Macrothink Institute™

Research in Applied Economics
ISSN 1948-5433
2013, Vol. 5, No. 4

Estimating Spatially Disaggregated Data by Entropy Econometrics: An Exercise of Ecological Inference for Income in Spain

Esteban Fernandez-Vázquez[1], Fernando Rubiera-Morollón[1,*] & Elizabeth Aponte-Jaramillo[2]

[1]REGIO*lab* - Regional Economics Laboratory, Business and Economics Faculty, University of Oviedo, Avda. del Cristo s/n, 33006 – Oviedo (Asturias), Spain

[2]Economics and Administration Sciences Faculty, Universidad Autonoma de Occidente, Santiago de Cali (Valle) Colombia

*****Corresponding author: E-mail: frubiera@uniovi.es

Received: October 3, 2013 Accepted: November 6, 2013 Published: November 20, 2013

doi:10.5296/rae.v5i4.4373 URL: http://dx.doi.org/10.5296/rae.v5i4.4373

Abstract

The availability of geographically disaggregated data, especially referred to the urban and metropolitan areas, is a growing need not only for academic studies in the field of economics but also for policy makers. However, in many cases the degree of disaggregation of official statistics does not allow to have information at a desirable level. In this paper a methodology to approximate highly-disaggregated data for the Spanish economy using entropy econometrics is proposed. The paper illustrates how the procedure works taking as empirical application the estimation of income for the Spanish municipalities classified according to their size. An evaluation of the estimates is presented by a simulation exercise and by comparing our results with previous estimates obtained by statistical agencies using more information-intensive estimation techniques. Our results suggest that entropy estimators could be considered as an alternative for recovering disaggregated economic data from aggregate figures, given that the errors seem relatively low.

Keywords: ecological inference; entropy econometrics; geographically disaggregated data and Spain

1. Introduction

One non-unusual limitation for empirical urban economics analysis is the lack of information at a highly disaggregated level. With some exceptions, such as the U.S. or Canada in America or France in Europe, disaggregated data of GDP or income are not normally available, which can lead to the so-called Ecological Fallacy problem(Note 1). This lack of information is special evident in some countries in which the statistical develop is still in process of improvement.

The non-availability of geographically disaggregated information prevents to obtain empirical evidence required to answer some relevant questions in the field of economics. For example: how agglomeration economies and diseconomies affects regional growth(Note 2), how much important is the structure of city (Duraton and Puga, 2002) or its local infrastructures (Elbers and McMiller, 1999), what are the effects of economic specialization or diversification, or what are the effects of local policies? (Thomas and Bromley, 2000). The theoretical literature has paid attention to these issues, but empirical analysis is often limited for the lack of data with a convenient spatial disaggregation.

The objective of this paper is to suggest an estimation procedure, based on entropy econometrics, which allows for inferring disaggregated information on income from more aggregated data. We illustrate the implementation of our proposal with an estimation exercise of income at municipal level in Spain.

The paper is divided into three further sections. The next section summarizes the entropy econometrics solution to the estimation problem and shows the main characteristics of the methodological proposal. In section 3 an application to estimate disaggregated income for a classification of municipalities for Spain in 2001 is presented. To evaluate the capacity of this estimation a Monte Carlo simulation exercise is proposed and we discuss the results obtained and compare them with other previous studies that applied different approaches. The main conclusions are summarizing in a last Section of conclusions and future research lines.

2. The Methodology: Ecological Inference by Entropy Econometrics

In this section, the basics of Entropy Econometrics will be introduced for estimate unknown probabilities in the context of *pure inverse problems(Note 3*).

2.1 The Maximum Entropy (ME) and Cross Entropy (CE) Solutions to Pure Inverse Problems

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:

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

that takes its maximum when $\boldsymbol{p}$ is a uniform distribution ($p_i = \frac{1}{K};\ i = 1,..,K$). 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 $\{x_1, x_2, \dots, x_K\}$ and variable *y* can take T values $\{y_1, y_2, \dots, y_T\}\{y_1, y_2, \dots, y_T\}$.

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:

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

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 (Golan, 2006; page 6), 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 the previous matrix can be rewritten in terms of a new matrix 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} 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 (note that in such a case, these p_{ij}elements can be seen as conditional probabilities to each column):

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

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

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 all the feasible probability distributions that fulfil (6) and (7).

So, the following constrained maximization problem is posed:

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

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:

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

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

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

where $\mathcal{L}$ 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:

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

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:

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

And the solutions are:

$$\tilde{p}_{ij} = \frac{q_{ij} exp\left[\mathcal{L} y_j\right]}{\sum_{i=1}^{K} q_{ij} exp\left[\mathcal{L}_i y_j\right]};\ \forall i = 1, \dots K; j = 1, \dots, T \quad (12)$$

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:

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

Or, more generally:

$$x = Py + \epsilon \quad (14)$$

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. 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 $q_{ij} = \frac{1}{K}\ \forall i = 1, .., K$.

A first step to estimate the p_{ij} probabilities is the reparametrization of the ε_i terms, given that the CE formulation is designed for dealing with elements that behave as proper probability distributions (condition fulfilled by the $p_{ij}s$ but not for the $\varepsilon_i s$). 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 ε_i as a discrete random variable with $J \geq 2$ possible outcomes contained in a convex set $\boldsymbol{v} = \{v_1, \ldots, v_J\}$, which for the sake of simplicity is assumed as common for all the ε_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. Under these conditions, each element ε_i can be defined as:

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

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

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

Or, more generally:

$$x = Py + Wv \tag{17}$$

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:

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

Subject to:

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

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

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

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 $\boldsymbol{\epsilon}$. The solutions of the GCE program are:

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

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

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 ($w_{ij}^0 = \frac{1}{J}\ \forall h = 1, .., J$), which is the natural (and most frequently applied) point of departure to reflect the high degree of uncertainty about $\boldsymbol{\epsilon}$.

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 recent works (King, 1997) implied a substantial development by proposing a methodology that conciliated and extended the approaches taken previously(Note 4).

Actually, in one of the chapters of that work, Judge et al. (2004) 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:

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

Subject to:

$$x = Py' \tag{25}$$

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

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 $\boldsymbol{W}$:

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

Subject to:

$$x = Py' + Wv \tag{28}$$

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

$$We_J = e_J \tag{30}$$

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

3. An Application to Estimate Urban Income in Spain According to City Size

3.1 Application to the Spanish Data

Spanish official data on income at a municipal level are not generally available, but and the subsequent estimation problem can be posed in similar terms to the matrix balancing described in the second section. 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(Note 5). Particularly interesting for our research, the longitudinal files containing the micro-data provide annual information about the personal income distribution according to the type of municipality. Table 4 shows this municipal classification.

Table 4: Classification on the Spanish municipalities on the continuous survey on household budgets

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

The information sources described above allow for obtaining the row and column margins represented by the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in Table 2. Vector $\boldsymbol{x}$, with dimension (6×1), contains the proportion of income by type of municipality and the (1×50) vector $\boldsymbol{y}$ with the provincial proportions of income. From these aggregate data, we will apply the entropy-based estimation strategy explained in previous sections to recover the allocation of provincial income according to 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, which provides information for specifying a natural a priori distribution $\mathbf{Q}$ based on the provincial distribution of labor by type of municipality. From this point of departure, let us assume a pure linear relationship between vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ to solve the following CE problem:

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

Subject to:

$$x_i = \sum_{j=1}^{T} p_{ij} y_j\,;\ \forall i = 1,\ldots,6 \tag{32}$$

$$\sum_{i=1}^{K} p_{ij} = 1\,;\ \forall j = 1,\ldots,50 \tag{33}$$

The solution to this CE program is presented 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 that the estimates have been divided by the respective population to provide results of income per capita (in thousands of Euros).

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

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

Given the lack of official geographically disaggregated data on income in Spain, there have been some previous attempts of estimating this variable. Usually, these estimates are based on some regression model that requires information on a set of regressors observable at the same level of disaggregation as the income itself. In order to check if our estimates were consistent with previous empirical findings, we would compare our CE estimates with these previous results. Although they differ in the specific region studied, the year of reference and/or the methodology used, the comparison with some of them could be still interesting. We opted for taking into consideration three different studies realized in years nearest as possible to the reference time period of our application. One of them is the work by Chasco and Lopez (2004), who estimated by means of spatial econometric models the income per capita of the municipalities located in Murcia for 2001. Besides, the Statistical Institute of Aragon made an estimation of income of municipalities of the region for 2000. Finally, SADEI (the Statistical Office of Asturias) made the same exercise for 2000. Table 6 presents the results for those municipalities that we are able to identify, given the classification used in this paper. The comparison with our estimates suggest that the results obtained by the CE approach are in line with these previous studies, given the reduced size of the differences (around 10% for the region of Murcia, but ranging between 5% and 7% for the cases of Asturias and Aragon).

Table 6: Income per capita previous studies and comparisons with our results (thousands €/person)

Province	**Previous studies**		**Our estimations**
Municipality	Year	Result	**(2001)**
Asturias			
Oviedo	2000	11,468	12,371
Gijón		10,378	10,534
Avilés		9,191	9,883
Aragón			
Huesca	2001	15,383	15,326
Teruel		13,343	14,093
Zaragoza		12,788	13,653
Murcia			
Murcia	2001	9,076	10,771
Cartagena		10,090	8,799

3.2 Testing the methodology by a numerical experiment

Even when 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), and when our estimation results seem to be close to previous estimates, 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 previous section 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 (6×50)

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 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 x is obtained in each trial, and together with the observations of vector 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 (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 $q_{ij} = 10^{-8}$):

$$\left.\begin{aligned} q_{ij} &= (p_{ij}) \cdot (u_{ij});\ \forall i = 1, \dots, 5;\ \forall j = 1, \dots, 50 \\ q_{6j} &= 1 - \sum_{i=1}^{5} p_{ij};\ \ \forall j = 1, \dots, 50 \end{aligned}\right\} \qquad (31)$$

where $\boldsymbol{u} \sim N(1, \sigma)$ and being σ a scalar. Note that if $= 0$, then $p_{ij} = q_{ij}$ for all the cells of both matrices. The bigger the value of σ, 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 σ: 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 $RMSE = \sqrt{\frac{1}{50\times 6}\sum_{i=1}^{6}\sum_{j=1}^{50}\left(\tilde{p}_{ij} - p_{ij}\right)^2}$, and the mean absolute error (MAE), defined as $MAE = \frac{1}{50\times 6}\sum_{i=1}^{6}\sum_{j=1}^{50}\left|\left(\tilde{p}_{ij} - p_{ij}\right)\right|$, where $\tilde{p}_{ij}$ stands for both the CE and GCE estimates. The Table 7 shows the results of these error measures.

Table 7: 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 σ.

These outcomes give a rough idea on the size of the error that presumably our empirical application on section 3 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 (50×1) 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.

4. Final Remarks and Future Research Lines

The availability of geographical disaggregated data, especially referred to the urban and metropolitan areas, is a growing need not only for academic studies but also for policy makers. Nevertheless, in most of the cases the degree of disaggregation of official statistics does not allows having information at that level. This paper proposes a methodology based on Entropy Econometrics to estimate data from aggregated information.

We apply it to the Spanish economy in which disaggregated local income data are not available. The results obtained for the 2001 year are in line with previous works applied for some specific cities of Spain. It is also checked using a Monte Carlo simulation that shows that this procedure do not make significant errors in the estimations.

Some basic ideas could already be observed in the Spanish economy with the obtained data. There exists, for example, important differences inside the regions among the urban and rural areas and, even, among different cities. Normally we can contrast that when the bigger is the city the higher is the aggregate income. From this estimations are now quite easy obtain

series of income desegregated and analyze the evolution. It is also possible to calculate the labor productivity and observe how it changes with the presence of agglomeration economies. The procedure could also be applied to other relevant data and have an opportunity to check the geographical economy in cases like Spain in which local information is not normally available.

References

Chasco, C., & López, F. (2004). 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), 605-629.

Dumedah, G., Schuurman, N., & Wanhong, Y. (2008). Minimizing effects of scale distortion for spatially grouped census data using rough sets. *Journal of Geographical Systems, 10*, 47–69. http://dx.doi.org/10.1007/s10109-007-0056-y

Duncan, O. D., & Davis B. (1953). An Alternative to Ecological Correlation. *American Sociological Review, 18*, 665–666. http://dx.doi.org/10.2307/2088122

Duque, J.C., Artís, M., & Ramos, R. (2006). The ecological fallacy in a time series context: evidence from Spanish regional unemployment rates. *Journal of Geographical Systems, 8*, 391–410. http://dx.doi.org/10.1007/s10109-006-0033-x

Duraton, G., & Puga, D. (2002). Diversity Specialization in Cities: Why, Where and Does it Matter. McCann, P. (Ed.): *Industrial Localization Economics*. Cheltenham. 151-186.

Eberts, R. W., & McMillen, D. P. (1999). Agglomeration economies and urban public infrastructure, in P. C. Cheshire & E. S. Mills (ed.), *Handbook of Regional and Urban Economics, 3*(38), 1455-1495.

Fujita, M., & Thisse J. F. (2002). *Economics of Agglomeration*. Cambridge University Press.

Golan, A. (2006). Information and Entropy Econometrics. *A review and synthesis, Foundations and Trends in Econometrics, 2*, 1-145.

Golan, A., Judge, G., & Miller, D. (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*, 663–666. http://dx.doi.org/10.2307/2088121

Henderson, J.V., & Thisse, J.F. (2004). *Handbook of Regional and Urban Economics*. North Holland, Amsterdam.

Judge, G., Miller, D. J., & Cho W. T. K. (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, 162-187.

Kapur, J. N., & Kesavan H. K. (1992). *Entropy Optimization Principles with Applications.* Academic Press. New York.

King, G. (1997). *A solution to the Ecological Inference Problem: Reconstructing individual behavior from aggregate data*. Princeton, Princeton University Press. http://dx.doi.org/10.1017/CBO9780511510595

King, G., Rosen, O., & Tanner M. A. (2004). *Ecological Inference: New Methodological Strategies*. Cambridge University Press. Cambridge, UK.

Peeters, L., & Chasco, C. (2006). Ecological inference and spatial heterogeneity: an entropy-based distributionally weighted regression approach. *Papers in Regional Science, 85*(2), 257-276. http://dx.doi.org/10.1111/j.1435-5957.2006.00082.x

Robinson W.S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, *15*, 351–357. http://dx.doi.org/10.2307/2087176

Thomas, C. J., & Bromley, R. D. F. (2000). City centre revitalization: problems of fragmentation and fear in the evening and night-time city. *Urban Studies, 37*, 1403-1429. http://dx.doi.org/10.1080/00420980020080181

Notes

Note 1. For an introduction to the problem see Robinson (1950). Recent examples of empirical analysis were the consequences of this problem are explicitly studied can be found, among many others, in Duque et al. (2006) or Dumedah et al. (2008).

Note 2. See Henderson and Thisse (2004) or, among many others, Fujita and Thisse (2002).

Note 3. More extensive introductions can be found in Kapur and Kesaban (1992), Golan et al. (1996) or Golan (2006)

Note 4. An extensive survey of recent contributions to the field can be found in King et al. (2004).

Note 5. More detailed information on these surveys can be found in www.ine.es.

Copyright Disclaimer