ECOLOGICAL INFERENCE AND SPATIAL HETEROGENEITY:
A NEW APPROACH BASED ON ENTROPY ECONOMETRICS

Ludo Peeters
Faculty of Applied Economics,
Limburg University Centre
3590 Diepenbeek-Hasselt (Belgium)
Tel. ++32 11 268635
Fax ++32 11 268700
Email: ludo.peeters@luc.ac.be

Coro Chasco-Yrigoyen
Department of Applied Economics
Autonoma University of Madrid
Campus de Cantoblanco
28049 Madrid (Spain)
Tel. ++34 91 4974266
Fax ++34 91 4973943
Email: coro.chasco@uam.es

Paper to be presented at the 45th European Congress of the Regional Science Association, Amsterdam (NL), Aug. 23-27, 2005.

Abstract

In this note we compare two approaches to ecological modeling using test data. The first one is the "traditional" approach based on Ordinary Least Squares (OLS), assuming constancy of parameters across disaggregated spatial units (spatial homogeneity) – an assumption that is rarely tenable. The second one is a new approach based on Generalized Cross-Entropy (GCE), assuming varying parameters (spatial heterogeneity). These two approaches are tested in two real-world applications or cases. The first case is based on aggregate data on per-capita GDP for the 17 regions in Spain. The second case is based on aggregate data on per-capita taxable income for the five provinces in the region of Flanders in Belgium. The performance of each approach is assessed by examining its capability in tracking the real – but "unobserved" (by the analyst) – data for the 50 provinces in Spain and the 22 districts in Flanders, respectively. The results clearly indicate that the GCE varying-parameter approach outperforms the OLS approach in both cases. In addition, the ecologically inferred values from GCE are even closer to the known truth than the fitted values from applying OLS directly to the disaggregated data.

Key words: Ecological inference, Spatial prediction, Cross-Entropy, Disposable income, Spatial heterogeneity

1. Introduction

Situations where the only available data are aggregated at a level other than the level of interest are common. Despite such inauspicious conditions, some "real-world" applications require the use of ecological estimation and inference models.

However, most efforts to recover disaggregate information from aggregate data generally result in "ill-posed" inverse problems, which yield a multitude of feasible solutions, due to the lack of sufficient information (JUDGE *et al.*, 2003). Specifically, ill-posed problems are fundamentally indeterminate, because there are more unknowns than data points.

The purpose of the present note is to compare the performances of two alternative approaches to ecological inference. 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 (e.g., CHO, 2001). The second one is a new approach based on Generalized Cross-Entropy (GCE), assuming varying – i.e., individual or subgroup-specific – parameters (spatial heterogeneity). In other words, the GCE approach does not take the usual "constancy assumption". The two approaches will be compared in two real-world applications or cases, using a testing procedure.

2. The ecological inference problem

Ecological inference is the process of drawing conclusions about individual- or subgroup-level behavior from aggregate- or group-level (historically labeled "ecological") data, when no individual- or subgroup data are available.

A fundamental difficulty with such inferences 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; SCHUESSLER, 1999). In the absence of individual- or subgroup-level measurement (for example, in the form of survey data), such information needs to be inferred. Another difficulty is that inferences may be subject to the so-called "ecological fallacy".

3. Two alternative approaches to ecological modeling

3.1 OLS assuming homogeneity across space

First, we run a simple OLS regression of y_i on $\mathbf{x}_i$ at the group (regional) level, based on available aggregate data:

$$y_i = \alpha + \boldsymbol{\gamma}'\mathbf{x}_i + u_i \tag{1}$$

where y_i is the observed, aggregate, per-capita income indicator for group (region) *i*, and $\mathbf{x}_i$ is a vector of explanatory variables for group (region) *i.*

Then, we predict the per-capita incomes at the subgroup (sub-regional) level, taking some available covariates $\mathbf{z}_{ij}$ at the level of the subgroups (sub-regions):

$$\hat{y}_{ij} = \hat{\alpha} + \hat{\boldsymbol{\gamma}}'\mathbf{z}_{ij} \tag{2}$$

where y_{ij} is the "unobserved", disaggregate, per-capita income indicator for subgroup (sub-region) *j* in group (region) *i*, and $\mathbf{z}_{ij}$ is the vector of explanatory variables for subgroup (sub-region) *j* in group (region) *i*.

A major problem with this approach is the possible *aggregation bias,* due to the (implicit) assumption of constancy or homogeneity of parameters across the spatial units (e.g., CHO, 2001).

2.1 GME assuming heterogeneity across space

2.1.1 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 to test possible determinants of the variation in the underlying subgroup indicators.

To be more precise, BIDANI and RAVALLION (1997) 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:

$$y_i = \sum_{j=1}^{M_i} y_{ij}\,\eta_{ij} \tag{3}$$

where y_i is the aggregate indicator for region *i*, y_{ij} is the indicator of the *j*th sub-region in region *i*, η_{ij} is the population share of sub-region *j* in *i*, with $\sum_{j=1}^{M}\eta_{ij} = 1$, and where *i* = 1,…,*N* denotes the regions and *j* = 1,…,M_i denotes the number of sub-regions in region *i*.

The sub-regional indicators are not observed, but the y_i's and η_{ij}'s are. If we also observe a vector of explanatory variables for region *i*, $\mathbf{x}_i$, and a vector of explanatory variables for sub-region *j* in region *i*, $\mathbf{z}_{ij}$, we get

$$y_{ij} = \alpha_{ij} + \boldsymbol{\beta}'_{ij}\mathbf{x}_i + \boldsymbol{\gamma}'_{ij}\mathbf{z}_{ij} + \varepsilon_{ij} \tag{4}$$

which, on substituting into (3), yields the following regression:

$$y_i = \sum_{j=1}^{M_i} (\alpha_{ij} + \boldsymbol{\beta}'_{ij}\mathbf{x}_i + \boldsymbol{\gamma}'_{ij}\mathbf{z}_{ij})\eta_{ij} + u_i \tag{5}$$

where $u_i = \sum_{j=1}^{M_i} \varepsilon_{ij}\eta_{ij}$ is a "composite" error term, which is heteroskedastic.

Using the regression in (5), we can obtain estimates of the (unobserved or latent) sub-regional indicators as

$$\hat{y}_{ij} = \hat{\alpha}_{ij} + \hat{\boldsymbol{\beta}}'_{ij}\mathbf{x}_i + \hat{\boldsymbol{\gamma}}'_{ij}\mathbf{z}_{ij} \tag{6}$$

2.1.2 Generalized Cross-Entropy estimation

Basically, the regression in (5) amounts to a standard random-coefficients model (e.g., HILDRETH and HOUCK, 1968; SWAMY and TAVLAS, 1995), which can be estimated by using Generalized Least Squares (GLS).

However, instead of using GLS we prefer to use the Generalized Cross-Entropy (GCE) method (e.g., GOLAN *et al.*, 1996). GCE has some important advantages over the "classical" techniques: unique estimates; reformulation of the fundamentally "ill-posed" problem into a "well-posed" problem; etc.[1]

The implementation of GCE requires that the parameters of the model 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 minimization problem, where the objective function, specified in the equation (7) below, consists of the joint cross-entropy. (Note that we do not consider any explanatory variables at the aggregate/regional level, $\mathbf{x}_i$.)

[1] Note that GCE does not require the assumption of random drawings from a particular distribution (as, for example, in FREEDMAN et al., 1998). Also, GCE is different from the Bayesian approach (e.g., ROSEN et al., 2001) or the switching-regression approach (e.g., CHO, 2001).

Specifically, we define sets of unknown probability vectors $\mathbf{p}'_{\alpha,ij} = [\, p_{\alpha,ij}(1),..., p_{\alpha,ij}(K)\,]$, $\mathbf{p}'_{\gamma,ij} = [\, p_{\gamma,ij}(1),..., p_{\gamma,ij}(K)\,]$ ($K \geq 2$), and $\boldsymbol{\mu}'_{ij} = [\, \mu_{ij}(1),..., \mu_{ij}(G)\,]$ ($G \geq 2$), and choose the corresponding support vectors $\mathbf{s}'_{\alpha} = [\, s_{\alpha}(1),..., s_{\alpha}(K)\,]$, $\mathbf{s}'_{\gamma} = [\, s_{\gamma}(1),..., s_{\gamma}(K)\,]$, and $\mathbf{e}' = [\, e(1),..., e(G)\,]$, for the parameters α_{ij}, γ_{ij}, and the residual terms u_{ij}, respectively, where $\alpha_{ij} = \mathbf{s}'\mathbf{p}_{\alpha,ij}$, $\gamma_{ij} = \mathbf{s}'\mathbf{p}_{\gamma,ij}$, and $\varepsilon_{ij} = \mathbf{e}'\boldsymbol{\mu}_{ij}$. In addition, prior information is included through specifying the prior probability vectors $\mathbf{p}^{\mathrm{o}}_{\alpha,ij}$, $\mathbf{p}^{\mathrm{o}}_{\gamma,ij}$ and $\boldsymbol{\mu}^{\mathrm{o}}_{ij}$, reflecting subjective information, informed “guesses”, or any other sample and pre-sample information. In the empirical applications below, we use as prior information the OLS estimates at the aggregate level, e.g., $\hat{\gamma}_i^{OLS} = \mathbf{s}'\mathbf{p}^{0}_{\gamma,ij}$, etc.

After the appropriate re-parameterization, the complete GCE optimization problem for the ecological model, described by the expressions in (3) through (6), can be formulated as

$$\underset{\mathbf{p},\boldsymbol{\mu}}{\mathrm{Min}}\, CE = \sum_{i=1}^{N}\sum_{j=1}^{M_i} \mathbf{p}'_{\alpha,ij} \ln\left(\frac{\mathbf{p}_{\alpha,ij}}{\mathbf{p}^{\mathrm{o}}_{\alpha,ij}}\right) + \sum_{i=1}^{N}\sum_{j=1}^{M_i} \mathbf{p}'_{\gamma,ij} \ln\left(\frac{\mathbf{p}_{\gamma,ij}}{\mathbf{p}^{\mathrm{o}}_{\gamma,ij}}\right) + \sum_{i=1}^{N}\sum_{j=1}^{M_i} \boldsymbol{\mu}'_{ij} \ln\left(\frac{\boldsymbol{\mu}_{ij}}{\boldsymbol{\mu}^{\mathrm{o}}_{ij}}\right) \qquad (7)$$

subject to

$$y_i = \sum_{j=1}^{M_i} (\mathbf{s}'_{\boldsymbol{\alpha}}\mathbf{p}_{\alpha,ij} + \mathbf{s}'_{\gamma}\mathbf{p}_{\gamma,ij}\mathbf{z}_{ij} + \mathbf{e}'\boldsymbol{\mu}_{ij})\eta_{ij} \quad \forall i \qquad (8)$$

$$\begin{gathered}\sum_{k=1}^{K} p_{\alpha,ij}(k) = 1; \ \sum_{k=1}^{K} p_{\gamma,ij}(k) = 1 \quad \forall i, j \\ \sum_{g=1}^{G} \mu_{ij}(g) = 1 \quad \forall i, j\end{gathered} \qquad (9)$$

Equation (7) denotes the cross-entropy objective, which is subject to the data-consistency constraints in (8). The constraints in (9) ensure that all unknown probabilities or weights add up to one.

4. Two real-world applications

To illustrate the GCE approach and to compare its results with its OLS counterpart, we consider two “real-world” applications, using test data for 2000. The first application employs aggregate data for the 17 regions (“autonomous communities”) and disaggregate data for the 50 provinces in Spain. The second application uses aggregate data for the 5 provinces and disaggregate data for the 22 districts in the region of Flanders, Belgium.

We aggregate the data at the group level (i.e., for the 17 Spanish regions or the five Flemish provinces), deliberately “losing” (for the moment) information at the subgroup level (i.e., for the 50 Spanish provinces or the 22 Flemish districts), and then use the two proposed methods (i.e., OLS versus GCE) to make ecological inferences. Subsequently, the inferences being made, are compared with the “truth” – i.e., the real data at the subgroup level.[2]

4.1 Case 1: Spain

In the case of Spain, the dependent variable is Gross Domestic Product (*GDP*) per capita. We consider the following explanatory variables: (1) the level of the so-called Economic Activity Tax (*TAX*); (2) the number of high-speed telephone lines for the Internet (voice & data);[3] and (3) the population density (*POPDENS*). The tax ratio per capita in each region/sub-region represents the total taxes paid by companies, self-employed and artists for the economic activity developed in the corresponding region/sub-region. GDP data provided by the Spanish Institute for Statistics (INE) in the Regional Accounts.

[2] The role of spatial effects in ecological inference (e.g., ANSELIN and WHO, 2002) is not considered relevant for our cases.

[3] This kind of telephone lines have been installed – at a higher percentage – in those places with high-tech firms (and sometimes in residential places, not only for domestic use, but also for the self-employed).

4.2 Case 2: Flanders (Belgium)

In the case of Flanders, Belgium, the dependent variable is per-capita taxable income (*INC*). We consider the following explanatory variables: (1) the level of educational attainment, measured as the percentage of the labor force having attained higher (tertiary) education in 1991 (*EDUC*); and (2) the population density (*POPDENS*).

5. Empirical results

5.1 Parameter estimates from OLS and GCE

For each method (OLS and GCE), we have derived two inferences: (1) the "ecological" inference, and (2) the "correct" inference.

The ecological inference consists of estimating the model at the regional (in the case of Spain) or the provincial (in the case of Flanders) level, and then apply the estimated coefficients to the provincial (in the case of Spain) or district (in the case of Flanders) covariates, to obtain the corresponding information at the disaggregate level. Table 1 shows of the model parameters for the 17 regions in Spain and the 5 provinces in Flanders.

The "correct" inference, on the other hand, consists of estimating the model at the provincial level (in the case of Spain) or the district level (in the case of Flanders), using the real, disaggregated, data. The OLS estimates of the parameters of the "correct" model are presented in Table 2.

Finally, Table 3 shows the results from the GCE ecological inference model; that is, the mean value of the 50 estimated varying parameters for Spain and the 22 estimated varying parameters for Flanders, along with the standard deviation of these individual parameter estimates.[4] The mean values are close to the OLS results. Also, there seems to be little variation in the sub-regional parameter estimates, except for γ_3, corresponding to the *POPDENS* variable.

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

5.2 Testing the performance of the models

In order to test the performance of OLS and GCE in correctly "predicting" (tracking) the sub-regional, disaggregated, data we compare the ecologically inferred estimates with the actual data for the corresponding sub-regions (not used in the estimation process).

We use two measures of accuracy: the Pseudo-R^2, and the Mean Absolute Percentage Error (*MAPE*). The Pseudo-R^2 is defined as the square of the simple correlation between y_{ij} and $\hat{y}_{ij}$. The *MAPE* is a relative error measure that is defined as:

$$MAPE = \frac{1}{\Sigma_i M_i} \sum_{i=1}^{N} \sum_{j=1}^{M_i} \left| y_{ij} - \hat{y}_{ij} \right| / y_{ij} \times 100 \tag{10}$$

In addition, we test for possible bias in the predictions, by looking at the (significance of the) mean prediction errors.

The test results are presented in Table 4. In other words, Table 4 shows for each data set which method comes closest to the truth. To provide a visual picture, the actual and fitted values are depicted in the Figures 1 and 2, for OLS en GCE, respectively. In terms of both the Pseudo-R^2 and the *MAPE*, the GCE model is outperforming the OLS model. A striking result is that GCE is also superior to the "correct" OLS model! The GCE model slightly underestimates the actual data, but the bias is not significant (at the 5% level).

6. Conclusions

We have tested two different approaches to ecological inference, where GDP/taxable income per capita for the 50 provinces in Spain and 22 districts, respectively, are predicted from aggregate data on the GDP per capita for the 17 Spanish regions ("autonomous communities") and the taxable income per capita for the five provinces in the region of Flanders, Belgium.

The two models are estimated by using OLS and GCE. Obviously, the results from the GCE-based model are "superior" (albeit, admittedly, only slightly) to those from the traditional OLS-based models, in terms of prediction accuracy. It is to be expected that the inclusion of additional explanatory variables, both at the group and subgroup level, would produce even more satisfying results.

References

ANSELIN, L. and CHO, W.K.T. (2002). "Spatial effect and ecological inference." *Political Analysis*, 10, 276-297.

BIDANI, B. and M. RAVALLION (1997). "Decomposing social indicators using distributional data" *Journal of Econometrics*, 77, 125-139.

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

FREEDMAN, D.A., KLEIN, S.P., OSTLAND, M., and ROBERTS, M. (1998). *On "solutions" to the ecological inference problem*. Mimeo.

GOLAN, A., G. JUDGE and D. MILLER (1996). *Maximum Entropy Econometrics: Robust Estimation with Limited Data*, New York: John Wiley & Sons.

HILDRETH, C. and HOUCK, J.P. (1968). "Some estimators for a linear model with random coefficients." *Journal of the American Statistical Association*, 63, 583-594.

JUDGE, G., D.J. MILLER and W.K. CHO (2003). *An Information Theoretic Approach to Ecological Estimation and Inference*, CUDARE Working Papers #946, University of California, Berkeley.

KING, G. (1997). *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior From Aggregate Data*. Princeton, NJ: Princeton University Press.

ROSEN, O., JIANG, W., KING, G., and TANNER, M.A. (2001). "Bayesian and frequentist inference for ecological inference: the $R\times C$ case." *Statistica Neerlandica*, 5, 134-156.

SCHUESSLER, A.A. (1999). "Ecological inference." *Proceedings of the National Academy of Science USA*, 96, 10578-10581.

SWAMY, P.A.V.B and TAVLAS, G. (1995). "Random coefficients models: theory and applications." *Journal of Economic Surveys*, 9, 165-197.

Table 1: Parameter estimates from OLS – aggregate (group) level

Spain – Regions		**Flanders (Belgium) – Provinces**	
Variables	Parameters estimates (Standard errors)	Variables	Parameters estimates (Standard errors)
Constant	2.460	*Constant*	3.418
	(0.940)		(0.894)
TAX	2.315	*EDUC*	0.273
	(0.344)		(0.040)
RSDI	0.125	*POPDENS*	0.004
	(0.070)		(0.001)
POPDENS	0.004		
	(0.002)		
Nobs	17	Nobs	5
SER	0.795	SER	0.196
R^2	0.93	R^2	0.98

Table 2: Parameter estimates from OLS – disaggregate (sub-group) level (“correct” model)

Spain – Provinces		**Flanders (Belgium) – Districts**	
Variables	Parameters estimates (Standard errors)	Variables	Parameters estimates (Standard errors)
Constant	2.962 (0.703)	*Constant*	6.173 (0.806)
TAX	2.316 (0.281)	*EDUC*	0.194 (0.040)
RSDI	0.136 (0.062)	*POPDENS*	0.002 (0.001)
POPDENS	-0.0001 (0.0015)		
Nobs	50	Nobs	22
SER	1.132	SER	0.552
R^2	0.83	R^2	0.74

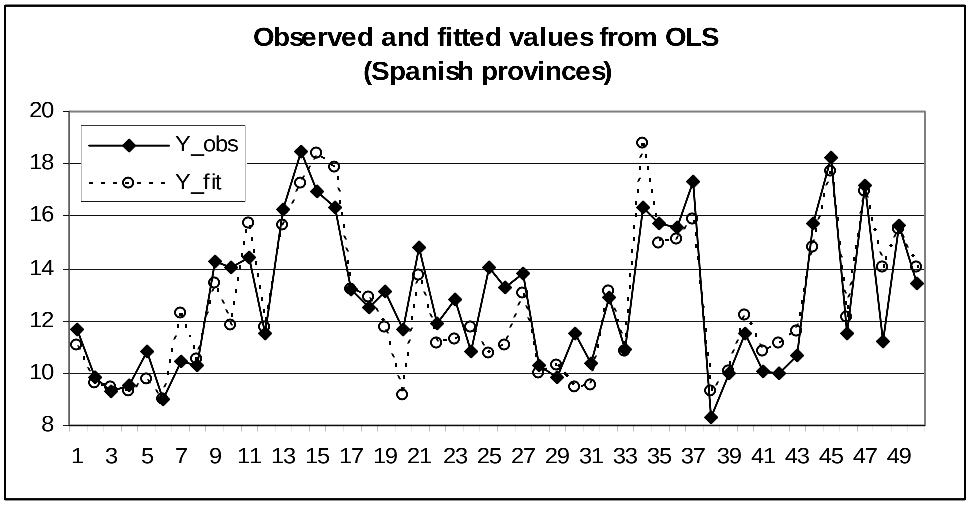
Table 3: Parameter estimates from GCE – disaggregate (sub-group) level

Spain – Provinces		**Flanders (Belgium) – Districts**	
Variables	Average estimates (Standard deviations)	Variables	Average estimates (Standard deviations)
Constant	2.461	*Constant*	3.418
	(0.000)		(0.000)
TAX	2.315	*EDUC*	0.273
	(0.001)		(0.000)
RSDI	0.126	*POPDENS*	0.004
	(0.003)		(0.0005)
POPDENS	0.006		
	(0.011)		
Nobs	17	Nobs	5

Table 4: A comparison of the two methods (OLS versus GCE) for making ecological inference, in a situation where the truth is known

	Spain – Provinces			**Flanders (Belgium) – Districts**		
	OLS	GCE	"Correct" OLS	OLS	GCE	"Correct" OLS
Pseudo-R^2	0.80	0.86	0.83	0.73	0.82	0.74
Mean error	0.169	0.234	0.000	0.052	0.225	0.000
MAPE	7.5%	5.9%	6.9%	5.9%	4.5%	5.9%
Stdev of mean error	0.175	0.146	0.155	0.171	0.128	0.112
t-value	0.965	1.608	0.000	0.305	1.759	0.000
Critical *t*-value (5%)	2.009	2.009	2.009	2.074	2.074	2.074

Figure 1
Observed and ecologically inferred values from OLS and GCE – Spain



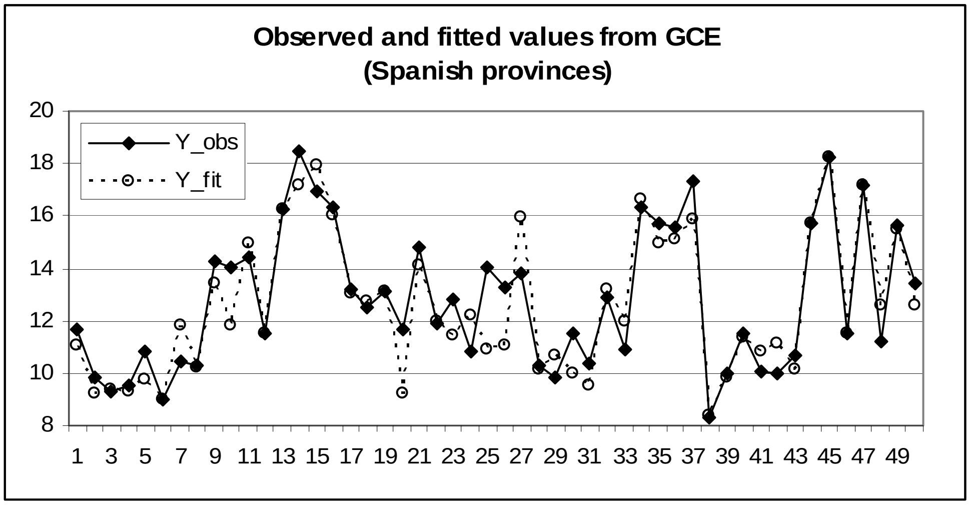
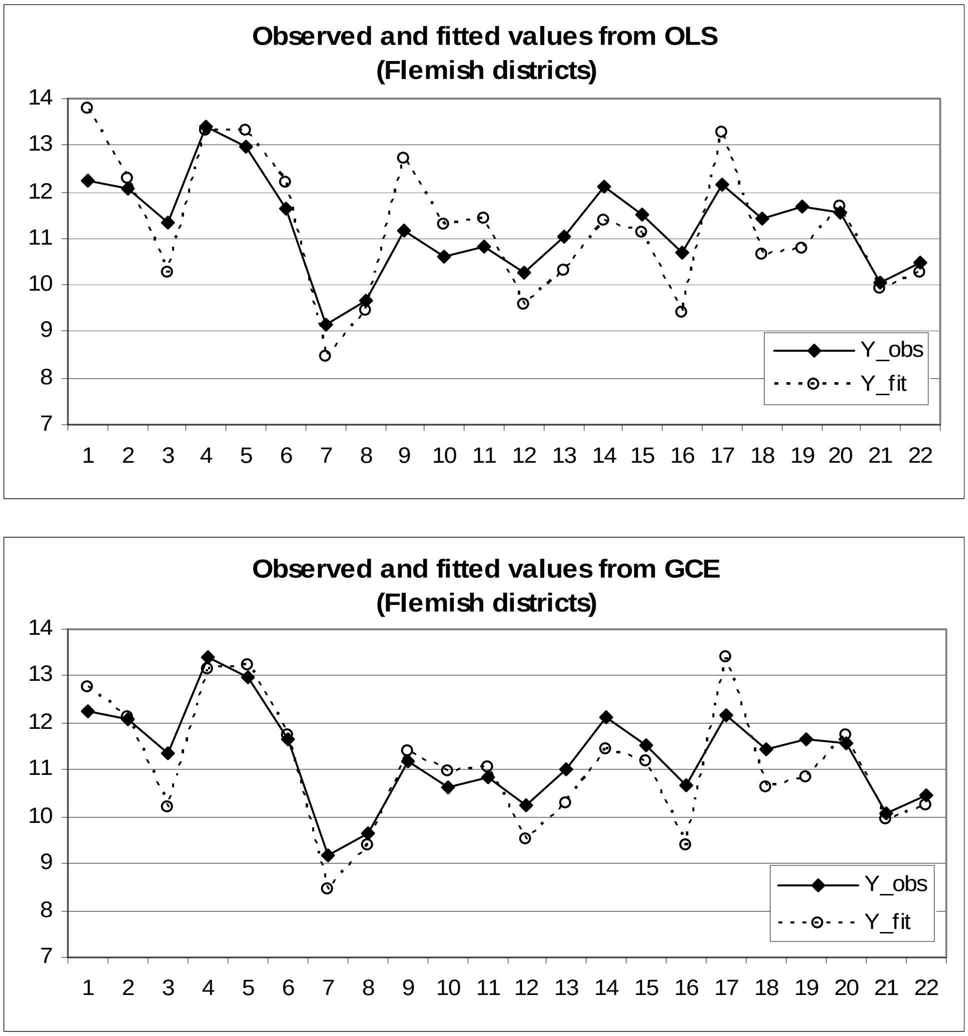


Figure 2
Observed and ecologically inferred values from OLS and GCE – Flanders


Observed and fitted values from OLS
(Flemish districts)
Y_obs
Y_fit
Observed and fitted values from GCE
(Flemish districts)
Y_obs
Y_fit