for this additional partition is a classification of the municipalities that configure the country, obtaining *K* different types of municipalities. In such a context, the objective would be to estimate how a variable is distributed among the regions according to the classification of municipalities, using as information aggregate data. Graphically, this estimation problem can be represented by a grid with the same structure as Table 2.

Table 3.
**A spatial division across regions and type of municipality.**

| | | Regions | | | | |
|---|---|---|---|---|---|---|
| | | $y_1$ | … | $y_j$ | … | $y_T$ |
| Type of municipality | $x_1$ | $p_{11}$ | … | $p_{1j}$ | … | $p_{1T}$ |
| | … | … | | … | | … |
| | $x_i$ | $p_{i1}$ | … | $p_{ij}$ | … | $p_{iT}$ |
| | … | … | | … | | … |
| | $x_K$ | $p_{K1}$ | … | $p_{Kj}$ | … | $p_{KT}$ |

Each one of the $p_{ij}$'s is now defined as the (unknown) proportion of the variable that is allocated in the municipalities of type *i* situated in the region *j*, forming a $(K \times T)$ matrix $\boldsymbol{P}$ with *T* unknown probability distributions. The $(1 \times T)$ row vector *y* represents the regional proportions of the variable and the $(K \times 1)$ column vector *x* shows the national allocation of the variable according to the type of municipality. Note that these two vectors contain the aggregate data existing for the researcher, which our EI estimation will be based on. If an *a priori* set of probability distributions $\boldsymbol{Q}$ is also available, the cross entropy procedures outlined previously can be directly applied.

Note that both the CE technique for pure inverse problem as well as a GCE program that include the presence of a random term are applicable in this context, and it is a decision to be made by the researcher to follow one specific approach. In the first case, we will assume that there is a pure linear relationship between the row and column margins of our matrix, and the following CE program would have to be solved:

$$\underset{\boldsymbol{P}}{\text{Min}}\, D(\boldsymbol{P} \| \boldsymbol{Q}) \tag{24}$$

Subject to:

$$\boldsymbol{x} = \boldsymbol{P}\boldsymbol{y}' \tag{25}$$

$$\boldsymbol{e}'_K\boldsymbol{P} = \boldsymbol{e}'_K \tag{26}$$

Where $\boldsymbol{e_K}$ stands for an appropriate (column) vector of ones. Alternatively, if it seems realistic the inclusion of a random term that affects the observations of vector ***x***, it would be necessary to solve the following GCE program and estimate jointly matrices $\boldsymbol{P}$ and $\boldsymbol{W}$ :

$$\underset{\boldsymbol{P},\boldsymbol{W}}{\text{Min}}\, D(\boldsymbol{P},\boldsymbol{W} \| \boldsymbol{Q},\boldsymbol{W}^0) \tag{27}$$

Subject to:

$$\boldsymbol{x} = \boldsymbol{P}\boldsymbol{y}' + \boldsymbol{W}\boldsymbol{v} \tag{28}$$

$$\boldsymbol{e}'_K\boldsymbol{P} = \boldsymbol{e}'_K \tag{29}$$

$$\boldsymbol{W}\boldsymbol{e}_J = \boldsymbol{e}_J \tag{30}$$

Being $\boldsymbol{e_J}$ the corresponding column vector of ones.

# 3. A FIRST APPLICATION: ESTIMATING 2001 URBAN INCOME IN SPAIN.

## 3.1. Estimation procedure.

Spanish official data on income at a municipal level are not generally available (, so an estimation procedure is necessary.

Spanish municipalities can be posed in similar terms to the matrix balancing problems described in previous sections. Spain is administratively divided in 50 provinces for which data on income is available in the Regional Accounts annually elaborated by the Spanish Statistical Institute (INE). Additionally, from 1998 to 2004 the INE also produced the Continuous Survey on Household Budgets (ECPF), where one can find information of income and expenditure characteristics from a quarterly sample of approximately 8.000 Spanish families[3]. Particularly interesting for our research, the longitudinal files containing the microdata provide annual information about the personal income distribution depending on the type of municipality where the household lived at the time of being surveyed. Specifically, this municipal classification is as appear in Table 4.

Table 4.
**Classification on the Spanish municipalities on the Continuous Survey on Household Budgets.**

| Type of municipality | Description |
|---|---|
| $m_1$ | Capital city of the province (independently on its population) |
| $m_2$ | Municipality with more than 100,000 inhabitants |
| $m_3$ | Municipality with a population between 50,000 and 100,000 |
| $m_4$ | Municipality with a population between 20,000 and 50,000 |
| $m_5$ | Municipality with a population between 10,000 and 20,000 |
| $m_6$ | Municipality with less than 10,000 inhabitants |

Note that this partition of the Spanish municipalities does not correspond exactly with the population size given that the category "capital city" does not reflect exactly the population size. Even so, this classification can be seen as a good indicator of the spatial distribution of income according to the size of the municipalities, given that there is a little number of provinces (Asturias, Cadiz, Pontevedra and Toledo are the only exceptions) where the capital is smaller than some other city on the same province.

[3] More detailed information on these surveys can be found in www.ine.es.

The information sources described above allow for obtaining the row and column margins represented by the vectors ***x*** and ***y*** in Table 3. Vector ***x***, with dimension ($6 \times 1$), contains the proportion of income per type of municipality and the ($1 \times 50$) vector ***y*** with the provincial proportions of income. From these aggregate data, we will apply the entropy-based estimation strategies explained in previous sections to recover the allocation of provincial income depending on the type of municipality for 2001. We have chosen this specific year because this is also the reference year of the most recent census elaborated in Spain[4], which provides information for specifying a natural *a priori* distribution $\boldsymbol{Q}$ based on the provincial distribution of labor per type of municipality. From this point of departure, two parallel estimation procedures have been applied.

Let us first assume that we can pose a pure linear relationship between vectors ***x*** and ***y*** to solve the following CE problem:

$$\underset{\boldsymbol{P}}{\text{Min}}\, D(\boldsymbol{P}\|\boldsymbol{Q}) = \sum_{i=1}^{6}\sum_{j=1}^{50} p_{ij}\, ln\left(\frac{p_{ij}}{q_{ij}}\right) \qquad (31)$$

Subject to:

$$x_i = \sum_{j=1}^{T} p_{ij} y_j \,;\ \forall i = 1, \dots, 6 \qquad (32)$$

$$\sum_{i=1}^{K} p_{ij} = 1 \,;\ \forall j = 1, \dots, 50 \qquad (33)$$

The solution to this CE program is reported in Table 5 for all the Spanish provinces. The income values have been obtained as the respective estimate of $p_{ij}$ multiplied by the total income of province *j*. Note also that, instead of showing the income value, the estimates have been divided by the respective population sizes and we provide results of income per capita (in thousands of Euros). The last column of the table, which is shaded in grey, shows the proportions of income per province. Similarly, the last row is shaded in grey as well, and contains the proportion of income per type of municipality. Note that these proportions correspond to vectors ***y*** and ***x*** respectively, the aggregate information used for the estimation, although rows and column of the

[4] For details about the Spanish Census, see http://www.ine.es/censo2001/infotec.htm.

matrix they have been transposed from their usual position in previous sections in order to fit the tables into the dimension of the pages.

Another possibility is to include a stochastic term in the linear model that relates $\boldsymbol{x}$ and $\boldsymbol{y}$ and transform the pure linear inverse relationship in a more general linear model. This can be justified by the fact that the observations of the proportions are obtained from samples, which implies the possibility that these observations might have been affected by some measurement error. In general, it may be unrealistic to assume that the $\boldsymbol{x}$ and $\boldsymbol{y}$ vectors are perfectly observed, so it seems plausible to consider a model like (17) with both systematic and stochastic components. It turns the CE problem into the following GCE program:

$$\underset{\boldsymbol{P},\boldsymbol{W}}{\text{Min}}\, D\left(\boldsymbol{P},\boldsymbol{W}\middle\|\boldsymbol{Q},\boldsymbol{W}^0\right) = \sum_{i=1}^{6}\sum_{j=1}^{50} p_{ij}\, ln\left(\frac{p_{ij}}{q_{ij}}\right) + \sum_{i=1}^{6}\sum_{h=1}^{3} w_{ih}\, ln\left(\frac{w_{ih}}{w_{ih}^0}\right) \quad (34)$$

Subject to:

$$x_i = \sum_{j=1}^{T} p_{ij} y_j + \sum_{h=1}^{3} w_{ih} v_h\,;\ \forall i = 1,\dots,6 \quad (35)$$

$$\sum_{i=1}^{K} p_{ij} = 1\,;\ \forall j = 1,\dots,50 \quad (36)$$

$$\sum_{h=1}^{3} w_{ih} = 1\,;\ \forall i = 1,\dots,6 \quad (37)$$

The support vector $\boldsymbol{v}'$ contains the possible values contains $J = 3$ elements centred on 0 and is defined as $\boldsymbol{v}' = [-\alpha, 0, \alpha]$. If we perfectly knew the variability present on $\boldsymbol{x}$, a reasonable rule for $\alpha$ is the three-standard deviation rule (Pukelsheim, 1994). However, given our incomplete knowledge, we will follow the proposal made in Golan et al. (1997), where the entropy econometrics techniques are applied to linear models with multinomial data on the dependent variable, and we will use the sample variance of $\boldsymbol{x}$ as an estimate for $\alpha$. The a priori probabilities $\boldsymbol{W^0}$ for the error have been fixed as uniform, as explained before. Table 6, which presents the same structure of rows and columns as Table 5, shows the solutions to this GCE estimation problem[5].

Note that the estimates reported on both tables are very similar, which suggests that the outcomes are relatively robust to changes in the specification of the estimation procedures.

### 3.2. Results discussion.

The results obtained for the Spanish case fit the basic economic theoretical expectations. The highest estimates of GDP's per capita are obtained for the large urban areas, confirming the relevance of agglomeration economies. The bigger the city, the larger the GDP per capita. This is especially clear for the largest cities of the country. The differences between the two main metropolises (Madrid and Barcelona) and the rest of the municipalities including other large cities are remarkable. Small cities and rural areas present lower GDP's per capita with some exceptions for the places located very close to a large metropolis.

To be able to interpret and analyse this results we can apply to these data a typical classification of the territory based on the Coffey and Polèse (1988), Polèse and Champagne (1999), Polèse and Shearmur (2004) and Polèse, Rubiera and Shearmur (2007) approaches, that take into account the *size* and *distance* effects (maybe the two most important effects in business localization and regional growth). First, we can distinguish all the spaces that can be considered as large metropolis. All these studies show a strong tendency of higher grow rates, particularly in strategic economic sectors such as knowledge intensive business services, in and around cities, and more specifically, in and around large metropolitan areas. Then, we can classify the remaining territories according with their distance to one large metropolis as central or peripheral areas. Central and peripheral areas could be classified also taking in account their size. First we can distinguish between urban and rural areas and the first one, urban, could be classified according with their size in different levels. As a result we have five types of regions: (i) Metropolitan Areas (MA), (ii) Urban Central Areas (UCA), (iii) Urban Peripheral Areas (UPA), (iv) Rural Central Areas (RCA) and (v) Rural Peripheral Areas (RPA), taking in account that the UCA and UPA could be classified in different types according with their sizes.

[5] The blank cells in both tables correspond with a type of municipalities that does not exist in a specific province.

Table 5.
**CE estimates of income per type of municipality.**
(thousands €/person)

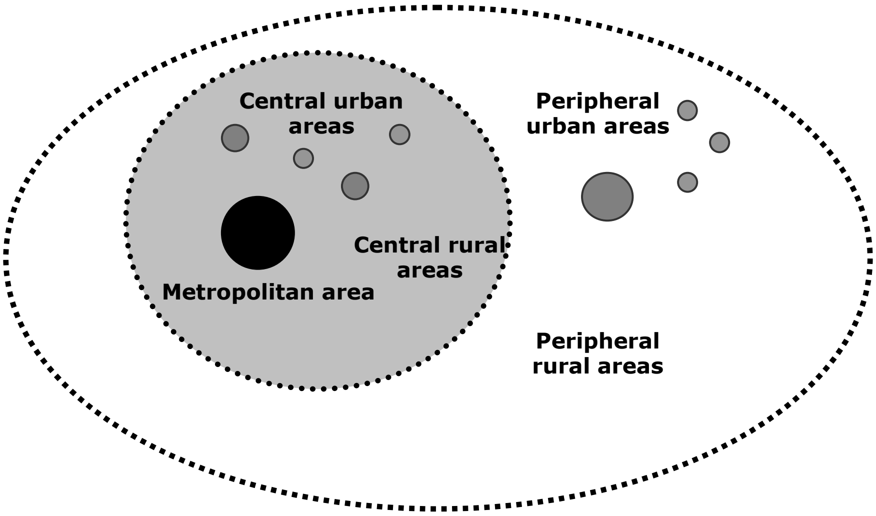
| | $m_1$ | $m_2$ | $m_3$ | $m_4$ | $m_5$ | $m_6$ | $\boldsymbol{y}$ |
|---|---|---|---|---|---|---|---|
| Almeria | 14.36 | | 18.22 | 16.13 | 16.12 | 13.77 | 0.0123 |
| Cádiz | 13.68 | 12.84 | 11.85 | 11.98 | 12.42 | 10.44 | 0.0207 |
| Cordoba | 11.42 | | | 10.83 | 10.75 | 9.67 | 0.0122 |
| Granada | 12.09 | | 12.16 | 8.09 | 12.78 | 10.36 | 0.0134 |
| Huelva | 14.00 | | | 12.55 | 13.18 | 11.81 | 0.0089 |
| Jaen | 12.79 | | 9.75 | 10.52 | 10.93 | 10.06 | 0.0103 |
| Málaga | 13.19 | 12.74 | 5.84 | 13.84 | 11.19 | 10.71 | 0.0240 |
| Sevilla | 16.24 | 12.20 | 11.21 | 7.19 | 10.56 | 10.52 | 0.0319 |
| Huesca | 19.86 | | | | 14.09 | 17.48 | 0.0053 |
| Teruel | 20.34 | | | | 19.78 | 15.51 | 0.0035 |
| Zaragoza | 17.92 | | | | 16.02 | 15.55 | 0.0222 |
| Asturias | 16.75 | 14.26 | 13.36 | 11.88 | 12.76 | 13.76 | 0.0221 |
| Baleares | 21.22 | | | 16.36 | 16.89 | 20.17 | 0.0258 |
| Las Palmas | 15.89 | | 14.76 | 15.95 | 14.76 | 18.33 | 0.0222 |
| Tenerife | 14.21 | 14.61 | 14.68 | 12.58 | 15.68 | 14.62 | 0.0186 |
| Cantabria | 15.87 | | 14.34 | 16.39 | 15.46 | 15.69 | 0.0125 |
| Avila | 14.90 | | | | | 11.87 | 0.0031 |
| Burgos | 18.64 | | | 17.19 | | 17.25 | 0.0093 |
| León | 14.89 | | 13.40 | 16.04 | 9.59 | 13.64 | 0.0100 |
| Palencia | 15.63 | | | | | 14.51 | 0.0039 |
| Salamanca | 14.59 | | | | 13.34 | 12.63 | 0.0070 |
| Segovia | 17.01 | | | | | 15.48 | 0.0035 |
| Soria | 16.89 | | | | | 15.18 | 0.0021 |
| Valladolid | 16.95 | | | 14.83 | 18.98 | 16.39 | 0.0124 |
| Zamora | 13.23 | | | | 12.61 | 10.81 | 0.0035 |
| Albacete | 13.41 | | | 12.29 | 12.16 | 10.84 | 0.0067 |
| Ciudad Real | 15.11 | | 10.95 | 13.81 | 13.63 | 12.25 | 0.0093 |
| Cuenca | 13.99 | | | | 13.53 | 11.84 | 0.0037 |
| Guadalajara | 16.04 | | | 17.31 | | 12.51 | 0.0038 |
| Toledo | 14.75 | | 12.42 | | 17.13 | 12.08 | 0.0104 |
| Barcelona | 28.26 | 16.35 | 13.12 | 13.13 | 13.94 | 22.43 | 0.1423 |
| Girona | 18.36 | | | 18.28 | 19.11 | 20.70 | 0.0173 |
| Lleida | 20.86 | | | | 19.95 | 19.75 | 0.0110 |
| Tarragona | 22.10 | | 19.71 | 19.96 | 19.71 | 20.66 | 0.0191 |
| Alicante | 16.44 | 14.71 | 11.83 | 10.55 | 15.96 | 16.50 | 0.0322 |
| Castellon | 19.14 | | | 18.00 | 18.80 | 17.55 | 0.0135 |
| Valencia | 17.65 | | 13.29 | 12.96 | 15.84 | 15.98 | 0.0523 |
| Badajoz | 12.65 | | 12.27 | 10.96 | 10.90 | 9.21 | 0.0103 |
| Cáceres | 11.90 | | | 10.73 | 7.92 | 10.44 | 0.0064 |
| Coruña | 14.46 | | 12.07 | 13.31 | 12.39 | 12.95 | 0.0215 |
| Lugo | 14.19 | | | | 12.37 | 11.56 | 0.0066 |
| Orense | 13.49 | | | | 12.43 | 10.68 | 0.0060 |
| Pontevedra | 13.80 | 12.98 | | 10.90 | 14.14 | 12.60 | 0.0175 |
| Madrid | 28.74 | 14.15 | 11.73 | 10.63 | 13.45 | 18.05 | 0.1779 |
| Murcia | 14.96 | 12.19 | 13.90 | 13.13 | 12.96 | 12.24 | 0.0244 |
| Navarra | 22.18 | | | 20.50 | 16.10 | 20.07 | 0.0172 |
| Alava | 23.02 | | | | 18.90 | 21.55 | 0.0097 |
| Guipúzcoa | 22.09 | | 20.00 | 20.30 | 20.40 | 21.97 | 0.0213 |
| Vizcaya | 20.59 | | 16.76 | 17.73 | 19.33 | 20.23 | 0.0318 |
| La Rioja | 18.96 | | | 17.74 | 17.42 | 17.39 | 0.0076 |
| $\boldsymbol{x}'$ | 0.4244 | 0.0804 | 0.0679 | 0.1159 | 0.1000 | 0.2115 | |

Table 6.
**GCE estimates of income per type of municipality.**
(thousands €/person)

| | $m_1$ | $m_2$ | $m_3$ | $m_4$ | $m_5$ | $m_6$ | $y$ |
|---|---|---|---|---|---|---|---|
| Almeria | 14.34 | | 18.28 | 16.17 | 16.13 | 13.74 | 0.0123 |
| Cádiz | 13.63 | 12.81 | 11.89 | 12.01 | 12.42 | 10.40 | 0.0207 |
| Cordoba | 11.41 | | | 10.86 | 10.77 | 9.66 | 0.0122 |
| Granada | 12.07 | | 12.21 | 8.12 | 12.81 | 10.35 | 0.0134 |
| Huelva | 13.99 | | | 12.58 | 13.20 | 11.81 | 0.0089 |
| Jaen | 12.78 | | 9.78 | 10.54 | 10.95 | 10.04 | 0.0103 |
| Málaga | 13.15 | 12.74 | 5.87 | 13.90 | 11.21 | 10.68 | 0.0240 |
| Sevilla | 16.20 | 12.20 | 11.31 | 7.24 | 10.61 | 10.49 | 0.0319 |
| Huesca | 19.85 | | | | 14.10 | 17.48 | 0.0053 |
| Teruel | 20.34 | | | | 19.79 | 15.51 | 0.0035 |
| Zaragoza | 17.91 | | | | 16.09 | 15.55 | 0.0222 |
| Asturias | 16.70 | 14.24 | 13.43 | 11.93 | 12.79 | 13.72 | 0.0221 |
| Baleares | 21.16 | | | 16.45 | 16.94 | 20.11 | 0.0258 |
| Las Palmas | 15.84 | | 14.83 | 16.01 | 14.78 | 18.26 | 0.0222 |
| Tenerife | 14.17 | 14.60 | 14.75 | 12.62 | 15.71 | 14.58 | 0.0186 |
| Cantabria | 15.85 | | 14.39 | 16.44 | 15.48 | 15.67 | 0.0125 |
| Avila | 14.90 | | | | | 11.87 | 0.0031 |
| Burgos | 18.63 | | | 17.23 | | 17.24 | 0.0093 |
| León | 14.88 | | 13.44 | 16.08 | 9.60 | 13.63 | 0.0100 |
| Palencia | 15.63 | | | | | 14.51 | 0.0039 |
| Salamanca | 14.59 | | | | 13.36 | 12.63 | 0.0070 |
| Segovia | 17.01 | | | | | 15.48 | 0.0035 |
| Soria | 16.89 | | | | | 15.18 | 0.0021 |
| Valladolid | 16.94 | | | 14.88 | 19.02 | 16.39 | 0.0124 |
| Zamora | 13.23 | | | | 12.62 | 10.81 | 0.0035 |
| Albacete | 13.40 | | | 12.31 | 12.17 | 10.83 | 0.0067 |
| Ciudad Real | 15.09 | | 10.98 | 13.84 | 13.64 | 12.23 | 0.0093 |
| Cuenca | 13.99 | | | | 13.54 | 11.84 | 0.0037 |
| Guadalajara | 16.04 | | | 17.32 | | 12.51 | 0.0038 |
| Toledo | 14.74 | | 12.46 | | 17.16 | 12.07 | 0.0104 |
| Barcelona | 27.89 | 16.35 | 13.64 | 13.54 | 14.19 | 22.13 | 0.1423 |
| Girona | 18.33 | | | 18.35 | 19.15 | 20.67 | 0.0173 |
| Lleida | 20.85 | | | | 19.99 | 19.75 | 0.0110 |
| Tarragona | 22.05 | | 19.80 | 20.03 | 19.75 | 20.61 | 0.0191 |
| Alicante | 16.36 | 14.68 | 11.91 | 10.60 | 15.99 | 16.42 | 0.0322 |
| Castellon | 19.11 | | | 18.04 | 18.83 | 17.52 | 0.0135 |
| Valencia | 17.54 | | 13.46 | 13.09 | 15.92 | 15.88 | 0.0523 |
| Badajoz | 12.64 | | 12.30 | 10.99 | 10.92 | 9.20 | 0.0103 |
| Cáceres | 11.90 | | | 10.74 | 7.93 | 10.44 | 0.0064 |
| Coruña | 14.42 | | 12.13 | 13.36 | 12.41 | 12.91 | 0.0215 |
| Lugo | 14.19 | | | | 12.38 | 11.56 | 0.0066 |
| Orense | 13.48 | | | | 12.45 | 10.68 | 0.0060 |
| Pontevedra | 13.77 | 12.98 | | 10.93 | 14.16 | 12.57 | 0.0175 |
| Madrid | 28.50 | 14.27 | 12.42 | 11.15 | 13.87 | 17.90 | 0.1779 |
| Murcia | 14.91 | 12.17 | 13.98 | 13.18 | 12.99 | 12.19 | 0.0244 |
| Navarra | 22.17 | | | 20.59 | 16.15 | 20.05 | 0.0172 |
| Alava | 23.02 | | | | 18.94 | 21.55 | 0.0097 |
| Guipúzcoa | 22.03 | | 20.10 | 20.38 | 20.44 | 21.90 | 0.0213 |
| Vizcaya | 20.49 | | 16.87 | 17.82 | 19.37 | 20.13 | 0.0318 |
| La Rioja | 18.96 | | | 17.78 | 17.45 | 17.39 | 0.0076 |
| $x'$ | 0.4244 | 0.0804 | 0.0679 | 0.1159 | 0.1000 | 0.2115 | |

Figure 1 presents a schematic representation for an idealized national space economy. The reader will undoubtedly note the resemblance with the classic idealized economic landscapes of Christaller, Lösch, and Von Thünen, all of which posit one metropolis or marketplace at the centre. Thus, Figure 1 shows one metropolis at the centre, but also other smaller "central" urban areas of different population sizes (urban areas close to the metropolis) as well as "central" rural areas (close to the metropolis). Other analogous territories are labelled as "peripheral" urban areas, located at some distance from the metropolis, surrounded by corresponding rural places. It is implicitly assumed that urban areas are distributed in accordance with the rank-size rule.

Figure 1.
**Schematic Representation of the Classification of Spatial Units.**



The result of the Tables 5 and 6 have been plotted in Figure 2, focusing only on the GCE estimates6. The Metropolitan Areas type 1 (MA1) are Madrid and Barcelona. The Metropolitan Areas type 2 (MA2) are the rest of large cities with more than five hundred thousand inhabitants. These cities have been identified knowing the province and the types of municipalities of tables 5 and 6. The rest of the municipalities are classified in UA1 and UA2. The UA1 are cities bigger than one hundred thousand inhabitants and UA2 cities bigger than ten thousand inhabitants but smaller than one hundred thousand. Finally rural areas (municipalities with less than ten thousand inhabitants) are labeled as RA. The solid line represents the values of the areas located close (less than one hour by driving) to a large metropolis (MA1 or MA2), and they are labeled as central areas (CUA1, CUA2 and CRA). The dotted line represents the values of the cities located far away (approximately more

[6] Results for the CE estimates are very similar; any relevant conclusion does not change.

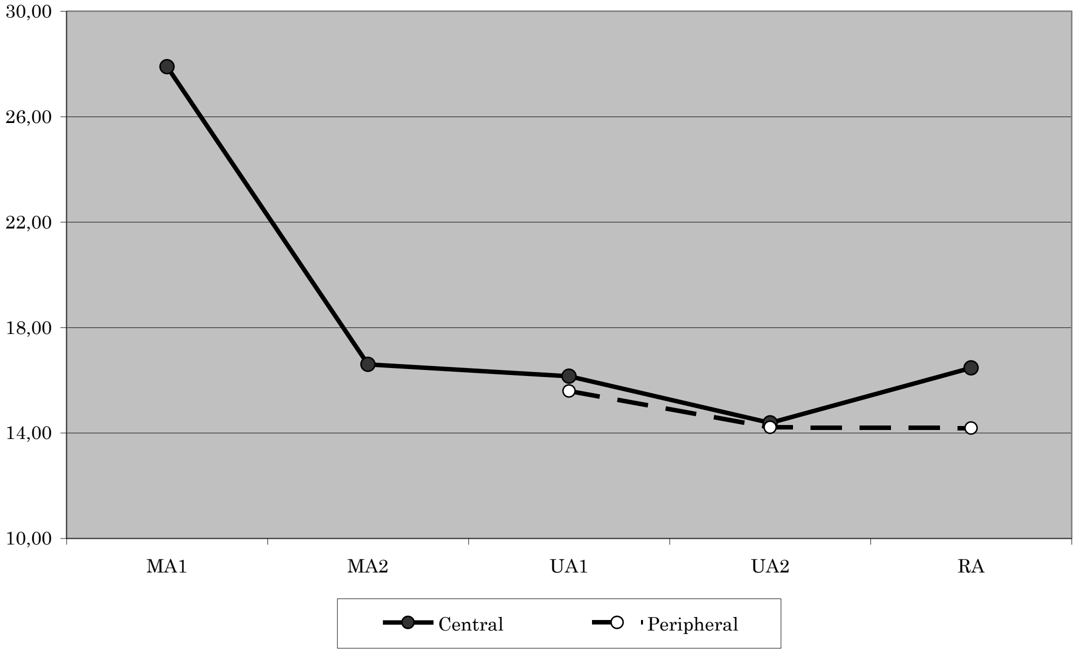
than one hour by driving7), and they are labeled as peripheral areas (PUA1, PUA2 and PRA).

It seems that the most relevant factor to explain spatial differences in GDP per capita is the size of the main urban center of the territory. The main metropolises present the highest values by far (approximately two times the average of the rest of the country). The differences between big and medium size cities are less relevant and being almost irrelevant to be located close or far from a large metropolis. The continuous line is always above the dotted one. This indicates that being a medium size city closely located to a large metropolis allows the attraction of part of the growth of the metropolis, due to the expulsion from the large cities of space intensive activities like manufacturing plants. However, the differences explained by the distance are especially clear in rural areas. The peripheral rural areas have a significantly lower GDP per capita than the central rural areas, given that they are usually residential areas with many commercial services while the first ones are based on agricultural activities.

These first results confirm the relevance of agglomeration economies, not only to the large cities also to the medium and small size cities and even rural areas locates close to a large metropolis. Nevertheless, more precise and extended in time analyses are needed to confirm and estimate the real role of size and distance. In any case these analyses could be possible with the extension of the methodology proposed in this paper to estimate the GDP for several years and other relevant variables in the empirical regional studies.

[7] In Polese, Rubiera and Shearmur (2007) a delimitation of all the Spanish territory as central or peripheral was made at a local level. Most of the municipalities can be classified knowing only the size and province in which they are located. Although it is an approximation because some cases are miss-classified, in any case they are urban areas with a relevant size. In consequence, no fundamental change is expected in the basic results due to these cases that we are not able to identify clearly.

Figure 2.
**GCE estimates of income per type of municipality averages representation.**
Thousands €/person



## 4. Conclusions and Future Research.

One of the main problems of the regional studies is related with the difficulty of work with spatial classifications with economic sense. Normally the data are not available a local level and researchers economics must deal with databases structured with political or administrative criteria. Certain aggregations of the information are possible but usually it is not enough to construct sets of regions economically meaningful.

This paper proposes a methodology based on the Entropy Econometrics to estimate data at a local level according with the size of each basic spatial unit. This estimation exercise allows the inference of information of regions grouped following a clear economic criterion: the urban size. Particularly, we propose a specific classification that considers jointly the urban size an distance from the main metropolis. This allows the measurement of agglomeration economics and location effects.

The methodology proposed is applied to Spain obtaining data at local level of GDP per capita for the 2001 year. The results obtained are in line with previous works by other authors to particular cases (see, for example, Chasco (2003) and Chasco and López (2004)): the larger the city the higher, the local GDP per capita. The size effect is especially clear in the biggest cities, Madrid and Barcelona. The position seems quite relevant too. The cities located close to a large metropolis present higher GDP's per person than the ones located far away; which also holds for the case of rural areas.

The methodology is tested using a Monte Carlo simulation experiment which concludes that this way of estimate local data are reasonably reliable.

This methodology opens wide possibilities that could be explored in following researches. In this paper we focused on the proposal of the methodology and tested their possibilities with a real world example. to the inclusion of a temporal dimension, where this estimation exercise could be carried out for several years would allow not only the estimation of GDP differences, but the evolution and growth of the different areas. Convergence analysis will be possible too using different regional classifications; which  would be useful to test if the convergence trends identified for the Spanish regions are maintained when regions are constructed according with not an administrative but an economic criterion.

## 5. REFERENCES.

BEHRENS, K. AND THISSE, J.F. (2007): "Regional Economics: a New Economic Geography Perspective", *Regional Science and Urban Economics*, 37, pp. 457-465.

COFFEY, W. AND M. POLÈSE (1988): *Locational Shifts in Canadian Employment, 1971-1981: Decentralization versus decongestion*, *Geographica*, *The Canadian Geographer/Le géographe.canadien,* 32 (3), pp. 248-256.

DUNCAN, O. D. AND B. DAVIS (1953): "An Alternative to Ecological Correlation," American Sociological Review, 18, pp. 665–666.

CHASCO, C. (2003): *Predicción–extrapolación espacial de datos microterritoriales,* Tesis Doctoral, Consejería de Economía e Innovación Tecnológica Comunidad de Madrid.

CHASCO, C. Y LÓPEZ, F. (2004b): "Modelos de regresión espacio temporales en la estimación de la renta municipal: el caso de la región de Murcia", *Estudios de Economía Aplicada*, 22 (3), pp. 605-629.

GOLAN, A. (2006): "Information and Entropy Econometrics. A review and synthesis", Foundations and Trends in Econometrics, 2, pp. 1-145.

GOLAN, A. JUDGE, G. AND D. PERLOFF (1997): "Estimation and inference with censored and ordered multinomial response data", Journal of Econometrics, 79, pp. 23-51.

GOLAN, A., JUDGE, G. AND D. MILLER, (1996): Maximum Entropy Econometrics: Robust Estimation with Limited Data, New York, John Wiley & Sons.

GOODMAN, L. (1953): "Ecological Regressions and the Behavior of Individuals," American Sociological Review, 18, pp. 663–666.

HENDERSON, J.V. AND THISSE, J.F. (2004): *Handbook of Regional and Urban Economics*, 4, North Holland, Amsterdam.

JUDGE, G., MILLER, D. J. AND W. T. K. CHO (2004): "An Information Theoretic Approach to Ecological Estimation and Inference", in King, G., Rosen, O. and M. A. Tanner (Eds.): Ecological Inference: New Methodological Strategies, Cambridge University Press, pp. 162-187.

KAPUR, J. N. AND H. K. KESAVAN, (1992); Entropy Optimization Principles with Applications. Academic Press. New York.

KING, G., ROSEN, O. AND M. A. TANNER (2004): Ecological Inference: New Methodological Strategies, Cambridge University Press. Cambridge, UK.

MELO, P.C., GRAHAM, D.J. AND NOLAND, R.B. (2009): "A meta-analysis of estimates of urban agglomeration economies" Regional Science and Urban Economics, 39(3), pp. 332-342.

PEETERS, L. AND CHASCO, C. (2006): "Ecological inference and spatial heterogeneity: an entropy-based distributionally weighted regression approach" Papers in Regional Science, 85(2), pp. 257-276, 06.

POLÈSE, M. Y CHAMPAGNE E. (1999): "Location matters: comparing the distribution of economic activity in the Mexican and Canadian urban systems", *International Journal Science Review*, 22, pp. 102-132.

POLÈSE, M. Y SHEARMUR, R. (2004): "Is distance really dead? Comparing industrial location patterns over time in Canada", *International Regional Science Review*, 27 (4), pp. 1-27.

POLÉSE M., SHEARMUR, R. AND RUBIERA, F. (2006): "Observing regularities in location patters. An analysis of the spatial distribution of economic activity in Spain", *European Urban and Regional Studies*, 14 (2), pp. 157-180.

PUKELSHEIM, F. (1994): "The three sigma rule", The American Statistician, 48 (2), pp. 88-91.

SHANNON, C.E., (1948): "A Mathematical Theory of Communication", Bell System Technical Journal, Vol. 27, pp. 379-423.

## APPENDIX: TESTING THE RESULTS BY A NUMERICAL EXPERIMENT.

Although the general properties of the CE-GCE estimators have been largely studied in the literature (see for example Golan et al., 1996, or Golan, 2006), some doubts about the accuracy of the specific estimates reported in the paper might emerge. In order to test if the entropy-based techniques applied in the section 5 of the paper perform well in such conditions, a simple numerical experiment has been carried out. The goal of this exercise is to get some empirical evidence on the performance of the CE and CGE approaches to estimate a unknown ( $6 \times 50$ ) matrix $\boldsymbol{P}$ of probabilities from aggregate data and some *a priori* matrix $\boldsymbol{Q}$ .

Our Monte Carlo experiment will depart from the actual vector $\boldsymbol{y}$ of proportions of income for the Spanish provinces in 2001 and it is kept fixed along the simulations. Additionally, in each trial of the simulation a randomly generated matrix $\boldsymbol{P}$ is obtained; which is composed by elements $p_{ij}$ that have been drawn from a uniform distribution as $p_{ij}\sim U[0,0.2];\ i = 1, \dots, 5;$ and $p_{6j} = 1 - \sum_{i=1}^{5} p_{ij}$ in order to assure that they behave as a set of proper (column) probability distributions. Based on the linear relationship $\boldsymbol{x} = \boldsymbol{P}\boldsymbol{y}'$, vector $\boldsymbol{x}$ is obtained in each trial, and together with the observations of vector $\boldsymbol{y}$, it represents the aggregate data to obtain the estimates of the (now assumed) unknown matrix $\boldsymbol{P}$ . Another important piece in the estimation process is the choice of the matrix $\boldsymbol{Q}$ . To reflect the idea that the specification of this *a priori* matrix can be more or less similar to the matrix $\boldsymbol{P}$ , in our experiment the cells of $\boldsymbol{Q}$ have been generated from $\boldsymbol{P}$ and a random disturbance $\boldsymbol{u}$ in the following way[8]:

$$\left.\begin{aligned} q_{ij} = (p_{ij}) \cdot (u_{ij});\ \forall i = 1, \dots, 5;\ \forall j = 1, \dots, 50 \\ q_{6j} = 1 - \sum_{i=1}^{5} p_{ij};\ \ \forall j = 1, \dots, 50 \end{aligned}\right\} \qquad (38)$$

[8] This approach is based on the experiment carried out in Golan et al. (1996, pages 63 and 64). To avoid undesirable negative values on $q_{ij}\ \ \forall i = 1, \dots, 5;$ where the number generation obtained a negative, it has been replaced by $q_{ij} = 10^{-8}$ .

Where $u \sim N(1,\sigma)$ and being $\sigma$ a scalar. Note that if $\sigma = 0$ , then $p_{ij} = q_{ij}$ for all the cells of both matrices. The bigger the value of $\sigma$ , the larger the divergence between matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$, and consequently, the smaller the expected accuracy of the estimation. This consequence is rather logical, given that a good specification of the $\boldsymbol{Q}$ matrix (close to the real $\boldsymbol{P}$ matrix) will be helpful in the estimation process. On the contrary, if the $\boldsymbol{Q}$ chosen differs significantly from the actual $\boldsymbol{P}$ the data observed in the sample (the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ ) will have more difficulties to lead the estimates to solutions close to the real values.

In the experiment six different scenarios have been simulated for several values of the scalar $\sigma$ : 0.1, 0.2, 0.25, 0.35, 0.4 and 0.5. Both the CE and the GCE (applying in this last case the three-sigma rule for the support of the error term) solutions have been obtained under these levels of divergence between $\boldsymbol{P}$ and $\boldsymbol{Q}$ . In each one of these six scenarios 1,000 trials have been carried out and the average of two overall measures of error have been computed: the root of the mean squared error (RMSE), which has been obtained as

$$RMSE = \sqrt{\frac{1}{50 \times 6}\sum_{i=1}^{6}\sum_{j=1}^{50}(\tilde{p}_{ij} - p_{ij})^2}$$

, and the mean absolute error (MAE), defined as

$$MAE = \frac{1}{50 \times 6}\sum_{i=1}^{6}\sum_{j=1}^{50}|(\tilde{p}_{ij} - p_{ij})|$$

, where $\tilde{p}_{ij}$ stands for both the CE and GCE estimates. The Table A1 shows the results of these error measures.

Table A1.
**Error measures in the Monte Carlo simulation.**

| CE estimation | $\sigma = 0.5$ | $\sigma = 0.4$ | $\sigma = 0.35$ | $\sigma = 0.25$ | $\sigma = 0.2$ | $\sigma = 0.1$ |
|---|---|---|---|---|---|---|
| RMSE | 0.005 | 0.003 | 0.003 | 0.001 | 0.001 | 0.000 |
| MAE | 0.049 | 0.040 | 0.035 | 0.025 | 0.020 | 0.010 |
| GCE estimation | $\sigma = 0.5$ | $\sigma = 0.4$ | $\sigma = 0.35$ | $\sigma = 0.25$ | $\sigma = 0.2$ | $\sigma = 0.1$ |
| RMSE | 0.072 | 0.059 | 0.052 | 0.037 | 0.030 | 0.015 |
| MAE | 0.050 | 0.040 | 0.036 | 0.026 | 0.021 | 0.010 |

As expected, the error measure are (slightly) larger in all cases if we apply a GCE estimation program compared with the estimates obtained a CE approach. This result is not surprising, given that the GCE allows for the presence of an error term that prevents an exact match between the row and column margins through the estimate of matrix $\boldsymbol{P}$. Moreover, the deviations between real and estimated $p_{ij}$ elements increase as the divergence between the a priori $\boldsymbol{Q}$ and the real matrix $\boldsymbol{P}$ get bigger. Although the RMSE measure seems more sensitive to the specification choice between a pure CE or a GCE estimation program, both error measures RMSE and MAE kept in moderate levels even for considerably big values of the scalar $\sigma$ .

These outcomes give a rough idea on the size of the error that presumably our empirical application on section 5 can present. If we compare the distribution of income per province with the provincial distribution of labor in the census (both taken in 2001) by means of a quotient, which is similar to the $\boldsymbol{u}$ disturbance considered in the Monte Carlo experiment, we obtain a ($50 \times \mathbf{1}$) vector that behaves approximately as a normal distribution and with a sample standard deviation of 0.19. This result suggests that the estimates obtained for the local per capita income, based on the estimates of the unknown $p_{ij}$ elements, for the case of Spain can be taken as reasonably reliable.

## AUTHORS

### Esteban Fernández Vázquez

Profesor contratado doctor en el Departamento de Economía Aplicada de la Universidad de Oviedo. Obtuvo el grado doctor en Economía por esa Universidad en octubre de 2004 con mención de doctorado europeo como consecuencia de su estancia pre-doctoral en la Universidad de Groningen (Holanda), centro en el que ha disfrutado también de estancias postdoctorales. Sus líneas de investigación se centran en la estimación y explotación de modelos económicos, particularmente en el campo de la econometría aplicada a la modelización regional y espacial y en el análisis input-output. Ha publicado artículos en revistas como *Economic Systems Research, International Regional Science Review, Annals of Regional Science, Energy Economics,* Revista de economía Aplicada, Revista de Economía Mundial, entre otras.

### Fernando Rubiera Morollón

Profesor titular de universidad en el Departamento de Economía Aplicada de la Universidad de Oviedo. Doctor por dicha universidad en 2003, premio extraordinario, ha realizado varias estancias entre la que destacan la realizada en el *SAREL* (*Spatial Analysis and Regional Economics Laboratory*) de Montreal (Canadá). Ha obtenido diversos premios de investigación, becas y ayudas. Sus líneas de investigación se centran en la economía regional y urbana y el análisis de la influencia en el desarrollo de los territorios de los servicios avanzados y las grandes ciudades, campos en los que es autor de varios libros y artículos en revistas como *European Urban and Regional Science, The Services Industries Journal,* Investigaciones regionales, Revista de economía, Papeles de economía española, entre otras. Recientemente ha publicado el libro "Economía regional y urbana: introducción a la geografía económica" (Thomson-Civitas) junto con Mario Polèse.

### Elizabeth Aponte Jaramillo

Profesora titular en la Facultad de Ciencias Económicas y Administrativas de la Universidad Autónoma de Occidente (Colombia). Magíster en Economía de la Universidad Nacional de Colombia (1999). Las líneas de investigación se relacionan con la Teoría y la Política Económica, el Desarrollo Regional y el Comercio Internacional. Derivado de los diferentes proyectos de investigación realizados, durante los años más recientes las publicaciones obedecen a documentos, artículos, libros y capítulos de libro en la Universidad Autónoma de Occidente, el Banco de la República (Colombia), la Cámara de Comercio de Cali (Colombia), la Red Latinoamericana de Cooperación Universitaria (Argentina) y la Universidad Austral de Chile. Actualmente es becaria del Instituto Colombiano para el Desarrollo de la Ciencia y la Tecnología (Colciencias) y *Academic and Professional Programs for the Americas* (*Laspau*), organismo afiliado con *Harvard University*, para la realización del Doctorado Economía y Sociología de la Globalización en la Universidad de Oviedo (España).