A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 6 (309), 2014

Georgi Shinkov Zabunov[*]
Georgi Georgiev Penchev[**]

REAL ESTATE PRICES – URBAN SECURITY RELATIONSHIPS: SPATIAL ANALYSES AND DEPENDENCIES

1. INTRODUCTION

The free market for housing real estate has existed in Bulgaria for a relatively short time – since the 1990s. The political and economic reforms in the country that accompanied the transition from centralised planning to a free market economy furthered the development of this business as well. Although the history of the real estate market in our country is barely longer than 20 years, it is extremely rich in events – there have already been two peaks and two falls. Currently, the market has frozen at a low and all experts have been looking forward to an upturn. In this situation, there are only major players with long-term goals on the market. Those who are interested in fast speculative profits have left the market. There are more attractive opportunities for them in other fields.

The present report is the result of the collaborative project work of the University of National and World Economy in Sofia and Yavlena – one of the leading Bulgarian real estate agencies. The title of the project is R&D Research 1-16/2013, "*Real Estate Price Indices and Ratings as an Instrument for Investment Risk Management*". The ultimate goal of the project is to offer real estate trade organisations a suitable model for providing information to their business. Two organisations of this kind have been established – the National Real Estate Association (*NREA*) and *FIABCI*-Bulgaria. All major real estate organisations are members of *NREA*. Estate agents (natural persons) are members of *FIABCI*. These organisations are partners in the joint efforts to establish a sustainable and predictable real estate market in Bulgaria, which is in the common interest of all stakeholders in the market. What is more,

[*] University of National and World Economy – Sofia.
[**] University of National and World Economy – Sofia.

considering the fact that the market is still new and market institutions are not mature enough, these trade organisations are in an on-going dialogue with the public authorities.

Due to the relatively complex research and in order to clarify its specifics, its goal, tasks and stages are presented in a separate part below.

2. GOAL, TASKS AND STAGES OF THE RESEARCH

Taking into account the project goal, we have defined the objective of this study as to propose an appropriate and approved method for the evaluation and prediction of real estate prices that will be suitable both for businesses and regulatory services from the municipal and state administration.

The study focuses on the urban security impact on real estate prices. On the one hand, urban security is a major determinant in customers' choices, and its value could seriously influence prices in a certain area. On the other hand, taking into consideration the level of urban security in urban environment could benefit citizens and businesses, as well as the administration dealing with urban planning. The urban security concept and its possible dimensions are discussed further in this study.

To accomplish the goal, the following tasks were set:

- selection of a model to analyse real estate prices,
- selection of a data source and creation of a dataset for this study,
- implementation and evaluation of the selected method and suggestions for its use.

In fact, the Bulgarian experience lacks a unified data for analysis of the real estate market. The lack of such standardised information is a negative fact, but it gives big opportunities for experiments in research. Hence, it was decided to use the best available real estate database. After identifying the most important indices necessary for the implementation of the selected method, they could be used for analysis in the Bulgarian real estate market.[1]

Real estate prices and security contain an important spatial component. Therefore, geographic reference to prices, property characteristics and security are a mandatory part of the information. The city of Denver, Colorado, was selected following a detailed search. This choice is discussed further in the study.

[1] This suggests the assumption for similarity between the preferences of participants in different markets, which the authors consider plausible.

The choice of a method to a large extent is based on the, Handbook on Residential Property Price Indices (EUROSTAT 2013). The selected method, Hedonic regression, was supplemented by a study on spatial dependencies between indices.

Due to the large number of variables used in the study, and the positive tests for co-linearity between them, one of the most popular methods for decreasing this impact was applied – the Principal Component Analysis. Having selected the indices for the model implementation, its potential for prognostication of house prices was assessed.

This study uses one of the largest and most detailed datasets for house price analysis, including over 191 000 units of real estate. Usually, studies are based on a selection of 10-15 000 units. The use of such a big database caused serious problems during the application of the selected method. Analysing the difficulties, however, helped to give useful recommendations for the selection and approach to aggregation of indices, especially from the spatial perspective.

3. REAL ESTATE PRICE INDICES (*REPI*)

Part of the report focuses on real estate price indices because of the ultimate goal of the project. This task is quite complicated, and our research does not aim at tackling it. Yet, the suggested model for information provision of the trade and town authorities should also provide information for *REPI* construction.

The reasons why it is difficult to create *REPI* are twofold. First, property depreciates over time (the depreciation problem), and property may have had major renovations done to it between the two time periods under consideration (the renovations problem) (Diewert et al. 2008). The popular short formulation of this problem is that each estate is unique and there are no two identical estates. At minimum, they are situated in different places on the earth, on different floors, with a different exposure, etc.

Therefore, in order to measure the change in pure estate prices, it is necessary to introduce corrections reflecting the differences in the quality, i.e. the features of the various real estates. Various methods of *REPI* construction are applied for this purpose. One of them is the repeated sales method, which assesses the influence of a number of features on prices through the comparison of estates sold more than once within a particular period of observation. If a comparison is made between estates that were sold repeatedly over a given period, then, obviously, the prices are comparable and the quality of the estates remained the same. The standard repeated sales method is based

on a regressive model that combines the observations from all finalised deals within itself. The main disadvantage of this model is the obligatory periodical data update. The addition of new periods and units to the initial sample changes the values of the previously calculated indices.

Hedonic regression methods are the ones most frequently used in order to avoid these disadvantages and restrictions. This is the reason why we used such a method in the present study. Time is also important for the calculation of *REPI*. In order to use it in the Hedonic regression model, it is necessary to implement the time dummy variable method. It was not applied here, however, because the aims of the study are limited to the analysis of spatial, not time, correlations. We will most likely have to focus on *REPI* closely at a certain stage of our project work. Then, the results from the research on spatial correlations will be complemented by some research into time correlations.

4. HEDONIC REGRESSION AND HEDONIC GEOGRAPHICALLY WEIGHTED REGRESSION

Generally, in hedonic modelling the price p_n^t of the estate *n* at the particular period of time *t* is the function of a certain number of features, for example *K*, and each of them has a value z_{nk}^t. If the number of time periods is *T*+1, beginning with period 0 and going on until period *T*, we would have:

$$p_n^t = f\left(z_{n1}^t, \ldots, z_{nK}^t, \varepsilon_n^t\right), \quad t = 0, \ldots, T, \tag{1}$$

where: ε_n^t is a random error term. In order to define the contribution of each estate price characteristic through standard regression techniques, the above-mentioned equation is presented as a parametric model. The most common models are the fully linear model and the logarithmic-linear model. The linear model can be presented through a dependency of the type:

$$p_n^t = \beta_0^t + \sum_{k=1}^{K} \beta_k^t z_{nk}^t + \varepsilon_n^t. \tag{2}$$

On the other hand, the logarithmic- linear model can be presented with the expression:

$$\ln p_n^t = \beta_0^t + \sum_{k=1}^{K} \beta_k^t z_{nk}^t + \varepsilon_n^t . \quad (3)$$

Here β_0^t and β_k^t correspond to the intercept term and the characteristics parameters to be estimated.

The Hedonic model is suitable for the analysis of spatial dependencies if it is transformed into a geographically weighted regression (*GWR*) model. If the main form of our regression model is:

$$p = \beta_0 + \sum_{k=1}^{K} \beta_k z_k + \varepsilon , \quad (4)$$

converting it into *GWR* we will receive an expression of type:

$$p_i = \beta_{0i} + \sum_{k=1}^{K} \beta_{ki} z_{ki} + \varepsilon_i . \quad (5)$$

Here $i = 1, \ldots, I$ indicates that there are a certain number of the coefficient I that have to be estimated for every observation in our primary data set. This step is necessary when the regression parameters based on OLS do not satisfy the requirements of statistical tests. After mapping the model residuals and calculating Moran's *I* and Local Moran's *I,* we found if there was a spatial correlation in our data.

The literature overview shows many publications that aim to estimate the performance of different spatial approaches and methods (Anselin, Rey 1991). The *GW* methods are compared with moving windows interpolation and regression methods, kriging and co-kriging methods – both global and local types (Páez et al. 2008). The prevailing view is that *GW* methods show excellent predicting abilities. Based on this, the current research is focused only on the use of these methods (Fotheringham et al. 2002).

5. URBAN SECURITY AND THE HEDONIC PRICE MODEL

Security and safety in urban environments are broadly discussed topics. They are complex and closely related to city planning and development. The safety of buildings, traffic and noise levels are among many urban planning and management issues. In cities, there are other processes related to security,

the most important of which are crime levels and counter-crime activities that should be considered in their complexity. In an urban environment these problems have other dimensions – prevention, planning and management. For example, a beautiful park with big trees and distant quiet places can be a nightmare for security guards at night due to the lack of visibility and many dark places.

Currently, the concept of Urban Security has evolved to ensuring human security and quality of life in urban environment (UN-HABITAT 2007). The European Forum for Urban Security (*EFUS*),[2] adopted the Manifesto of Aubervilliers and Saint-Denis – Security, Democracy and Cities in 2012 (*EFUS* 2014). The Manifesto states 19 thematic recommendations for long-term policies toward Urban Security. Security and hedonic values are undisputedly positively related. Thus, many of the above mentioned recommendations can be measured as place characteristics and can be used as variables for Hedonic Price Model.

After detailed consideration of all 19 topics from Manifesto, the following 10 topics were selected as measurable:

- Shared public spaces – measured as the distance to parks, roads and the central business district;
- Sport and prevention, art, culture and prevention, tourism and security – as the distance to recreational facilities, historic places, cinemas, theatres, etc.;
- Addiction and drugs, cities and organised crime, violence at school and school dropout, collective violence – as the level of crime near the property;
- The city at night – can be measured both as crime rate at night and as distance to night-life areas;
- Urban risk management – as the distance to different problems such as trash issues, sewer floods, etc.

The measures are not ideal, but to some degree they can provide an overview of the relation between property price, security and place satisfaction (or hedonic pleasure). Future, more comprehensive research can use not only distance measures, but also measures based on resident interviews, benchmarking, etc. in order to include all 19 recommendations.

[2] *EFUS* was established in 1986 as a forum for cities and local planning authorities, and has a consultative role to the UN and EU working on areas as crime prevention, safety in cities, drug preventions, mediation to central governments and others.

6. THE DATASET

Many data sources were searched in order to find the most complete and appropriate data for the purpose of the research. The search for data began with Europe. National statistics and international sources ranging from the countries' statistical offices and EUROSTAT to sites and databases of real estate agencies were examined. Many of the examined sources were complete and well established, but did not contain the spatial component – there was no geographical reference to the properties. Other sources had spatial references that were too vague – they referred to relatively big administrative neighbourhoods.

The search for data was expanded to include US sources, which gave better results. Many US cities or even states provide spatially referenced public information. The information is usually referenced to US Census Blocks, Block Groups or Tracts, but in some cities the real estate properties information is referenced to the property itself or to the parcel on which it was built. This openness is supported by government and civil-society initiatives. There are many sites devoted to cities' housing properties, life quality and conditions, city policies and regulations on national or regional levels. After a close look at these sites, we chose the datasets provided by the City of Denver, as follows:

– Parcels – vector polygon data set, containing parcels with their spatial references'
– Real Property Residential Characteristics – land and property characteristics, e.g. sq. feet, bedrooms, bathrooms – full and half, stories, units, etc., vector point data set, containing also the key column to Parcels;
– Blocks, Block Groups and Tracts of the City of Denver;
– Zoning – geographically referenced zones – vector polygon file, description of Denver zones and their permits are described in a separate file. The main idea behind zones (a highly topical issue in US) is to state both the purpose of the zone – industrial, commercial, housing, mixed, etc. and the minimum building area in sq. feet;
– Recreation Centres, Historic Buildings and Places, Parks, Schools, After-School Program (places for children's after-school activities), Streets Centerline – vector files referencing the places and some characteristics;
– 911 calls and 311 calls – point vector files referencing the calls to 911 and 311 (city risk management system and phone), which also contains the type of crime, the time of the crime, graffiti problems and trash recycling issues;

– Denver Police Department Precincts and Denver Fire Departments – the vector polygon files with places of police precincts and fire departments.

In order to find distances to other amenities such as restaurants, bars and clubs, an extract from Open Street Data for Denver was used. The Parcel dataset contains a variable "*TOTAL VALUE*" for each property which is likely estimated by the administration price for taxation purposes. Last but not least, the dataset Sales Book 2013 was used. It contains the sales of real estate properties for 2013 and consists of about 24 000 records. This dataset can be used both as controlling data for consistency of estimated administrative prices and as a yearly correction to the Hedonic Price Model.[3] The list of all variables can be found in Appendix 1.

7. INITIAL SET OF VARIABLES

From the above-described datasets, three types of variables can be derived – structural variables, variables closely related to the security environment, and variables showing the distances to different amenities and other places. Variables related to ethnic and income structure are not included. Many analyses of real estate prices use these kinds of variables as an indicator for submarket identification. The phenomena of segregation (Bischoff, Reardon 2013) is inevitable even within one ethnic group and can be based on social and income differences (Bjerk 2006), but in the case of Denver the data from the Census Blocks shows very low levels of spatial dependencies between ethnic groups. On the other hand, the income distribution variable can be resultant rather than independent variable in case of zoning based on the minimum construction area.

Zoning regulations are among the most controversial and discussed issues in US urban planning. In case of Denver, it suggests 8 levels of minimum building areas in sq. feet for each of the residential zones. As a building's area is one of the most important variables in price formation, it forms the affordability level and expected income for the property buyer (or renter). The income (and prices) distribution will probably have 8 clusters. Therefore, the zoning rules can be treated as nested characteristics of a submarket (Chen et al. 2007; Goodman, Thibodeau 1998), and it is better to use other variables showing socially stable structures (Chakrabarti and Roy 2012) rather than income distribution (Hoesli et al. 2002).

[3] All files can be found in Denver Open Data Catalog site (DODC 2014).

The following variables were included in the initial set of variables:

– Structural – the land, building area, basement area – in sq. feet; number of bedrooms, bathrooms, stories and units, year of construction, and improvements to property as amount of money;

– Environmental – crime, crime at night, day crime, number of calls for graffiti, number of calls for trash recycling issues;

– Amenities and others – distances to parks, sport and recreation facilities, historical sites, restaurants, clubs and bars, health clinics and offices, schools (for all grades) and after-school activities programmes, Denver Fire Department and Police Department Precincts, as well as distance to the Central Business District, Commercial and Industrial zones as proxies to an important hedonic variable – the distance to work.

The structural variables were used without any further calculations. For each property, the 911 calls were queried in a 1 km diameter distance around the property. The type of crime has different levels of severity and cost to society. The types of crime were weighted by the Canada Severity Crime Index (Statistics Canada 2014) and averaged by months. The crimes were also divided by day and night. Some authors claim that not all crime influences the level of residents' insecurity – just that part of crime which is close and visible (Buonanno et al. 2013). This was the reason to look at graffiti and trash issues, measured as numbers in a 1 km diameter around each property and averaged by month.

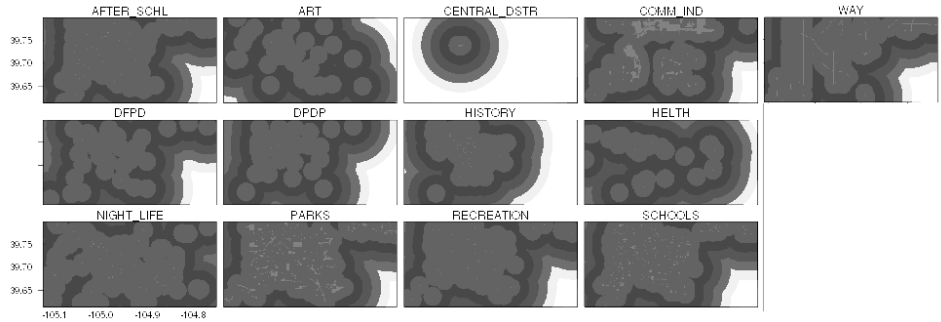

Figure 1. Raster files with interpolated distances

Source: own calculations and geographical reference.

The distances to amenities and other places were measured in two steps. First, the points, lines and polygons were rasterised in raster file with extent of housing data set and then distances were interpolated on this raster.

Second, the raster was queried with the point (centroid) of each property, and the distance between place and property was measured. This approach was less computationally complex and provides stable information. Of course, it is better to measure distances more accurately using streets' centre lines

or with other related variables such as average time for travelling to work, but the former method is extremely computationally complex for large datasets and the latter is not available in current datasets.

8. INITIAL DATA SET, TRANSFORMATION AND AGGREGATION

The initial dataset consists of all types of real estate properties, with more than 191 000 records and 24 variables in columns. The 1 588 outliers were detected based on the examination of value and structural variables with the R's "*OutlierDm*" package.[4] There are no missing values in the dataset, therefore there was no need for data imputation or resampling.

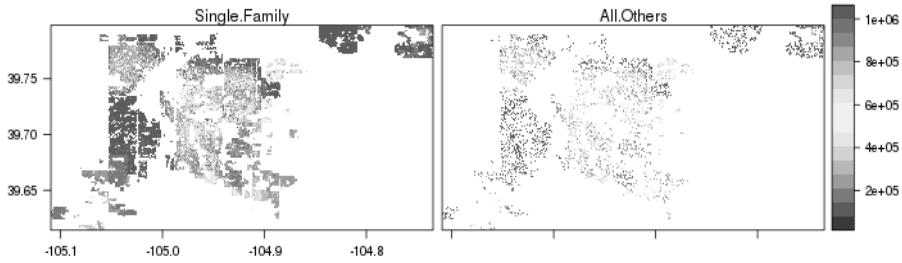

Figure 2. Single-family houses and other types of real-estate properties

Source: own calculations and geographical reference.

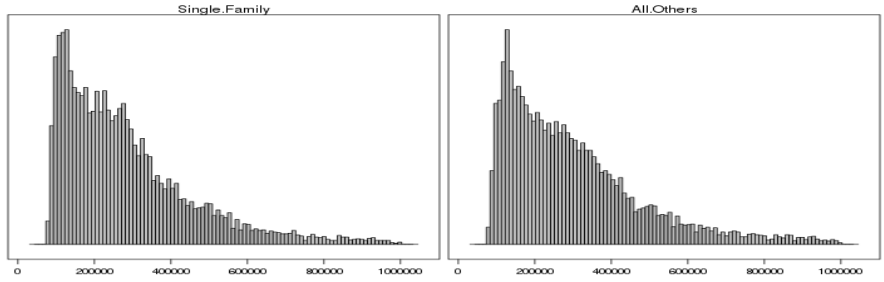

Figure 3. Histograms of values of single-family houses and other types of real-estate properties

Source: own calculations and geographical reference.

[4] Package "*OutlierDm*" – http://cran.at.r-project.org/web/packages/OutlierDM/OutlierDM.pdf.

Subsetting different types of real estate properties shows that family houses are spatially clustered as opposed to other types of properties (Figure 2). Their values vary from 50 000 to 8 000 000 USD, but a more reasonable interval from 80 000 to 1 000.000 USD was chosen (Figure 3).

The dataset was initially transformed to the Spatial Point Data Frame (R "*sp*" package format).[5] The regression with the dependent variable "*TOTAL VALUE*" and all 24 independent variables described above was calculated according to Formula 5.

The dependent variable and residuals (spatially referenced) were added to the initial point's data frame. In order to apply the spatial dependencies test, the point data frame was rasterised.[6] Moran's *I* test showed a positive spatial dependence of 0.671981. Therefore, the application of a Geographically Weighted Regression (*GWR*) or other approach to address this issue should be used. All independent variables have a very high statistical significance.

The dataset, with 122 587 entries, is far beyond the computational abilities of the current hardware and software. One possible method is to take just a fraction from the dataset on the basis of a random sampling procedure. Usually, 10% of the whole sample is used, but in the current case it means more than 12 000 records, which is still a large sample. Another possible approach is to aggregate points in predefined regions. In this case, the region's definition should be clarified.

A reasonable approach is to aggregate housing characteristics according to submarket regions – areas with the same influence of variables and relatively stable price levels.

The idea is to group prices in 8 clusters (according to the 8 residential zoning types) and then to define the area with the majority of cluster representatives. In the literature, delineation of submarkets is based on hierarchical clustering, but the computational complexity of this and many other clustering algorithms do not permit such operations on a large dataset. The only possibility is to use a *k*-mean clustering algorithm.

The procedure used here for submarket delineation has the following steps. First, the point dataset was rasterised in a high-resolution raster (about 10 meters) and median function for aggregation. Then, cluster on the raster with *k*-mean algorithm with 8 predefined clusters was carried out. After clustering, a focal majority filter was applied. Finally, the resulting raster was transformed into a vector file representing regions that encompass neighbours of the same cluster. The vector file consists of more than 150

[5] Package "*sp*" – http://cran.r-project.org/web/packages/sp/sp.pdf.

[6] Calculating Moran's *I* test on the dataset was impossible because of its size. The solution was to rasterise the point vector file and to apply Moran's *I* test on the resultant raster file with the functions from R's package "*raster*".

polygons (regions). All prices and characteristics were aggregated within these regions and the result is shown in Figure 4.

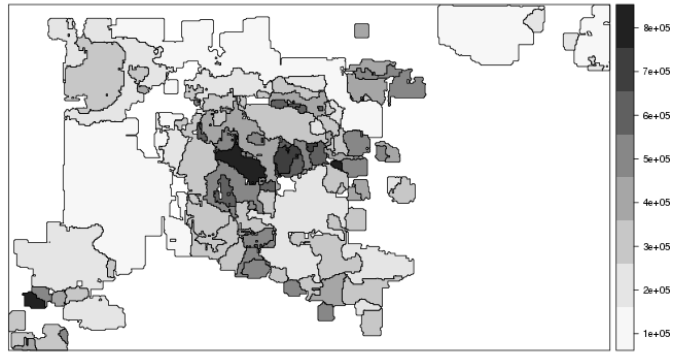

Figure 4. Regional extraction and price aggregation

Source: own calculations and geographical reference.

Initially, the procedure looks promising, but further analysis shows serious problems. An attempt to apply any type of spatial models leads to a computational singularity error. This error is well-known and is the result of a very high level of correlation within the variables. Thus, computing a regression model on such a dataset is impossible. Aggregation rules, the relatively small number of regions and their dissimilarity in size probably led to the error.

An experimental aggregation based on Denver's predefined building zones (about 300 polygons) again led to the same error. This is an additional argument for the explanation of the errors. Therefore, the aggregation should be performed on regions with definition (delineation), based on reasons unrelated to real estate property price formation.

9. TWO TYPES OF VARIABLES AGGREGATIONS AND LOCAL CO-LINEARITY

The Census Block Groups and Block vector polygon datasets are used in order to aggregate data with different types of regional delineation. Census Blocks are defined mainly on the basis of the geographical position of streets rather than on other factors. The Block Groups are a simple geographical union of Blocks based on the nearest neighbour and size.

After the aggregation, polygons without any data in "*TOTAL VALUE*" were deleted. The aggregated file of Block Groups contains 450 residential regions and the Blocks file contains 7 238 residential polygons.

Dependencies between the variables were considered and some variables were removed. From among the structural variables, the number of bathrooms and bedrooms were excluded because they can be considered as a function of the most important variable of the floor area (*AREA_ABG*). The variable land area is kept in a list because of its importance according to the EUROSTAT Handbook Residential Property Prices Indices (*RPPI*s) (EUROSTAT 2013).

Some of the variables show interesting interdependencies. Year of construction, improvement, and distances to the Central Business District (*CBD*) and historical sites have a very high correlation with almost all other variables. One possible explanation is that historical places and the oldest houses are built in close proximity to the *CBD* and these older houses need improvement. The *CBD* and historical places attract tourists and city residents, thus making these areas more attractive for establishing restaurants, bars, cinemas and other amenities. As a result, six variables were excluded from the initial set of variables.

10. CENSUS BLOCKS AND GEOGRAPHICALLY WEIGHTED PRINCIPAL COMPONENT ANALYSIS

As a remedy both to local co-linearity and to related computational troubles, the *GW* Principal Analysis was chosen. The main purpose of Principal Component Analysis (*PCA*) is to reduce the number of explanatory variables in a regression function, without losing useful information and avoiding multi co-linearity issues (Demšar et al. 2013).

In the *GW PCA*, we have one additional matrix of the coordinates (u, v) of each location i of variables x_i. The local covariance matrix is:

$$\sum(u,v) = X^{\mathrm{T}} W(u,v) X \,, \tag{6}$$

where: X is data matrix with n observation in rows and m variables in columns. The W is a diagonal matrix of geographical weights. The local principal components in location i can be formulated as:

$$L(u_i, v_i)V(u_i, v_i)L(u_i, v_i)^{\mathrm{T}} = \sum(u_i, v_i). \quad (7)$$

where: L is the matrix of local eigenvectors and V is diagonal matrix of local eigenvalues. Thus, for each location *GW PCA* with m variables has m eigenvalues, the same number of components, and a set of loadings and scores for each observed locations. (Gollini et al. 2013: 12-19).

The first step was to calculate global *PCA* and to find the number of principal components (*PC*s), which together accumulated enough percentage of total variance. Usually, *PC*s are selected until they form about 80% of the total variance. In our case, the first 9 *PC*s were selected from a model of 18 variables. The large number of *PC*s indicates a relatively low correlation between variables.

The bandwidth for the *GW PCA* was selected with a cross-validation test and the basic *GW PCA* was calculated. Based on the absolute value of the biggest loading (portion of variable included in principal component), the "*winner*" variable for the principle component was defined. The "*winner*" can also be defined as the most influential variable for the principal component (Gollini et al. 2013).

In the case with the *GW PCA*, in every observed location there is one set of component loadings and therefore a local winner can be chosen. Figure 5 shows the local winner for each of the 7 238 locations in Census Blocks areas.

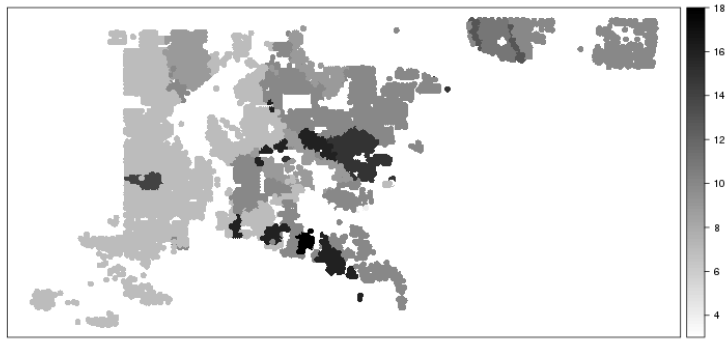

Figure 5. Local PCs "*winners*" for 18 independent variables

Source: own calculations and geographical reference.

It is obvious that the "*winner*" variable forms regions of influence. The most spread-out winner variable is the Trash issue (2 852), followed by Fire Department (2 360), Distance to Commercial and Industrial Zones (806), Land in Sq. Feet (369), Recreation (296), Parks (263), Police Department (144), Health Facilities (55), Nightlife (53), Highway Distance (37) and Floor Area (3).

Similarly, winning variables can be defined for all *PC*s in the model. It is a useful approach to identify positive and negative influences from Urban Security point of view. It can also be simulated to some degree in an appropriate resampling framework. Such experiments can give a reasonable foundation for making a decision related to city planning and management.

11. PREDICTION AND CORRECTION OF ESTIMATED REAL PROPERTIES PRICES

The research is based on the study of models in which dependent variables present estimated values of real estate properties. Having one additional set of market prices from the 2013 Sales Book, we can study how close the estimated prices are related to market prices. If we can define this relation and address its local values, we can predict missing market values or correct estimation prices in order to improve the estimation process.

The Sell Book has more than 24 000 records, about 14 000 of these which were selected and extracted as single-family houses. The new dataset was spatially referenced as point data frame and aggregated with the Census Blocks vector polygon file. The polygons with missing sell prices were deleted and the resultant file has 5 244 polygons.

The global correlation coefficient is above 0.876 and local correlation coefficients are shown in Figure 6.

Figure 6. Local correlation coefficients of estimated and market prices

Source: own calculations and geographical reference.

The figures show strong global and local dependencies between two variables where both local and global regressions can be estimated.

If the regression has the type *Estimated Price ~ Sell Price*, we can use local (or global) coefficients to improve the estimation process for administrative purposes. If the used model is *Sell Price ~ Estimated Price*, we can use estimated prices as a predictor for Sell Prices.

As Figure 7 shows, the predicted values are uniformly distributed and have a few points which can be treated as outliers.

The results of this model show a good fit between the real estate prices estimation process and market prices.

Figure 7. Difference between predicted values of real estate prices and market prices recorded by the 2013 Sales Book

Note: the average difference is about 15 000 USD.

Source: own calculations and geographical reference.

12. CONCLUSIONS

In conclusion, it could be stated that this study has achieved its goals. As a result of the data search, a detailed dataset was created with geo-references about prices and structural features for over 191 000 real estate properties in Denver, Colorado. This information was supplemented with more indices to take into account the so-called hedonic characteristics of entries. Special attention was paid to the security of properties, as an important component of customers' choice. Urban Security was used as a basic concept to define indices referring to security.

The choice of the Geographically Weighted Hedonic Regression method for analysis of real estate properties proved to be appropriate; however, in the course of its implementation, the following considerations emerged:

First, the scope of the application of this method is limited by the hardware's computational capabilities. Its use with a relatively large dataset could be expensive and time-consuming. Therefore, advance data aggregation is necessary before applying the method.

Second, the data aggregation shall be performed in a way which excludes interdependency between the indices and rules for aggregation. The analysis has shown that it is suitable to use aggregation in regions defined by a geographic (or demographic) principle, e.g. the preliminary set statistic regions Census Blocks in Denver.

Third, there is a high probability for co-linearity when using many indices. The indices could turn out to be related in a different, initially unexpected, way. Therefore, analysis of the presence of co-linearity between indices is mandatory. At the same time, some methods such as Principal Component Analysis could be used to remove the impact of co-linearity. They could bring additional benefits such as defining the most important indices with the highest impact in different regions.

This research is focused on creating an analytical database to study the Bulgarian real estate market. The use of data from another market – the city of Denver, Colorado – has shown some significant specifics of the implementation of the method selected which need to be accounted for and will be recommended in a future database for analysing the Bulgarian market.

REFERENCES

Anselin L., Rey S. (1991), *The Performance of Tests for Spatial Dependence in a Linear Regression, Report 91-13*, National Center for Geographic Information and Analysis, Santa Barbara, CA, http://www.ncgia.ucsb.edu/Publications/Tech_Reports/91/91-13.pdf.

Bischoff K., Reardon S. F. (2013), *Residential Segregation by Income, 1970–2009*, Russell Sage Foundation, American Communities Project of Brown University. Research Paper, http://cepa.stanford.edu/sites/ default/files/report10162013.pdf (access: May 2, 2014).

Bivand R., Bivand M. R., Brunsdon M. C., Fortheringham S. (2013), Package "*spgwr*", "R Software Package", http://ftp.iitm.ac.in/cran/web/packages/spgwr/spgwr.pdf (access: June 30, 2014).

Bjerk D. J. (2006), *The Effect of Segregation on Crime Rates In American Law & Economics Association Annual Meetings*, The Berkeley Electronic Press, http://law.bepress.com/cgi /viewcontent.cgi?article=1693&context=alea (access: July 2, 2014).

Buonanno P., Montolio D., Raya-Vílchez J. M. (2013), *Housing Prices and Crime Perception*, "Empirical Economics", vol. 45, no. 1, pp. 305-321.

Chakrabarti R., Roy J. (2012), *Housing Markets and Residential Segregation: Impacts of the Michigan School Finance Reform on Inter-and Intra-District Sorting*, Staff Report, Federal Reserve Bank of New York, http://www.econstor.eu/handle/10419/62940 (access: June 27, 2014).

Chen Z., Cho S., Poudyal N., Roberts R. K. (2007), *Forecasting Housing Prices under Different Submarket Assumptions*, American Agricultural Economics Association Annual Meeting,

Portland, OR, http://ageconsearch.umn.edu/bitstream/9689/1/sp07ch04.pdf (access: May 6, 2014)

Demšar U. et al. (2013), *Principal Component Analysis on Spatial Data: An Overview*, "Annals of the Association of American Geographers", vol. 103, no. 1, pp. 106-128.

Diewert W. E., Nakamura A. O., Nakamura L. I. (2008), *The Housing Bubble and a New Approach to Accounting for Housing in a CPI*, Social Science Research Network, Rochester, NY. SSRN Scholarly Paper, http://papers.ssrn.com/abstract=2274933 (access: June 30, 2014).

DODC (2014), Denver Open Data Catalog, http://data.denvergov.org/ (access: May 28, 2014).

EFUS (2014), *The Manifesto of Aubervilliers and Saint-Denis*, "European Forum for Urban Security", http://efus.eu/en/resources/publications/efus/3779/ (access: June 30, 2014).

EUROSTAT (2013), *Handbook on Residential Property Price Indices*, Publications Office of the European Union, Luxembourg.

Fotheringham A. S., Brunsdon C., Charlton M. (2002), *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*, Wiley.

Gollini I. et al. (2013), *GWmodel: An R Package for Exploring Spatial Heterogeneity Using Geographically Weighted Models*, "Cornell, University Library", http://arxiv.org/abs/1306.0413 (access: July 26, 2014).

Goodman A. C., Thibodeau T. G. (1998), *Housing Market Segmentation*, "Journal of Housing Economics", vol. 7, no. 2, pp. 121-143.

Hoesli M., Bourassa S. C., Peng V. S. (2002), *Do Housing Submarkets Really Matter?*, Social Science Research Network, Rochester, NY. SSRN Scholarly Paper, http://papers.ssrn.com/abstract=372160 (access: June 30, 2014).

Páez A., Long F., Farber S. (2008), *Moving Window Approaches for Hedonic Price Estimation: An Empirical Comparison of Modelling Techniques*, "Urban Studies", vol. 45, no. 8, pp. 1565-1581.

Statistics Canada (2014), *Measuring Crime in Canada: Introducing the Crime Severity Index and Improvements to the Uniform Crime Reporting Survey: Table 1 — Examples of Weights for the Crime Severity Index*, http://www.statcan.gc.ca/pub/85-004-x/2009001/t001-eng.htm (access: June 30, 2014).

UN-HABITAT (2007), *Enhancing Urban Safety and Security: Global Report on Human Settlements 2007*, United Nations Human Settlements Programme, Earthscan, London; Sterling, VA.

ABSTRACT

The purpose of this article is to assess the possibilities for creating an analytical database to study real estate prices. To a large extent, the article presents some of the findings of a joint project with leading Bulgarian real estate agencies. Using a suitable analytical approach and standardising the information would bring substantial benefits to the dynamic Bulgarian market. Due to the lack of tools and experience, it was necessary to select an appropriate method and to apply it to the largest possible database created with the help of information from other markets. The study focused on the impact of urban security on real estate prices. On the one hand, this is a basic determinant for customers' choice, and on the other hand, information about security rating could be used in urban planning and management.

As a result, a georeferenced dataset was created with information about the characteristics of over 191 000 properties in Denver, Colorado. The application of the selected method – the Geographically Weighted Hedonic Regression – for this dataset showed a number of issues related to hardware and software restrictions of the application, the manner of data aggregation and the presence of co-linearity between indices. The application of the Geographically Weighted

Principal Analysis as a means of solving the problem of co-linearity has shown other advantages such as defining the impact of various indices in smaller urban regions.

Despite using data from other markets, this research has made some important conclusions regarding the definition, collection and study of data necessary for the creation of a suitable database to analyse the Bulgarian real estate market.

RELACJA CENY NIERUCHOMOŚCI A BEZPIECZEŃSTWA MIEJSKIEGO: ANALIZY PRZESTRZENNE I BADANIA ZALEŻNOŚCI

ABSTRAKT

Celem niniejszego artykułu jest ocena możliwości utworzenia bazy danych analitycznych do badania cen nieruchomości. W dużej mierze, artykuł przedstawia niektóre z ustaleń wspólnego projektu, prowadzonego wspólnie z wiodącymi bułgarskimi agencjami nieruchomości. Za pomocą odpowiedniego podejścia analitycznego i ujednolicenia informacji możliwe jest osiągnięcie znacznych korzyści dla bułgarskiego dynamicznego rynku. Ze względu na brak narzędzi i doświadczenia, zaistniała konieczność wyboru właściwej metody i zastosowania jej do największej bazy danych utworzonej z informacji pochodzących z innych rynków. Badanie koncentruje się na wpływie bezpieczeństwa miejskiego na ceny nieruchomości. Z jednej strony, jest to podstawowym wyznacznikiem wyboru klientów, a z drugiej strony, informacje o ocenie bezpieczeństwa mogą być wykorzystane w planowaniu przestrzennym i zarządzaniu.

W rezultacie powstał zbiór danych georeferencyjnych zawierający informacje o cechach ponad 191 000 nieruchomości w Denver, Kolorado. Zastosowanie wybranej metody – Ważonej Geograficznie Regresji Hedonicznych – na zbiorze danych wykazało na szereg kwestii związanych z ograniczeniem sprzętowym i oprogramowania, sposobu agregacji danych i obecności współliniowość pomiędzy indeksami. Zastosowanie zasad analiz geograficznego ważenia, jako sposób rozwiązania problemu współliniowości wykazały również zalety, takie jak określenie wpływu różnych wskaźników w mniejszych obszarach miejskich.

Pomimo użycia danych z innych rynków, badania umożliwiły wyciągnięcie istotnych wniosków dotyczących definicji, gromadzenia i opracowania danych niezbędnych do stworzenia odpowiedniej bazy danych do analizy bułgarskiego rynku nieruchomości.

Appendix 1. The list of variables

Variable name	Variable Description	Source
Structural Variables		
LAND_SQFT	Land of property in sq. feet	DODC
AREA_ABG	Area of the building in sq. feet	DODC
BED_RMS	Number of bedrooms	DODC
FULL_B	Number of bathrooms	DODC
CCYRBLT	Year of built	DODC
IMPROVEMENT	Year of improvement	DODC
Quality and Security of Life Issues		
CRIME	Number of crimes within 1 km radius around the property	911 calls, DODC
CRIME_Day	CRIME registered between 6:00 and 21:00	911 calls, DODC
CRIME_Night	CRIME registered between 21:01 and 5:59	911 calls, DODC
GRAFFITI	Number of complains about graffiti within 1 km radius around the property	311 calls, DODC
TRASH	Number of complains about trash issues within 1 km radius around the property	311 calls, DODC
Distances to Amenities or to Security Institutions:		
AFTER_SCHL	Afterschool Denver Program	DODC
ART	Cinema and Theatres	OpenStreetMap
CBD	Central Business District	DODC
COMM_IND	Commercial and Industrial Districts	DODC
DFD	Fire Departments	DODC
DPD	Police Departments	DODC
HISTORY	Historical (tourism) sites	DODC
HEALTH	Clinics and Cabinets	OpenStreetMap
NIGHT_LIFE	Restaurants and Bars	OpenStreetMap
PARKS	Parks and Recreation Facilities	DODC
SCHOOLS	Schools (all grades)	DODC
WAYS	Highways and Boulevards	DODC

Source: own elaborations.