Incorporating Spatial Structures in Ecologiacal Inference: an Information Theoretic Approach

R. Bernardini Papalia,
Department of Statistical Sciences, University of Bologna, Italy
rossella.bernardini@unibo.it

Abstract: *This paper proposes a maximum entropy (ME) – based method for modeling economic aggregates and estimating their sub-group (sub-area) decomposition when no individual or sub-group data are available. This method also offers a tractable framework for modeling the underlying variation in sub-group indicators. A basic ecological inference problem which allows for spatial heterogeneity is presented with the aim of estimating the model at the aggregate level and then employing the estimated coefficients to obtain the sub-group level indicators.*

Keywords: Generalized Cross Entropy Estimation, Ecological Inference, Spatial Heterogeneity

1. Introduction

This paper proposes a maximum entropy (ME) – based method for modeling economic aggregates and estimating their sub-group (sub-area) decomposition when no individual or sub-group data are available. This method also offers a tractable framework for modeling the underlying variation in sub-group indicators. A basic ecological inference problem which allows for spatial heterogeneity is presented with the aim of estimating the model at the aggregate level and then employing the estimated coefficients to obtain the sub-group level indicators. The latent sub-group indicators are treated as random coefficients in a regression model in which the observed aggregates are regressed on the explanatory variables both at the group and sub-group level by taking as a point of departure the approach presented in Bernardini Papalia 2008.

Our approach uses an estimation criterion based on an entropy measure of information and provides an effective and flexible procedure for reconciling micro and macro data. The maximum entropy procedures (Golan, Judge and Miller, 1996) give also 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. A 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).

2. Ecological inference assuming heterogeneity and dependence across space

2.1 The ecological inference problem

Ecological inference is the process of drawing conclusions about individual (or subgroup) level behaviour from aggregate (or group level) data, when no individual (or subgroup) data are available. The problem is that many different possible relationships at the individual (or subgroup) level can generate the same observations at the aggregate (or group) level (King 1997). In the absence of individual (or subgroup) level measurement (in the form of survey data), such information need to be inferred. The traditional approach to ecological inference is based the spatial 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).

2.2 Models with heterogeneity across space

In developing an alternative approach to ecological inference which assumes heterogeneity across space, as a point of departure we deal with the problem of decomposing aggregate indicators for various sub groups of a population by introducing unknown individual-specific effects into the model specification. This approach allows to test possible determinants of the variation in the underlying subgroup indicators. The basic idea is to treat the latent sub-group values as random coefficients in a regression model in which the observed aggregates are regressed on the population distribution by sub-groups. We define the aggregate indicator for group i, y_i, as a weighted *geometric* mean of the latent sub-group indicator y_{ij} in group i: $y_i = \prod_{j=1}^{J_i} \left(y_{ij}\right)^{\theta_{ij}}$, that is:

$$\ln y_i = \sum_{j=1}^{J_i} \left(\ln y_{ij}\right)\theta_{ij} \tag{1}$$

where y_{ij} is the indicator of the jth sub-group/sub-region in region i, θ_{ij} is the weight (as population/employees share) of sub-group/sub-region j in i, with $\sum_{j=1} \theta_{ij} = 1$, and where i=1,..,N denotes the regions and j=1,..,J_i denotes the number of sub-group/sub-regions in i. The sub-regional indicators are not observed, but the y_i 's and θ_{ij} 's are. In addition, by introducing an observed vector of explanatory variables for group i, x_i, and an observed vector of explanatory variables for sub-group/sub-region j in region i, z_{ij}, the latent sub-group indicators (values) are specified in a multiplicative form, which is consistent with a Cobb-Douglas type production function as:

$$y_{ij} = \alpha_{ij} \prod_{k=1}^{K} z_{ij,k}^{\beta_{ij,k}} \prod_{h=1}^{H} x_{i,h}^{\gamma_{ij,h}} e^{\varepsilon_{ij}} \tag{2}$$

where $z_{ij,k}$ (k=1,   K) are the covariates observed at the level of sub-group/sub-region j within the region i, and $x_{i,h}$ (h=1,..H) are the covariates observed only at the level of region i.
By substituting Eq. (2) into Eq. (1), we can obtain the following model:

$$\ln y_i = \sum_{j=1}^{J_i} \left(\ln \alpha_{ij} + \sum_{k=1}^{K} \beta_{ij,k} \ln z_{ij,k} + \sum_{h=1}^{H} \gamma_{ij,h} \ln x_{i,h} + \varepsilon_{ij} \right) \theta_{ij}, \text{ or}$$

$$\ln y_i = \sum_{j=1}^{J_i} \left(\ln \alpha_{ij} + \sum_{k=1}^{K} \beta_{ij,k} \ln z_{ij,k} + \sum_{h=1}^{H} \gamma_{ij,h} \ln x_{i,h} \right) \theta_{ij} + u_i \tag{3}$$

where $u_i = \sum_{j=1}^{J_i} \varepsilon_{ij}\theta_{ij}$ 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 region. It is important to point out that we assume: (i) unit specific coefficients for the sub-groups/sub-regions (parameter heterogeneity); (ii) a parametric specification of the unobserved spatial effects (spatial heterogeneity) through ε_{ij} 's, which can be positive or negative.

Using the estimated coefficients in Eq (3) we can obtain estimates of the unobserved or latent sub-regional indicators as follows:

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

2.3 Models with spatial dependence

In order to take into account the correlation between neighbouring areas (groups/regions) we introduce two alternative spatial model specifications: the spatial Lag and spatial error models.

When the spatial autocorrelation is modeled by a SPATIAL LAG MODEL, SPATIAL AUTOREGRESSIVE MODEL – SAR MODEL, the previous model specification (3) can be generalized by introducing a spatial-lag term $\rho \ln wy_i$ into the model. The resulting latent sub-group indicators (values) are specified in a multiplicative form as follows:

$$\ln y_i = \sum_{j=1}^{J_i} \left(\ln \alpha_{ij} + \sum_{k=1}^{K} \beta_{ij,k} \ln z_{ij,k} + \sum_{h=1}^{H} \gamma_{ij,h} \ln x_{i,h} + \rho \ln wy_i + \varepsilon_{ij} \right) \theta_{ij} \tag{5}$$

where ρ is a spatial lag coefficient (the parameter associated to the spatially lagged dependent variable, $\ln wy$), *w* is a proximity matrix of order N. This 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) regions through the spatial transformation.

In alternative, by assuming a spatial dependence is the error structure (that is a first order spatial autoregressive process), the resulting SPATIAL ERROR MODEL (SEM) specification is derived:

$$\ln y_i = \sum_{j=1}^{J_i} \left(\ln \alpha_{ij} + \sum_{k=1}^{K} \beta_{ij,k} \ln z_{ij,k} + \sum_{h=1}^{H} \gamma_{ij,h} \ln x_{i,h} + \left(\lambda w \varepsilon_{ij} + \tau_{ij} \right) \right) \theta_{ij} \tag{6}$$

where λ is a spatial autoregressive coefficient, *w* is a proximity matrix of order N, as previously defined. 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 region affects performances in that region and additionally impacts all the other regions through the spatial transformation but it measures the joint effect of misspecification, omitted variables, and spatial autocorrelation.

3. An Information theoretic approach

An entropy-based estimation approach (Golan, Judge, and Miller, 1996) is suggested as an adequate solution in the present context. More specifically, the Generalized Cross-Entropy, GCE, and the Composite Generalized Cross-Entropy, CGCE, (Bernardini Papalia, 2008) methods present some useful advantages over classical estimation techniques (as Generalized Least Squares, GLS) that refer to the possibility to (i) reformulate "ill-posed" or "under-determined" problems into "well-posed" problem, (ii) to allow for the estimation of each individual parameter *directly*; and (iii) to deal with the problem of collinearity and endogeneity arising in spatial models. The implementation of these methods require that the parameters and errors of model in Equations (5) and (6) are specified as linear combinations of some predetermined and discrete support values and unknown probabilities (weights). The estimation problem is converted into a constrained minimization problem, where the objective functions is specified through the Kullback-Leibler entropy criterion (Kullback, 1959). For the parameters: $\alpha_{ij}, \beta_{ij}, \gamma_{ij}, \rho, \theta_{ij}$, assuming: $\alpha_{ij} = s_\alpha ' p_{\alpha,ij}, \beta_{ij} = s_\beta ' p_{\beta,ij}, \gamma_{ij} = s_\gamma ' p_{\gamma,ij}, \rho = s_\rho ' p_{\rho,ij}, \varepsilon_{ij} = s_\varepsilon ' p_{\varepsilon,ij}$, we choose the support vectors $s_\alpha = (s_1^\alpha ,..s_M^\alpha)', s_\beta = (s_1^\beta ,..s_M^\beta)', s_\gamma = (s_1^\gamma ,..s_M^\gamma)', s_\rho = (s_1^\rho ,..s_M^\rho)', s_\varepsilon = (s_1^\varepsilon ,..s_R^\varepsilon)'$ and we define the corresponding unknown probability (weights) vectors as: $p_{\alpha,ij} = (p_{ij,1}^\alpha ,..p_{ij,M}^\alpha)', p_{\beta,ij} = (p_{ij,1}^\beta ,..p_{ij,M}^\beta)', p_{\gamma,ij} = (p_{ij,1}^\gamma ,..p_{ij,M}^\gamma)', p_{\rho,ij} = (p_{ij,1}^\rho ,..p_{ij,M}^\rho)', p_{\varepsilon,ij} = (p_{ij,1}^\varepsilon ,..p_{ij,R}^\varepsilon)'$, with $M,R \geq 2$, respectively. In addition, prior information is included through specifying the prior probability vectors: $\tilde{p}_{\alpha,ij}, \tilde{p}_{\beta,ij}, \tilde{p}_{\gamma,ij}, \tilde{p}_{\rho,ij}, \tilde{p}_{\varepsilon,ij}$ reflecting subjective information or any other sample and pre-sample information. 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:

$$H = \sum_i \sum_j \left(p_{\alpha,ij}\right)' \ln\left(\frac{p_{\alpha,ij}}{\tilde{p}_{\alpha,ij}}\right) + \sum_i \sum_j \left(p_{\beta,ij}\right)' \ln\left(\frac{p_{\beta,ij}}{\tilde{p}_{\beta,ij}}\right) + \sum_i \sum_j \left(p_{\gamma,ij}\right)' \ln\left(\frac{p_{\gamma,ij}}{\tilde{p}_{\gamma,ij}}\right)$$
$$+ \sum_i \sum_j \left(p_{\rho,ij}\right)' \ln\left(\frac{p_{\rho,ij}}{\tilde{p}_{\rho,ij}}\right) + \sum_i \sum_j \left(p_{\varepsilon,ij}\right)' \ln\left(\frac{p_{\varepsilon,ij}}{\tilde{p}_{\varepsilon,ij}}\right) \qquad (7)$$

subject to:
i) data consistency conditions:

$$\ln y_i = \sum_{j=1}^{J_i} \left(s_\alpha' p_{\alpha,ij} + \sum_{k=1}^{K} \left(s_\beta' p_{\beta,ij}\right) \ln z_{ij,k} + \sum_{h=1}^{H} \left(s_\gamma' p_{\gamma,ij}\right) \ln x_{i,h} + \left(s_\rho' p_{\rho,ij}\right) \ln wy_i + \left(s_\varepsilon' p_{\varepsilon,ij}\right) \right) \theta_{ij} \qquad (8)$$

ii) adding-up constraints for probabilities.
The data consistency conditions corresponding to Equation (6), can be reformulated, as:

$$\ln y_i = \sum_{j=1}^{J_i} \left(s_\alpha' p_{\alpha,ij} + \sum_{k=1}^{K} \left(s_\beta' p_{\beta,ij}\right) \ln z_{ij,k} + \sum_{h=1}^{H} \left(s_\gamma' p_{\gamma,ij}\right) \ln x_{i,h} \right) \theta_{ij} + u_i$$
$$u_i = \sum_{j=1}^{J_i} \left(s_\varepsilon' p_{\varepsilon,ij}\right) \theta_{ij} \, . \qquad (9)$$

The optimal solutions depend of 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.

4. Concluding remarks

The disaggregation of economic data permits economic analysis at the most disaggregated level especially when high quality, more detailed data are lacking. In the present paper we present a ME-based disaggregation method capable of yielding disaggregate data consistent with prior information, resulting from the data generation process, and with the aggregate data. Our method uses an estimation criterion based on an entropy measure of information and as such provides an effective, flexible way of reconciling micro and macro data. The maximum entropy method also takes into account out-of-sample information which may be introduced either as an additional constraint on the optimization problem or by specifying particular priors for parameters and errors. An unique optimum solution may also be obtained if there are more parameters to be estimated than available moment conditions and the problem is ill-posed. If additional non-sample information from theory and/or empirical evidence exists beyond that contained in the consistency and adding-up constraints, with regard to the unknown probabilities, this information may be introduced in the form of known probabilities by means of cross-entropy formalism.

Bibliography

Barker, T., and Pesaran M.H. (1990), An introduction in disaggregation in econometric modelling. In T. Barker and H Pesaran (eds), *Disaggregation in Econometric Modelling*. London, 1-14.

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

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

Chow G., and Lin A.L. (1971), Best linear unbiased Interpolation distribution and extrapolation of time series by related series, *The Review of Economics and Statistics*, 53, 4, 372-375.

Golan, A., Judge, G., Miller, D. (1996), Maximum entropy econometrics: robust estimation with limited data. Wiley, New York.

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

Kullback, J. (1959). *Information theory and statistics.* Wiley, New York, NY.