ZNSSII3 - An information theoretic approach to ecological inference in presence of spatial dependence

Rosa BERNARDINI-PAPALIA
University of Bologna, Via Belle Arti 41, Bologna, Italy

Abstract : This paper introduces Information Theoretic – based methods for estimating a target variable in a set of small geographical areas, by exploring spatially heterogeneous relationships at the disaggregate level. Controlling for spatial effects means introducing models whereby the assumption is that values in adjacent geographic locations are linked to each other by means of some form of underlying spatial relationship. This method offers a flexible framework for modeling the underlying variation in sub-group indicators, by addressing the spatial dependency problem. A basic ecological inference problem, which allows for spatial heterogeneity and dependence, is presented with the aim of first estimating the model at the aggregate level, and then of employing the estimated coefficients to obtain the sub-group level indicators. The IT-based disaggregation procedure is applied to Italian data.

Keywords: Generalized Cross Entropy Estimation, Ecological Inference, Spatial Dependence.

1. Introduction

This paper introduces an Information Theory (IT)-based method for modeling economic aggregates and for obtaining estimates for small area (sub-group) or subpopulations when no sample units or limited data are available. The proposed approach offers a tractable framework for modeling the underlying variation in small area indicators in particular when data set contains outliers and in presence of collinearity among regressors since the maximum entropy estimates are robust with respect to the outliers and also less sensitive to a high condition number of the design matrix. A basic ecological inference problem which allows for spatial heterogeneity and dependence is presented with the aim of estimating small area/sub-group indicators by combining all available information at both macro and micro data level.

The latent small area indicators may be treated as random coefficients or modeled as a parametric function in the unit level model in which the observed aggregate is regressed on the explanatory variables both at the group and sub-group level.

By taking as a point of departure the approach presented in Johnston and Pattie (2000) in Judge, Miller; Cho (2004); Peeters, and Chasco, 2006 and Bernardini Papalia (2010a,b), the basic idea is to introduce an estimator based on an entropy measure of information which provides an effective and flexible procedure for reconciling micro and macro data. The maximum entropy (ME) procedures (Golan, Judge and Robinson, 1994; Golan, Judge and Miller, 1996; Golan, 2008) give the possibility to take into account out-of-sample information which can be introduced as additional constraints in the optimization program or by specifying particular priors for parameters and errors. An unique optimum solution can be achieved also if there are more parameters to be estimated than available moment conditions and the problem is ill-posed. If there exists additional non-sample information from theory and/or empirical evidence, over that contained in the consistency and adding-up constraints, for the unknown probabilities, it may be introduced in the form of known probabilities, by means of the cross-entropy formalism (Kullback, 1959).

The paper is structured as follows. In Section 2 an introduction to the traditional ecological inference (EI) problems is presented. Alternative approaches to ecological modeling that account for spatial heterogeneity and spatial dependence problems, are also introduced. Section 3 provides the formulation of the proposed information theoretic approaches incorporating both spatial heterogeneity and dependence. In Section 4, the IT-based disaggregation procedure is applied to Italian data. Finally, the last section provides concluding remarks and outlines some direction for further research.

2. Ecological Inference and Dependence across Space

The traditional approach to ecological inference is based on the homogeneity across space hypothesis which assumes constancy of parameters across the disaggregate 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). It is recognized that observations at an aggregate level of analysis do not necessarily provide useful information about lower levels of analysis, particularly when spatial heterogeneity is present. Moreover, the objective of recovering disaggregate information from aggregate data may produce “ill-posed” or “undetermined” inverse problems given that there are more unknowns than data points. In EI it is also important to deal with the “modifiable area unit problem” which refers to (i) the scale effect or aggregation effects, and (ii) the grouping effect or zoning effect. In the first case the resulting aggregation bias may produce different results when data (or individuals)

are grouped into increasingly larger areal units. In the second case, the resulting specification bias is connected to the variability in results due to alternative formulations of the areal units leading to differences in unit shape at the same or similar scales and arises when there is a non linear relationship that is not properly accounted for in the specification of the aggregated model. Many different possible relationships at the individual (or subgroup) level can generate the same observations at the aggregate (or group) level (King, 1997; King, Rosen and Tanner, 2004). In the absence of individual (or subgroup) level measurement (in the form of survey data), such information need to be inferred. Estimates of the disaggregated values for the variable of interest can be inferred from aggregate data by using appropriate statistical techniques. However, in many situations, given that micro-data of interest are not available, the accuracy of any predicted value cannot be verified.

Moreover, in presence of spatial structures, (i) absolute location effects (that refer to the impact—for each unit—of being located at a particular point in space), and (ii) relative location effects (that consider relevant the position of an unit relative to other units, Spatial dependence), have to be considered.

The absolute location effects can be introduced by assuming: (i) slope heterogeneity across spatial units, implying that parameters are not homogeneous over space but vary over different geographical locations; (ii) the presence of cross-sectional correlation due to the presence of some common immeasurable or omitted factors.

The relative location effects are traditionally introduced by incorporating: a spatial autoregressive process in the error term, and/or a spatially lagged dependent variable. A Spatial Error Model specification assumes that the spatial autocorrelation is modeled by a spatial autoregressive process in the error terms. It follows that: spatial effects are assumed to be identical within each unit, but all the units are still interacting spatially through a spatial weight matrix. The presence of spatial dependence is then associated with random shocks (due to the joint effect of misspecification, omitted variables, and spatial autocorrelation). In alternative, a Spatial Autoregressive Model specification, (Spatial Lag Model) assumes that all spatial dependence effects are captured by the lagged term. The spatial autocorrelation is then modeled by including a spatially lagged dependent variable. Global and local measures of spatial autocorrelation are computed to determine whether the data exhibit spatial dependence and a series of test statistics based on the Lagrange Multiplier (LM) or Rao Score (RS) principle are used to determine whether the variables in the model sufficiently capture the spatial dependence in the data. If the variables do not fully model the dependence, the diagnostics indicate whether the researcher should estimate a model with a spatially lagged dependent variable, a

spatially lagged error term, or both. The LM/RS principle can also be extended to more complex spatial alternatives, such as higher order processes, spatial error components and direct representation models, and to discrete choice models. Paralleling and complementing the theoretical motivation may represent a useful guide for modelling the spatial dependence.

The objectives of the paper are (i) to formulate an informational-theoretical approach to estimate small area /sub group variables or indicators in the presence of spatial structure and limited/incomplete information; (ii) to provide an empirical application to real data.

As a first task, a functional relationship between the variable to be disaggregated and a set of variables/indicators at area level is specified by combining different macro and micro data sources. The model at the aggregate level is then estimated and the sub-group level variables/indicators are obtained by employing these parameter estimates. Different model specifications extended to include spatial effects are also introduced with the aim of testing the hypothesis of: (i) parameters homogeneity/heterogeneity; (ii) uniform/varying spatial dependence.

We start by defining the aggregate indicator for group/region i, y_i, as a weighted geometric mean of the latent small area or sub group indicator $\mathbf{y}_{ij}$ in group/region i: $\mathbf{y}_i = \prod_{j=1}^{J_i} (\mathbf{y}_{ij})^{\boldsymbol{\theta}_{ij}}$, that is:

$$\ln \mathbf{y}_i = \sum_{j=1}^{J_i} (\ln \mathbf{y}_{ij}) \theta_{ij} \tag{1}$$

where $\mathbf{y}_{ij}$ is the indicator of the jth small area (sub group/region) in group/region i, $\boldsymbol{\theta}_{ij}$ is the weight of small area (sub group) j in i, with $\sum_{j=1} \boldsymbol{\theta}_{ij} = 1$, and where i = 1,..,N denotes the groups/regions and j = 1,..,Ji denotes the number of small areas (sub groups/regions) in *i*.

The small area/sub-regional indicators are not observed, but the $\mathbf{y}_i$'s and $\boldsymbol{\theta}_{ij}$'s are. In addition, by introducing an observed vector of explanatory variables for group/region i, x_i, an observed vector of explanatory variables for small area (sub group/region) j in group/region i, z_{ij}, the latent small area/sub-group indicators are expressed in a multiplicative form as follows:

$$\mathbf{y}_{ij} = \boldsymbol{\alpha}_{ij} \prod_{k=1}^{K} \mathbf{z}_{ij,k}^{\boldsymbol{\beta}_{ij,k}} \prod_{h=1}^{H} \mathbf{x}_{i,h}^{\boldsymbol{\gamma}_{ij,h}} e^{\boldsymbol{\varepsilon}_{ij}} \tag{2}$$

where z_{ij},k (k = 1, K) are the covariates observed at the level of small area/ sub group j within the group/region i, $x_{i,h}$ (h = 1,..H) are the covariates observed only at the level of group/region i, $\boldsymbol{\alpha}_{ij}$ are unobserved fixed effects, and $\boldsymbol{\varepsilon}_{ij}$ are error terms.

By substituting Equation (2) into Equation (1), we can obtain the following model:

$$\ln \mathbf{y}_i = \sum_{j=1}^{J_i} \left(\ln \boldsymbol{\alpha}_{ij} + \sum_{k=1}^{K} \boldsymbol{\beta}_{ij,k} \ln \mathbf{z}_{ij,k} + \sum_{h=1}^{H} \boldsymbol{\gamma}_{ij,h} \ln \mathbf{x}_{i,h} + \boldsymbol{\varepsilon}_{ij} \right) \boldsymbol{\theta}_{ij}$$

or

$$\ln \mathbf{y}_i = \sum_{j=1}^{J_i} \left(\ln \boldsymbol{\alpha}_{ij} + \sum_{k=1}^{K} \boldsymbol{\beta}_{ij,k} \ln \mathbf{z}_{ij,k} + \sum_{h=1}^{H} \boldsymbol{\gamma}_{ij,h} \ln \mathbf{x}_{i,h} \right) \boldsymbol{\theta}_{ij} + \mathbf{u}_i \tag{3}$$

where $\mathbf{u}_i = \sum_{j=1}^{J_i} \boldsymbol{\varepsilon}_{ij} \boldsymbol{\theta}_{ij}$ is a "composite" error term, which is heteroskedastic.

This model implies some kind of weighted regression, capturing "distributional effects" by using data on weights for each small area/sub group. It is important to point out that we assume: (i) unit specific coefficients for the small areas/sub groups (parameter heterogeneity); (ii) a parametric specification of the unobserved spatial effects (spatial heterogeneity) through $\boldsymbol{\varepsilon}_{ij}$'s, which can be positive or negative.

Using the estimated coefficients in Equation (3) we can obtain estimates of the unobserved or latent small area/sub group indicators as follows:

$$\hat{\mathbf{y}}_{ij} = \hat{\boldsymbol{\alpha}}_{ij} \prod_{k=1}^{K} \mathbf{z}_{ij,k}^{\hat{\boldsymbol{\beta}}_{ij,k}} \prod_{h=1}^{H} \mathbf{x}_{i,h}^{\hat{\boldsymbol{\gamma}}_{ij,h}} e^{\hat{\boldsymbol{\varepsilon}}_{ij}} \tag{4}$$

As proxies for the ignorance of the sources of spatial dependence, statistically significant parameters on dummy variables for geographic areas merely indicate that behaviours differ for units in these particular areas in contrast to the reference category (Anselin, 1988). Such an approach cannot indicate whether the spatial dependence is consistent with diffusion or with the spatial clustering of the behaviour's sources. Spatial diffusion occurs because units' behaviour is directly influenced by the behaviour of "neighbouring units." This diffusion effect corresponds to a positive and significant parameter on a spatially lagged dependent variable capturing the direct influence between neighbours. In the diffusion case, neighbors influence the behavior of their neighbors and vice versa.

If one is unable to fully model the sources of spatial dependence in the data generating process (DGP), the spatial dependence in the error terms between neighboring locations is assumed. This spatial error dependence can be modeled via a spatially lagged error term. It is also possible to hypothesize that spatial dependence is produced both by the

diffusion and by the independent adoption of behaviors by neighbors. This joint spatial dependence can be modeled by incorporating both a spatially lagged dependent variable and a spatial error term, with proper identifying restrictions imposed.

When the spatial autocorrelation is modeled by a Spatial Lag Model, Spatial Autoregressive Model (SAR Model), the previous model (3) can be generalized by introducing a spatial-lag term into the model. The resulting latent small area/sub group indicators are specified in a multiplicative form as follows:

$$\ln \mathbf{y}_i = \sum_{j=1}^{J_i} \left(\ln \boldsymbol{\alpha}_{ij} + \sum_{k=1}^{K} \boldsymbol{\beta}_{ij,k} \ln \mathbf{z}_{ij,k} + \sum_{h=1}^{H} \boldsymbol{\gamma}_{ij,h} \ln \mathbf{x}_{i,h} + \rho \ln \mathbf{w}\mathbf{y}_i + \boldsymbol{\varepsilon}_{ij} \right) \boldsymbol{\theta}_{ij} \tag{5}$$

where ρ is a spatial lag coefficient (the parameter associated to the spatially lagged dependent variable, $\ln \mathbf{wy}$), $\mathbf{w}$ is a proximity matrix of order N.

The definition of neighbors for each observation via a spatial weights matrix is a critical decision in modeling spatial autocorrelation. In empirical applications, it is common practice to derive spatial weights from the location and spatial arrangements of observation by means of a geographic information system. In this case, units are defined 'neighbors' when they are within a given distance of each other, ie w ij=1 for d ij ≤δ and i≠j, where dij is the distance function chosen, and δ is the critical cut-off value.

More specifically, a spatial weights matrix w* is defined as follow:

$$w_{ij}^{*} = \begin{cases} 0 & \text{if} \quad i = j \\ 1 & \text{if} \quad d_{ij} \le \delta, i \ne j \\ 0 & \text{if} \quad d_{ij} > \delta, i \ne j \end{cases} \tag{6}$$

and the elements of the row-standardized spatial weights matrix **w** (with elements of a row sum to one) result:

$$w_{ij} = \frac{w_{ij}^{*}}{\sum_{j=1}^{N} w_{ij}^{*}}, \qquad i, j = 1, .., N. \tag{7}$$

The SAR model assumes that all spatial dependence effects are captured by the lagged term by showing how the performance of the dependent variable impacts all the other (neighbor) groups/regions through the spatial transformation.

In alternative, by assuming a spatial dependence is the error structure (in terms of a first order spatial autoregressive process), the resulting Spatial Error Model (SEM Model) specification relative to model (3) is derived as follows:

$$\ln \mathbf{y}_i = \sum_{j=1}^{J_i} \left(\ln \boldsymbol{\alpha}_{ij} + \sum_{k=1}^{K} \boldsymbol{\beta}_{ij,k} \ln \mathbf{z}_{ij,k} + \sum_{h=1}^{H} \boldsymbol{\gamma}_{ij,h} \ln \mathbf{x}_{i,h} + \left(\lambda \mathbf{w} \boldsymbol{\varepsilon}_{ij} + \boldsymbol{\tau}_{ij} \right) \right) \boldsymbol{\theta}_{ij} \tag{8}$$

where λ is a spatial autoregressive coefficient, $\mathbf{w}$ is a proximity matrix of order N, as previously defined, and $\boldsymbol{\tau}_{ij}$ are the usual stochastic error terms.

The Spatial Error Model leaves unchanged the systematic component and assumes spatially autocorrelated errors. In this respect, it is observed how a random shock in a small area/sub group affects performances in that small area/sub group and additionally impacts all the other small areas/sub groups through the spatial transformation. This model specification measures the joint effect of misspecification, omitted variables, and spatial autocorrelation.

3. The Information Theoretic Formulation

The application of Maximum Entropy methods and Information Theoretic techniques has been explored within the context of ecological. The first use of entropy-maximizing models concerned the application of gravity models and transportation flows. Recently, applications of Information Theoretic methods have focused on the analysis of spatial patterns of voting at the individual level (King, Rosen, and Tanner, 2004)

However, the present study extends the IT approach to the case of Ecological Inference incorporating Spatial Dependence. Past studies have given little weight to the role of spatial effects in ecological inference analysis, and so this present study is going to introduce a basic framework for EI in the presence of spatial heterogeneity and dependence. It also deals with the specification of models that explicitly control for spatial effects, interpretation and IT-based formulation.

An Information Theoretic technique (Golan, Judge, and Miller, 2006; Peeters, and Chasco, 2006; Bernardini Papalia, 2010a,b) is suggested as an adequate solution in the present context since it provides an effective and flexible procedure for reconciling micro and macro data and for addressing problems related to spatial structures.

Implementation of these methods requires that the parameters and errors of the model in Equations (5) and (8) are specified as linear combinations of some predetermined and discrete support values and unknown probabilities (weights). Thus, all coefficients $\boldsymbol{\alpha}_{ij}, \boldsymbol{\beta}_{ij}, \boldsymbol{\gamma}_{ij}, \rho, \lambda$ and unknown errors ε_{ij}, $\boldsymbol{\tau}_{ij}$ in Equations 5 and 8, are reparameterized and expressed in terms of proper probabilities. For each parameter, a set of M support points (with $2 \leq M < \infty$) has been chosen:

$\mathbf{s}_\alpha = \left(\mathbf{s}_1^\alpha,..,\mathbf{s}_M^\alpha\right)'$, $\mathbf{s}_\beta = \left(\mathbf{s}_1^\beta,..,\mathbf{s}_M^\beta\right)'$, $\mathbf{s}_\gamma = \left(\mathbf{s}_1^\gamma,..,\mathbf{s}_M^\gamma\right)'$, $\mathbf{s}_\rho = \left(\mathbf{s}_1^\rho,..,\mathbf{s}_M^\rho\right)'$, $\mathbf{s}_\lambda = \left(\mathbf{s}_1^\lambda,..,\mathbf{s}_M^\lambda\right)'$, and the corresponding unknown probabilities defined on these support spaces $\mathbf{p}_{\alpha,ij} = \left(\mathbf{p}_{ij,1}^\alpha,..,\mathbf{p}_{ij,M}^\alpha\right)'$, $\mathbf{p}_{\beta,ij} = \left(\mathbf{p}_{ij,1}^\beta,..,\mathbf{p}_{ij,M}^\beta\right)'$, $\mathbf{p}_{\gamma,ij} = \left(\mathbf{p}_{ij,1}^\gamma,..,\mathbf{p}_{ij,M}^\gamma\right)'$, $\mathbf{p}_{\rho,ij} = \left(\mathbf{p}_{ij,1}^\rho,..,\mathbf{p}_{ij,M}^\rho\right)$, $\mathbf{p}_{\lambda,ij} = \left(\mathbf{p}_{ij,1}^\lambda,..,\mathbf{p}_{ij,M}^\lambda\right)$. Similarly, the errors ε_{ij}, $\boldsymbol{\tau}_{ij}$, are treated as unknowns, and a set of R support points $\mathbf{s}_\varepsilon = \left(\mathbf{s}_1^\varepsilon,..,\mathbf{s}_R^\varepsilon\right)'$, $\mathbf{s}_\tau = \left(\mathbf{s}_1^\tau,..,\mathbf{s}_R^\tau\right)'$, has been chosen, with $2 \le j < \infty$ with reference to the unknown probabilities $\mathbf{p}_{\varepsilon,ij} = \left(\mathbf{p}_{ij,1}^\varepsilon,..,\mathbf{p}_{ij,R}^\varepsilon\right)'$, $\mathbf{p}_{\tau,ij} = \left(\mathbf{p}_{ij,1}^\tau,..,\mathbf{p}_{ij,R}^\tau\right)'$.

For the sake of simplicity, the above support spaces are constructed as discrete, bounded entities. It is possible to construct unbounded and continuous supports within the same framework (Golan, 2008).

The support points are chosen on the basis of a priori information as discussed in Golan, Judge and Miller 2006). However, such knowledge is not always available, and symmetric parameter supports around zero are generally used in the presence of scarce prior information about each parameter. With regard to errors, in most cases where the underlying distribution is unknown, one conservative way of choosing the error supports $\mathbf{s}_\varepsilon, \mathbf{s}_\tau$, is to employ the "three-sigma rule" established by Pukelsheim.

Under the GCE framework, the full distribution of each parameter and of each error (within their support spaces) is simultaneously estimated under minimal distributional assumptions. More specifically, the parameters $\boldsymbol{\alpha}_{ij}, \boldsymbol{\beta}_{ij}, \boldsymbol{\gamma}_{ij}, \rho, \lambda$ and errors ε_{ij}, τ_{ij} are reparameterized as:

$$\boldsymbol{\alpha}_{ij} = \mathbf{s}_\alpha'\mathbf{p}_{\alpha,ij}, \quad \boldsymbol{\beta}_{ij} = \mathbf{s}_\beta'\mathbf{p}_{\beta,ij}, \quad \boldsymbol{\gamma}_{ij} = \mathbf{s}_\gamma'\mathbf{p}_{\gamma,ij}, \quad \rho = \mathbf{s}_\rho'\mathbf{p}_\rho, \quad \lambda = \mathbf{s}_\lambda'\mathbf{p}_\lambda \quad \boldsymbol{\varepsilon}_{ij} = \mathbf{s}_\varepsilon'\mathbf{p}_{\varepsilon,ij}, \quad \tau_{ij} = \mathbf{s}_\tau'\mathbf{p}_{\tau,ij} \tag{9}$$

with support vectors for parameters $\boldsymbol{\alpha}_{ij}, \boldsymbol{\beta}_{ij}, \boldsymbol{\gamma}_{ij}, \rho, \lambda$ and errors ε_{ij}, τ_{ij} given by:

$$\begin{aligned} &\mathbf{s}_\alpha = \left(\mathbf{s}_1^\alpha,..,\mathbf{s}_M^\alpha\right)', \quad \mathbf{s}_\beta = \left(\mathbf{s}_1^\beta,..,\mathbf{s}_M^\beta\right)', \quad \mathbf{s}_\gamma = \left(\mathbf{s}_1^\gamma,..,\mathbf{s}_M^\gamma\right)', \quad \mathbf{s}_\rho = \left(\mathbf{s}_1^\rho,..,\mathbf{s}_M^\rho\right)', \quad \mathbf{s}_\lambda = \left(\mathbf{s}_1^\lambda,..,\mathbf{s}_M^\lambda\right)' \\ &\mathbf{s}_\varepsilon = \left(\mathbf{s}_1^\varepsilon,..,\mathbf{s}_R^\varepsilon\right)', \quad \mathbf{s}_\tau = \left(\mathbf{s}_1^\tau,..,\mathbf{s}_R^\tau\right) \end{aligned} \tag{10}$$

and corresponding unknown probabilities given by:

$$\begin{aligned} &\mathbf{p}_{\alpha,ij} = \left(\mathbf{p}_{ij,1}^\alpha,..,\mathbf{p}_{ij,M}^\alpha\right)', \quad \mathbf{p}_{\beta,ij} = \left(\mathbf{p}_{ij,1}^\beta,..,\mathbf{p}_{ij,M}^\beta\right)', \quad \mathbf{p}_{\gamma,ij} = \left(\mathbf{p}_{ij,1}^\gamma,..,\mathbf{p}_{ij,M}^\gamma\right)', \quad \mathbf{p}_{\rho,ij} = \left(\mathbf{p}_{ij,1}^\rho,..,\mathbf{p}_{ij,M}^\rho\right) \\ &\mathbf{p}_{\lambda,ij} = \left(\mathbf{p}_{ij,1}^\lambda,..,\mathbf{p}_{ij,M}^\lambda\right)', \quad \mathbf{p}_{\varepsilon,ij} = \left(\mathbf{p}_{ij,1}^\varepsilon,..,\mathbf{p}_{ij,R}^\varepsilon\right)', \mathbf{p}_{\tau,ij} = \left(\mathbf{p}_{ij,1}^\tau,..,\mathbf{p}_{ij,R}^\tau\right)' \end{aligned} \tag{11}$$

with M,R ≥ 2.

In addition, prior information reflecting subjective information or any other sample and pre-sample information is introduced by specifying the priors for all parameters and errors: $\tilde{\mathbf{p}}_{\alpha,ij}, \tilde{\mathbf{p}}_{\beta,ij}, \tilde{\mathbf{p}}_{\gamma,ij}, \tilde{\mathbf{p}}_{\rho,ij}, \tilde{\mathbf{p}}_{\varepsilon,ij}, \tilde{\mathbf{p}}_{\tau,ij}$. These priors may come from prior data, theory, and/or other experiments.

The GCE optimization problem for the ecological spatial model corresponding to Equation (5) can be reformulated by minimizing the following objective function H(.) as follows:

$$H = \sum_i \sum_j \left(\mathbf{p}_{\alpha,ij}\right)' \ln\left(\frac{\mathbf{p}_{\alpha,ij}}{\tilde{\mathbf{p}}_{\alpha,ij}}\right) + \sum_i \sum_j \left(\mathbf{p}_{\beta,ij}\right)' \ln\left(\frac{\mathbf{p}_{\beta,ij}}{\tilde{\mathbf{p}}_{\beta,ij}}\right) + \sum_i \sum_j \left(\mathbf{p}_{\gamma,ij}\right)' \ln\left(\frac{\mathbf{p}_{\gamma,ij}}{\tilde{\mathbf{p}}_{\gamma,ij}}\right) + \sum_i \sum_j \left(\mathbf{p}_{\rho,ij}\right)' \ln\left(\frac{\mathbf{p}_{\rho,ij}}{\tilde{\mathbf{p}}_{\rho,ij}}\right) + \sum_i \sum_j \left(\mathbf{p}_{\varepsilon,ij}\right)' \ln\left(\frac{\mathbf{p}_{\varepsilon,ij}}{\tilde{\mathbf{p}}_{\varepsilon,ij}}\right) \tag{12}$$

subject to:

(i) data consistency conditions:

$$\ln \mathbf{y}_i = \sum_{j=1}^{J_i} \left(\mathbf{s}_\alpha' \mathbf{p}_{\alpha,ij} + \sum_{k=1}^{K} \left(\mathbf{s}_\beta' \mathbf{p}_{\beta,ij}\right) \ln \mathbf{z}_{ij,k} + \sum_{h=1}^{H} \left(\mathbf{s}_\gamma' \mathbf{p}_{\gamma,ij}\right) \ln \mathbf{x}_{i,h} + \left(\mathbf{s}_\rho' \mathbf{p}_{\rho,ij}\right) \ln \mathbf{wy}_i + \left(\mathbf{s}_\varepsilon' \mathbf{p}_{\varepsilon,ij}\right) \right) \boldsymbol{\theta}_{ij} \tag{13}$$

(ii) adding-up constraints for probabilities.

$$\sum \mathbf{p}_{\alpha,ij} = \sum \mathbf{p}_{\beta,ij} = \sum \mathbf{p}_{\gamma,ij} = \sum \mathbf{p}_{\rho,ij} = \sum \mathbf{p}_{\varepsilon,ij} = 1 \quad \forall i, j$$

$$\sum \hat{\mathbf{p}}_{\alpha,ij} = \sum \hat{\mathbf{p}}_{\beta,ij} = \sum \hat{\mathbf{p}}_{\gamma,ij} = \sum \hat{\mathbf{p}}_{\rho,ij} = \sum \hat{\mathbf{p}}_{\varepsilon,ij} = 1 \quad \forall i, j$$

Analogously, the GCE optimization problem for the ecological spatial model corresponding to Equation (8) can be reformulated by minimizing the following objective function H(.) as follows:

$$H = \sum_i \sum_j \left(\mathbf{p}_{\alpha,ij}\right)' \ln\left(\frac{\mathbf{p}_{\alpha,ij}}{\tilde{\mathbf{p}}_{\alpha,ij}}\right) + \sum_i \sum_j \left(\mathbf{p}_{\beta,ij}\right)' \ln\left(\frac{\mathbf{p}_{\beta,ij}}{\tilde{\mathbf{p}}_{\beta,ij}}\right) + \sum_i \sum_j \left(\mathbf{p}_{\gamma,ij}\right)' \ln\left(\frac{\mathbf{p}_{\gamma,ij}}{\tilde{\mathbf{p}}_{\gamma,ij}}\right) + \sum_i \sum_j \left(\mathbf{p}_{\lambda,ij}\right)' \ln\left(\frac{\mathbf{p}_{\lambda,ij}}{\tilde{\mathbf{p}}_{\lambda,ij}}\right) + \sum_i \sum_j \left(\mathbf{p}_{\varepsilon,ij}\right)' \ln\left(\frac{\mathbf{p}_{\varepsilon,ij}}{\tilde{\mathbf{p}}_{\varepsilon,ij}}\right) + \sum_i \sum_j \left(\mathbf{p}_{\tau,ij}\right)' \ln\left(\frac{\mathbf{p}_{\tau,ij}}{\tilde{\mathbf{p}}_{\tau,ij}}\right) \tag{14}$$

subject to:

(i) data consistency conditions:

$$\ln \mathbf{y}_i = \sum_{j=1}^{J_i} \left(\mathbf{s}_\alpha' \mathbf{p}_{\alpha,ij} + \sum_{k=1}^{K} \left(\mathbf{s}_\beta' \mathbf{p}_{\beta,ij} \right) \ln \mathbf{z}_{ij,k} + \sum_{h=1}^{H} \left(\mathbf{s}_\gamma' \mathbf{p}_{\gamma,ij} \right) \ln \mathbf{x}_{i,h} + \left(\mathbf{s}_\lambda' \mathbf{p}_{\lambda,ij} \right) w\left(\mathbf{s}_\varepsilon' \mathbf{p}_{\varepsilon,ij} \right) + \left(\mathbf{s}_\tau' \mathbf{p}_{\tau,ij} \right) \right) \boldsymbol{\theta}_{ij} \tag{15}$$

(ii) adding-up constraints for probabilities:

$$\sum \mathbf{p}_{\alpha,ij} = \sum \mathbf{p}_{\beta,ij} = \sum \mathbf{p}_{\gamma,ij} = \sum \mathbf{p}_{\lambda,ij} = \sum \mathbf{p}_{\tau,ij} = 1 \qquad \forall i, j$$

$$\sum \hat{\mathbf{p}}_{\alpha,ij} = \sum \hat{\mathbf{p}}_{\beta,ij} = \sum \hat{\mathbf{p}}_{\gamma,ij} = \sum \hat{\mathbf{p}}_{\lambda,ij} = \sum \hat{\mathbf{p}}_{\tau,ij} = 1 \qquad \forall i, j$$

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

In order to determine whether additional information in the data, expressed in the form of constraints, produce a departure from the condition of total uncertainty and a consequent reduction of uncertainty related to the phenomenon, the standard normalized entropy measure can be used (Golan, Judge and Miller, 1996).

Note that one can simultaneously consider the choice of the model, that is the functional relationship linking the variable to be disaggregated and a set of variables/indicators at area level, and the choice associated with the macro and micro data sources.

4. An Empirical Application

We present the application of the GME formulation introduced in section 2 to the case of an Italian data set. The GME-based formulation is used to disaggregate the value-added of Umbria's local labour systems (LLS) in macro-sectors of manufacturing industry, in the year 2001. Nine manufacturing sectors are dealt with: 1. Food, beverages and tobacco; 2. Textiles and clothing; 3. Wood products; 4. Paper, printing and publishing; 5. Coke and refined petroleum products, chemicals; 6. Non-metallic mineral products; 7. Basic metals, fabricated metal products; 8. Machinery, computing, precision medical instruments, transport; 9. Rubber, plastic and other manufacturing sectors.

The case study is particularly suitable to represent the usefulness of our approach to study the local labour systems. The Umbria region assumes the character of the region-not region, that is, a political-administrative unit dominated by centripetal and

centrifugal forces which thus tend to enhance linkages and integration with neighboring regions. The different areas are caracterized by specific features: (i) the rural high Valnerina area (Norcia and Cascia) projected to enhance the economic potential of cultural and environmental specificities; (ii) Città di Castello and Umbertide characterized by an territorial organization of district type, (iii) the area of Tevere's valley, re-organized into several spatial components (the rural Todi, the area relative to Perugia, Deruta, and an area of small and medium enterprises with a significant systemic organizational structure, Marsciano) and (iv) the territories of the Lake Trasimeno, Orvieto, those of the Valle Umbra (Assisi, Foligno), and so on (the Terni, in the Gubbio area Gualdese), each with its own characteristics and distinct growth path characterized by distinctive specifities.

The basic formulation assumes that: (i) the GME estimates of the value-added of Umbria's LLS, disaggregated by sector, are consistent with the total value-added observed at the regional level; (ii) the value-added of the LLS are measured with error.

By introducing the baseline statistical model, $y_i = \sum_{j=1}^{J_i}\left(\alpha_{ij} + \sum_{K=1}^{K}\beta_k z_{j,k} + \sum_{h=1}^{H}\gamma_{j,h}x_{i,h} + \rho_i w y_j + \varepsilon_{ij}\right)\theta_{ij}$ we estimate the total value-added for each sector at the level of Umbria's LLS, by employing all available information, that is: sub-area (LLS) level information about K explanatory variables Z_j, that refer to: total value-added of manufacturing's local labour systems, employment rate, ER; Job placement rate, JPR, but also refer to measures of spatial externalities. The sectors' shares of the total number of manufacturing firms here is used for $\boldsymbol{\theta}_{ij}$. From the macro perspective, the total value-added for each sector within Umbria is a known quantity, and is regarded as a fixed regional total.

Spatial dependence of the LLS's value added is confirmed; specifically, Moran's I and Geary's C tests cannot accept the null hypothesis of global spatial independence (0.0593; p-value: 0.061 for the former; 0.0493; p-value: 0. 0003 for the latter). In our analysis, the weight matrix is computed by means of the distance of each LLS from Perugia, where the critical cut-off value is given by the first quartile of the distance's distribution as well as by means of weights based on contiguity measures of LLS. Results produced by different weight matrices are robust for all model specifications. Alternative specifications, also related to spatial LAG model and spatial Error model have been the objective of a preliminary analysis.

The ME principle is used to yield the most uninformed distribution in keeping with the observed sample data, with minimal assumptions made regarding the underlying

distribution generating the data. We choose symmetric parameter supports around zero, given that we have very little prior information about each parameter, and M=5 support points for each parameter, since estimation is not improved by choosing more than about five support points. We choose j=3 support points for each error, and we specify error supports according to Pukelsheim's "Three Sigma Rule". The estimation procedure is implemented using the GAMS software and a nonlinear solver, CONOPT2.

Using the measure of normalized entropy (NE) (Golan, Judge and Miller, 1996) relative to different scenarios, alternative formulations are compared with the aim of choosing the model specification that, conditional to the information available, incomplete and limited, contributes in reduction of uncertainty concerning the phenomenon of interest. The NE of the Model 8 is the smallest one (Table 1), indicating that it has the lowest uncertainty of all models considered. These results show the sensitivity of variable selection relative to the data generation process.

Results of the selected model (see Table 2) seem to be relatively robust with respect to the parameter supports: the GME parameter estimates do not vary a great deal as parameter supports are modified. The choice of support vectors for the parameters, within the intervals (−100,100) and (−20,20), has a negligible effect on the coefficients. The asymptotic standard errors are calculated using the method proposed by Golan, Judge and Miller, 1996.

The distribution of the value-added of Umbria's 16 LLS, disaggregated by sector for 2001, seems to be quite heterogeneous. Our analysis validates the hypothesis of spatial heterogeneity across the LLSs, as well as the contribution of the indicator chosen as weight for the small area latent indicators that is the share of the total number of firms operating in each sector h and located in local labour system j.

5. Conclusions

In this paper we have tackled the problem of providing reliable estimates of a target variable in a set of small geographical areas, by exploring spatially relationships at the disaggregate level. Controlling for spatial effects means introducing models whereby the assumption is that values in adjacent geographic locations are linked to each other by means of some form of underlying spatial relationship. Given researchers' uncertainty about spatial data sampling processes and error-correlation structures, it seems reasonable to explore estimation and inference frameworks more flexible that reduce the

assumptions about some or all of these features while, at the same time, allowing them to incorporate knowledge about the spatial structure in a sample.

In certain cases, in order to account for spatial dependency we need to grasp the spatial variations in the regression coefficients, since empirical predictions based on global parameters may be biased, and thus misrepresent local behavior. This is particularly problematic in the case of regional analysis, where locally representative regression coefficients are required for micro-level policy decisions to be taken.

We have discussed the importance of taking into account individually- and spatially-correlated small area level variations, and we have recommended the use of Information Theoretic-based methods for the estimation of variables within the small groups of interest.

The proposed ME-based methods of disaggregation are capable of yielding disaggregate data consistent with prior information, resulting from different sources of data in the absence of high quality and detailed data as well as in the presence of problems of collinearity and endogeneity, without imposing strong distributional assumptions. Within this framework, we have shown how partial information at the disaggregated level can be combined with aggregated data to provide estimates of latent variables or indicatorrs which are of interest at the small area/sub group level.

Two interesting points emerge here. Firstly, the ME-based formulation has the advantage of being consistent with the underlying spatial dependence in the data-generating process, and eventually with the restrictions implied by certain non-sample information, or by previous empirical experience. Compared to traditional estimation methods, this approach is characterized by its robustness to ill-conditioned designs, and by its ability to fit over-parametrized models such as those pertaining to data disaggregation problems and small area estimation. It is also particularly effective to deal with problems of skewed distributions and outliers and also represent a good choice in presence of collinearity and endogeneity problems.

Secondly, within a ME-based framework, the informative contribution in reduction of uncertainty of the phenomenon under study, made by each restriction and by each variable included in the basic problem formulation can be verified simultaneously.

The GME formulation has been employed in relation to an Italian data set in order to compute the value-added of Umbria's local labour systems in 2001 for nine manufacturing sectors which are consistent with the total regional value-added per

sector, and by formulating a suitable set of constraints for the optimization problem in the presence of errors in the aggregates at sub-area level.

The results show that this approach provides a flexible, powerful data-disaggregation method, since it enables us to: (i) consider prior knowledge introduced by adding linear and nonlinear inequality constraints, errors in equations, and error in variables; (ii) allow for the efficient use of information from a variety of sources; (iii) reconcile data at different levels of aggregation within a coherent framework.

Further work should be done in order to explore IT methods by considering (i) small area parameters which are a non linear functions of the small area total variable (small rates and proportions) in presence of spatial structures; and (ii) temporal dependence. Possible extensions of the proposed procedure include estimation using composite IT methods incorporating both GME and GCE estimators (Bernardini Papalia, 2008) that can be used when some of the small areas have no sample units.

Acknowledgements

This paper has been completed within the BLU-ETS project "Blue-Enterprises and Trade Statistics", a small or medium-scale focused research project funded by the Seventh Framework Programme of the European Commission, FP7-COOPERATION-SSH (Cooperation Work Programme: Socio-Economic Sciences and the Humanities).

References

Anselin L. (1988) Spatial Econometrics: Methods and Models, Kluwer, Boston.

Bernardini Papalia, R. (2008) A Composite Generalized Cross Entropy formulation in small samples Estimation. Econometric Review, 27 (4-6), 596–609.

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

Bernardini Papalia R. (2010b) Data Disaggregation Procedures within a Maximum Entropy Framework, Journal of Applied Statistics, 37, 11, 1947-1959.

Cho, W.K.T. (2001) Latent groups and cross-level inferences. Elect. Stud, 20, 243–263.

Golan, A. (2008) Information and Entropy Econometrics: A review and Synthesis. Foundations and Trends® in Econometrics: Vol. 2: No 1–2, Now publishers: Hanover, USA, 1-145.

Golan, A.; Judge, G.; Miller, D. (1996) Maximum Entropy Econometrics: Robust Estimation with Limited Data. Wiley: New York, NY, USA.

Golan A., Judge G., and S. Robinson (1994) Recovering Information from Incomplete or Partial Multisectoral Economic Data, The Review of Economics and Statistics, 76, 3, 541-549.

Johnston, R., and Pattie, C. (2000) Ecological inference and entropy-maximizing: An alternative estimation procedure for split-ticket voting“, Political Analysis, 8, 333-345.

Judge, G., Miller, D., and Cho, W.K.T. (2004) An information theoretic approach to ecological inference. In: Ecological Inference: New Methodological Strategies. King, G., Rosen, O. and Tanner, M.A. eds), Cambridge University Press, 162-187.

King, G. (1997) A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data, Princeton University Press.

King, G., Rosen, O. and Tanner, M.A. (2004) Ecological Inference: New Methodological Strategies,Cambridge University Press, 162-187.

Kullback, J. (1959) Information Theory and Statistics. Wiley, New York, NY.

Levine, R.D. (1980) An information theoretical approach to inversion problems, Journal of Physics A, 13,91-108.

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.

Shannon C.E. (1948) A mathematical theory of communication. Bell System Technical Journal 27, 379-423.

TABLE 1. Comparison of alternative model specifications in terms of Normalized Entropy Measures

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**	**Model 8**
Explanatory Variables								
LLS_Employment Rate: ER	X	X	X			X	X	X
LLS _Job Placement Rate: JPR	X			X	X	X		
LLS_ValueAdded: VA	X	X	X	X	X	X	X	X
LLS_Spatial-Lag ValueAdded: WVA	X		X		X		X	X
Spatial fixed effects		X	X	X	X	X	X	
Weight of sub group indicator: θ_{ij}	Sectors' shares of the total number of manufacturing firms							
Normalized Entropy Measure	0.5401	0.5235	0.5401	0.5243	0.5403	0.5402	0.5543	0.5231

TABLE 2 . Estimates of the value added of Umbria's LLS disaggregated by manufacturing sector for the year 2001

	Manufacturing Sectors								
Local Labour Systems (Umbria region)	Food, beverages tobacco	Textiles and clothing	Wood products	Paper, printing publishing	Coke, chemicals	Non-metallic mineral products	Basic metals, metal products	Machinery, computing, transport;	Rubber, plastic, other manufactu-ring sectors.
ASSISI	32.95	68.65	0.00	14.29	0.00	30.34	36.28	33.60	31.29
CASCIA	1.43	0.00	0.00	0.00	0.00	0.38	1.05	0.53	1.08
CASTIGLIONE DEL LAGO	13.59	14.31	0.41	2.27	0.00	4.13	18.23	18.41	10.43
CITTA' DI CASTELLO	10.71	21.68	0.84	44.14	0.00	8.45	21.62	25.82	32.55
FOLIGNO	53.06	33.45	0.00	20.31	24.33	20.53	45.82	49.19	28.16
GUALDO TADINO	15.00	7.94	6.55	7.23	12.99	57.01	16.82	29.36	9.50
GUBBIO	26.27	20.24	0.61	6.70	12.05	30.50	18.43	12.89	13.21
MARSCIANO	9.52	16.38	2.08	4.59	8.25	8.35	14.57	9.81	11.56
NORCIA	10.02	0.90	0.00	0.27	0.00	1.32	4.14	1.16	3.09
PERUGIA	69.02	148.96	6.78	56.15	67.25	136.21	95.00	111.98	69.65
SPOLETO	31.35	21.62	0.65	5.37	12.86	13.03	22.71	19.89	15.67
TODI	28.08	22.99	0.00	7.61	13.68	9.24	28.99	22.78	16.67
UMBERTIDE	10.77	27.25	0.00	4.38	0.00	7.08	30.88	9.98	10.23
FABRO	3.87	1.97	0.00	0.80	0.00	1.94	2.37	1.37	2.54
ORVIETO	16.46	10.38	0.38	3.15	7.55	19.11	9.78	9.88	10.12
TERNI	151.99	79.18	0.00	51.52	74.04	49.99	165.62	176.14	94.74
Regional VA	484	496	18	229	233	398	532	533	360