
Departamento de Economía Aplicada
Universidad de Oviedo

DPAE/13/01

Discussion Papers on Applied Economics.

Department of Applied Economics. University of Oviedo.

Geographical disaggregation of labor market indicators by distributionally weighted regressions and GCE-GME techniques

Celia Alonso
Esteban Fernández Vázquez
Begoña Cueto
(Universidad de Oviedo)

Abstract:

Information for constructing labor market indicators is often observed with a higher level of geographical aggregation than it would be desired. In such a situation, an Ecological Inference (EI) exercise is required to disaggregate information from aggregated data. In this article we propose an estimator based on Entropy Econometrics approach that will be applied to distributionally weighted regressions. An empirical application to Spanish data is presented, where unemployment rates at NUTS-III level are estimated basing on (i) the aggregate rate for the whole country and a regressor constructed for registers of unemployed, and (ii) affiliated workers to Social Security, which is a variable observable at NUTS-III scale that will be the regressor in our equation. The approach proposed here does not require parameter homogeneity across space, which allows for capturing potential spatial heterogeneity in regional labor markets

Keywords: Ecological inference, Generalized cross-maximum entropy, distributionally weighted regression, regional labor markets indicators, Spanish provinces.

1. Introduction.

One relatively frequent limitation for empirical economics is the lack of data available at an appropriate spatial scale. To overcome this problem, a process of Ecological Inference (EI) is applied in order to recover the information at the required spatial scale. EI can be defined as the process of estimating disaggregated information from data reported at aggregate level. The foundations of EI were introduced in the works by Duncan and Davis (1953) and Goodman (1953). More recently, the work of King (1997) introduced a methodology that reconciled and extended previously adopted approaches. The methodological framework where EI is based on is generically known as "small area estimation" in the statistical literature (Ghosh and Rao, 1994; or more recently, You and Rao, 2003; and Toto and Nadram, 2010).

Within the set of techniques used for EI problems,[1] the estimation procedures based on entropy econometrics are gaining weight. Recent applications can be found in Judge et al. (2004), Peeters and Chasco (2006) or Bernardini Papalia (2010). On this background, our proposal is based on a specific type of distributionally weighted regressions. This type of techniques requires of disaggregated observations on the regressors included in the equations. In such a situation, we propose to approach the Ecological Inference by relaxing the spatial homogeneity hypothesis of parameters.

The paper is divided into three further sections. In section two a Distributionally Weighted Regression (DWR) is proposed as a way to estimate disaggregated data. The use of entropy econometrics for DWR estimators allows for introducing parameter heterogeneity, which usually

[1] An extensive survey of recent contributions to the field can be found in King, Rosen and Tanner (2004).

implies more accurate estimates of the spatially disaggregated data. Section four presents an empirical application with actual NUTS-III Spanish data. The last section presents the main conclusions.

2. Distributionally weighted regressions for Ecological Inference

Consider a geographical area (a country or region) for which we have $i = 1,..,T$ observations of the indicator of interest. These T data points can refer to T observations along time of our indicator or to smaller spatial sub-areas where it is observable. Further, suppose that there is a more detailed geographical disaggregation, which is contained in the classification into $j = 1,...,K$ different sub-areas (counties or municipalities, for example), on which we would like to observe the indicator of interest. The objective of the estimation problem would be to recover the values of the variable disaggregated by K sub-areas. This is an exercise of EI that will allow obtaining $K \times T$ estimates from the aggregate information we have in the T observable data points.

The traditional approaches to EI based on some DWR of the type proposed in Bidani and Ravallion (1997), are based on the homogeneity across space hypothesis and assume constancy of parameters across the disaggregated spatial units. This assumption is rarely tenable, since the aggregation process usually generates macro-level observations across which the parameters describing individuals may vary (Cho, 2001). Several solutions to deal with this kind of problems have been proposed (Calvo and Escolar 2003, Judge, et al. 2004; Bernardini Papalia 2010), but the approach we follow in this paper is closely related to the idea suggested by Peeters and Chasco (2006).

We start by paying attention to some aggregate indicator observable for each data point i, $\psi_{i\cdot}$. In the context of a DWR, $\psi_{i\cdot}$ is usually defined as a weighted sum of a function of the latent sub-area indicators ψ_{ij}, i. e.:

$$\psi_{i\cdot} = \sum_{j=1}^{K} \psi_{ij}\theta_{ij}\,;\ \forall i = 1, \dots, T \qquad 1$$

where θ_{ij} is the (observable) weight give to area j at point i. Very often the estimation objective is a sub-area indicator, obtained as a function of the sub-area value of the target variable. This is usually defined as the value of one variable of interest (i.e., number of unemployed people) by unit of other variable (i.e., potentially active population).

Additionally, by including an observed explanatory variable (or several) for the sub-areas j on each data point i, x_{ij}, the latent sub-group values can be specified as follows:

$$\psi_{ij} = \alpha_{ij} + \beta_{ij}x_{ij} + \varepsilon_{ij} \qquad 2$$

Note that the linear model in equation (2) also includes an idiosyncratic effect (α_{ij}) at the sub-area level. The estimation of DWR models like this can be based on the use of GCE for estimating linear models, even when it departs from a technique originally designed to estimate probabilities.

It is clear that the elements in equation (2) do not behave as probabilities, however. The noise components, for example, can be either positive or negative and do not necessarily add-up to one. In a GCE framework, we represent our uncertainty about the realizations of the errors treating each element ε_{ij} as a discrete random variable with $L \geq 2$ possible outcomes

contained in a convex set $\boldsymbol{v'} = \{v_1, \dots, v_L\}$, which for the sake of simplicity will be assumed as common for all the ε_{ij}. We also assume that these possible realizations are symmetric around zero ($-v_1 = v_L$). 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 ε_{ij} can be defined as:

$$\varepsilon_{ij} = \sum_{l=1}^{L} w_{ijl} v_l \; ; \; \forall i = 1, \dots, T; \; \forall j = 1, \dots, K \qquad 3$$

where w_{ijl} is the unknown probability of the outcome v_l for the sub-area j in point i.

The parameters to be estimated (α_{ij} and β_{ij}) are treated in a similar way and they are assumed as discrete random variables that can take values considered in some supporting vectors with $M \geq 2$ possible values ($\boldsymbol{b^{\alpha}}$ and $\boldsymbol{b^{\beta}}$) with respective unknown probabilities ($\boldsymbol{p^{\alpha}}$ and $\boldsymbol{p^{\beta}}$). For the sake of simplicity, the support spaces are constructed as discrete, bounded entities. The support points are chosen on the basis of *a priori* information. [2]

Under this GCE framework, the full distribution of each parameter and of each error (within their support spaces) is simultaneously estimated under minimal distributional assumptions, by means of the following program:

$$\underset{p^{\alpha}, p^{\beta}, p^{\gamma}, W}{Min} D(p^{\alpha}, p^{\beta}, W \| q^{\alpha}, q^{\beta}, W^{0}) = \sum_{m=1}^{M} \sum_{i=1}^{T} \sum_{j=1}^{K} p_{mij}^{\alpha} \, ln\left(\frac{p_{mij}^{\alpha}}{q_{mij}^{\alpha}}\right) + \qquad 4$$

[2] The choice of M, and the choice of continuous support spaces and different priors, is discussed in Golan, Judge and Miller, (1996). It is also possible to construct unbounded and continuous supports within the same framework (Golan, Judge and Miller, 1996).

$$\sum_{m=1}^{M}\sum_{i=1}^{T}\sum_{j=1}^{K} p_{mij}^{\beta}\, ln\left(\frac{p_{mij}^{\beta}}{q_{mij}^{\beta}}\right) + \sum_{l=1}^{L}\sum_{i=1}^{T}\sum_{j=1}^{K} w_{ijl}\, ln\left(\frac{w_{ijl}}{w_{ijl}^{0}}\right)$$

Subject to:

$$\sum_{m=1}^{M} p_{mij}^{\alpha} = \sum_{m=1}^{M} p_{mij}^{\beta} = \sum_{l=1}^{L} w_{ijl} = 1; \qquad 5$$

$$\forall i = 1, \dots, T;\ \forall j = 1, \dots, K$$

$$\sum_{j=1}^{K}\left(\sum_{m=1}^{M} p_{mij}^{\alpha} b_{m}^{\alpha} + \sum_{m=1}^{M} p_{mij}^{\beta} b_{m}^{\beta}\, x_{ij} + \sum_{l=1}^{L} w_{ijl} v_{l}\right)\theta_{ij} = \psi_{i.}; \qquad 6$$

$$\forall i = 1, \dots, T$$

It is important to point out that we assume: (i) unit specific coefficients for the sub-areas (parameter heterogeneity); (ii) a parametric specification of the unobserved spatial effects (spatial heterogeneity) through ε_{ij} errors, which can be positive or negative. Once estimated the coefficients in equation (), the estimates of the sub-area indicators will be given by:

$$\hat{\psi}_{ij} = \hat{\alpha}_{ij} + \hat{\beta}_{ij} x_{ij} + \hat{\varepsilon}_{ij} \qquad 7$$

The optimal solutions depend on the prior information and the data. If the priors are specified such that each choice is equally likely to be selected (uniform distributions), then the GCE solution reduces to the Generalized Maximum Entropy (GME) one. As with the GME estimator, numerical optimization techniques should be used to obtain the GCE solution.

3. How the method works: disaggregating unemployment rates in Spain at NUTS-III level from information in registers

In this section we try to find some empirical evidence on the performance of the GCE-based DWR technique to estimate a set of $(T \times K)$ disaggregated latent indicators. Specifically, we estimate the quarterly rates of unemployment in Spain along the period 2008-2012 at a NUTS-III level of geographical disaggregation. Spain is administratively divided into 50 provinces for which quarterly data on unemployment rates are published in the Labor Force Survey by the National Statistical Institute (INE). We would assume that the only observable information is the national unemployment rate (aggregate $\psi_{i.}$) and by applying the GCE-GME approach described before we estimate the provincial rates.

Equations like (2) require a regressor (x_{ij}) observable at the desired geographical level and a set of observable weights for each sub-area (θ_{ij}). We have chosen as regressor a proxy for the unemployment rates that can be observed at NUTS-III level in the registers of unemployed people and the affiliates to the Social Security system. By summing up the number of registered unemployed and the total affiliates to the Social Security, we have a rough approximation to the potentially active population by province. This information is used in order to define the weights of each province j for each time period i. Furthermore, by dividing the number of registered unemployed by this number we obtain a pseudo-rate of unemployment at provincial level that will be taken as our regressor x_{ij}.

The parameters in (2) will be estimated by the GCE program described in equations (4) to (6). The equal supporting vectors for the β parameter has been set as (0.75 ,1, 1.25) with $M = 3$. This indicates that in absence of additional information, we would expect the pseudo-rate of unemployment

(x_{ij}) to be equal to the actual but unobservable rate (ψ_{ij}), although it could vary by 25% around this central point. The idiosyncratic term α_{ij} is assumed to have a more complex structure, containing a quadratic trend (i and i^2) and a dummy for each quarter (d) as:

$$\alpha_{ij} = \delta_{ij} i + \pi_{ij} i^2 + \lambda_{ij} d; \forall i = 1, \dots, T; \forall j = 1, \dots, K \qquad 8$$

And δ_{ij}, π_{ij} and λ_{ij} are parameter to be estimated. The specific supporting vectors for this set of parameters has been set as wide as (-1 ,0, 1). Finally, for the error terms the support with $L = 3$ values has been chosen again, now applying the three-sigma rule with uniform a priori weights. The a priori probability distributions taken for all the coefficients are uniform as well, so the CGE estimation becomes a GME program.

A comparison between the actual and the estimated rates is possible and the (unweighted) mean deviation in percentage for the whole set of 50 provinces along the 20 quarters from 2008 to 2012 is reported in Table 1. Additionally, Figure 1 summarizes this information by plotting the simple mean of the errors only for the fifteen most populated provinces in Spain:

<<Insert Table 1 about here>>

<<Insert Figure 1 about here>>

As a first indicator of the accuracy of the GME inference, the error in the estimation of unemployment rates a NUTS-III level seem to be low. Considering the largest provinces according to their population size, the errors oscillate between $\pm 3\%$, with the exception of the provinces of Malaga and Coruña, where these are slightly larger. In general, the larger errors are concentrated in the provinces with a high concentration of agriculture activities (Almeria or Jaén, for example) and most of them are concentrated in the southern regions of Spain (mainly in Andalusia). A possible explanation for this spatial pattern is that in such provinces the information

contained in the registers of unemployed and affiliated workers are less reliable than for the average of the country.

4. Concluding remarks.

In this paper an Entropy-based approach with to Ecological Inference (EI) with distributionally weighted regressions (WR) and spatial heterogeneity of parameters is formulated throughout a real data application. The results observed suggest that a DWR based on a GCE-GME estimator can be useful to recover geographically disaggregated indicator of labor markets, since the deviations between the estimates and the actual values in our empirical application are moderate in most cases.

References

Bernardini Papalia R., 2010. Incorporating spatial structures in ecological inference: an information theoretic approach, Entropy, 12, 10, 2171-2185.

Bidani B. and Ravallion M., 1997. Decomposing social indicators using distributional data. Journal of Econometrics 77: 125–139.

Calvo, E. and Escolar, M., 2003. The local voter: a Geographically Weighted Approach to Ecological Inference, American Journal of Political Science, 47, 1, 189-204.

Cho, W.K.T. 2001. Latent groups and cross-level inferences, Electoral Studies, 20, 243-263.

Duncan, O. D. and Davis B., 1953. An Alternative to Ecological Correlation, American Sociological Review, 18, pp. 665–666.

Ghosh, M. and Rao J.N.K., 1994. Small area estimation: an appraisal. Statistical Sciences, 9, pp. 55-93.

Golan, A., Judge, G. and 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, pp. 663–666.

Judge, G., Miller, D. J. and 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, pp. 162-187).

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

King, G., 1997. A solution to the Ecological Inference Problem: Reconstructing individual behavior from aggregate data. Princeton, Princeton University Press.

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.

Robinson W.S., 1950. Ecological correlations and the behavior of individuals. American Sociological Review,15, pp. 351–357.

Toto M.C.S. and Nandram B., 2010. A Bayesian predictive inference for small area means incorporating covariates and sampling weights. Journal of Statistical Planning and Inference, 140, pp. 2963–2979.

You Y. and Rao J.N.K., 2003. Pseudo hierarchical Bayes small area estimation combining unit level models and survey weights. Journal of Statistical Planning and Inference, 111, pp. 197–208

Figure 1. Mean deviation 2008-2012 (%): estimates - actual unemployment rates, fifteen most populated provinces

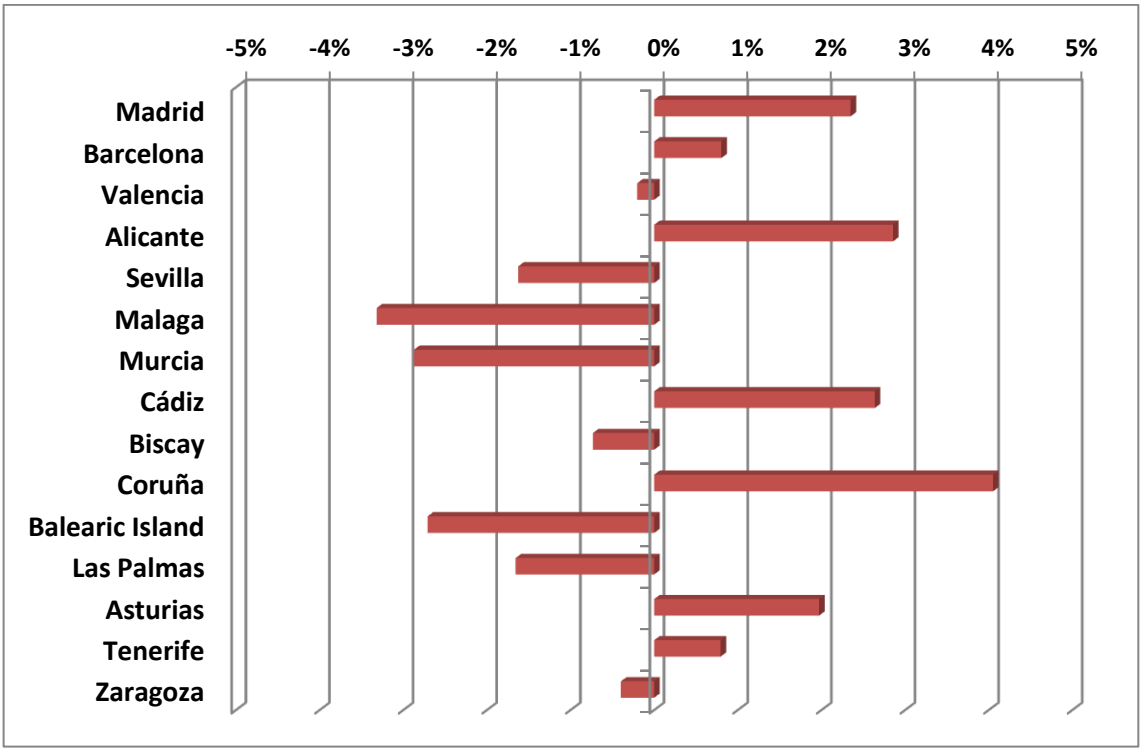

Table 1. Mean deviation 2008-2012 (%): estimates - actual unemployment rates

Province	Error	Province	Error
Madrid	2.35%	Navarra	1.71%
Barcelona	0.80%	Castellon	-2.38%
Valencia	-0.20%	Cantabria	3.00%
Alicante	2.86%	Valladolid	2.10%
Sevilla	-1.63%	Ciudad Real	0.76%
Malaga	-3.32%	Huelva	-6.55%
Murcia	-2.87%	Leon	2.02%
Cádiz	2.64%	Lleida	-0.42%
Biscay	-0.73%	Caceres	-0.39%
Coruña	4.05%	Albacete	0.38%
Balearic Island	-2.71%	Burgos	0.45%
Las Palmas	-1.66%	Salamanca	3.51%
Asturias	1.97%	Lugo	3.70%
Tenerife	0.79%	Ourense	4.83%
Zaragoza	-0.40%	La Rioja	0.08%
Pontevedra	2.52%	Alava	1.02%
Granada	-6.04%	Guadalajara	2.18%
Tarragona	-0.93%	Huesca	0.54%
Cordoba	-6.15%	Cuenca	-0.03%
Girona	-4.31%	Zamora	2.20%
Gipuzkoa	5.40%	Avila	-0.39%
Toledo	1.64%	Palencia	0.36%
Almeria	-8.10%	Segovia	-0.21%
Badajoz	0.02%	Teruel	1.01%
Jaen	-7.98%	Soria	-0.19%