Spatial Analysis of the Flat Market in Prague

Petr Hrobař[1], Vladimír Holý[2]

Abstract. Our study aims to examine the effect of spatial dependency in the Prague flat market. Following the Tobler's first law of geography: near things are more related than distant things, we allow the flat's price to be not only the function of its own characteristics but also function of its neighbourhood unit characteristics. We also evaluate different spatial dependence matrices for analysis of estimated results. The findings are as follows. First, based on the positive parameter of spatial autocorrelation, we confirm spatial dependency in our dataset, i.e. prices of flats tend to form spatial clusters. Second, once controlling for spatial dependency, we are able to evaluate flat's prices more accurately. Third, based on the residual distribution across space, we identify "grandiose" clusters in which the price of estate can be multiple times higher than outside the cluster simply due to the location factor.

Keywords: *Spatial Econometrics, Spatial Autocorrelation, Moran Test, Real Estate*

JEL Classification: C33, C38, R31

1 Introduction

Prices of flats or generally any estates tend to form spatial clusters. As a result of this process, the location characteristics are key determinants to accurate evaluation of estates. When some form of spatial autocorrelation occurs in data generating process (DGP), using ordinary least square regression (OLS) can lead to inaccurate results and estimation bias. Spatial econometric framework allows us to account for various spatial processes and provide much more accurate analysis and estate evaluation.

Even though usual spatial analyses are applied to regional macroeconomic data, e.g., regional unemployment rates, GDPs, etc., applications to non-regional data can also be utilized. Many studies contributing to hedonic price modeling of flats have been conducted, e.g. [1]. To our best knowledge, latest study that contributes to spatial analysis of flat market in Prague, using spatial econometrics framework, is study [2] from 2016. We follow and expand the very study by applying different set of regressors, by using contemporary data and by evaluating spatial stability and estimation robustness for multiple structures of spatial dependency.

Our study is structured as follows: Section 2 describes all models and methods used in this study as well as fundamental literature sources, Section 3 discusses the dataset and the selected methodology and provides an illustrative examples of spatial methods applications. Section 4 shows illustrative evaluation of spatial stability. Lastly, Section 5 concludes our study.

2 Methods for Cross-Sectional Data

In this section we briefly describe non-spatial and spatial models used in our study. We provide fundamental description of methods and used models as well as references to corresponding literature.

2.1 Linear Regression

First model used in our study is simple linear regression given as

$$y = X\beta + \varepsilon, \tag{1}$$

where y is a $(n \times 1)$ vector of dependent variable, X represents a $(n \times k + 1)$ matrix of exogenous regressors, β is a $(k + 1 \times 1)$ vector of regression coefficients to be estimated and ε represents vector of error elements. The most used estimation method is the ordinary least square (OLS) method. For a quick recapitulation of the OLS method as well as all statistical assumptions for errors ε, see eg. [5].

[1] University of Economics, Prague, Faculty of Informatics and Statistics, Department of Econometrics, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic, hrop00@vse.cz

[2] University of Economics, Prague, Faculty of Informatics and Statistics, Department of Econometrics, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic, vladimir.holy@vse.cz

Model 1 does not take spatial dependency into account. In order to partially allow for spatial dependency, we propose simple feature based on statistical learning clustering algorithms. Since we are working with spatial data, we have coordinates available although we do not recommend including coordinates into linear model directly.

However, various types of clustering methods can be applied. We use widely known *k-means* clustering method (see e.g. [3]). Then we include dummy variables representing each cluster into linear regression model. Assuming that all/some dummy variable/s representing cluster are statistically significant, we can conclude that there are some spatial dissimiarities present in the DGP. Since we are using both variants of the linear regression model – with and without dummy cluster variables, we shall be able to evaluate efficiency of both models.

2.2 Spatial Models

As mentioned in [5] the standard approach is to start by using no spatial models at all, then considering various types of spatial interactions. Generally, three main different types of spatial interactions can arise:

1. Dependent variable y of unit i affects/is affected by dependent variable y of unit j.
2. Dependent variable y of unit i affects/is affected by independent variable x of unit j.
3. Error term ϵ_i of unit i affects/is affected by ϵ_j of unit j.

Using a model to cover all interaction effects above, we get model that takes form

$$\begin{aligned} y &= \rho W y + X\beta + WX\theta + u, \\ u &= \lambda W u + \epsilon, \end{aligned} \tag{2}$$

where y is $(n\times 1)$ vector of dependent variable, W is a $(n\times n)$ weight matrix that represent a spatial dependency among observed units, X is a $(n \times k+1)$ matrix of independent variables, β is a $(k+1\times 1)$ vector of regression coefficient to be estimated and ρ, θ, λ are spatial autocorrelation parametrs to be estimated. Wy denotes the endogenous interactions among dependent variable. Similarly, WX represents interaction effect among independent variables. Lastly, Wu stands for interaction effect among error terms.

For the purpose of our analysis we shall not use entire model as described above but rather use two derived specifications described down below.

Spatial Lag Model

Spatial lag model can be obtained from 2 when θ and λ are both equal to 0. Under this assumption derived model allows for accounting for spatial dependency in dependent variable of neighbourhood units. Therefore, model can be written as

$$y = \rho W y + X\beta + \varepsilon. \tag{3}$$

Under the assumption of existing inverse of $(I_N - \rho W)$, we can obtain a reduced form given as

$$y = (I_N - \rho W)^{-1}(X\beta + \varepsilon), \tag{4}$$

where 4 is usually described as a DGP equation for y as emphasised by [4]. To obtain regression parameters β and ρ, the maximum likelihood approach is commonly used. Paper [4] offers an overview of spatial econometric models as well as detailed maximum likelihood (ML) estimation for model 3 and 5.

Spatial Error model

Second spatial model used in our study is spatial error model. Given the model 2 we can obtain spatial error model form when ρ and θ are both equal to 0. Hence, using this model specification we can account for spatial interactions among the error terms. Formal model form can be described as

$$\begin{aligned} y &= X\beta + u, \\ u &= \lambda W u + \varepsilon. \end{aligned} \tag{5}$$

Selection for model 5 can imply that some (spatially distributed) independent variable was not included in the model. Once again, parameters β, λ are estimated by the ML, see e.g. [4]. Unlike the spatial lag model, coefficients from spatial error model can be directly interpreted as marginal effects.

Table 1 Descriptive statistics.

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Price	2,984	8,495,653.000	5,696,973.000	80,000	4,990,000	9,990,000	79,000,000
Rooms	2,984	2.804	1.128	1	2	4	6
Meters	2,984	79.634	40.968	15	52	96	435
KK	2,984	0.792	0.406	0	1	1	1
Maisonette	2,984	0.027	0.163	0	0	0	1
Panel	2,984	0.132	0.338	0	0	0	1
Balcony/Terrace	2,984	0.514	0.500	0	0	1	1
New Estate	2,984	0.291	0.454	0	0	1	1

2.3 Moran *I* test

Most importantly, the use of spatial model should always be confirmed, i.e. formal statistical test should be implemented. In Section 2.1 we propose simple clustering approach to (moderately) account for spatial dependency. However, more sophisticated tests are at disposal. The most common test is the Moran I test, see e.g. [4].

The formal testing statistic takes form given by

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \times \frac{\varepsilon^{'} W \varepsilon}{\varepsilon^{'} \varepsilon}, \tag{6}$$

where N is the number of (spatial) observations, w_{ij} are elements of spatial weight matrix W and ε is a $(n \times 1)$ vector of residuals obtained from the OLS model. Generally, the Moran I statistic value is between $[-1; 1]$. However, values outside the interval can also arise. Under the null of spatial independence, statistic take form as

$$E(I) = \frac{-1}{N-1}. \tag{7}$$

Not only is the Moran the most broadly used test for spatial dependency but also simulations showed that moran test is the most accurate test for spatial dependency [5].

Spatial dependency structure

Since W matrix is needed to perform tests for spatial dependency, it is necessary to correctly specify spatial W matrix and spatial dependency structure in general. Main obstacle in spatial framework is the fact that spatial dependency structure is not estimated but predefined and thus one needs to evaluate spatial dependency thoroughly.

Spatial Weight matrix W is derived from neighbor structure. As far as neighborhood dependency goes multiple approaches are common. Since we are not working with data on regional level we shall not consider methods based on polygons such as *queen and rook* methods.

Since we are evaluating flat estates we assume that estates that are more distant tend to effect prices slightly, but on the other hand estates closer to each other tend to influence one another more significantly. Thus using neighborhood dependency based on the number of neighborhoods units seems as a good starting point. Another approach considered in our study is method based on the maximum distance of units. However, this approach would be very computationally demanding and time consuming. Therefore we create neighborhood dependency based on the number of nearest neighbor units. This dependency is then transformed in W matrix in order to perform Moran test and estimate spatial models described in Section 2.2.

3 Dataset and Results

Flats estates are retrived from Czech estates site Sreality.cz which contains various estates that are available to rent or buy. We suppose that Prague flats listed here are credible representation of the real flats market of the city, i.e. the listed estates follow the same DGP as the estates not listed and thus not included in the dataset.

In our study, more than 4 000 flats profiles were collected data were collected over a one month period starting on March 20, 2020. The data collection process was done in PYTHON programming language using webscrapping approach. After the data collection, a filtering processes were implemented to obtain credible dataset. Data are cross-sectional hence no individual effects nor time effects need to taken into account. First step was to select

Table 2 Estimated models.

	Dependent variable:			
	log(price)			
		kmeans	Spatial lag	Spatial Error
	(1)	(2)	(3)	(4)
Rooms	0.061***	0.077***	0.084***	0.085***
	(0.010)	(0.009)	(0.008)	(0.007)
log(Meters)	0.808***	0.749***	0.685***	0.691***
	(0.021)	(0.019)	(0.017)	(0.017)
Maisonette	−0.003	−0.016	−0.062**	−0.065***
	(0.031)	(0.028)	(0.025)	(0.023)
KK	0.117***	0.155***	0.171***	0.178***
	(0.016)	(0.014)	(0.012)	(0.012)
Panel	−0.324***	−0.255***	−0.127***	−0.118***
	(0.016)	(0.015)	(0.013)	(0.014)
Balcony/Terrace	−0.007	0.039***	0.055***	0.075***
	(0.011)	(0.010)	(0.009)	(0.009)
New Estate	−0.011	0.016	0.060***	0.077***
	(0.011)	(0.011)	(0.009)	(0.010)
$Kmean_2$		0.344***		
		(0.019)		
Constant	12.145***	12.097***	2.432***	12.412***
	(0.065)	(0.062)	(0.223)	(0.075)
ρ			0.636***	
λ				0.933***
Observations	2,984	2,984	2,984	2,984

Note: *p<0.1; **p<0.05; ***p<0.01

Number of cluster (5) selected based on AIC.

We do not show additional 3 dummy coefficients for cluster.

relevant variables for analysis. Thus, variables preserved for analysis are as follows: *Price* – price of estate, *Rooms* – number of rooms of flat, *Meters* – number of square meters of flat, *Maisonette* – an apartment occupying multiple floors *KK* – kitchenette. Also variables describing estate building type were retained, i.e. *Panel* - if an apartment is in a high-rise building, *Balcony/Terrace* - if an estate has balcony or terrace, and lastly, *New Estate* - if a flat is new-build property.

Additionally, since we are interested in spatial modeling, *coordinates*, i.e., *longitude and latitude*, were also kept. Last steps taken in data cleaning process were to delete all observations that do not have information about variables described above available and therefore were removed. Eventually, we have dataset that has 2 984 observations which should allow us for proper estimation and authentic results. Descriptive statistics are in Table 1.

Lastly, we investigate correlation between all independent variables. Particularly, we were concerned about correlation between *Meters* and *Rooms*. However, this correlations turns out not to be excessively high and we shall keep both variables in all models. We also considered whether variable *Rooms* should be dummy or not. We decided not to use dummy since we would be adding another five variables in model. However, we are fully aware that since variable *Rooms* is discrete we should not interpret associated coefficient, nevertheless coefficient sign and statistical significance can be interpreted. The final model is selected as

$$\begin{aligned}\ln(Price) = \beta_0 + \beta_1 Rooms + \beta_2 \ln(Meters) + \beta_3 Maisonette + \beta_4 KK + \beta_5 Panel \\ + \beta_6 Balcony/Terrace + \beta_7 NewEstate + \varepsilon.\end{aligned} \tag{8}$$

After the OLS model estimation we calculated Moran I statistic to identify whether spatial autocorrelation is present in flats prices. The results of Moran test are as follows: *I-statistic = 0.43, p-value = 0.000.*

Based on the significant results given by Moran test we shall apply spatial models. Interestingly enough, the result also shows that flats prices are not randomly distributed in space, i.e., prices of flats tend to form spatial clusters. Estimations of all presented models are summarized in Table 2. Model comparison metrics are in Table 3.

Model selection is be done using the metrics in Table 3. We also include pseudo R^2 calculated as $corr(y, \hat{y})^2$. Firstly, when comparing baseline OLS model and OLS model with clustering approach we can conclude that once

Table 3 Models metrics.

	OLS	*OLS Kmeans*	*Spatial Error*	*Spatial Lag*
AIC	583.827	75.441	-931.300	-688.538
Log-like.	-282.913	-24.720	475.650	354.269
R^2	0.748	0.788	0.854	0.837

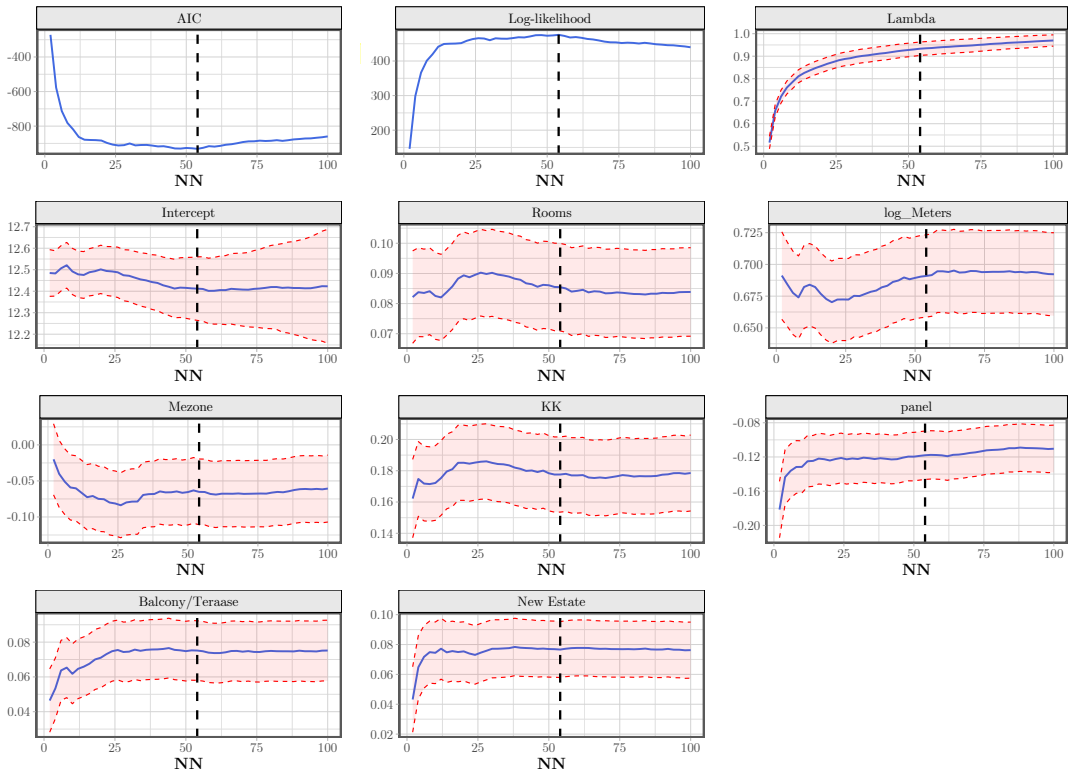

Figure 1 Estimated model parameters and the *AIC* statistics for different numbers of neighbor units.

(partially) controlling for spatial heterogeneity estimated results, tend to be more steady. Additionally, model with clustering still allows for nice interpretability. For example, we can identify cluster in which prices of states are (on average) 34% higher (supposedly) due to unobserved location factor.

However, spatial models particularly spatial error model seems to be the best model for DGP rendering. Therefore, spatial error model shall be used for statistical inference. Lastly, we tested error model residuals for homoscedasticity using the modified Breusch-Pagan test.

4 Model Specification Robustness

An important step is to evaluate model stability for various types of spatial dependency. Study describing importance of such step is e.g. [6]. Also, since we are using neighborhood dependency based on the number of nearest neighbor units we do not know the optimal number that should be used. Therefore, evaluation of spatial stability using information criteria should serve as a fair indicator. Results of this "spatial cross validation" are in Figure 1.

Spatial dependency structure and derived spatial weight matrix W that minimizes *AIC* is for 54 nearest neighbors. Models in Table 2 are using such spatial dependency structure. Lastly, we take OLS model residuals and investigate their distribution across space. Since we are using the logarithmic transformation we need to use correct transformation of residuals rather than using the simple exponential transformation. After this step we discretize residual values and inspect residuals distribution in Figure 2.

Clearly, we can identify cluster with high prices in the historical part of Prague. Those findings are in compliance with [2], where after applications of different sets of variables we still see significant clustering. This allows us to

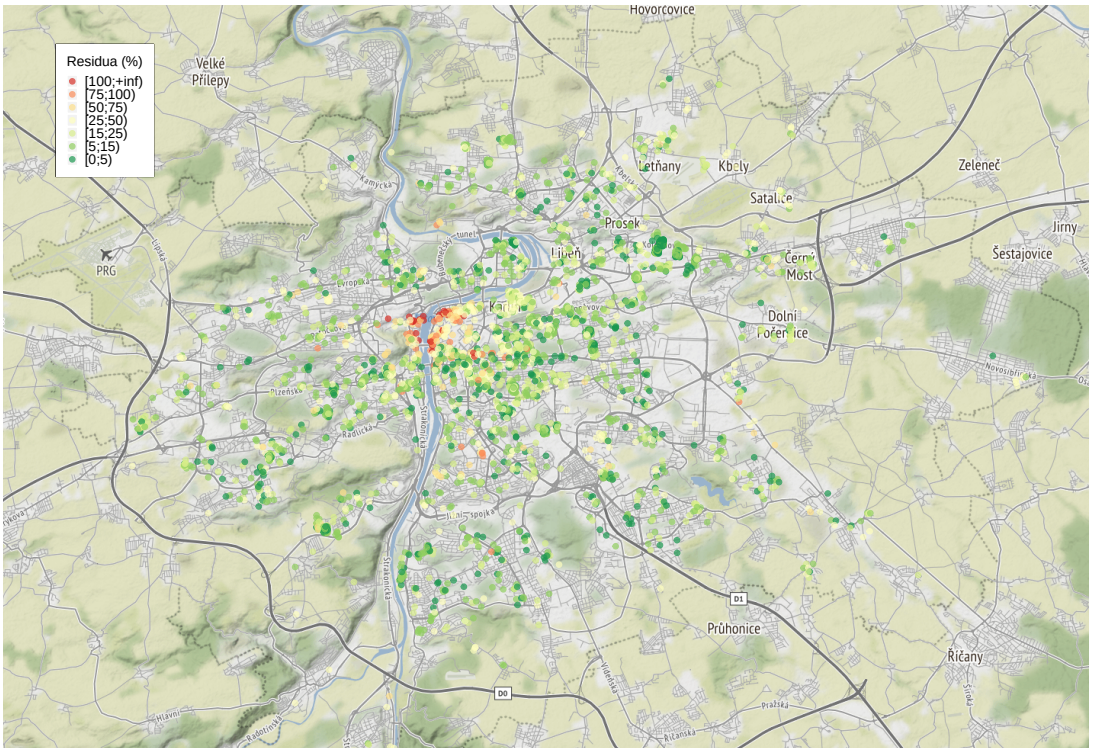

Figure 2 Residual distribution in space.

believe that residual noise is (presumably) entirely due to the factor of location.

5 Conclusions

Spatial econometrics framework was utilized to investigate spatial dependency of flat market. To account for spatial heterogeneity we used simple clustering approach. After confirming spatial dependency using moran *I* test, we used two straightforward models, i.e., *spatial lag* and *spatial error*. Based on the models fit metrics we found that spatial models are more suitable for flats analysis. Lastly, we investigate residuals from OLS model which do no account for spatial dependency and investigate their distribution in space. Using this approach we are able to identify "grandiose" clusters witch tends to increase estate price simply because of location factor.

Acknowledgements

The work on this paper was supported by the grant No. F4/27/2020 of the Internal Grant Agency of University of Economics, Prague.

References

[1] Dubin, Robin, A *Spatial autocorrelation and neighborhood quality*. Regional science and urban economics (1992).

[2] Lipán, M. *Spatial approaches to hedonic modelling of housing market: Prague case.* Bachelor's thesis, Charles University, Faculty of Social Sciences, Institute of Economic Studies (2016).

[3] Hastie, T., Tibshirani, R., & Friedman, J. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media. (2009).

[4] J. P. LeSage and R. K. Pace. *Introduction to spatial econometrics*. Boca Raton: CRC Press, Taylor Francis Group, 2009

[5] Anselin, Luc. *Spatial econometrics: methods and models. Vol. 4.* Springer Science & Business Media, 2013.

[6] Formánek, T., & R. Hušek. *On the stability of spatial econometric models: Application to the Czech Republic and its neighbors.* Mathematical Methods in Economics (2016): 213-218.