Ecological inference and spatial heterogeneity: an entropy-based distributionally weighted regression approach

Ecological inference and spatial heterogeneity: an entropy-based distributionally weighted regression approach*

Ludo Peeters[1], **Coro Chasco**[2]

[1] KIZOK – Research Centre for Entrepreneurship and Innovation, Hasselt University, 3590 Diepenbeek, Belgium (e-mail: ludo.peeters@uhasselt.be)
[2] Department of Applied Economics, Autonomous University of Madrid, 28049 Madrid, Spain (e-mail: coro.chasco@uam.es)

Abstract. In this article we compare two competing approaches to ecological modelling using test data. The first approach is based on the "traditional" method of Ordinary Least Squares (OLS), assuming constancy of parameters across disaggregated spatial units (spatial homogeneity). The second (new) approach is based on the method of Generalised Cross-Entropy (GCE), assuming varying parameters (spatial heterogeneity). The latter approach is designated as entropy-based "distributionally weighted regression" (DWR). The two approaches are tested in a real-world application, using data on per-capita GDP for the 17 regions and some covariates for the 50 provinces of Spain. Specifically, the performances of the two approaches are assessed by examining their capability in tracking the actual per-capita GDP data for the provinces (while treating them as if they were not observed by the econometrician), and in showing evidence of spatial heterogeneity. Our findings indicate that the GCE varying-parameter approach outperforms the OLS approach in terms of predictive power. Specifically, we find that the GCE predictions make efficient use of the lower-level information that is available. In addition, it is shown that entropy-based DWR has some potential as a useful technique for investigating spatially heterogeneous relationships at the lower level of analysis that might otherwise be overlooked.

JEL classification: C21, C51, C53, O18, R12, R15

Key words: Ecological inference, spatial prediction, generalised cross-entropy, spatial heterogeneity, spatial non-stationarity

1 Introduction

Situations where the only available data are aggregated at a level other than the level of interest are quite common. This is the typical setting for "ecological inference" or EI (e.g., Freedman et al. 1998; Schuessler 1999) or "cross-level inference" (e.g., Achen and Shively 1995; Cho 2001), which can be roughly described as making inferences about individual behaviour drawn from data about aggregates. Clearly, observations at an aggregated level of analysis do not necessarily provide useful information about lower levels of analysis, particularly when spatial heterogeneity (non-stationarity) is present.

In his seminal paper, Robinson (1950) introduced the term "ecological correlation" in a situation where the unit of analysis is a group or an aggregate of people, which may be entirely different from the correlation at the individual or micro-level. This is a well-known methodological problem, designated by sociologists as the "ecological fallacy" (King 1997). Basically, the ecological fallacy consists of thinking that relationships observed for groups necessarily hold for individuals.[1]

One perspective on the EI problem is that it may be impossible to solve, because the properties of the predicted values remain unverifiable. Given that the micro-data we are interested in are not available, the accuracy of any predicted value simply cannot be verified. Accordingly, most efforts to recover disaggregate information from aggregate data generally result in "ill-posed" or "underdetermined" inverse problems (because there are more unknowns than data points), which yield a multitude of feasible solutions, due to the lack of sufficient information (Judge et al. 2004). In other words, many different possible relationships at the individual or subgroup level can generate the same observations at the aggregate or group level (King 1997; Schuessler 1999).

Despite such inauspicious conditions, some "real-world" applications require the use of EI to recover disaggregate information from aggregate data. For that very purpose, several solutions have been formulated to overcome the main obstacle of EI, the problem of confounding and aggregation bias, which is said to occur when the parameters in a regression model are correlated with the regressors. In this sense, EI can be seen as an example of spatial heterogeneity – that is, the phenomenon whereby a specific relationship (e.g., its parameters, functional specification, error specification, and so on) is not constant across spatial observations (Anselin 1990).

[1] EI is an "old" and familiar methodological problem, known to geographers as the "modifiable areal unit problem" (Openshaw and Taylor 1979; Arbia 1989), to earth scientists as the "change-of-support problem" (Chilès and Delfiner 1999), and to statisticians as "small-area estimation" (Rao 2003).

The purpose of the present article is to propose a new approach to EI, based on Generalised Cross-Entropy (GCE), which assumes varying individual or subgroup-specific parameters. Specifically, we compare the performances of two alternative approaches to EI. The first one is the traditional approach based on Ordinary Least Squares (OLS), assuming constancy of parameters across the disaggregated spatial units (spatial homogeneity) – an assumption that is rarely tenable, since the aggregation process usually generates macro-level observations across which the parameters describing individuals may vary (Cho 2001). The second one is GCE, which does not take the "constancy assumption", assuming spatial heterogeneity. The two approaches will be compared in two real-world applications or cases, using a testing procedure.

The remainder of the article is organised as follows. Section 2 introduces the theoretical framework of EI or cross-level regressions as a spatial-heterogeneity problem. Section 3 presents two alternative approaches to ecological modelling: the "classical" OLS estimation, and the new GCE estimation. In Section 4, we test the performance of these two approaches by applying them to a "real-world" data set for Spain. The last section provides concluding remarks and outlines some directions for further research.

2 The ecological inference problem

EI is the process of drawing conclusions about individual or subgroup-level behaviour from aggregate or group-level (historically labelled "ecological") data, when no individual or subgroup data are available. Ecological and micro-area correlations are certainly not equal. This phenomenon should question the – possibly fallacious – results of studies in which conclusions on micro-area behaviour has been drawn from grouped (macro or meso-area) data. The "ecological fallacy" or "ecological bias" occurs when analysis based on grouped data lead to conclusions different from those based on micro-data. The root of the problem lies on the so-called "aggregation bias" due to the differential distribution of confounding variables created by grouping (Morganstern 1982).

Although the term EI is typically used in social sciences, it is isomorphic to the "modifiable areal unit problem" (MAUP) in geography (King 1997; Gotway and Young 2004). The MAUP occurs when inference based on data aggregated to a particular set of geographical regions changes if the same data are aggregated to a different set of geographical regions (Openshaw and Taylor 1979; Arbia 1989). The MAUP involves actually two inter-related problems: (1) the scale or aggregation effect, which produces different results and inferences when data are grouped into increasingly larger areal units, and (2) the grouping or zoning effect, which reflects the variability in results due to alternative formations of the areal units. Both problems are closely connected to the aggregation bias in EI.

The aggregation bias (EI), or the scale/aggregation effect (MAUP), consists basically in a smoothing effect, which is similar to that of a spatial filter that results from the averaging outcome of aggregation. In effect, as heterogeneity among

units decreases through aggregation, the uniqueness of each unit and the dissimilarity among units is reduced. As Openshaw (1984) rightly pointed out, "(. . .) whether the ecological fallacy problem exists or not depends on the nature of the aggregation. A completely homogeneous grouping system would be free of this problem". Besides, spatial autocorrelation can be another mitigation factor when it is positive (variability is moderated in this case), but exacerbated when it is negative (e.g., Arbia 1986; Cressie 1993a).

The aggregation bias or scale/aggregation effect can be seen as an example of spatial heterogeneity, where the parameters in the ecological regression model are correlated with the regressors. In this case, the standard estimation techniques, such as OLS, are not valid. Suppose one is investigating the relationship between household income and education in a sub-region level using regional data. The data exhibit no aggregation bias if this relationship is constant, that is, the same in every region. Freedman et al. (1991) call this condition the "constancy assumption". In this context, the constancy assumption means that a determined educational level tends to provide households with the same income regardless of the region of residence. If this assumption holds, then aggregate (regional) data analysis is straightforward to estimate sub-region data using, for instance, OLS. Parameters that are constant will not be correlated with any set of regressors, and so cross-level inferences are simple (as firstly proposed by Goodman 1953). A strong assumption underlying this model, then, is that people with a determined educational level have the same income regardless of which region they live in. In real data, this assumption is generally false as, for example, salaries are not the same everywhere.

Consequently, when the aggregation bias is present in an EI, it is a typical case of extreme spatial heterogeneity or incidental parameter problem, that is a different parameter for each spatial unit (Anselin and Cho 2002). In this case, a "solution" is to impose spatial or geographical structure on the nature of the variation of the individual coefficients across observations. The typical spatial heterogeneity specifications are only a partial solution (Anselin 1990, 2000), in the sense that the parameters to be estimated are not incidental, and they must to be constrained to vary either continuously as a function of a small set of "hyperparameters" (e.g., trend-surface and expansion models, Bayesian hierarchical modelling) or in a discrete fashion by being constant across (spatial) subsets of the observations (spatial ANOVA and spatial regimes model).

Recently, new approaches have been proposed to overcome aggregation bias in EI, such as the Geographically Weighted Regression approach or GWR (Calvo and Escobar 2003). GWR is a relatively simple technique that extends the traditional regression framework by allowing local rather than global parameters to be estimated (Fotheringham and Brunsdon 1999). In the calibration of this model, one different parameter is estimated for each observation for the relationship between each independent variable and the dependent variable. Hence, this relationship is not assumed to be constant across the study region. In GWR, an observation is weighted in accordance with its proximity to point *i*, so that the weighting of an observation varies with *i*. Data from observations close to *i* are weighted more than data from observations further away.

3 Two alternative approaches to ecological modelling

In this section, we present two approaches to EI: (1) the "ecological (OLS) regression", assuming homogeneity across space, and (2) the entropy-based "Distributionally Weighted Regression" (DWR) technique, which is considered here as a way to at least partly solve the problem of aggregation bias, by incorporating spatial heterogeneity (varying parameters) in a hierarchical or multilevel model. Also, the method of Generalized Cross-Entropy (GCE) will be described.

3.1 Ecological inference assuming homogeneity across space

A frequently used model in EI is that of a simple linear "ecological regression", which can be estimated by Ordinary Least Squares (OLS). Specifically, one may run a simple OLS regression of y_i on a set of covariates $\mathbf{x}_i$, both defined at the *group* (regional) level, of the form:

$$y_i = \alpha + \sum_{k=1}^{K} \beta_k x_{i,k} + u_i, \quad i = 1, \ldots, N, \tag{1}$$

where y_i is an observed (aggregate) indicator, say, GDP per capita, for region i, $x_{i,k}$ ($k = 1, \ldots, K$) are explanatory variables for region i, and u_i are error terms that are generally assumed to be independently and normally distributed with zero mean and common variance σ_u^2.

Then, the per-capita GDP indicator at the sub-regional level can be predicted by taking the corresponding (available) covariates $z_{ij,k}$ (=$x_{ij,k}$) at the level of the *subgroups* (sub-regions) $j = 1, \ldots, M_i$, contained in the larger region i (i.e., each region i comprises a total of M_i sub-regions):

$$\hat{y}_{ij} = \hat{\alpha} + \sum_{k=1}^{K} \hat{\beta}_k z_{ij,k}, \quad i = 1, \ldots, N;\ j = 1, \ldots, M_i, \tag{2}$$

where $\hat{y}_{ij}$ is the predicted per-capita income indicator for sub-region j in region i, and $z_{ij,k}$ ($k = 1, \ldots, K$) are explanatory variables for sub-region j in region i.

However, individual behaviour can only be inferred from aggregate data under very restrictive assumptions (Goodman 1953, 1959). Specifically, a problem of "local" estimation bias may arise due to a false assumption of constancy of the parameters across the spatial units (spatial homogeneity) within each region i (e.g., Cho 2001), because of the non-zero correlation that is assumed to exist between the (unobserved) y_{ij} and the associated u_{ij} (e.g., Holt et al. 1996). This is the problem usually referred to as the "ecological fallacy" (an unfortunate effect of aggregation and confounding) mentioned above.

Varying-parameter model. In developing our alternative approach to ecological inference, we take Bidani and Ravallion (1997) as a point of departure. In their paper, they are dealing with the problem of decomposing aggregate (health) indicators using a random-coefficients model in which the aggregates are regressed on the population distribution by sub-groups, taking into account the statistical properties of the error terms. Their approach allows testing possible determinants of the variation in the underlying subgroup indicators. More precisely, they are dealing with the problem of retrieving indicators for various sub-groups of a population. The latent sub-group values are treated as random coefficients in a regression of the observed aggregates on the distributional data.

To illustrate their approach, consider the following identity in which the index $i = 1, \ldots, N$ denotes the regions, and index $j = 1, \ldots, M_i$ denotes the sub-regions in each of the regions i:

$$y_i = \sum_{j=1}^{M_i} y_{ij}\eta_{ij}, \quad i = 1, \ldots, N. \tag{3}$$

In other words, the group i values are treated as a weighted *arithmetic* mean of the latent sub-group values j in group i, where y_i is the aggregate per-capita indicator for region i, y_{ij} is the per-capita indicator of the j-th sub-region in region i, η_{ij} is the share of the population of sub-region j in the total population of region i, with $0 \leq \eta_{ij} \leq 1$, for all i, j, and $\Sigma_{j=1}^{M_i}\eta_{ij} = 1$, for all i. Obviously, this model implies a weighted regression, capturing distributional "effects" by using data on population shares.

The sub-regional indicators y_{ij} are not observed, whereas the y_i's and η_{ij}'s are. Now, if we can observe covariates for each sub-region j in region i, and possibly also covariates for each region i, we attain:

$$y_{ij} = \alpha_{ij} + \sum_{k=1}^{K}\beta_{ij,k}z_{ij,k} + \sum_{h=1}^{H}\gamma_{ij,h}x_{i,h} + u_{ij}, \quad i = 1, \ldots, N;\ j = 1, \ldots, M_i, \tag{4}$$

where $z_{ij,k}$ ($k = 1, \ldots, K$) are the covariates observed at the level of sub-region j within region i, and $x_{i,h}$ ($h = 1, \ldots, H$) are the covariates observed only at the level of region i.

On substituting (4) into (3), yields the following regression equation:

$$y_i = \sum_{j=1}^{M_i}\left(\alpha_{ij} + \sum_{k=1}^{K}\beta_{ij,k}z_{ij,k} + \sum_{h=1}^{H}\gamma_{ij,h}x_{i,h}\right)\eta_{ij} + \varepsilon_i, \quad i = 1, \ldots, N, \tag{5}$$

where $\varepsilon_i = \Sigma_{j=1}^{M_i} u_{ij}\eta_{ij}$ is a composite error term, which is heteroskedastic. Essentially, this model implies some kind of weighted regression, capturing "distributional effects" by using data on population shares for each region.

Next, using the regression in (5), we can predict the unobserved (latent) sub-regional indicators as:

$$\hat{y}_{ij} = \hat{\alpha}_{ij} + \sum_{k=1}^{K} \hat{\beta}_{ij,k} z_{ij,k} + \sum_{h=1}^{H} \hat{\gamma}_{ij,h} x_{i,h}, \quad i = 1, \ldots, N;\ j = 1, \ldots, M_i. \tag{6}$$

In contrast with Bidani and Ravallion (1997), though, we define y_i now as the weighted g*eometric* mean of the y_{ij}'s within region *i*:

$$y_i = \prod_{j=1}^{M_i} \left(y_{ij}\right)^{\eta_{ij}}, \quad i = 1, \ldots, N, \tag{7}$$

or

$$\ln y_i = \sum_{j=1}^{M_i} \eta_{ij} \ln y_{ij}, \quad i = 1, \ldots, N. \tag{8}$$

Furthermore, the latent sub-group values are specified in a multiplicative form, which is consistent with a Cobb-Douglas type of production function:

$$y_{ij} = \alpha_{ij} \prod_{k=1}^{K} z_{ij,k}^{\beta_{ij,k}} \prod_{h=1}^{H} x_{i,h}^{\gamma_{ij,h}} e^{\theta_{ij}}, \quad i = 1, \ldots, N;\ j = 1, \ldots, M_i. \tag{9}$$

Then, on substituting (9) into (8), we arrive at:

$$\ln y_i = \sum_{j=1}^{M_i} \left(\ln \alpha_{ij} + \sum_{k=1}^{K} \beta_{ij,k} \ln z_{ij,k} + \sum_{h=1}^{H} \gamma_{ij,h} \ln x_{i,h} + \theta_{ij} \right) \eta_{ij}, \quad i = 1, \ldots, N. \tag{10}$$

Two remarks on Equation (10) are in order: (i) the model assumes *unit-specific* coefficients for the sub-regions, thereby allowing for (continuous) parameter variation; (ii) we use a *parametric* specification of the unobserved effects, through the θ_{ij}'s, which can be positive or negative.

Although this model is clearly underdetermined (i.e., the number of unknown parameters is larger than the number of observations), it can be estimated by using the "non-classical" maximum-entropy method, which will be discussed later in

this article. The unobserved (latent) sub-regional per-capita indicators can be predicted as:

$$\hat{y}_{ij} = \hat{\alpha}_{ij} \prod_{k=1}^{K} z_{ij,k}^{\hat{\beta}_{ij,k}} \prod_{h=1}^{H} x_{i,h}^{\hat{\gamma}_{ij,h}} e^{\hat{\theta}_{ij}}, \quad i = 1, \ldots, N; j = 1, \ldots, M_i. \tag{11}$$

Although the model we propose belongs to a class of other models that are particularly designed for estimating varying parameters,[2] our approach is remarkably different at least in two major respects. Firstly, the *hierarchical* (two-level) structure of the model allows to usefully exploiting lower-level information for modelling relationships at the sub-regional level. Secondly, the model adopts a *parametric* specification of spatial heterogeneity, which means that each individual coefficient at the sub-regional level is treated as a fixed or unique (unknown) value.

Generalized cross-entropy estimation. Although the regression in (10) can be treated as a "classical" random-coefficients models (Bidani and Ravallion 1997) and can be estimated by using, say, Generalized Least Squares (GLS), we prefer to use the Generalized Cross-Entropy (GCE) method, which is based on the well-known Kullback-Leibler entropy criterion (Golan et al. 1996).

GCE has some useful advantages over the classical estimation techniques, where the most important advantage, in the present context, is that it allows to reformulate the fundamentally "ill-posed" or "under-determined" problem into a "well-posed" problem, given that the number of parameters to be estimated, $(2 + K + H)M$, is larger than the number of observations available (N on the dependent variable, y_i, and M and N on the covariates $z_{ij,k}$ and $x_{i,h}$, respectively). This, in turn, allows for the estimation of each individual parameter *directly*, rather than "predicting" them, as is usually done in a classical random-coefficients modelling framework.

The practical implementation of the GCE method requires that the parameters of the model in Equation (10) are specified as linear combinations of some predetermined and discrete support values and unknown probabilities (weights). Furthermore, the estimation problem is converted into a constrained minimisation problem, where the objective function, specified in the Equation (12) below consists of the joint cross-entropy.

Specifically, we define unknown probability (weight) vectors $\mathbf{p}_{\alpha,ij} = [p^{\alpha}_{ij,1}, \ldots, p^{\alpha}_{ij,Q}]'$, $\mathbf{p}_{\beta,ij} = [p^{\beta}_{ij,1}, \ldots, p^{\beta}_{ij,Q}]'$, $\mathbf{p}_{\gamma,ij} = [p^{\gamma}_{ij,1}, \ldots, p^{\gamma}_{ij,Q}]'$, and $\mathbf{p}_{\theta,ij} = [p^{\theta}_{ij,1}, \ldots,$

[2] Other noticeable approaches in EI for coping with varying parameters are, among others, the expansion method (e.g., Cassetti and Jones 1992), the method of spatial adaptive filtering (e.g., Gorr and Olligschlaeger 1994), (mixed) geographically weighted regression (e.g., Brunsdon et al. 1996; Calvo and Escobar 2003; Mei et al. 2004), switching regression (Cho 2001), random-coefficients modelling (e.g., King 1997; Greene 2004), or multi-level modelling (e.g., Goldstein 1987). Other solutions to this same problem are the geostatistics "kriging" methods (Cressie 1993b, Chapter 3), the multi-scale and hierarchical modelling (e.g., Goldstein 1987; for a recent review, see Gotway and Young 2002), as well as some model-dependent small area estimation methods (Ghosh 2001; Rao 2003).

$p_{ij,R}^{\theta}\big]'$, with $Q, R \geq 2$, and choose the corresponding (common) support vectors $\mathbf{s}_{\alpha} = \left[s_1^{\alpha}, \ldots, s_Q^{\alpha}\right]'$, $\mathbf{s}_{\beta} = \left[s_1^{\beta}, \ldots, s_Q^{\beta}\right]'$, $\mathbf{s}_{\gamma} = \left[s_1^{\gamma}, \ldots, s_Q^{\gamma}\right]'$, and $\mathbf{s}_{\theta} = \left[s_1^{\theta}, \ldots, s_R^{\theta}\right]'$, for the parameters α_{ij}, β_{ij}, γ_{ij}, and θ_{ij}, respectively, where $\alpha_{ij} = \mathbf{s}_{\alpha}'\mathbf{p}_{\alpha,ij}$, $\beta_{ij} = \mathbf{s}_{\beta}'\mathbf{p}_{\beta,ij}$, $\gamma_{ij} = \mathbf{s}_{\gamma}'\mathbf{p}_{\gamma,ij}$, and $\theta_{ij} = \mathbf{s}_{\theta}'\mathbf{p}_{\theta,ij}$. In addition, prior information is included through specifying the prior probability vectors $\tilde{\mathbf{p}}_{\alpha,ij}$, $\tilde{\mathbf{p}}_{\beta,ij}$, $\tilde{\mathbf{p}}_{\gamma,ij}$, and $\tilde{\mathbf{p}}_{\theta,ij}$, reflecting subjective information, informed "guesses", or any other sample and pre-sample information.

After the appropriate re-parameterisation, the complete GCE optimization problem for the ecological model, corresponding to Equation (10), can be formulated as:

$$\begin{aligned} \underset{\mathbf{p}}{\text{Min}}\, CE = & \sum_{i=1}^{N}\sum_{j=1}^{M_i}\left(\mathbf{p}_{\alpha,ij}\right)'\ln\left(\frac{\mathbf{p}_{\alpha,ij}}{\tilde{\mathbf{p}}_{\alpha,ij}}\right) + \\ & \sum_{i=1}^{N}\sum_{j=1}^{M_i}\left(\mathbf{p}_{\beta,ij}\right)'\ln\left(\frac{\mathbf{p}_{\beta,ij}}{\tilde{\mathbf{p}}_{\beta,ij}}\right) + \sum_{i=1}^{N}\sum_{j=1}^{M_i}\left(\mathbf{p}_{\gamma,ij}\right)'\ln\left(\frac{\mathbf{p}_{\gamma,ij}}{\tilde{\mathbf{p}}_{\gamma,ij}^{\gamma}}\right) + \\ & \sum_{i=1}^{N}\sum_{j=1}^{M_i}\left(\mathbf{p}_{\theta,ij}\right)'\ln\left(\frac{\mathbf{p}_{\theta,ij}}{\tilde{\mathbf{p}}_{\theta,ij}}\right) \end{aligned} \tag{12}$$

subject to:

$$\ln y_i = \sum_{j=1}^{M_i}\eta_{ij}\left[\mathbf{s}_{\alpha}'\mathbf{p}_{\alpha,ij} + \sum_{k=1}^{K}\left(\mathbf{s}_{\beta}'\mathbf{p}_{\beta,ij}\right)\ln z_{ij,k} + \sum_{h=1}^{H}\left(\mathbf{s}_{\gamma}'\mathbf{p}_{\gamma,ij}\right)\ln x_{i,h} + \mathbf{s}_{\theta}'\mathbf{p}_{\theta,ij}\right] \forall i, \tag{13}$$

$$\sum_{q=1}^{Q} p_{ij,q}^{\alpha} = \sum_{q=1}^{Q} p_{ij,q}^{\beta} = \sum_{q=1}^{Q} p_{ij,q}^{\gamma} = \sum_{r=1}^{R} p_{ij,r}^{\theta} = 1 \quad \forall i, j, \tag{14}$$

$$\sum_{q=1}^{Q} \tilde{p}_{ij,q}^{\alpha} = \sum_{q=1}^{Q} \tilde{p}_{ij,q}^{\beta} = \sum_{q=1}^{Q} \tilde{p}_{ij,q}^{\gamma} = \sum_{r=1}^{R} \tilde{p}_{ij,r}^{\theta} = 1 \quad \forall i, j. \tag{15}$$

Equation (12) denotes the cross-entropy objective, which has to be minimised subject to the data-consistency constraints in (13). The "normalisation" constraints in (14) and (15) ensure that all unknown and prior probabilities, respectively, add up to one.

The principle of minimum CE means that, given the various constraints, we are choosing the estimates of the unknown $\mathbf{p}_{\alpha,ij}$, $\mathbf{p}_{\beta,ij}$, $\mathbf{p}_{\gamma,ij}$, and $\mathbf{p}_{\theta,ij}$ that can be discriminated from the priors $\tilde{\mathbf{p}}_{\alpha,ij}$, $\tilde{\mathbf{p}}_{\beta,ij}$, $\tilde{\mathbf{p}}_{\gamma,ij}$, and $\tilde{\mathbf{p}}_{\theta,ij}$ with a minimum of difference (Golan et al. 1996, p. 11). In other words, we are looking for the "least informative" (i.e., most close to the uniform) probability distributions that are consistent with the

data and other constraints, and with the prior information reflected in the support ranges and the prior probabilities.

From the solution of this optimisation programme, the varying coefficients and the residual terms can be calculated as follows: $\hat{\alpha}_{ij} = \mathbf{s}'_{\alpha}\hat{\mathbf{p}}_{\alpha,ij}$, $\hat{\beta}_{ij} = \mathbf{s}'_{\beta}\hat{\mathbf{p}}_{\beta,ij}$, $\hat{\gamma}_{ij} = \mathbf{s}'_{\gamma}\hat{\mathbf{p}}_{\gamma,ij}$, and $\hat{\theta}_{ij} = \mathbf{s}'_{\theta}\hat{\mathbf{p}}_{\theta,ij}$. Hence, a total of $(1 + K + H)MQ + MR$ unknown probabilities have to be estimated with only N observations on the dependent variable (y_i).

4 Real-world application

In this section, the entropy-based DWR technique will be applied to regional and sub-regional data on per-capita GDP for Spain, for the year 2000. Specifically, our intention is to predict GDP per capita at the (sub-regional) level of the 50 provinces, assuming that GDP data (used in measuring the dependent variable) are only available at the (regional) level of the 17 autonomous communities.

4.1 Definition of variables and data sources

A relationship is posited between GDP per capita (*gdp*), on the one hand, and the primary inputs labour (*lab*) and capital (*cap*), corrected for R&D (*r&d*). The variable *lab* is defined as the employment rate, while *cap* is defined as the real capital stock (base year 1990) per capita, and *r&d* is defined as R&D expenditures per capita. In addition, we take into account the existence of spatial externalities or agglomeration economies (i.e., scale/localisation and scope/urbanisation economies),[3] which are assumed to be driven by the population density (*pop*) in each of the provinces. For convenience, we treat population agglomeration as an exogenous source of spatial externalities,[4] and the effects of *r&d* and *pop* on GDP per capita are assumed to be Hicks neutral (see also, e.g., Henderson 2003).

Data on labour, capital, and population density are available at the sub-regional (province) level, whereas data on R&D expenditures are available only at the regional (autonomous community) level. This disadvantage is dealt with by assuming that R&D expenditures per capita are identical in all provinces within the same community. More details on the variables are provided in Table 1.

---

[3] A better but longer name would be *net* agglomeration economies. This terminology would make explicit that the effect on GDP per capita depends on (positive) externalities, on the one hand, and congestion, on the other.

[4] This contrasts with Ciccone (2002, p. 214), who treated agglomeration effects as endogenous and, accordingly, used an instrument for regional employment density. However, here it can reasonably be expected that correcting for endogeneity (due, for example, to migration from relatively "poor" (low per-capita GDP) provinces to relatively "rich" (high per-capita GDP) provinces) would only have a minor or negligible effect on the estimation results, given the fact that the population-density variable (*pop*) is quite persistent or "sticky" over time. Investigating this potential endogeneity problem (e.g., in a dynamic panel-data framework) is also far beyond the scope of the present article.

Table 1. Variables used in the empirical application and summary statistics[a]

Variable	Numerator	Denominator	Mean	Standard deviation	Coefficient of variation (%)
Gdp	Gross Domestic Product (1,000 Euros)	Population			
– community			13.547	2.781	20.5
– province			12.787	2.685	21.0
Lab	Employment (1,000 units)	Population			
– community			0.380	0.036	9.6
– province			0.372	0.043	11.7
Cap	Real Capital stock (1,000 Euros)	Population			
– community	(base = 1990)		0.274	0.044	16.0
– province			0.273	0.057	20.9
Pop	Population (1,000 units)	Area (km^2)			
– community			0.141	0.153	108.2
– province			0.113	0.144	128.2
r&d	R&D expenditures (1,000 Euros)	Population			
– community			1.170	0.755	64.6

[a] INE, Instituto Nacional de Estadística (*gdp*, *lab*, *pop*, *r&d*); Fundación BBVA, Instituto Valenciano de Investigaciones Económicas (*cap*).

We aggregate the per-capita GDP data at the level of the 17 autonomous communities (regions), deliberately "losing" information at the level of the 50 provinces (sub-regions), and then use the two proposed methods (OLS versus GCE) to make ecological inferences about GDP per capita at the level of the provinces. We present results for the ecological OLS model ("Eco-OLS") corresponding to Equation (1) and the ecological GCE model ("Eco-GCE") corresponding to Equation (10).

4.2 Ecological OLS model

The linear regression model, designated as the "Eco-OLS" model, is given by:

$$\ln gdp_i = \alpha + \beta_1 \ln lab_i + \beta_2 \ln cap_i + \beta_3 \ln pop_i + \gamma \ln r\&d_i + u_i. \qquad (16)$$

The results of the OLS estimation are reported in column [1] of Table 2. Overall, the OLS results are quite satisfactory. All the slope coefficients have a positive sign (consistent with prior expectations) and all of them (but one) are statistically significant at the 10% level. However, the OLS model assumes "fixed" coefficients and, as a result, the estimates may suffer from aggregation bias. In addition, the estimated coefficients do not reveal any variation in the relationship across the provinces.

Interestingly, though, the value for the estimated *pop* coefficient is equal to 0.041 (i.e., estimate of the agglomeration effects is 4.1%), which implies that a doubling of the population density would lead to an increase in per-capita GDP of roughly 4.0%, ceteris paribus. This value is very close to the results (around 5.0%) obtained by Ciccone and Hall (1996) and Ciccone (2002), for the U.S. counties and some European *Nuts 3*-regions, respectively.

4.3 Ecological GCE model

The varying-coefficients "Eco-GCE" model is as follows:

$$\begin{aligned} \ln gdp_i = \sum_{j=1}^{M_i} \big(\alpha_{ij} + \beta_{1,ij} \ln lab_{ij} + \beta_{2,ij} \ln cap_{ij} + \beta_{3,ij} \ln pop_{ij} + \\ \gamma_{ij} \ln r\&d_i + \theta_{ij}\big)\eta_{ij} \quad \forall i. \end{aligned} \qquad (17)$$

For implementing GCE, we have to choose appropriate support vectors for the unknown parameters. Given the uncertainty concerning the values of the estimates, we choose $\mathbf{s}_\alpha = \mathbf{s}_\beta = \mathbf{s}_\gamma = (-100, -50, -10, 0, 10, 50, 100)'$, with $Q = 7$, and $\mathbf{s}_\theta = (-100,\ 100)'$, with $R = 2$. Although somewhat arbitrarily defined, these

Table 2. Estimation results from OLS and GCE with standard errors in parentheses[a]

	Eco-OLS [1]	Eco-GCE [2]
constant	4.106 (0.175)***	4.106 [0.006]
Min.		4.082
Max.		4.120
C.V.%		0.1%
lab (employment rate)	0.456 (0.282)	0.456 [0.006]
Min.		0.443
Max.		0.478
C.V.%		1.3%
cap (capital stock)	0.748 (0.156)***	0.748 [0.008]
Min.		0.726
Max.		0.779
C.V.%		1.1%
r&d (R&D expenditures)	0.074 (0.035)*	0.074 [0.002]
Min.		0.062
Max.		0.080
C.V.%		3.2%
pop (population density)	0.041 (0.019)**	0.042 [0.010]
Min.		0.021
Max.		0.080
C.V.%		23.0%
$\hat{\theta}_{ij}$ (unobserved effects)		0.00001 [0.00219]
Min.		–0.00839
Max.		0.00448
R^2	0.936	
SER	0.061	
N	17	17 (50)

[a] The symbols *, **, *** indicate significance at the 10, 5, and 1% levels, respectively. In column [2], the standard deviations of the varying parameter estimates are reported between square brackets. *SER* is standard error of regression.

support ranges are expected to be wide enough to include all "possible" values. In addition, as prior information we use the obtained OLS estimates at the level of the autonomous communities (regions). For example, $\mathbf{s}'_{\alpha}\tilde{\mathbf{p}}_{\alpha,ij} = \hat{\alpha}_{\text{OLS}} = 4.106$ (see column [1] in Table 2), and so on.

Column [2] of Table 2 shows the results from the GCE-based model. Specifically, the mean values of the 50 estimated varying parameters, along with the estimated standard deviations are reported.[5] It should be noted that the GCE procedure ensures (by construction) that the means of the individual coefficients are equal to the corresponding OLS estimates reported in column [1] of Table 2.

[5] The GCE method is implemented by using the GAMS software package (CONOPT3 solver).

Table 3. Prediction accuracy measures

	Eco-OLS [1]	Eco-GCE [2]
Pseudo-R^2	0.834	0.884
RMSE	1.217	1.008
MAPE	7.1%	5.5%
Spearman's *rho*	0.914	0.942

From the reported standard deviations (coefficients of variation) in column [2] of Table 2, it can be seen that the coefficients associated with *lab*, *cap*, and *r&d* are quite stable across the provinces, whereas the coefficients associated with *pop* display a (relatively) wide variation. In other words, spatial externalities seem to be markedly different across the provinces. Finally, the estimated unobserved effects are, captured by the estimated θ_{ij}'s, turn out to be negligible in size (i.e., the minimum and maximum values are −0.008 and 0.004, respectively), and, thus, hardly affecting the predicted $\hat{y}_{ij}$'s.

5 Predicting province level GDP per capita

In order to evaluate the predictive performances of the EI models, we examine their ability of "tracking" the actual GDP per capita data for the Spanish provinces, which have not been used in the estimation process.

Traditional measures of prediction accuracy are presented in Table 3. The statistics reported are the pseudo-R^2, the mean absolute percentage error (*MAPE*), and the root mean squared error (*RMSE*). The pseudo-R^2 is simply defined as the squared correlation between the actual values, y_{ij}, and the predicted values, $\hat{y}_{ij}$. The *RMSE* is defined as:

$$RMSE = \sqrt{\frac{1}{\Sigma_i M_i} \sum_{i=1}^{N} \sum_{j=1}^{M_i} \left(y_{ij} - \hat{y}_{ij}\right)^2}, \tag{18}$$

while the *MAPE* is defined as:

$$MAPE = \frac{1}{\Sigma_i M_i} \sum_{i=1}^{N} \sum_{j=1}^{M_i} \left|y_{ij} - \hat{y}_{ij}\right| \Big/ y_{ij} \times 100. \tag{19}$$

By all measures, the GCE-based predictions are more accurate than the OLS-based predictions, in terms of pseudo-R^2, *MAPE*, and *RMSE*. Most noticeable is the 1.6% point reduction in the *MAPE* compared to OLS.

Finally, we present Spearman's (*rho*) rank correlations, to test the correspondence in the rankings between the predicted and observed values at the level of the province. Clearly, the GCE-based predictions are more closely matching the ranking of the observed values.

6 Spatial heterogeneity

Since the OLS model of EI assumes "fixed" parameters, the estimated coefficients do not provide any information about possible variations across the spatial units (regions and/or provinces). This is a major drawback of the OLS model in the context of EI. On the other hand, the GCE model is designed to provide some evidence of spatial heterogeneity; that is, to produce maps of coefficient variation over the provinces within each autonomous community.

One limitation of the present analysis, however, is that without direct observations on the individual parameters, there is no way to verify the spatial heterogeneity.

6.1 Measures of dispersion

The results from GCE presented in column [2] of Table 2 are indicative of the variability of the coefficients across spatial units. The dispersion is measured as the standard deviation and the coefficient of variation.

Obviously, the estimated coefficients of the *pop* variable exhibit a much wider dispersion than the other coefficients, with coefficients of variation of 23.0%. It is not clear, though, why this is so. On the other hand, our results are broadly in line with Ciccone's (2002) findings for some European countries, where the estimated agglomeration (density) effect ranges from 4.8% for Germany to 8.0% for the UK. However, given the fact that we use a total of 50 observations, taken at a more disaggregated geographical level, we find a wider dispersion of the agglomeration effects across Spanish provinces ranging from a minimum of only 2.1% (Las Palmas) to a maximum of 8.0% (Valencia), with an average value of 4.2%.

In order to provide a more informative picture of the variations across the provinces, Figure 1 presents the kernel density plot for the distribution of the estimated, unit-specific *pop* coefficients.[6] A cursory look at this plot reveals the existence of (at least) four distinct segments or "clusters" in the distribution of the *pop* coefficients. Furthermore, the clusters can be characterised by estimating a four-component finite normal-mixture model (see also, e.g., Tsionas 2000). The means and standard deviations for these four clusters are shown in Table 4. They reveal no or only minimal overlap. A simple analysis of variance (one-way ANOVA) further indicates that the mean coefficients are significantly different across the four clusters ($F_{3,46} = 81.65$, with $p < 0.0001$). Thus, the variability in the

[6] The kernel density is calculated using a Gaussian kernel with an adaptive bandwidth parameter set as in Silverman (1986).

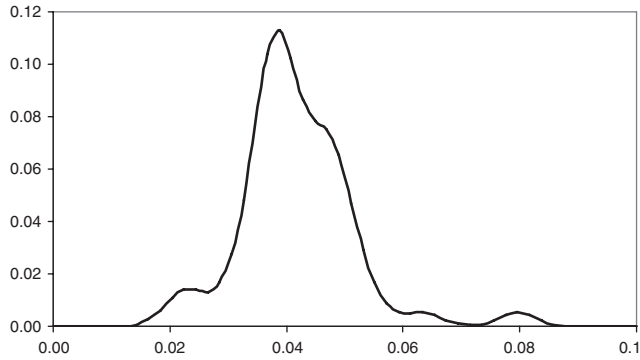

Fig. 1. Kernel density of estimated agglomeration effects from GCE (individual *pop* coefficients)

Table 4. Means and standard deviations for the four identified clusters of *pop* coefficients

Clusters c	Means μ_c	Standard deviations σ_c	Probabilities p_c	Number of provinces n_c
1	0.030	0.006	0.164	8
2	0.038	0.002	0.424	21
3	0.047	0.003	0.362	18
4	0.067	0.011	0.050	3
Total	0.042	0.010	1.000	50

workings of spatial externalities cannot reasonably be dismissed. Finally, it can be seen that the distribution is strongly (positively) skewed, with only three provinces situated at the right tail. In other words, the agglomeration effects seem to be quite similar in most of the Spanish provinces (4.0%, on average), whereas they are very high (6.7%, on average) in only three provinces.

6.2 Mapping spatial heterogeneity

Differences in the effect of spatial externalities may arise from a large number of factors. They can differ because of differences in the industry mix across provinces, such as services (in densely populated provinces) vs. manufacturing (in less densely populated provinces), "old" vs. "new" sectors of the economy at varying stages in the product life-cycle, small vs. large firms (varying firm-size distributions), and so on. Also, spatial externalities can be of a different type across provinces: (i) localisation economies, (ii) urbanisation economies, and (iii) "activity-complex" economies, originating from strong input-output linkages (Parr 2002).

While a thorough analysis (explanation) of these differences is beyond the scope of the present article, it is interesting to look at the geographical distribution

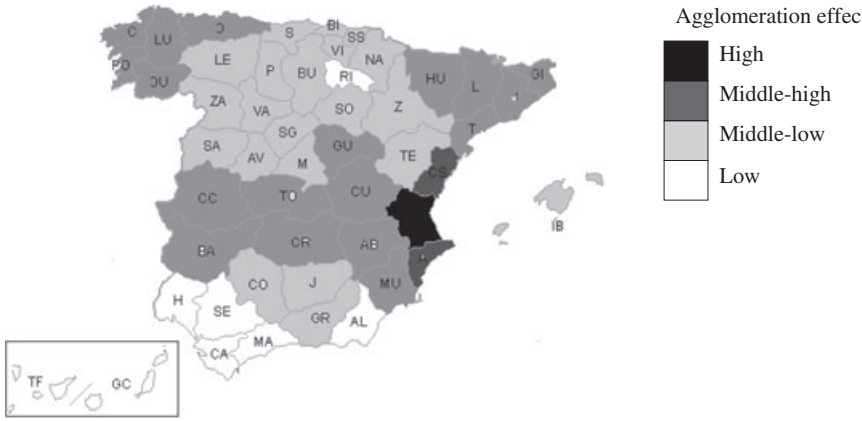

Fig. 2. Geographical distribution of estimated levels of agglomeration effect from GCE (*pop* coefficients), based on four-components normal-mixture model

of the estimated agglomeration effects as shown in Figure 2. From this map it can be seen that the effects of spatial externalities are the highest in: (i) the provinces of the Valencian Community and Murcia (east), Catalonia (north-east), which are areas with high-tech industries and services (mainly urbanisation and/or "activity-complex" economies), and (ii) the provinces of Castile-La Mancha (south/east of Madrid) and Extremadura (west), which are predominantly rural (farm-based) areas, benefiting to some extent from the urban sprawl of Madrid, and (iii) Galicia and Asturias (north-west), which are areas with traditional, heavy industries (mainly localisation economies). The leading position of Valencia is consistent with earlier observations (OECD 2000) of the existence of "local pockets" in many countries, exhibiting extremely high levels of entrepreneurial activity and benefiting from strong "information spillovers" (e.g., Valencia in Spain, Arezzo and Modena in Italy, Nüremberg in Germany, and so on). The relatively strong effect of spatial externalities in the rural provinces of (ii) may be attributable to the absence of (negative) congestion effects in these sparsely populated areas. The same can be said about Valencia, with a low level of congestion, and Barcelona and Cantabria, with moderate or decreasing levels of congestion (see also Brañas-Garza and Alcalá-Olid 2000). On the other hand, spatial externalities seem to be of relatively minor importance in the north-central part of Spain, in particular in the provinces of Castile and Leon, Cantabria, the Basque Country, Navarre, Aragon, where in most areas economic activity is based on heavy industries that are not highly labour demanding, as well as the northern provinces of Andalusia, where the population mainly consists of small farmers, generating a lower per-capita GDP. Finally, spatial externalities seem to be minimal in the province of La Rioja, as well as in the southern provinces of Andalusia and on the Canary and Balearic Islands. In the case of La Rioja, there has been a major influx of people from neighbouring regions recently, whereas in the cases of southern Andalusia and the Islands, the population is increasing mainly with non-active people (European "jubilees").

7 Conclusions and directions for further research

In this article we estimated a varying-parameter EI model by regressing aggregate GDP per capita for the (17) Spanish autonomous communities against data on the distribution of the population at the level of the (50) Spanish provinces, allowing for differential effects of employment rate, capital stock per capita, R&D expenditures per capita, and population density. Individual coefficients and per-capita GDP predictions were then retrieved for each of the Spanish provinces. The estimation was performed using the method of Generalised Cross-Entropy (GCE).

Subsequently, we evaluated the performance of this entropy-based "distributionally weighted regression" (DWR) by comparing its predictions with those from a simple OLS-based EI regression. We found that the predictions from the GCE-based model are "superior" to those from the OLS-based model, in terms of accuracy. Furthermore, it was shown that the entropy-based DWR approach is a useful tool for exploring spatially heterogeneous relationships at the disaggregate level that might otherwise be overlooked or missed. In particular, our analysis demonstrated that differences in GDP per capita at the level of the Spanish provinces can largely be explained by differences in the effects of spatial externalities.

An important question requiring further research is to investigate how the (predicted) differences in spatial externalities across provinces are related to industry structure. In particular, spatial externalities may be stronger in some industries than in others. In densely populated areas one may find highly productive industries.

Some other directions for future research, related to the methodology, are to use different "mixed" models (for identifying the most appropriate mix of regional and sub-regional covariates), to test the role of spatial effects in ecological inference, to examine alternative ways of incorporating distributional effects, and to assess the sensitivity of the GCE results to the particular choice of informative priors and support ranges.

References

Achen CH, Shively WP (1995) *Cross-level inference*. Chicago University Press, Chicago

Anselin L (1990) Spatial dependence and spatial structural instability in applied regression analysis. *Journal of Regional Science* 30: 185–207

Anselin L (2000) The alchemy of statistics, or creating data where no data exist. *Annals of the Association of American Geographers* 90: 93–155

Anselin L, Cho WKT (2002) Spatial effect and ecological inference. *Political Analysis* 10: 276–297

Arbia G (1986) The modifiable areal unit problem and the spatial autocorrelation problem: towards a joint approach. *Metron* 44: 391–407

Arbia G (1989) *Spatial data configuration in statistical analysis of regional economics and related problems*. Kluwer Academic Publishers, The Netherlands

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

Brañas-Garza P, Alcalá-Olid F (2000) Different paths of urban agglomeration in Spanish regions: Evidence from 1960–2000. *Research Bulletin* 24, Globalization and World Cities Study Group and Network

Brunsdon C, Fotheringham AS, Charlton ME (1996) Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis* 28: 281–298

Calvo E, Escobar M (2003) The local voter: a geographically weighted approach to ecological inference. *American Journal of Political Science* 47: 189–204

Cassetti E, Jones JP (1992) *Applications of the expansion method*. Routledge, London

Chilès JP, Delfiner P (1999) *Geostatistics: Modeling spatial uncertainty*. John Wiley, New York

Cho WKT (2001) Latent groups and cross-level inferences. *Electoral Studies* 20: 243–263

Ciccone A (2002) Agglomeration effects in Europe. *European Economic Review* 46: 213–227

Ciccone A, Hall RE (1996) Productivity and the density of economic activity. *American Economic Review* 86: 54–70

Cressie N (1993a) *Statistics for spatial data*. John Wiley, New York

Cressie N (1993b) Aggregation in geostatistical problems. In: Soares A (ed) *Geostatistics troia 1992*, vol. 1. Kluwer Academic Publishers, Dordrecht

Fotheringham AS, Brunsdon C (1999) Local forms of spatial analysis. *Geographical Analysis* 31: 340–358

Freedman DA, Klein SP, Ostland M, Roberts MR (1998) Review of "A solution to the ecological inference problem". *Journal of the American Statistical Association* 93: 1518–1522 (with discussion in vol. 94, 1999: 352–357)

Freedman DA, Klein S, Sacks J, Smyth C, Everett C (1991) Ecological regression and voting rights. *Evaluation Review* 15: 673–711

Golan A, Judge G, Miller D (1996) *Maximum entropy econometrics: Robust estimation with limited data*, John Wiley & Sons, New York

Goldstein H (1987) *Multilevel models in educational and social research*. Oxford University Press, London

Goodman LA (1953) Ecological regressions and the behavior of individuals. *American Sociological Review* 18: 663–666

Goodman LA (1959) Some alternatives to ecological correlation. *American Journal of Sociology* 64: 610–625

Gorr WL, Olligschlaeger AM (1994) Weighted spatial adaptive filtering: Monte Carlo studies and application to illicit drug market models. *Geographical Analysis* 26: 67–87

Ghosh M (2001) Model-dependent small area estimation – Theory and practice. In: Lehtonen R, Djerf K (eds) *Lecture notes on estimation for population domains and small areas*. Statistics Finland, Helsinki

Gotway CA, Young LJ (2002) Combining incompatible spatial data. *Journal of the American Statistical Association* 97: 632–648

Gotway CA, Young LJ (2004) A spatial view of the ecological inference problem. In: King G, Rosen O Tanner MA (eds) *Ecological inference. New methodological strategies*. Cambridge University Press

Greene W (2004) Interpreting estimated parameters and measuring individual heterogeneity in random coefficients models. Department of Economics, Stern School of Business, New York University, May 6, 2004 (available from www.stern.nyu.edu/~wgreene)

Henderson JV (2003) Marshall's scale economies. *Journal of urban Economics* 53: 1–28

Holt D, Steel DG, Tranmer M, Wrigley N (1996) Aggregation and ecological effects in geographically based data. *Geographical Analysis* 28: 244–261

Judge G, Miller DJ, Cho WKT (2004) An information theoretic approach to ecological estimation and inference. In: King G, Rosen O, Tanner MA (eds) *Ecological inference. New methodological strategies*. Cambridge University Press

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

Mei CL, He SY, Fang KT (2004) A note on the mixed geographically weighted regression model. *Journal of Regional Science* 44: 143–157

Morganstern H (1982) Uses of ecologic analysis in epidemiologic research. *American Journal of Public Health* 72: 1336–1344

OECD (2000) Small and medium-sized enterprises: local strength, global reach. *OECD Observer*, June 2000, Organisation for Economic Co-operation and Development, Paris

Openshaw S (1984) *The modifiable areal unit problem*. Geobooks, Norwick, UK
Openshaw S, Taylor P (1979) A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In: Wrigley N (ed) *Statistical methods in the spatial sciences*. Pion, London
Parr JB (2002) Missing elements in the analysis of agglomeration economies. *International Regional Science Review* 25: 151–168
Rao JNK (2003) *Small area estimation*. Wiley, New York
Robinson WS (1950) Ecological correlations and the behavior of individuals. *American Sociological Review* 15: 351–357
Schuessler AA (1999) Ecological inference. *Proceedings of the National Academy of Science USA* 96: 10578–10581
Silverman BW (1986) *Density estimation for statistics and data analysis*. Chapman & Hall, New York
Tsionas EG (2000) Regional growth and convergence: evidence from the United States. *Regional Studies* 34: 231–238