Interpolation Methods for Geographical Data: Housing and Commercial Establishment Markets

Authors Jose M. Montero and Beatriz Larraz

Abstract

The estimation of commercial property prices in a touristic city can be explored through spatial interpolation methods, but in the presence of small sample sizes, auxiliary stochastic processes that are correlated with the prices of commercial establishments are needed. The aim of this paper is to compare the various estimates of commercial establishment prices in Toledo (Spain) provided by methods based on inverse distance weighting, 2-D shape functions for triangles, kriging, and cokriging (the housing prices being the auxiliary stochastic process). The results indicate that kriging improves the classical interpolation methods and that cokriging has a clear advantage over kriging.

The literature on estimating real estate prices offers no shortage of well-known interpolation methods or algorithms for the relevant geographical data. Trend surfaces, Fourier series, splines, inverse distance weighting, and kriging or cokriging are all commonly employed. Shape functions and n-D Delaunay tessellation can also be included as spatial interpolation methods (and are both popular in engineering applications).[1] Among these, only cokriging methods (Chica-Olmo, 2007; Montero, Larraz, and Paez, 2009) are able to take advantage of the information provided by an auxiliary variable, such as other property prices correlated with the price of the main real estate variable. Cokriging also overcomes the important obstacle of dealing with a small sample of data.

In the real estate literature, research on the commercial property market (in addition to the traditional emphasis on housing prices) is becoming increasingly necessary (Geltner and Ling, 2006), particularly given the current economic crisis (Hardin, Johnson, and Wu, 2009 or Des Rosiers, Thériault, and Lavoie, 2009, among others). Following Montero, Larraz, and Paez (2009), four points—the importance of commercial establishments, spatial autocorrelation, a general lack of available information, and the correlation between two variables—have provided a foundation for proposing cokriging as a methodology for estimating the values of commercial establishments when sample sizes are small. Cokriging allows for the estimation of a variable of interest at a specific location based on

data about the variable itself and about auxiliary variables in the neighborhood. This ability is an important advantage when the sample size for the variable of interest is small and other correlated auxiliary variables are available that provide abundant and reliable information on the former variable.

The aim of this paper is to contrast results obtained by classical interpolation methods (i.e., those that do not take into account spatial correlation), such as 2-D shape functions for triangles and methods based on inverse distance weighting with ordinary kriging and ordinary cokriging strategies, both of which crucially rely on the existing spatial dependence between observations. In other words, we want to remark on the importance of spatial dependencies in prediction tasks, as was suggested by Osland, Thorsen, and Gitlesen (2007). Our research also seeks to evaluate the importance of the use of an auxiliary process (housing prices) that is correlated with the principal one to improve the accuracy of the univariate estimates of commercial establishment prices. To summarize, this study explores the importance of taking into account the existing spatial dependencies in the prices of commercial establishments for prediction purposes. However, because information about the prices of commercial establishments has historically been relatively scarce, it is also extremely useful to take advantage of the abundant and reliable information about the prices of other properties (namely, houses), the prices of which are closely related to the prices of commercial establishments.

The paper proceeds as follows. Section 2 is devoted to methodological questions. Section 3 focuses on the study site, the database, and the variogram modeling. In Section 4, ordinary cokriging estimates are compared with the univariate estimates. Finally, the paper ends with some concluding remarks.

Statistical Methodology

Inverse distance weighting (IDW) methods and 2-D shape functions for triangles are simple interpolation methods with a weighting mechanism assigning more influence to the data points near the location where the estimation is being carried out. In other words, each measured point (or known value) has a local influence that diminishes with distance. Ordinary kriging (OK) shares the weighting mechanism of these methods, but it also takes into account the structure of the existing spatial correlation among the prices of commercial establishments (weights obtained through the kriging system using variograms are substituted for those based on an inverse distance approach). Ordinary cokriging (OCK) goes further still and incorporates information provided by other random functions correlated with the main one. Therefore, OK is a specific instance of OCK when interpolation is only based on one random (main) function (in our case, the price of commercial establishments). In other words, OCK reduces to OK when all OCK weights are zero except for the variable of interest.

Specifically, the weighting mechanism of IDW-based methods can be expressed (Johnston, Ver Hoef, Krivoruchko, and Lucas, 2001) as:

$$\lambda_i = \frac{\left(\frac{1}{d_i}\right)^p}{\sum_{j=1}^{N}\left(\frac{1}{d_j}\right)^p};\ i = 1, \ldots, N, \tag{1}$$

where all λ_i are the weights assigned to the sampled locations; d_i are the Euclidean distances between the sampled locations and the point where we are predicting the value; N is the sample size (in our case, there is a unique neighborhood); and p is an exponent that influences the weighting (in our case, $p = 1$ and $p = 2$). Estimates are obtained as a weighted average of the sampled commercial establishment prices, the weights being λ_i.

In the 2-D shape functions, on the other hand, the triangular meshes are made by dividing the total domain into a finite number of triangular domains, and the construction of the triangles uses ''Delaunay triangulation,'' (Li and Revesz, 2004). The function values inside the domain are obtained by a weighted average of the three values at the apices of the triangle, the weights being:

$$\lambda_i = \frac{A_i}{A};\ i = 1,2,3, \tag{2}$$

where A_i are the subtriangle areas and A is the area of the outside triangle.

Finally, in the multivariate case (OCK methodology) our framework will follow the joint intrinsic hypothesis, considering that the partial heterotopy case (some variables share some sampled locations) characterizes current real estate markets. A thorough presentation of the theory of multivariate geostatistics is available in textbooks (see Wackernagel, 2003). Consider $\mathbf{X} = (X_1, X_2, \ldots, X_m)^t$, a vector of intrinsic random functions. Under these conditions, the OCK estimator of a particular variable X_i at the point $\mathbf{s}_0$ is a weighted linear combination of the data values from the variables $X_j (j = 1, \ldots, m)$ located at sampled points in the neighborhood of $\mathbf{s}_0$:

$$X_i^*(\mathbf{s}_0) = \sum_{j=1}^{m}\sum_{\alpha=1}^{n_j} \lambda_\alpha^j X_j(\mathbf{s}_\alpha^j), \tag{3}$$

with $\{\mathbf{s}_\alpha^j,\ \alpha = 1, \ldots, n_j\}$ being the set of locations where variable X_j, for $j = 1, \ldots, m$, has been sampled.

The weights λ_α^j, $\alpha = 1,..., n_j$, and $j = 1,..., m$ are calculated to ensure that the estimator is optimal in the sense that it is unbiased with minimum error-variance. Choosing weights that sum up to one for the variable of interest and have zero sums for auxiliary variables guarantees against bias with a supposed constant mean (Montero, Larraz, and Paez, 2009). Then, the OCK system (4) is obtained by minimizing that variance with the constraint on weights:

$$\begin{cases} \sum_{k=1}^{m} \sum_{\beta=1}^{n_k} \lambda_\beta^k \gamma_{jk}(\mathbf{s}_\alpha^j - \mathbf{s}_\beta^k) + \omega_j = \gamma_{ji}(\mathbf{s}_\alpha^j - \mathbf{s}_0) \\ \quad \forall j = 1,..., m; \forall \alpha = 1,..., n_j \\ \sum_{\alpha=1}^{n_j} \lambda_\alpha^j = \delta_{ij} = \begin{cases} 1 & si\ i = j \\ 0 & si\ i \neq j \end{cases} \end{cases} \tag{4}$$

Cross- and direct variograms, γ_{ij} and γ_{ii}, respectively, are obtained in two steps. First, point estimates of the variograms are obtained using the classical variogram estimator based on the method-of-moments. The second step is to fit a theoretical variogram function to the sequence of average dissimilarities according to the linear model of coregionalization (this being the typical strategy to ensure a positive definite model).

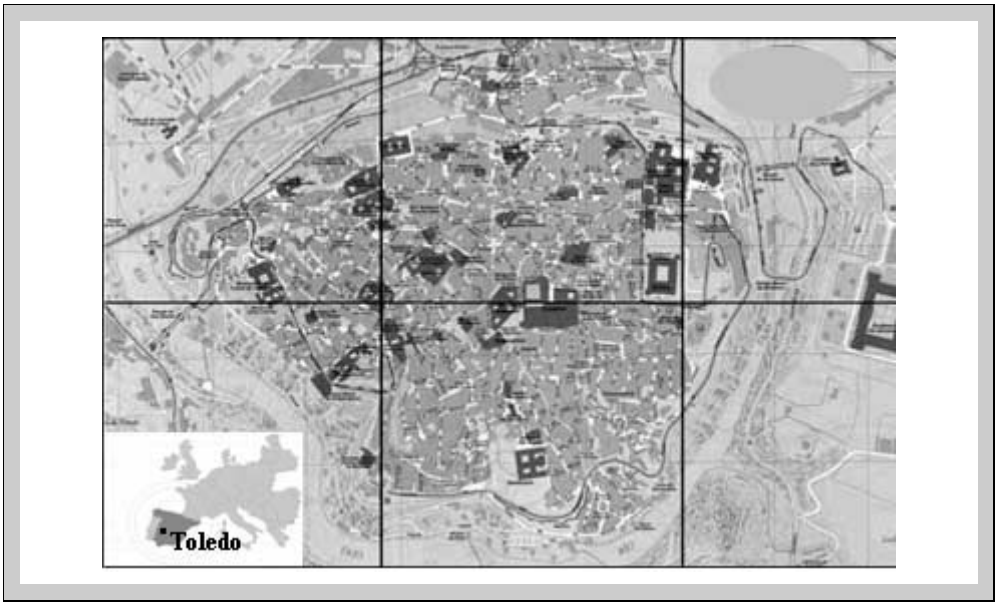
Case Study: Estimation of Commercial Establishment Prices in the Historic Area of Toledo

There are several reasons for having chosen this emblematic area: (1) it is known worldwide as a UNESCO World Heritage City; (2) it is an excellent area for exploring the commercial real estate market due to its popularity as a tourist destination; and (3) it is marred by neither geographical accidents nor artificial barriers inside the walls that could break down the spatial dependence structure. The study area and its position in Spain are depicted in Exhibit 1.

Our database contains information about 223 houses and 123 commercial establishments located in the historic part of Toledo. The information was provided by the real estate agencies that operate in this area, and the data refer to the market price, age, location, condition, and surface of the properties. In the case of the commercial establishments, we also know whether they have a basement,[2] and in the case of houses, we know if they have parking space.

Obviously, the age of a property usually has an important influence on its price, but in a historical part of a city like Toledo, the influence of this factor vanishes. Thus, age has not been considered in the analysis. Moreover, we have detected some deficiencies in the measurement of the surface and have consequently decided to consider it as a categorical variable. By reducing square meters to a

Exhibit 1 | Historic Toledo



categorical variable, it is only possible to adjust price for the average size within the size category, but this does not significantly affect the final results. Of course, there are also additional explanatory variables, but unfortunately, they are not provided by the real estate agencies for research purposes. The data correspond to commercial establishments and houses sold in the third quarter of 2007.

In order to isolate the spatial component of commercial establishment and home prices (the analysis being conducted in terms of price per square meter), we have proceeded to adjust for the housing and commercial establishment mix using the traditional hedonic model (Goodman, 1978). To illustrate the results of the homogenization procedure, the estimated coefficients of the hedonic model for the city of Toledo are depicted in Exhibit 2. The weighted least squares method is used, and estimated parameters are used to transform the original prices into other prices corresponding to a house or commercial property with the reference class's characteristics. From this point forward, the commercial establishment and home prices are quality-constant, as in Fotheringham, Brunsdon, and Charlton (2002). That is to say, the prices of houses only differ because of their spatial location, not because of their individual characteristics.

Next, we proceeded to represent the spatial dependence in both cases by the appropriate theoretical variogram model and to account for cross-dependence between the two processes. The values of the parameters are reported in Exhibit 3.

Exhibit 2 | Parameter Estimates of the Hedonic Model for the City of Toledo

	Variable	Estimate	*t*-Value
Houses[a]	Intercept	1,982.34	1,024.46*
Condition	New or completely renovated	1.1346	1.54
	Good condition	Reference class	
	Little renovation needed	0.8833	−1.73
	Complete renovation needed	0.7550	−8.85*
Surface	Less than 65 m^2	1.2245	11.61*
	Between 65 and 120 m^2	Reference class	
	More than 120 m^2	0.8847	−7.72*
Parking	Yes	1.1469	33.34*
	No	Reference class	
Commercial Properties[b]	Intercept	1,894.29	659.68*
Condition	Ready for business	Reference class	
	Some renovation needed	0.9230	−0.98
	Complete renovation needed	0.7614	−13.45*
	Unfinished	0.6717	−24.91*
Surface	Less than 50 m^2	1.3828	7.35*
	From 50 to 100 m^2	Reference class	
	From 100 to 200 m^2	0.9453	−1.90
	More than 200 m^2	0.7101	−3.84*
Basement	Yes	1.2611	27.80*
	No	Reference class	
R^2		0.64	
F-statistic		328.71	$P < 2.2e\text{-}16$

Notes:
[a]Dependent variable: housing price per square meter in euros/m^2.
[b]Dependent variable: commercial property price per square meter in euros/m^2.
*Significant at the 5% level.

Exhibit 4 reports the experimental direct and cross-variograms (drawn as a thin line) and the theoretical models that have been fitted to the experimental ones (drawn as a thick line). As expected, the existing spatial dependencies cannot be modeled using only one of the valid variograms provided by the literature. This is why we have determined the best linear combination of variograms that best fits the aforementioned dependencies. The optimal combination includes a spherical variogram, a pure nugget variogram (or nugget effect that captures a discontinuity at the origin) and a Gaussian model.

The proposed variograms have been checked for validity using the cross-validation or ''leave-one-out'' procedure. The neighborhood was a moving one with a radius

Exhibit 3 | Linear Model of Coregionalization

Model	Sill[a]		
	Commercial Establishment Prices Direct Variogram	House Prices Direct Variogram	Commercial Establishment Prices: House Prices Cross-variogram
Spherical (330 m. range[b])	340,978.332	142,783.006	70,505.189
Nugget effect	1	8,000	−85
Gaussian (165 m. range[b])	200,000	10,000	30,000

Note: Created by the authors using ISATIS software (ISATIS, 2001). Prices of houses and commercial properties are expressed in euros per square meter.
[a]Sill is the variance of the process.
[b]Range is the distance at which the correlation becomes zero.

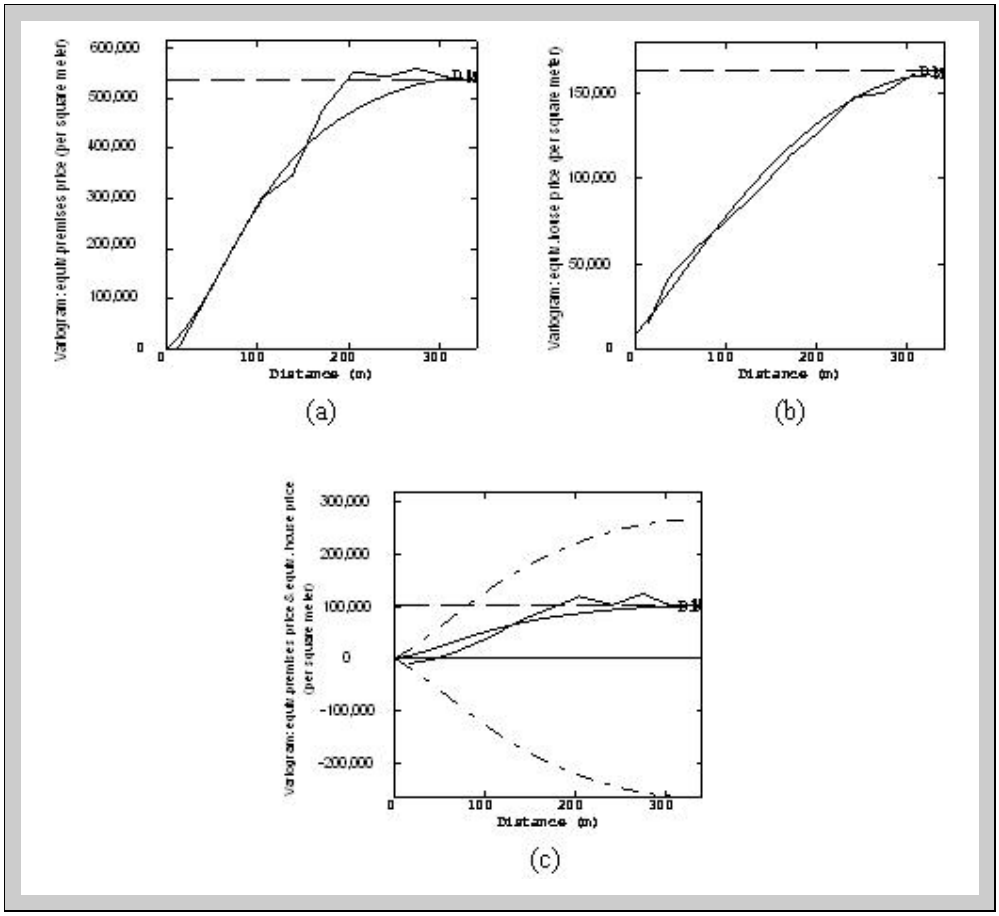
of 132 meters. Specifically, models from Exhibit 3 provide 119 robust estimates when estimating commercial establishment prices (96.7% from a total of 123) and 214 in the case of the house prices (96% from a total of 223), an estimate being robust when its standardized value belongs to the interval [−2.5; 2.5] (Emery, 2000, 117–18). This percentage of robust estimates (greater than 95%) leads us to consider the models from Exhibit 3 valid for cokriging estimation. Exhibits 5 and 6 report the cokriged estimation map and the standard deviation error map, respectively. As drawn in Exhibit 5, commercial property prices in light areas are around 4,000 euros per square meter, and as a rule, the darker the color, the lower the price. Exhibit 6 shows how black dots match up with sample locations due to the exact interpolation kriging characteristic and the way in which the standard deviation becomes higher far from locations where prices are known.

Results

We have estimated the price of commercial establishments in the old part of Toledo through OCK and OK, as well as by other classical univariate interpolation methods (shape functions and methods based on IDW) that do not take into account the spatial dependence among the prices of urban properties. OCK and OK estimates were obtained using ISATIS, a spatial statistical program. The aim is to compare the accuracy of these various methods.

The criteria for the comparison are the interpolation accuracies when carrying out a cross-validation procedure and when using a one-in-ten hold-out sample as a test set. In the first case, we averaged 123 errors, while only an average of 12 errors was made in the second case. Mean error (ME), mean absolute error

Exhibit 4 | Experimental and Fitted Variograms



Note: (a) Equivalent commercial establishment prices (per square meter) direct variogram, (b) Equivalent house prices (per square meter) direct variogram, (c) Equivalent commercial establishment prices (per square meter)-house prices (per square meter) cross-variogram.

(MAE), mean square error (MSE), and error variance (VarEs) were calculated to evaluate the forecasting accuracy of the models. Additionally, OCK and OK standardized errors $((x^*(s_0) - x^*(s_0))/\sigma^*(s_0))$ were determined, as well as their mean (MStE), mean square (MSStE), and variance (VarStE). The results are reported in Exhibit 7.

The procedures using variograms to represent the structure of the spatial dependence among the prices of properties lead to a lower mean and variance of the estimation errors (Exhibit 7). In particular, the models with a lower MAE and MSE are considered to be relatively superior models (OK and OCK). As expected, when comparing OK against OCK results, the latter outperforms the former (the

Exhibit 5 | Cokriged Estimation Map: Commercial Property Prices

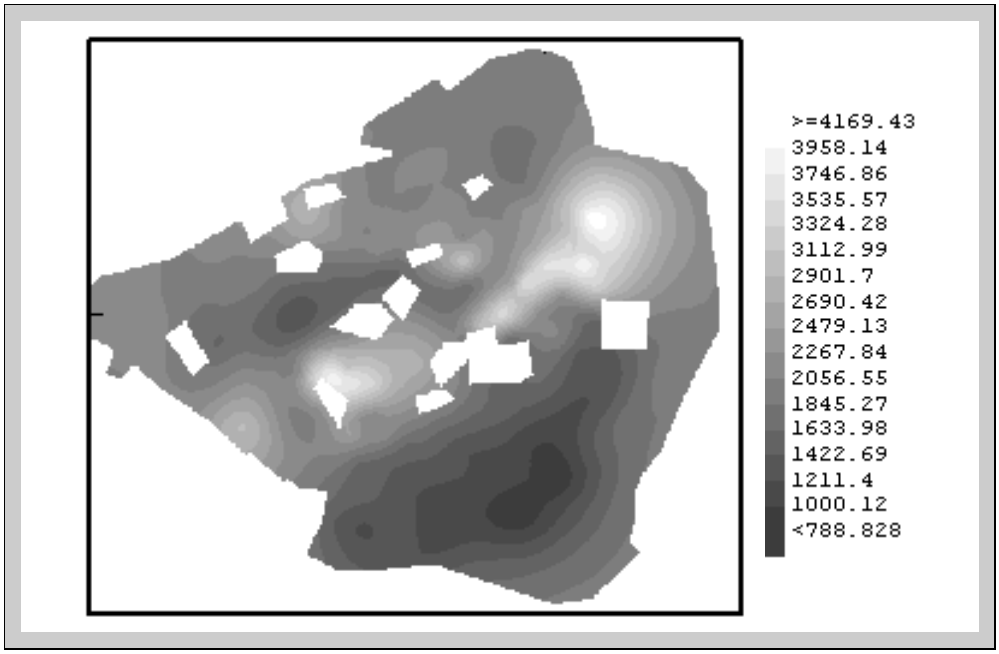


Exhibit 6 | Cokriged Standard Deviation Error Map: Commercial Property Prices

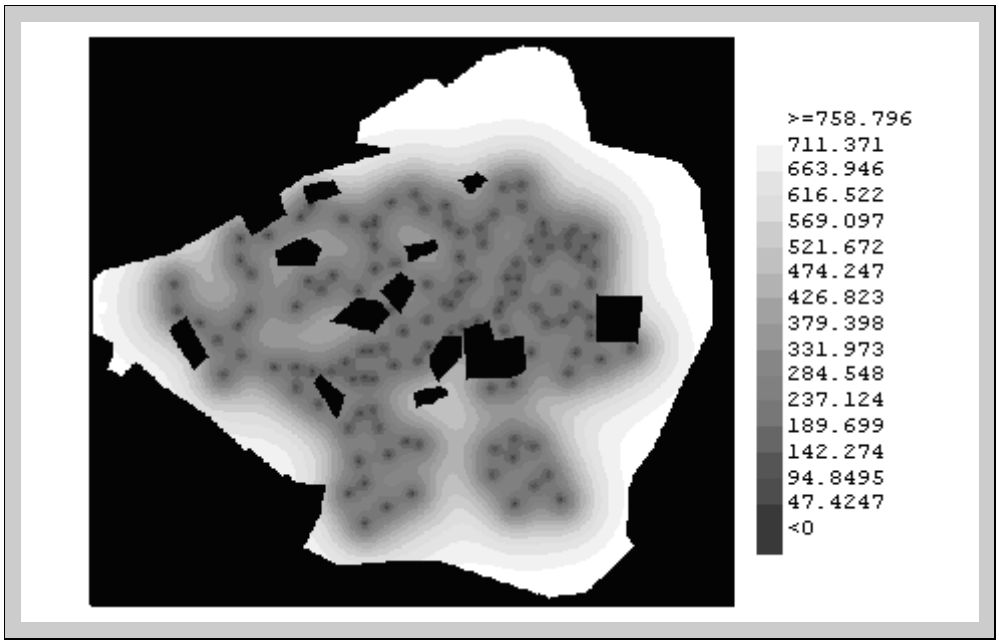


Exhibit 7 | Cross-validation and Hold-out Sample Results

	Interpolation Method				
	Inverse Distances ($p = 1$)	Squared Inverse Distances ($p = 2$)	2-D Shape Functions	Kriging	Cokriging
Cross-validation Error					
ME	−30.390	−18.240	−2.390	−1.672	1.501
MAE	295.418	280.560	250.762	243.319	222.183
MSE	107,235.970	98,111.718	91,700.056	91,189.029	80,623.699
VarE	106,312.420	97,779.020	91,694.344	91,186.233	80,621.447
MStE				−0.015	0.004
MSStE	—	—	—	1.097	0.922
VarStE	—	—	—	1.097	0.922
Hold-out Sample Error					
ME	−40.220	−23.150	−2.543	−1.928	1.597
MAE	335.239	291.334	264.739	248.771	228.987
MSE	119,165.660	99,768.549	92,350.154	91,859.552	82,495.364
VarE	117,548.010	99,232.627	92,343.687	91,855.835	82,492.814
MStE				−0.017	0.004
MSStE	—	—	—	1.101	0.934
VarStE	—	—	—	1.101	0.934

Note: Authors' own elaboration using R (R Development Core Team, 2008).

correlation coefficient between commercial establishment and house prices, computed with the 65 pairs of prices corresponding to the isotopic case, is $\rho = 0.696$). Focusing on cross-validation results, the OCK procedure has several advantages over OK: (1) the mean estimation error decreases by 10.2% (from −1.672 to 1.501) when compared to the OK result; (2) the mean error, in standardized terms, decreases by 70% (from −0.015 to 0.004); (3) the variance of the estimation errors decreases by 11.58% (from 91,186.233 to 80,621.447); and (4) both the mean square standardized error and the standardized error variance drop by 15.95% (from 1.097 to 0.922). Despite the fact that the hold-out sample consists of only 12 data points, the results for the test set are very similar.

Conclusion

In this paper, we show the importance of considering the structure of the spatial dependence among the prices of properties when estimating them. Furthermore, the existing correlation between the prices of different types of properties (in our case, houses and commercial establishments) is used to obtain more accurate

estimates, as available information about commercial establishment prices is usually more sparse than information about house prices.

We evaluate four geostatistical interpolation methods in terms of their ability to estimate the commercial establishment prices in the historic area of Toledo (Spain). First, we use three univariate interpolation methods (OK, IDW, and 2-D shape functions) to show the advantages of the OK procedure, which uses a variogram to represent the structure of the spatial correlation among the prices of properties (in our case, commercial establishments). Secondly, we consider the OCK methodology as an improvement to the OK estimates, which adds an auxiliary stochastic process corresponding to the house prices in the study area.

As expected, and in accordance with the theoretical specialized literature in geostatistics, the results show that OK improves the classical interpolation methods that do not take account the spatial correlation among the property prices (IDW and 2-D shape functions) and that OCK has a clear advantage over OK, which only uses the data of the target variable. Therefore, the results indicate that the use of an auxiliary variable improves the OK estimates, and this capability is crucial when the information on the main variable is lacking. This is precisely the case when estimating commercial establishment prices, as information on them is usually scarce.

Endnotes

[1] In addition to interpolation methods, several other estimation techniques are widely used in the real estate literature, such as hedonic models, pioneered by Rosen (1974) and recently updated by Malpezzi (2003), who made a selective revision of the hedonic models applied to the real estate valuation; neural networks (Worzala, Lenk, and Silva, 1995) and spatial econometrics (Osland, 2010). More recently, replication methods for estimating property values have been developed by Lai, Vandell, Wang, and Welke (2010).

[2] The basement being the lowest level of a structure partly or wholly below ground level often used for storage.

References

Chica-Olmo, J. Prediction of Housing Location Price by a Multivariate Spatial Method: Cokriging. *Journal of Real Estate Research*, 2007, 29:1, 91–114.

Des Rosiers, F., M. Thériault, and C. Lavoie. Retail Concentration and Shopping Center Rents—A Comparison of Two Cities. *Journal of Real Estate Research*, 2009, 31:2, 165–207.

Emery, X. Geoestadística Lineal. Universidad de Chile, 2000.

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

Geltner, D. and D.C. Ling. Considerations in the Design and Construction of Investment Real Estate Research Indices. *Journal of Real Estate Research*, 2006, 28:4, 411–44.

Goodman, A.C. Hedonic Prices, Price Indices and Housing Markets. *Journal of Housing Research*, 1978, 3, 25–42.

Hardin, W.G., K.H. Johnson, and Z. Wu. Brokerage Intermediation in the Commercial Property Market. *Journal of Real Estate Research*, 2009, 31:4, 397–420.

ISATIS. Isatis Software Manual Avon. v4.1.1. Geovariances and Ecole des Mines, 2001.

Johnston, K., J.M. Ver Hoef, K. Krivoruchko, and N. Lucas. *Using ArcGIS Geostatistical Analyst*. ESRI Press, 2001.

Lai, T-Y., K. Vandell, K. Wang, and G. Welke. Estimating Property Values by Replication. An Alternative to the Traditional Grid and Regression Methods. *Journal of Real Estate Research*, 2008, 30:4, 441–60.

Li, L. and P. Revesz. Interpolation Methods for Spatio-temporal Geographic Data. *Computational, Environment and Urban Systems*, 2004, 28:3, 201–27.

Malpezzi, S. Hedonic Pricing Models: A Selective and Applied Review. In: A. O'Sullivan and K. Gibb (eds.), *Housing Economics and Public Policy*. Blackwell, Oxford, 2003.

Montero, J.M., B. Larraz, and A. Paez. Estimating Commercial Property Prices: An Application of Cokriging with Housing Prices as Ancillary Information. *Journal of Geographical Systems*, 2009, 11:4, 407–25.

Osland, L. An Application of Spatial Econometrics in Relation to Hedonic House Price Modeling. *Journal of Real Estate Research*, 2010, 32:3, 289–320.

Osland, L., I. Thorsen, and J.P. Gitlesen. Housing Price Gradients in a Region with One Dominating Center. *Journal of Real Estate Research*, 2007, 29:3, 321–46.

R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org, 2008.

Rosen, S. Hedonic Prices and Implicit Markets. Product Differentiation in Pure Competition. *Journal of Political Economy*, 1974, 82, 34–55.

Wackernagel, H. *Multivariate Geostatistics: An Introduction with Applications*. Third edition. Springer-Verlag, 2003.

Worzala, E., M. Lenk, and A. Silva. An Exploration of Neural Networks and its Application to Real Estate Valuation. *Journal of Real Estate Research*, 1995, 10:2, 185–202.

The authors thank Hans Wackernagel and Xavier Emery for their valuable comments on the standard deviation of the estimation error provided by kriging and cokriging. This research was partially supported by the MICINN project CSO2009-11246. The research for this article would not have been possible without the financial support of The Junta de Comunidades de Castilla-La Mancha (Spain), the European Regional Development Fund, and the European Social Fund (POII10-0250-6976).

Jose M. Montero, Castilla-La Mancha University, Toledo, Spain or Jose.mlorenzo@uclm.es.

Beatriz Larraz, Castilla-La Mancha University, Toledo, Spain or Beatriz.Larraz@uclm.es.