ATTI DELLA XLV RIUNIONE SCIENTIFICA (2010)

XLV Riunione Scientifica
(Padova, 16-18 giugno 2010)

ISBN 978-88-6129-566-7
Editore: CLEUP – Padova

Plenary Session

James J. Heckman, *University of Chicago, USA*
Structural vis-à-vis reduced form estimation of program effects

George Casella, *University of Florida, USA*
Estimation in Dirichlet process random effect models

Paul Boyle, *University of St Andrews, UK*
Space, migration and health: learning from routine data

Enrico Giovannini, *ISTAT, Rome*
Statistics, knowledge and policy: towards a new paradigm

I	Enrico Rettore, *University of Padua*	James J. Heckman, *University of Chicago, USA*	Structural vis-à-vis reduced form estimation of program effects
II	Gabriella Salinetti, *University of Rome "La Sapienza"*	George Casella, *University of Florida, USA*	Estimation in Dirichlet process random effect models
III	Silvana Salvini, *University of Florence*	Paul Boyle, *University of St Andrews, UK*	Space, migration and health: learning from routine data
IV	Maurizio Vichi, *University of Rome "La Sapienza"*	Enrico Giovannini, *ISTAT, Rome*	Statistics, knowledge and policy: towards a new paradigm

Contributed Sessions

Small area estimation
Morales D., Pagliarella M.C., Salvatore R.
Influence analysis in small area estimation

Fabrizi Enrico, Pratesi Monica, Salvati Nicola, Tzavidis Nikos
Design consistent small area estimators based on M-quantile regression

Roberto Benedetti, Danila Filipponi, Federica Piersimoni
Multiple imputation in land cover EBLUP estimates

Giorgio E. Montanari, M. Giovanna Ranalli and Cecilia Vicarelli
A comparison of small area estimators of counts aligned with direct higher level estimates

Clustering methods: theory and applications

Measuring dependence

Anna Gottard
Stumbling across symmetries

M. Sofia Massa
Numerical studies on covariance selection problems

Silvia Angela Osmetti
Parametric, non-parametric estimate of bivariate survival functions: the copula approach

Paolo Giudici, Emanuela Raffinetti
Lorenz zonoids, dependence measures: a proposal

Statistical models for the analyis of demographic phenomena

Federico Benassi, Raffaele Ferrara
Regionalization with dynamically constrained agglomerative clustering, partitioning. An application on spatial segregation of foreign population in Italy at regional level

Elisabetta Santarelli
Economic resources, marital childbearing in Italy between 1994, 2004

Agnese Vitali, Bruno Arpino
Home leaving among second generation migrants in Spain: A Cross-classified Multilevel Analysis

Claudio Ceccarelli, Andrea Cutillo
The wage returns from internal relocation

Specialized Session

A non-random sample of Statistics in the U.K.

Alan Kimber, Ali Yousef
The sensitivity of the Triple Sampling Sequential Procedure to Departures from Normality

(Michael J Campbell, Suzanne Mason, Jon Nicholl)
Relative Risks, Odds Ratios, Cluster Randomised Trials

Robin Mitra, Jerome P. Reiter
An approach to impute missing covariates in observational studies when estimating treatment effects with propensity score matching

Statistical methods for evaluation of educational services

Enrico Gori
Accountability Systems in Education

Roberto Ricci
The Economic, Social, Cultural Background: a continuous index for the Italian Students of the fifth grade

Tommaso Agasisti
The determinants of students' achievement: a multilevel analysis of new Italian data

Statistical methods for business, industry

Narayanaswamy Balakrishnan
Generalized Gamma Frailty Model, Applications

Chiara Bocci, Alessandra Petrucci
Statistical Methods for the Analysis of Spatial Patterns: a Geoadditive Approach

Renato Salvatore
Technical efficiency estimators in linear mixed effects stochastic production frontier models

Nonparametric methods

Statistical Methods for the Analysis of Spatial Patterns: a Geoadditive Approach

Chiara Bocci and Alessandra Petrucci

Abstract The explosive growth of spatial data and widespread use of spatial databases emphasize the need for the discovery of spatial knowledge.
Nowadays, very rich databases of spatially referenced socio-economic data are available from local statistical offices and in the last few years the demand of spatially detailed statistical data is dramatically increased. Extracting interesting and useful patterns from spatial data sets is more difficult than extracting corresponding patterns from traditional numeric and categorical data due to the complexity of spatial data types, spatial relationships, and spatial autocorrelation.
The complexity of spatial data and intrinsic spatial relationships limits the usefulness of conventional techniques (i.e. data mining) for extracting spatial patterns.
Moreover, the area definition and the assignment of the data to appropriate areas can pose problems in the estimation process. This paper presents statistical methods which face these problems and analyze the geographical pattern of the spatially referenced socio-economic data by incorporating the spatial location as an additional covariate.

Key words: spatial statistics, semiparametric methods, socio-economic data

1 Introduction

The analysis of the regional spatial pattern of socio-economical processes has become a relevant area of economics. Since the early seventies, regional economics has been defined as the field concerned with the role of space, distance and regional differentiation in economics [15].

Chiara Bocci, Alessandra Petrucci
Department of Statistics “G. Parenti”, University of Florence, viale Morgagni 59 - 50134 Firenze
e-mail: bocci@ds.unifi.it, alessandra.petrucci@unifi.it

Several reasons support the importance of this subject: first, the spatial clustering of economic activities is a product of the regional differences and could reflect individual inequalities that are object of policies; second, the geographical pattern can have great influence on the results of economic policies; and third, exploring spatial clustering of economic activities is a relevant input to model economic theories at a regional scale.

Research in this area focuses on the specification and estimation of spatial effects in a theoretical economic model, and on the use of such estimates to obtain spatial interpolations and predictions of the study variables. The set of methodologies concerned with this target belongs to the field of spatial econometrics [1, 2], that is defined by [1] as the collection of techniques that deal with the peculiarities caused by space in the statistical analysis of regional science models.

The explosive growth of spatial data and widespread use of spatial databases have emphasized the need for the discovery - also automated - of spatial knowledge. Moreover, very rich databases of spatially referenced socio-economic data are available from local statistical offices and in the last few years the demand of spatially detailed statistical data is dramatically increased.

Nowadays, the fields of spatial statistics is broadly understood. In general, spatial statistics is concerned with statistical and mathematical descriptors of spatial structure and it focuses on the nature of space and spatial data. In this way it can face with problems which are characterized by the difficulties associated with assessing the importance of spatial dependence and spatial heterogeneity, the so-called "spatial effects" mentioned before.

Extracting interesting and useful patterns from spatial data sets is more difficult than extracting corresponding patterns from traditional numeric and categorical data due to the complexity of spatial data types, spatial relationships, and spatial autocorrelation. The complexity of spatial data and intrinsic spatial relationships limits the usefulness of conventional techniques (i.e. data mining) for extracting spatial patterns. Therefore, the area definition and the assignment of the data to appropriate areas can pose problems in the estimation process.

It is worth to stress the usefulness of the geographical location for the analysis of non stationary spatial phenomena. The "global" dependence models, such as the classical regression model, assume the independence of the data from the spatial location, generate spatially autocorrelated residuals and bring often to wrong conclusions. Thus, statistical models which take into account the spatial variability can help to understand the underlying phenomenon.

This study discusses a new approach to identify the spatial pattern using recent advances in semiparametric models that allow incorporation of spatial location as an additional component. Thus, the estimated spatial patterns reflect the propensity of the considered characteristic in a region, after controlling for other unit-level effects. In particular, the work focuses on socio-economic data collected by the World Bank program on Living Standard Measurement Study (LSMS) [6]. The program is designed to assist policy makers in their efforts to identify how policies could be designed and improved to positively affect outcomes in health, education, economic activitgies, housing and utilities, etc.

As a matter of fact the goal of many empirical studies in urban economics, regional science, and geography is to measure the effects of proximity.

A primary issue, also considering just a single-explanatory variable model, is the functional form: frequently the exact shape of the function is ultimately an empirical issue. Some studies attempt to isolate the effects of proximity to a site while controlling for the effects of other variables. While the results of such studies can be highly sensitive to functional form assumptions, multiple explanatory variable models are made more complicated by the necessity of controlling for the effects of other variables that may be highly correlated with the one of central interest.

In addition to functional form issues, another critical issue in spatial data analysis is that important variables are highly correlated and no study includes all relevant variables. Some variable is always missing no matter how long the variable list gets. The problem of spatially correlated missing variables is endemic and is not confined to direct measures of proximity. Thus, the model has a spatial dimension even though the analysis is not explicitly spatial at first glance.

Both the issues, functional form and spatially correlated missing variables, have largely been treated separately in the empirical literature. Functional form choice is typically addressed directly using series expansions or nonparametric estimation procedures. Spatially correlated missing variables are often considered only indirectly through tests for spatial autocorrelation.

The standard models used to address spatial autocorrelation are based on ad hoc specifications of a spatial weight matrix. A common approach is to begin with a simple functional form, test for spatial autocorrelation, and then to estimate a model that includes a spatially lagged dependent variable or that accounts directly for spatial autocorrelation in the error terms. After perhaps including some experimentation with different specifications of the spatial weight matrix, further specification testing typically stops.

The main problem with this approach is that it is likely to fail in identifying the root cause of the spatial autocorrelation. Functional form misspecification can itself cause residuals to be spatially correlated. More importantly, if the underlying problem is that a spatially correlated variable has been omitted from the regression, a misspecified spatial econometric approach may be accepted in place of the true model [11]. Unfortunately, the real solution of adding the omitted variable to the analysis may not be feasible if the data are not available. In this case, the role of further statistical testing is to assess the robustness of the results to alternative model specifications. We can have more confidence in the results if a variety of model specifications lead to similar results.

Nonparametric and semiparametric models are attractive alternatives to parametric alternatives because they admit at the start that the true model structure is unknown. However, nonparametric estimation suffers from the rapidly increase of the variance of the estimates with the number of variables. In this situation, semiparametric models become an effective alternative to full nonparametric estimation. The advantage of the semiparametric approach is that it imposes parametric structure where the structure may be reasonable, while leaving the structure of the modelunrestricted for another set of variables. Thus, the semiparametric approach is a particularly easy and flexible approach for modeling broad spatial trends while also permitting the effects of other explanatory variables to vary by location.

The paper is structured as follows. In the next section, the methodology is extensively discussed. The third section describes the datasets used in the analysis and the presents the empirical results. The final section summarizes the main findings and discusses the possible limitations of the analysis.

2 Methodology

Geostatistical methodologies are concerned with the problem of producing a map of a quantity of interest over a particular geographical region based on measurement taken at a set of locations in the region. The aim of such a map is to describe and analyze the geographical pattern of the phenomenon of interest.

These methodologies are born and apply in areas such as environmental studies and epidemiology, where the spatial information is traditionally recorded and available. However, in the last years the diffusion of spatially detailed statistical data is considerably increased and these kind of procedures - possibly with appropriate modifications - can be used as well in any statistical fields of application.

Basically, to obtain a surface estimate we can exploit the exact knowledge of the spatial coordinates (latitude and longitude) of the studied phenomenon by using bivariate smoothing techniques, such as kernel estimate or kriging [5, 16]. However, usually the spatial information alone does not properly explain the pattern of the response variable and we need to introduce some covariates in a more complex model.

Geoadditive models, introduced by [9], answer this problem as they analyze the spatial distribution of the study variable while accounting for possible linear or non-linear covariate effects. Under the additivity assumption they can handle such covariate effects by merging an additive model [7] - that accounts for the relationship between the variables - and a kriging model - that accounts for the spatial correlation - and by expressing both as a linear mixed model. The linear mixed model representation is a useful instrument because it allows estimation using mixed model methodology and software. Moreover, we can extend geoadditive model to include generalized responses, small area estimation, longitudinal data, missing data and so on [17].

Let s_i and t_i, $1 \leq i \leq n$, be continuous predictors of y_i at spatial location $\mathbf{x}_i$, $\mathbf{x} \in \Re^2$. A geoadditive model for such data can be formulated as

$$y_i = f(s_i) + g(t_i) + h(\mathbf{x}_i) + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma_\varepsilon^2), \tag{1}$$

where f and g are unspecified smooth functions of one variable and h is an unspecified bivariate smooth functions.

Considering two low-rank truncated linear splines for f and g and a low-rank thin plate spline for h, the model (1) can be written as a mixed model

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \varepsilon, \qquad \mathrm{Cov}\begin{bmatrix}\mathbf{u}\\ \varepsilon\end{bmatrix} = \begin{bmatrix}\sigma_s^2\mathbf{I}_{K_s} & 0 & 0 & 0\\ 0 & \sigma_t^2\mathbf{I}_{K_t} & 0 & 0\\ 0 & 0 & \sigma_x^2\mathbf{I}_{K_x} & 0\\ 0 & 0 & 0 & \sigma_\varepsilon^2\mathbf{I}_n\end{bmatrix} \tag{2}$$

where

$$\beta = \left[\beta_0, \beta_s, \beta_t, \beta_x^T\right], \qquad \mathbf{u} = \left[u_1^s, \ldots, u_{K_s}^s, u_1^t, \ldots, u_{K_t}^t, u_1^x, \ldots, u_{K_x}^x\right],$$

$$\mathbf{X} = \left[1, s_i, t_i, \mathbf{x}_i^T\right]_{1\le i\le n}$$

and $\mathbf{Z}$ is obtained by concatenating the matrices containing spline basis functions to handle f, g, and h, respectively

$$\mathbf{Z} = \left[\mathbf{Z}_s | \mathbf{Z}_t | \mathbf{Z}_x\right],$$

$$\begin{aligned}
\mathbf{Z}_s &= \left[(\mathbf{s_i} - \kappa_1^s)_+, \ldots, (\mathbf{s_i} - \kappa_{K_s}^s)_+\right]_{1\le i\le n},\\
\mathbf{Z}_t &= \left[(\mathbf{t_i} - \kappa_1^t)_+, \ldots, (\mathbf{t_i} - \kappa_{K_t}^t)_+\right]_{1\le i\le n},\\
\mathbf{Z}_x &= \left[C\left(\mathbf{x}_i - \kappa_k^x\right)\right]_{1\le i\le n, 1\le k\le K_x} \cdot \left[C\left(\kappa_h^x - \kappa_k^x\right)\right]_{1\le h,k\le K_x}^{-1/2},
\end{aligned}$$

where $C(\mathbf{r}) = \|\mathbf{r}\|^2 \log\|\mathbf{r}\|$ and $\kappa_1^s, \ldots, \kappa_{K_s}^s$, $\kappa_1^t, \ldots, \kappa_{K_t}^t$ and $\kappa_1^k, \ldots, \kappa_{K_x}^k$ are the knot locations for the three functions.

Let

$$\mathrm{Var}(\mathbf{u}) \equiv \Sigma_u = \begin{bmatrix}\sigma_s^2\mathbf{I}_{K_s} & 0 & 0\\ 0 & \sigma_t^2\mathbf{I}_{K_t} & 0\\ 0 & 0 & \sigma_x^2\mathbf{I}_{K_x}\end{bmatrix}$$

and

$$\mathrm{Var}(\mathbf{y}) \equiv \mathbf{V} = \mathbf{Z}\Sigma_u\mathbf{Z}^T + \sigma_\varepsilon^2\mathbf{I}_n$$

be the covariance matrices of $\mathbf{u}$ and $\mathbf{y}$. Following from the linear mixed model theory, representation (2) permits to fit model (1) simultaneously using mixed model methodology and software, to obtain the estimates $\hat{\beta}$ and $\hat{\mathbf{u}}$ by the EBLUPs (3) and (4)

$$\hat{\beta} = \left(\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{y}, \tag{3}$$

$$\hat{\mathbf{u}} = \hat{\Sigma}_u\mathbf{Z}^T\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}), \tag{4}$$

and $\hat{\sigma}_s^2$, $\hat{\sigma}_t^2$, $\hat{\sigma}_x^2$ and $\hat{\sigma}_\varepsilon^2$ by REML/ML estimation.

The addition of others explicative variables is straightforward: smoothing components are added in the random effects term $\mathbf{Z}\mathbf{u}$, while linear components can be incorporated as fixed effects in the $\mathbf{X}\beta$ term. Moreover, the mixed model structure provides a unified and modular framework that allows to easily extend the model to include various kind of generalization and evolution.

3 Data and Empirical Results

The 2002 Living Standard Measurement Study (LSMS), conducted in Albania by the INSTAT (Albanian Institute of Statistics), provides individual level and household level socio-economic data from 3,599 households drawn from urban and rural areas in Albania. The sample was designed to be representative of Albania as a whole, Tirana, other urban/rural locations, and the three main agro-ecological areas (Coastal, Central, and Mountain). The survey was carried out by the Albanian Institute of Statistics (INSTAT) with the technical and financial assistance of the World Bank.

Four survey instruments were used to collect information for the 2002 Albania LSMS: a household questionnaire, a diary for recording household food consumption, a community questionnaire, and a price questionnaire. The household questionnaire included all the core LSMS modules as defined in [6], plus additional modules on migration, fertility, subjective poverty, agriculture, and nonfarm enterprises. Geographical referencing data on the longitude and latitude of each household were also recorded using portable GPS devices [19].

In addition to the geographical location of each household, the covariates selected to fit the geoadditive model are chosen following prior studies on poverty assessment in Albania [3, 12]. We have selected the following household level covariates:

- *size of the household* (in term of number of components)
- *information on the components of the household*:
  - age of the householder,
  - marital status of the householder,
  - age of the spouse or husband of the household,
  - number of children 0-5years,
  - age of the first child,
  - number of components that work,
  - highest level of education in the household;
- *information on the house*:
  - building with 2-15 units,
  - built with brick or stone,
  - built before 1960,
  - number of rooms per person,
  - house surface < 40 m^2,
  - house surface $40 - 69^2$,
  - wc inside;
- *presence of facilities in the dwelling*:
  - TV,
  - parabolic,
  - refrigerator,
  - washing machine,
  - air conditioning,

  - computer,
  - car;
- *ownership of agricultural land*

The response variable is the logarithm of the household per-capita consumption expenditure. The use of the logarithmic transformation is typical for this type of data as it produce a more suitable response for the regression model (see the distributions presented in Figure 1).

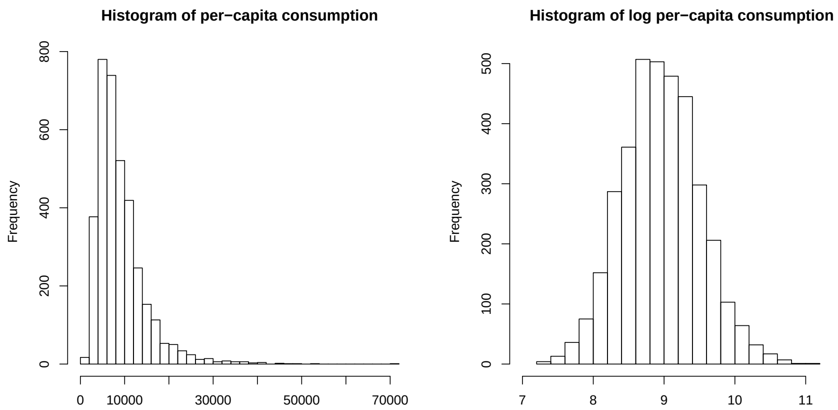

Fig. 1 Distribution of the household per-capita consumption expenditure, both in original scale and in logarithmic scale.

After the preliminary analysis of various combination of parametric and non-parametric specifications for the selected covariates, the chosen model is composed by a bivariate thin plate spline on the universal transverse Mercator (UTM) coordinates, a linear term for all the other variables and a random intercept component for the area effect. The spline knots are selected setting $K = 100$ and using the clara space filling algorithm of [10] that is available in the `R` package `cluster` (the resulting knots location is presented in Figure 2). The model is then fitted by REML using the `lme` function in the `R` package `nlme`.

The estimated parameters are presented in Table 1, along with their confidence interval at 95% and the p-values. With the exclusion of the intercept and the coordinates coefficients (that are required by the model structure), almost all the parameters are highly significant. The exceptions are the coefficients of 'marital status of the householder', 'number of children 0-5years' and 'built with brick or stone' that are significant at 5% level, and the coefficient of 'building with 2-15 units' that is significant at 10% level.

The resulting spatial smoothing of the log per-capita consumption expenditure is presented in Figure 3. From this map, it is evident the presence of a spatial dynamic

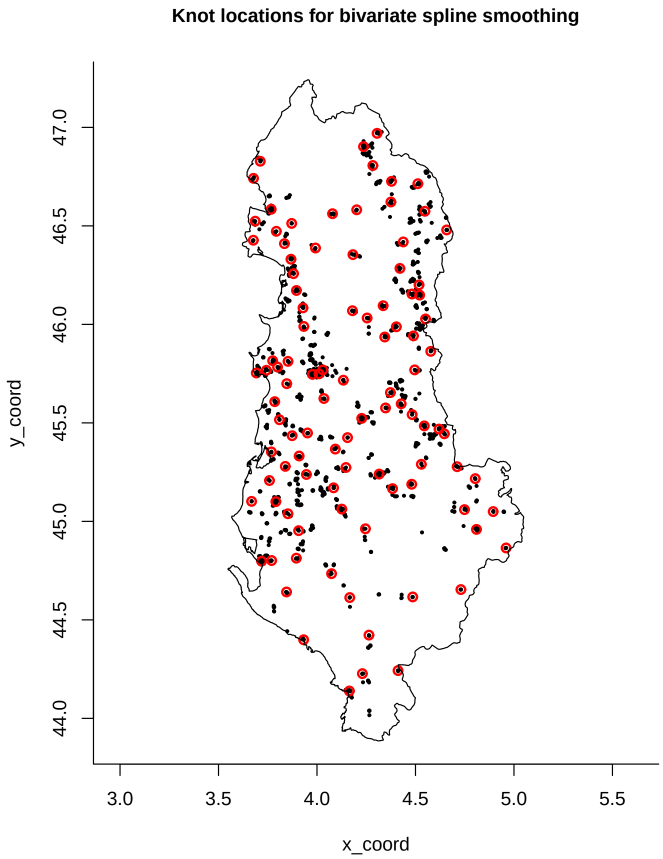

Fig. 2 Knots location (in red) for the thin plate spline component. Black dots indicate the locations of the LSMS sample.

in the Albanian consumption expenditure, even after controlling for all the descriptive household level covariates. The variable presents a clear geographical pattern, with the higher values in the south, south-west and north-east of the country and the lower value in the mountainous area (north and east). These results are consistent with previous applications on the same datasets presented in literature [12, 18].

4 Concluding remarks and open questions

The interest in the spatial data analysis is increased in every area of statistical research. Particular interest is given to the possible ways in which spatially referenced

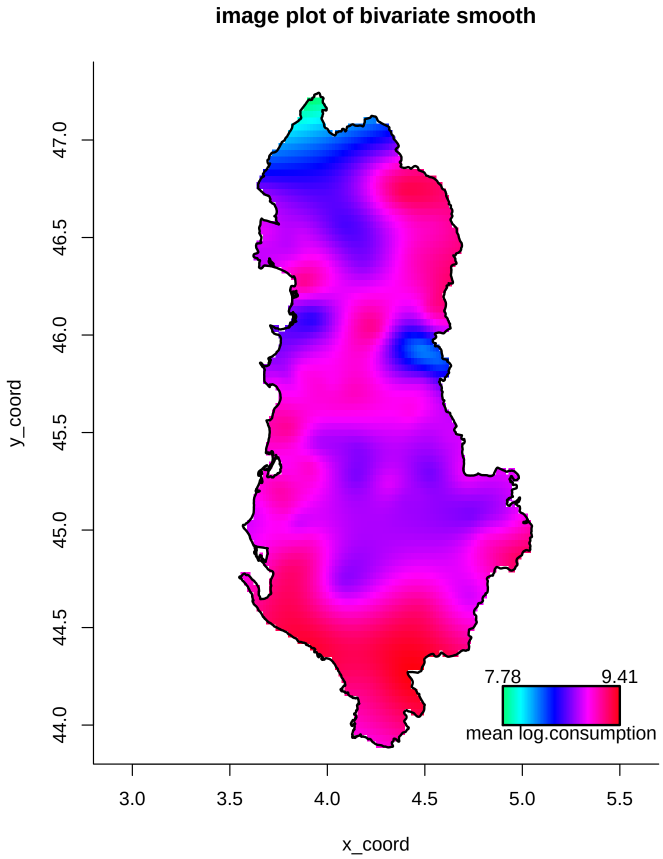

Fig. 3 Spatial smoothing of the household log per-capita consumption expenditure.

data can support local policy makers, especially in areas of social and economical interventions. The geographical information is frequently available in many areas of observational sciences, and the use of specific techniques of spatial data analysis can improve our understanding of the studied phenomena.

A large literature has emerged emphasizing the importance of location in influencing economic phenomenon, however the impact of location has generally been ignored in the majority of the studies. The findings of our analysis make a contributions in this direction introducing geoadditive models and developing an application in a fields of statistics that differ from their native research areas.

Moreover, the geoadditive model produces a continuous surface estimation over the entire area of study overcoming the modifiable area unit problem (MAUP) [8].

The empirical evidence suggests that, despite being overlooked in the previous studies, the spatial location is an important component to understand the distribu-

Table 1 Estimated parameters of the geoadditive model for the household log per-capita consumption expenditure.

Parameter	Estimate	Confidence Interval	p-value
		Fixed Effects	
Intercept	17.0828	(- 71.59 ; 105.75)	0.706
X coordinate	-0.2504	(-1.7389 ; 1.2382)	0.742
Y coordinate	-0.1650	(-2.1115 ; 1.7814)	0.868
household size	-0.0752	(-0.0890 ; -0.0612)	< 0.001
age of the householder	0.0031	(0.0016 ; 0.0046)	< 0.001
marital status of the householder	0.0754	(0.0008 ; 0.1500)	0.048
age of the spouse or husband	-0.0022	(-0.0035 ; -0.0009)	0.001
number of children 0-5years	-0.0221	(-0.0401 ; -0.0040)	0.017
age of the first child	-0.0025	(-0.0039 ; -0.0010)	0.001
number of components without work	-0.0674	(-0.0799 ; -0.0549)	< 0.001
high level of education	0.0930	(0.0662 ; 0.1198)	< 0.001
medium level of education	0.2410	(0.2015 ; 0.2804)	< 0.001
building with 2-15 units	0.0272	(-0.0028 ; 0.0572)	0.075
built with brick or stone	0.0295	(-0.0005 ; 0.0642)	0.095
built before 1960	-0.0425	(-0.0719 ; -0.0131)	0.005
number of rooms per person	0.1398	(0.1070 ; 0.1726)	< 0.001
house surface < 40 m^2	-0.0529	(-0.0944 ; -0.0113)	0.013
house surface $40 - 69^2$	-0.0361	(-0.0623 ; -0.0099)	0.007
wc inside	0.0520	(0.0194 ; 0.0846)	0.002
TV	0.1076	(0.0515 ; 0.1637)	< 0.001
parabolic	0.0803	(0.0506 ; 0.1100)	< 0.001
refrigerator	0.1161	(0.0802 ; 0.1519)	< 0.001
washing machine	0.1064	(0.0763 ; 0.1366)	< 0.001
air conditioning	0.2451	(0.1601 ; 0.3302)	< 0.001
computer	0.2398	(0.1656 ; 0.3139)	< 0.001
car	0.3269	(0.2878 ; 0.3659)	< 0.001
ownership of agricultural land	0.0624	(0.0279 ; 0.0969)	< 0.001
		Random Effects	
σ_γ	0.8999	(0.7228;1.1294)	< 0.001
σ_e	0.3307	(0.3230;0.3386)	< 0.001

tion of the consumption expenditure. In particular, the results from this paper show that the region morphology can explain, to some degree, the spatial patterns of the household per-capita consumption expenditure that remain after controlling for all the descriptive household level covariates effect.

A strictly connected method, which it is not presented in this work, is the spatial small area estimation (SAE) method that exploit the spatial information to borrow strength from the neighbour areas to produce more reliable estimations [14]. As both the SAE models and the geoadditive models are formulated as linear mixed models, it seems an obvious choice to merge the two models in a geoadditive SAE model to take advantage of the spatial information and produce estimates at small area level [13]. This development is broadly discussed in [4]. Here we recall that the condition for which the analysis is carried out is to knowledge of the location of all population units at the point level. This requirement is not so easy to be accomplished, especially if we work with socio-economic data. Usually is much more easy to know the areas to which the units belong to (i.e. census districts, blocks, municipalities, enumeration areas, etc.) and the classic approach is to refer the data with respect to the area centroids. Another aspect to be explored is the use of a more precise spatial location data: a measurement error approach which considers a more realistic hypothesis on spatial distribution. Definitely, further investigations should be done in this direction.

Acknowledgements This research was supported by the Italian Ministry for Education, University and Research (PRIN 2007RHFBB3_004).

References

1. Anselin, L.: Spatial Econometrics: Methods and Models. Kluwer Academic, Dordrecht (1988)
2. Arbia, G.: Spatial Econometrics: Statistical Foundation and Applications to Regional Convergence. Springer, Berlin (2006)
3. Betti, G. and Ballini, F. and Neri, L.: Poverty and Inequality Mapping in Albania, Final Report to the World Bank (2003)
4. Bocci, C.: Geoadditive Models for Data with Spatial Information. Ph.D. thesis, Ph.D. in Applied Statistics, Department of Statistics, University of Florence (2009)
5. Cressie, N.: Statistics for Spatial Data (revised edition). Waley, New York (1993)
6. Grosh, M., Glewwe, P.: A Guide to Living Standards Measurement Study Surveys. In: Household Accounting: Experience and Concepts in Compilation. Statistics Division. Department of Economic and Social Affairs. The United Nations (2000)
7. Hastie, T.J., Tibshirani, R.: Generalized Additive Models. Chapman & Hall, London (1990)
8. Holt, D., Steel, D.G., Tranmer, M.: Area homogeneity and the modifiable areal unit problem. Geogr Syst. **3**, 181–200 (1996)
9. Kammann, E.E., Wand, M.P.: Geoadditive Models. Appl. Stat. **52**, 1–18 (2003)
10. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (1990)
11. McMillen, D.P.: Issues in spatial data analysis. J. Reg. Sci. **50**, 119–141 (2010)
12. Neri, L., Ballini, F., Betti, G.: Poverty and inequality mapping in transition countries. Stat. Transit. **7**, 135–157 (2005)
13. Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G., Breidt, F.J.: Non-parametric small area estimation using penalized spline regression. J. R. Stat. S. ser. B **70**, 265–286 (2008)
14. Petrucci, A., Pratesi, M., Salvati, N.: Geographic Information in Small Area Estimation: Small Area Models ans Spatially Correlated Random Area Effects. Stat. Transit. **7**, 609–623 (2005)
15. Richardson, H.W.: Regional Economics. MacMillan, London (1970)
16. Ruppert, D., Wand, M.P., Carroll, R.J.: Semiparametric Regression. Cambridge University Press, Cambridge (2003)
17. Ruppert, D., Wand, M.P., Carroll, R.J.: Semiparametric regression during 2003–2007. Electron. J. Stat. **3**, 1193–1256 (2009)
18. Tzavidis, N., Salvati, N., Pratesi, M., Chambers, R.: M-quantile models with application to poverty mapping. Stat. Meth. Appl. **17**, 393–411 (2008)
19. World Bank and INSTAT: Albania Living Standard Measurement Survey 2002. Basic Information Document (2003). URL http://go.worldbank.org/IDTKJRT8Y0