Spatial Prediction of Human Population Change

Branislav Bajat[1], Nikola Krunić[2], Milan Kilibarda[1], Mileva Samardžić-Petrović[1]

[1] University of Belgrade, Faculty of Civil Engineering, Institute for Geodesy and Geoinformatics bajat@grf.bg.ac.rs, kili@grf.bg.ac.rs, mimas@grf.bg.ac.rs
[2]Institute for Architecture and Urban Planning of Serbia, nikola@iaus.ac.rs (corresponding author)

INTRODUCTION

In recent demographic researche, quantitative methods have been generally used for analysing changes in population growth without consideration of the spatial aspects of population distribution. The problem of population flows, the phenomena and processes by which it is determined, is often not sufficiently dealt with in traditional demographic analysis, especially on the strategic level, and in most cases, it is reduced to the simple statistic analysis. The links and relations that exist between certain population structures and natural-geographic characteristics of a region are not researched in such analyses. However, the population of a certain area, primarily by its activity, relations and links that are the result of these activities, largely defines the development, manner of utilization and organization of territory. Contemporary urbanization processes, which are expressed by the formation of centers and zones of (over)concentrated population in a specific geospace, by function and capital on one hand, and big parts of the territory that remain demographically, and thereby, functionally empty on the other, have the strong impact on the transformation of the specific national/regional geospace.

In the theory and practice of urban geography and spatial/regional planning, there are many models based on the center-periphery relation, in which the role of the nucleus play the poles of growth and development (urban settlements of varied hierarchical rank and importance), and are linked by axes of development (transport corridors). In these spatial models, the dynamics of population migration represent the key baseline of the entire system and have a great effect on the (un)predictability of its behavior.

Statistical data of population are available at the level of the spatial poligons, like census blocks in the United States, enumeration districts in the United Kingdom, or Census Designation Places (CDPs) in Serbia. The size of these polygons is very heterogeneous across the space, hence this level of spatial resolution may be insufficient in many cases for planning or modelling purposes (Gallego & Peedell, 2001). The objective is to downscale population data, taking into account the analysis of spatial correlation, the correlation

between population change index and chosen spatial indicators, such as digital terrain models (DTM), road networks, and slopes of terrain, as well as, CORINE 2000 entity classes. The feasibility of using stated indicators as auxiliary variables in spatial prediction of population change index was tested in the case study region of South Pomoravlje in Republic of Serbia, by three different techniques; multiple regression, geographically weighted regression and regression kriging.

MATERIALS AND METHODS

Population Change Index

This work is focused on spatial estimation of the **Population Change Index** (*PCI*), which represents the ratio of change in the number of inhabitants at certain location for an observed period between two censuses:

$$PCI = \frac{P_2}{P_1} \times 100\% \quad (1)$$

P_1 represents the population at the beginning of the observed period, and P_2 is the population at the end of the observed period. Index values range from 0 to ∞; values below 1 (i.e. less than 100%) are considered to be negative, i.e. the number of inhabitants has decreased during the observed period. Usually, the inventory of human populations has been drawn up through decennial censuses, which spatially refer to some administrative districts or the census block level, called Census Designation Places (CDPs) in Serbia. In this case study, the beginning (1961) can be considered as the year when a massive planned industrialization started. This was caused mainly by uncontrolled urbanization, which led to the formation of zones of intensive concentration of population, economy, transportation and capital (Krunić, et al., 2009). For the case study area, the index value ranges from 0 (for settlements that are population-wise totally desolate) to 6.32 (or 632%) for once suburban villages, and now parts of urban agglomerations and residential zones. The *PCI* for the period from 2002 to 2027 is based on the planned projection of the region's population, obtained by analytical method of components (Booth, (2006), O'Neill et al. (2001)).

The main geographic and socio-economic characteristics of the region

Region of South Pomoravlje (hereinafter, the region) includes 13 municipalities of the southeastern part of the Republic of Serbia. According to last census (year 2002), the region has over 468,500 inhabitants distributed in 699 settlements and its total surface is 6,289 km^2.

There is an evident center-periphery dichotomy, at the regional, as well as on the subregional levels. About 6.2% of the total population of Serbia lives in the region. The average density of population is 75 inh/km^2, which is less than the Serbian average (85 inh/km^2). The level of urbanization increases, but even at the current level of 41.4% in 2002, it is far less than the Republic average. The region's population increased from 1948 to 1981 by the average of 2.5% in every intercensus period, and the first decrease in the total number of inhabitants (by 1.6%) was between 1981 and 1991. During the latest

intercensus period (1991-2002), even a greater drop in the number of inhabitants was registered, of 5% (index 95.9 using the census methodology). (Petrić&Krunić, 2009; Petrić&Milijić, 2010)

Most village settlements are permanently losing their population, while the municipal centers and suburban villages are growing demographically. The process of expansion of urban influence of the urban nuclei onto the villages in their vicinity has started from the 1980s. Because of a lack of development land, insufficiently developed public-social and communal infrastructure, then suprastructure in the urban nucleuses, suburban villages become destinations for migrants. This has led to their demographic growth followed by intensive residential building and socio-economic transformation, marked by a drop in the share of agricultural population in the total and active population, by an increase in the number of non-agricultural households and combined sources of income.

The relations that exist between certain population structures, the impacts of the socio-economic and psychological factors, and especially natural-geographic characteristics of a region are omitted from the analyses.

Prediction models

The utilized prediction models are classified as statistical (probability) models, where the model parameters are commonly estimated in an objective way, following probability theory. The main disadvantage for these models is that input data set usually needs to satisfy strict statistical assumptions.

Multiple Regression

One of the mature prediction techniques based on the set of auxiliary variables is multiple regression.

A linear multiple regression models could be expressed as:

$$y_i = \beta_0 + \sum_{k=1}^{m} \beta_k x_{ik} + \varepsilon_i, \quad i = 1,...n \tag{2}$$

where y_i represents depended (target) variable, x_{ik} (k=1,…m) is the set of independent variables (predictors) and ε_i is the residual term, all at location i. (Bourennane et al. 2000).The regression coefficients β are usually determined by ordinary least-squares methods.

Geographically Weighted Regression

GWR is a method of modeling the relationships between variables using standard regression but allowing the regression coefficients to vary spatially. GWR represents an extension of the conventional multiple regression framework, by addressing the issue of non-stationary processes:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{m} \beta_k(u_i, v_i) x_{ik} + \varepsilon_i, \quad i = 1,...n \tag{3}$$

where (u_i, v_i) are the coordinates for i-th point, $\beta_k((u_i, v_i)$ are the realizations of continuous function $\beta_k((u, v)$ at the same location, x_{i1}, $x_{i2},\ldots,x_{im}$ are the explanatory variables at point i and ε_i are error terms.

$$\hat{\beta}(i) = (\mathrm{X}^T \mathrm{W}(i) \mathrm{X})^{-1} \mathrm{X}^T \mathrm{W}(i) \mathrm{y} \tag{4}$$

W(i) is a matrix of weights for particular location i such that observations nearer to i are given greater weight than observations further away.

$$\mathrm{W}(i) = diag[w_{i1}, w_{i2}, ..., w_{in}] \tag{5}$$

w_{in} is the weight related to data point n for the estimate of the local parameters at location i. Several types of parameterized weight functions may be used (Fotheringham *et al.* 2002). A common choice is the Gaussian curve that has the form:

$$w_{ij} = \exp\left[-\frac{d_{ij}^2}{2b^2}\right] \tag{6}$$

here, d_{ij} is the distance between observation point i and the realization points j, and $\boldsymbol{b}$ is the parameter called bandwidth which must be determined.

To calculate the parameters associated with a weight function, such as the bandwidth, the GWR methodology utilizes a calibration process. This calculates the parameter so as to form an appropriate trade-off between bias and standard error in the prediction of the overall model.

Regression Kriging

Kriging is a synonym for geostatistical methods for spatial prediction. One of the most commonly used types of Kriging techniques is ordinary Kriging (OK), which assumes weak stationarity of the data. In the case where data are with changing mean and the trend is considered as a function of spatial coordinates, the Universal Kriging (UK) is used as an extension of ordinary Kriging. UK represents a combination of multiple-linear regression and ordinary kriging. If the trend is defined as a linear function of auxiliary variables than prediction techniques such as Kriging with External Drift (KED) or Regression Kriging (RK) are usually used. Generally, KED and RK are applied to merge different sources of information: a target variable that is well known at sparsely sampled observations; and ancillary variables that are not precise but they are available everywhere in the spatial domain. Although those two methods are similar and provide same results, there are some differences in methodology (Hengl *et al.* 2003).

Let measured values of the target variable be symbolized as $Z(s_i)$, i=1…n, where s_i represents spatial location and n number of realized measurements. The system of equations that estimates values of target variables $\hat{Z}(s_0)$ at location s_0 is:

$$\begin{aligned} &\hat{Z}(s_0) = \hat{m}(s_0) + \hat{e}(s_0) \\ &\hat{Z}(s_0) = \sum_{k=0}^{p} \hat{\beta}_k \cdot q_k(s_0) + \sum_{i=1}^{n} w_i(s_0) \cdot e(s_i); \\ &q_0(s_0) = 1 \end{aligned} \tag{7}$$

where $\hat{m}(s_0)$ is the fitted deterministic part, $\hat{e}(s_0)$ is the interpolated residual, $\hat{\beta}_k$ are estimated deterministic model coefficient, w_i are ordinary kriging weights resolved by the spatial structure of residuals $e(s_i)$. Regression coefficients $\hat{\beta}_k$ could be obtained by some fitting method, like ordinary least squares (OLS) or, generalized least squares (GLS), which is more recommended:

$$\hat{\beta}_{GLS} = (\mathrm{q}^{\mathrm{T}} \cdot \mathrm{C}^{-1} \cdot \mathrm{q})^{-1} \cdot \mathrm{q}^{\mathrm{T}} \cdot \mathrm{C}^{-1} \cdot \mathrm{z} \tag{8}$$

hence $\hat{\beta}_{GLS}$ is the vector of estimated regression coefficients, **C** covariance matrix of residuals, **q** is matrix of predictors at the sampling location and **z** is the vector of measured values of target variable. The estimation of $\hat{\beta}_{GLS}$ coefficients presents, in essence, a special case of geographically weighted regression. Estimated variable value $\hat{Z}(s_0)$ at the location s_0, obtained by regression Kriging, is commonly written in matrix notation as:

$$\hat{Z}_{RK}(\mathrm{s}_0) = \mathrm{q}_0^{\mathrm{T}} \cdot \hat{\beta}_{GLS} + \lambda_0^{\mathrm{T}} \cdot \left(z - \mathrm{q} \cdot \hat{\beta}_{GLS} \right) \tag{9}$$

where $\mathbf{q_0}$ is the vector of $p+1$ predictors and λ_0 is the vector of $\boldsymbol{n}$ kriging weights used for interpolation of residuals (Hengl, 2009).

All utilized methods were implemented using the *open-source* **R** statistical computing environment (R Development Core Team, 2009) with ***gstat*** and ***spgwr*** packages (Bivand et al. 2008) intended for modeling and prediction, as well as ***sp*** package, which provides classes and methods for dealing with spatial data in **R** (Pebesma, 2004). The results obtained in **R** can easily be converted into any of the standard GIS formats, which afterwards enable the manipulation and analysis of results in commercial GIS packages.

There are also several *open source* or commercial software packages with associated GWR methods. Unfortunately, GWR is computationallly time

consuming procedure, especially in the case of large datasets, like census data for the whole country. However, it is possible to solve this problem by using Grid computing (Harris et al. 2010). The ***spgwr*** pacakage used in R environment has been also adopted for use on Grid based systems.

Data layers

Census Designation Places (CDPs). The number of CDPs in the region is 699, which is equal to the number of settlements that are followed by the Statistical Office of the Republic of Serbia (2003). CDPs are presented here as points (observed sites), even though they physically depicted as polygonal features. In the area of the region, there are over 1100 settlements that are of dispersed type, smaller or bigger hamlets, or groupings of houses, and they can be seen even on topographic map of the scale 1: 100,000. However, these small settlements „statistically" belong to one central settlement, that „statistic" settlement (in which are also included other small villages hamlets, and groupings of houses) has helped here with the identification of location and the adding PCI value. (Fig. 1, left)

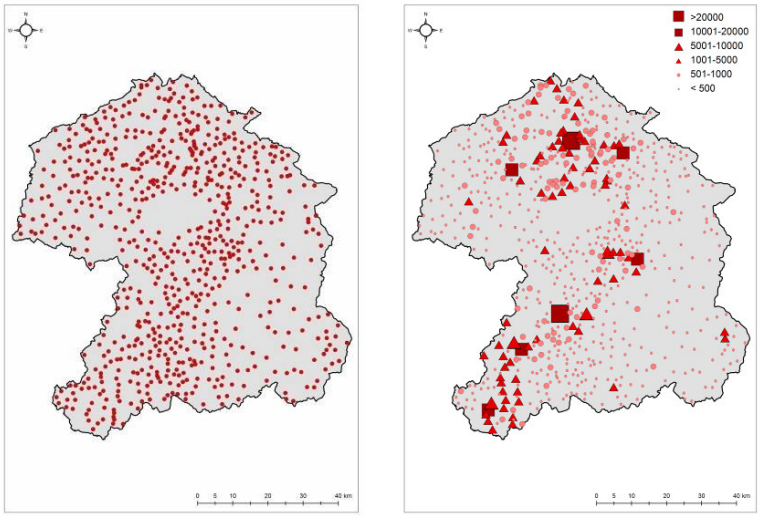

Figure 1*:* The position of CDPs and population size of settlements

Roads network. In the cover of the region, there are 5 roads of first category with an approximate total length of 306 km (including the motorway E-75, European corridor 10), then 26 roads of the 2nd category in the approximate length of 982 km. Roads are particularly indicative predictor because of their vital role in human settlements with or without other forms of traffic connections. Distances to 1st and 2nd categories roads networks were calculated independently and treated as separable environmental grids. (Fig.2)

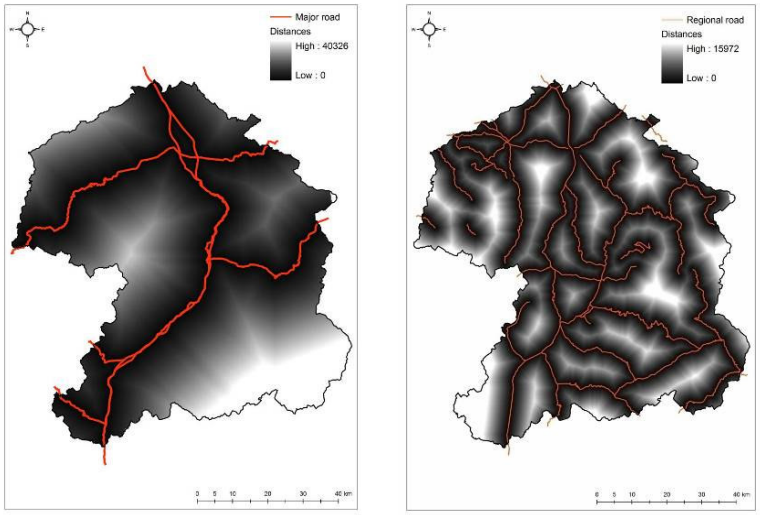

***Figure 2**:* Distances from roads of 1st and 2nd rank

Digital Elevation model (DEM) was produced based on digitalized contours from topographic maps of the scale 1:25000. The resolution of DEM is 200×200 m. (Fig. 3.) The most dominant landscape features are mountains and valleys. A dissected relief is noticeable, with altitudes from 195 m in the northern part up to about 1930 m in the eastern part. Same DEM was used to obtain slopes data layer.

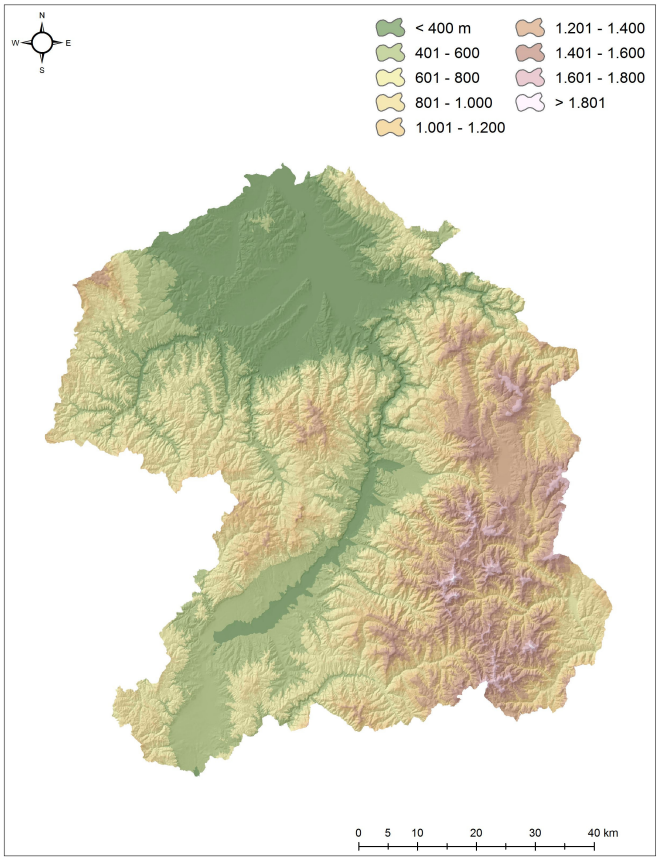

***Figure 3**:* DEM of South Pomoravlje in Republic of Serbia

Land cover. The data based on **Corine 2000** database are from the year 2000. The Corine Land Cover categories of land use were agregated in three layers (1. Artificial surfaces, 2. Agricultural land 3. Forest, water and other) in accordance with studies dealing with population distribution in Europe (Gallego & Peedell, 2001). The structure of land use of the Region is dominated by forest, water and other land with ca. 4,207 km^2 (67% of the total surface of the Region), then the agricultural land with ca. 2,030 km^2 (32%) and artificial surfaces with ca. 51 km^2 (less than 1%). (Fig. 4)

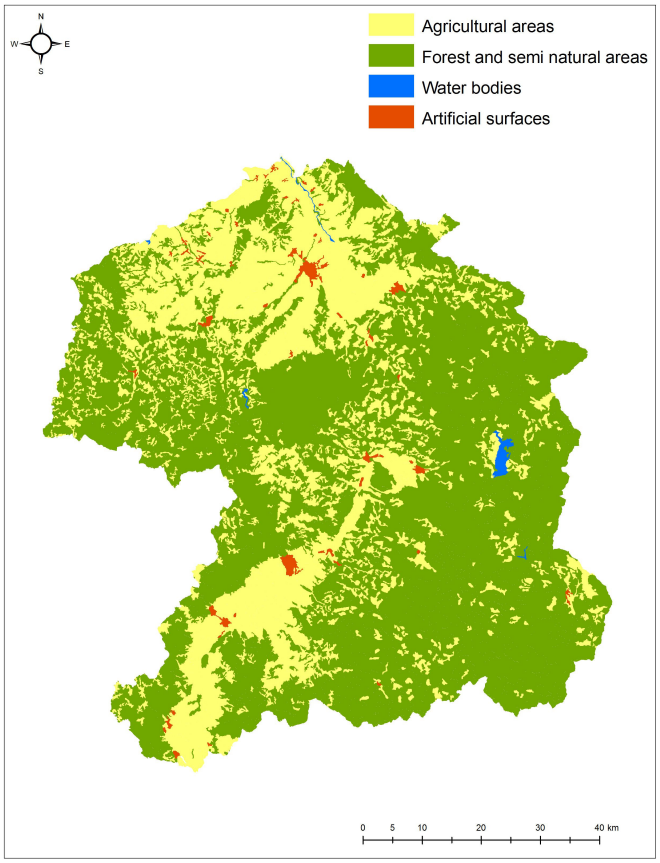

***Figure 4**:* Land cover classes of South Pomoravlje in Republic of Serbia

RESULTS AND DISCUSIONS

One of the basic assumptions before applying regression techniques is that data fulfill central distribution. The distribution of *PCI* follows the Poisson distribution (Fig. 5.) and log-transformation (Hengl, 2009) of input *PCI* values is necessary for both time intervals.

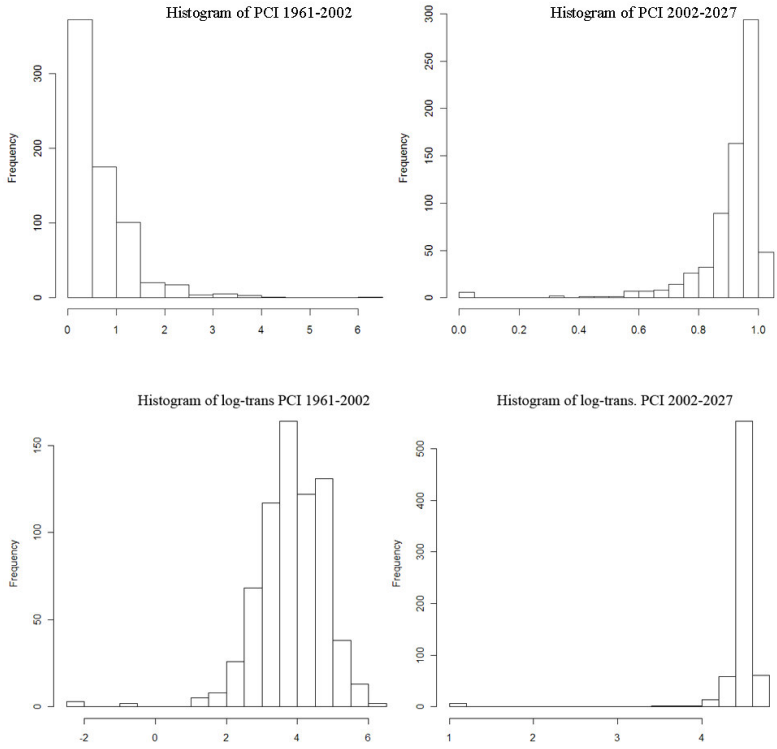

***Figure 5**:* Histograms of the index of population change (before and after lognormal transformation) for the periods 1961-2002 and 2002-2027.

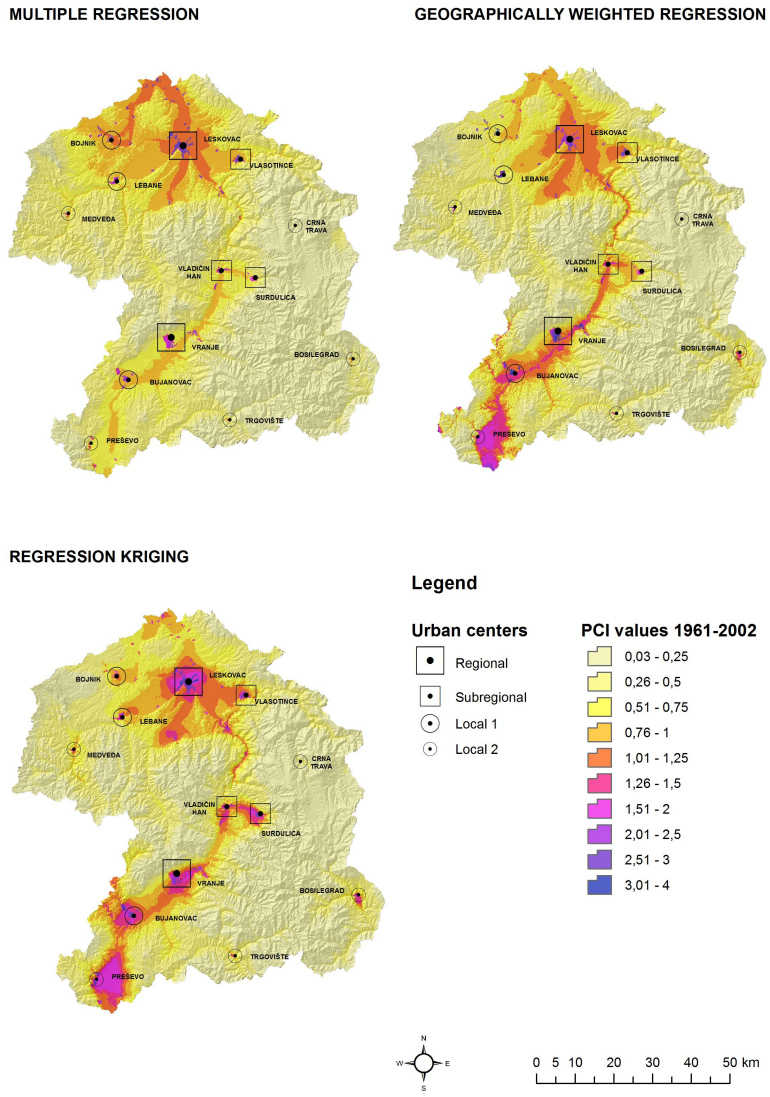

Figure 6: Predicted maps of PCI for the period 1961-2002.

The models of spatial predictions of *PCI* were carried out firstly for the data referring to the period 1961-2002. Predicted *PCI* values are mapped at Fig. 6. The number 1 represents trade-off value for *PCI* (values less than 1 indicate trends of depopulation, values bigger than 1 indicate population growth).

At first glance, the results look identical and therefore the performance of each prediction method was evaluated using *leave-one-out* cross-validation

which involves computing of the distribution of residuals $(\hat{Z}(s_i) - Z(s_i))$ for all data points, when each data point is successively left out and predicted from rest of the data (Burrough &McDonnell, 1997). Obtained residuals (Fig. 7) were used to calculate the mean error (ME) and the root mean square error (RMSE).

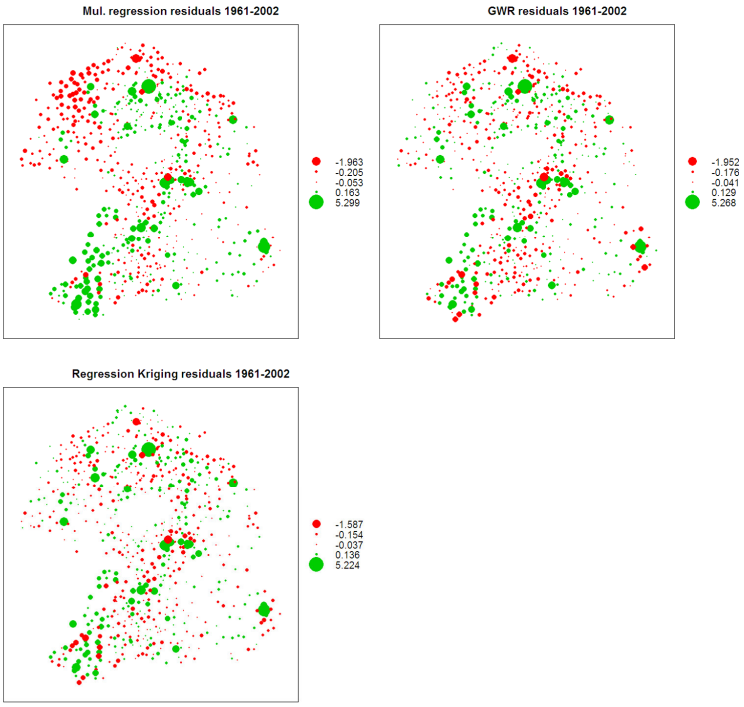

Figure 7: Residuals (1961-2002) obtained using cross-validation

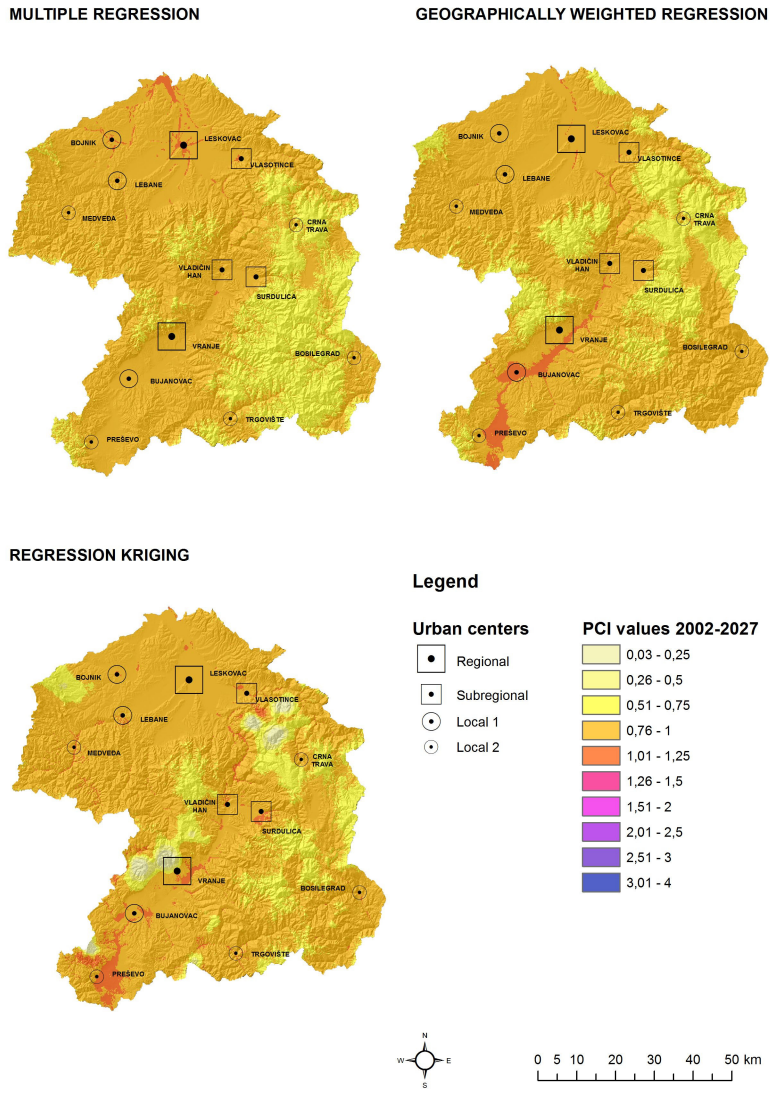

Figure 8: Predicted maps of PCI for the period 2002-2027.

Based on calculated values for ME and RMSE (Table 1.), RK method gives the best results.

	1961-2002		2002-2027	
	ME	RMSE	ME	RMSE
multiple regression	0.294	0.502	0.066	0.117
geog. weighted regression	0.253	0.458	0.058	0.110

regression kriging	0.240	0.443	0.066	0.117

Table 1: Calculated ME and RMSE for periods 1961-2002 and 2002-2027.

The same was done for the period 2002-2027; the mapped values of *PCI* show again the similar prediction results between different methods (Fig 8.). Spatially balanced values of PCI in this period resulted with lower value of residuals (Fig. 9.) and therefore ME and RMSE values are significantly lower. Based on the results given in Tab. 1, it can be concluded that regression kriging model performs better than other used models when abrupt changes are presented in input data set values, whereas GWR shows best performance in the case of spatially balanced data sets.

Comparison of prediction map for 2002-2007 with the map related to period 1961-2002 points to obvious depopulation trends, especially in areas that are distant from roads and regional centers.

Figure 9: Residuals (2002-2027) obtained using cross-validation

CONCLUSIONS

Nowadays, new methods of spatial data modeling are based on utilization of ancillary predictors, and that becomes more important since the layers with

ancillary predictors can be found in databases that are web accessible and free of charge. It is obvious that all methods have the same spatial uncertainty pattern, apparently at locations close to regional or subregional urban centers where residual values are at the highest. However, the estimation of spatial trends in population changes can be effectively done. As expected, GWR methods and regression kriging give better results because they take into consideration the locational component of observed values.

This analysis determines dependence between the settlement process and environmental factors with use of spatial regression models. Observed and projected values of PCI were used to model the process of urbanization and identification of its intensity and spatial locations of urban concentration zones (agglomerations). An effective visualization of complex and long urbanization process was achieved by creation of maps. In this way, complex demographic processes which are mapped become an important factor in defining strategic decisions at regional and national levels.

Acknowledgments

This work was supported by the Ministry of Science of the Republic of Serbia (Contracts No. III 47014 and TR 36035).

BIBLIOGRAPHY

Booth, H. (2006). Demographic Forecasting: 1980 to 2005 in Review. Working Papers in Demography No. 100. Canberra: The Australian National University.

Bivand, R., Pebesma, E., and Rubio, V. (2008). Applied Spatial Data Analysis with R. Use R Series. Springer, Heidelberg ISBN 978-0-387-78170-9. pp.374.

Bourennane, H., King, D., and Couturier, A. (2000). Comparison of kriging with external drift and simple linear regression for predicting soil horizon thickness with different sample densities. Geoderma 97, 255–271.

Burrough, A.P., and McDonnell, A.R. (2000) Principles of Geographical Information Systems,2nd ed. Spatial Information Systems and Geostatistics. Oxford University Press. pp. 333

Fotheringham, A.S., Brunsdon, C., and Charlton, M. (2002) Geographically Weighted Regression: The analysis of spatially varying relationships. Wiley, Chicester. ISBN 0-471-49616-2.pp.265.

Gallego F.J., Peedell S. (2001) Using CORINE Land Cover to map population density. Towards Agri-environmental indicators, Topic report 6/2001 European Environment Agency, Copenhagen, 92-103.

Harris, R., Singleton, A., Grose, D., Brunsdon, C., and Longley, P. (2010) Grid-enabling Geographically Weighted Regression: A Case Study of

Participation in Higher Education in England. Transactions in GIS, Vol. 14, no. 1, 43–61.

Hengl, T., Heuvelink G., and Stein A. (2003) Comparison of kriging with external drift and regression kriging. Technical report. International Institute for Geo-information science and Earth Obeservation (ITC) Enchede. http://www.itc.nl/library/Academic_output/

Hengl, T. (2009). A Practical Guide to Geostatistatistical Mapping. Office for Official Publications of the European Communities, Luxembourg, pp.270.

Krunić, N; Tošić D; Milijić, S (2009): Problems of spatial-functional organization of Južno pomoravlje Region`s network of settlements SPATIUM 19, 56-68.

O'Neill, B. C., Balk, D., Brickman, M., and Markos, E. (2001). A Guide to Global Population Projections. Demographic Research 4 , 203-288.

Pebesma, E., J., 2004. Multivariable geostatistics in S: the gstat package. Comp. Geosci. 30, 683–691.

Petrić, J., Krunić, N. (2009): Regional disproportion encapsulated – Case studies of Južno pomoravlje and Timočka krajina regions of Serbia, Conference Proceedings. Košice: Technical University of Košice, Faculty of Economics and Institute of Regional and Community Development, Košice, Slovakia; p. 694-703.

Petric, J., Milijić, S. (2010) Latest experience of regional planning in Serbia - case studies of Juzno Pomoravlje and Timocka krajina regions of Serbia, Territorial aspects of development of Serbia and neighbouring countries, International conference proceedings, Belgrade: University of Belgrade, Faculty of Geography, Serbia; p.143-149.

R Development Core Team, (2009) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Statistical Office of the Republic of Serbia (2003). 2002 Census of Populations, Households and Dwellings (in Serbian). Belgrade: Statistical office of the Republic of Serbia.