PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 7 (1207) — 2008
TAKSONOMIA 15
Klasyfikacja i analiza danych – teoria i zastosowania

Jose-Maria Montero, Beatriz Larraz, Gema Fernandez-Aviles

University of Castilla-La Mancha, Spain

ESTIMATING HOUSING PRICES: KRIGING AND COKRIGING AS AN ALTERNATIVE[1]

1. Introduction

The importance of space as the fundamental concept underlying the essence of social science is unquestioned. Since 1950s, a large number of spatial theories and operational models have been developed, which have gradually disseminated into the practice of urban and regional policy and analysis. However, this theoretical contribution has not been matched by a similar advance in the methodology for the econometric analysis of data observed in space.

The main idea when the spatial effects appear is how these effects can be measured. So, in this sense is very important to show that the spatial effects can be considered as special cases of general frameworks in standard econometrics, and to outline how they need a separate set of methods and techniques, encompassed within the field of spatial econometric. Nevertheless, the spatial effects have been also studied from a statistical point of view. The distinction between "spatial econometrics" and "spatial statistics" is far from being straightforward. One possible categorization can be extracted from Haining [1986] and Anselin [1988]. They refer to "the data driven orientation" in spatial statistics (Cressie [1993] and Wackernagel [2003] are two well known examples) and to "the model driven approach" in spatial econometrics (being Anselin [1988] one of the main pioneers).

This paper briefly introduces the "model-driven approach", presents kriging and cokriging strategies, which belong to the "data-driven orientation", and provides two applications of these spatial techniques to the real estate case. Some interesting recent contributions in this field are Gámez et al. [2000], Clapp et al. [2002], Case et al. [2004], Gelfand et al. [2004], Montero and Larraz [2006a] and

[1] The authors wish to thank Professor Witkowska for their helpful comments and suggestions. This research was partially supported by the FEDER PAI-05-021 project of the Junta de Comunidades de Castilla-La Mancha.

Anderson and West [2006]. In particular, the paper is organized as follows. After this introduction, section two investigates briefly into the traditional spatial econometric approach. Section three sets up methodological issues related with the spatial statistics, suggesting kriging and cokriging as an alternative. In section four two different applications are shown. Finally, section five summarizes the paper and concludes.

2. The traditional spatial econometric approach for valuating properties

The main characteristic of spatial econometrics is the way in which spatial effects (in our case, spatial autocorrelation) are taken into account. Typically, the use of a spatial weight matrix achieves that spatial models can be applied to many empirical contexts, provided that the spatial dependence is properly expressed in the weights.

The *simple* and *multiple linear regression models* (*SRM, MLRM*) have been widely used in the valuation of properties, but usually they did not take into account the location of properties. The introduction of the micro-localization factors and the spatial dependence within the regression models implies the definition of the concept of vicinity and the well-known matrix **W** of vicinities. There are two basic alternatives to introduce **W** in the model: including **WY** and/or **WX** as exogenous variables or including a spatial autoregressive disturbance, or both. These alternatives lead to the following models, widely used in the real estate case:

(i) the *first-order spatial autoregressive model* (*SAR* (1)): $\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{u}$; $\mathbf{u} \approx N(\mathbf{0}, \sigma^2\mathbf{I})$,

(ii) the *spatial lag model* (*SLM*): $\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\beta + \mathbf{u}$; $\mathbf{u} \approx N(\mathbf{0}, \sigma^2\mathbf{I})$,

(iii) the *spatial autoregressive model with spatial error dependence*: $\mathbf{y} = \rho\mathbf{W}_1\mathbf{y} + \mathbf{u}$; $\mathbf{u} = \lambda\mathbf{W}_2\mathbf{u} + \boldsymbol{\varepsilon}$; $\boldsymbol{\varepsilon} \approx N(\mathbf{0}, \sigma^2\mathbf{I})$,

(iv) the *spatial error model (SEM)*: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$; $\mathbf{u} = \lambda\mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}$; $\boldsymbol{\varepsilon} \approx N(\mathbf{0}, \sigma^2\mathbf{I})$, which can also be considered as a second order spatial autoregressive structure for the error term: $\mathbf{u} = \lambda_1\mathbf{W}_3^1\mathbf{u} + \lambda_2\mathbf{W}_3^2\mathbf{u} + \boldsymbol{\varepsilon}$; $\boldsymbol{\varepsilon} \approx N(\mathbf{0}, \sigma^2\mathbf{I})$, or a moving average structure: $\mathbf{u} = \theta\mathbf{W}_3\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}$; $\boldsymbol{\varepsilon} \approx N(\mathbf{0}, \sigma^2\mathbf{I})$,

(v) the *general spatial regression model* (*GSRM*): $\mathbf{y} = \rho\mathbf{W}_1\mathbf{y} + \mathbf{X}\beta_1 + \mathbf{W}_2\mathbf{R}\beta_2 + \mathbf{u}$; $\mathbf{u} = \lambda\mathbf{W}_3\mathbf{u} + \boldsymbol{\varepsilon}$; $\boldsymbol{\varepsilon} \approx N(\mathbf{0}, \sigma^2\mathbf{I})$, where **R** is the matrix of the exogenous variables, which could entirely match up with the matrix **X** or not. The number of specifications can be duplicated by allowing heteroskedasticity of a specific form.

These specifications have been widely used in the scientific literature related to the valuation of properties (some of them successfully), but there is a room for some statistical schemes such as kriging and cokriging, included in the data-driven orientation, which take into account the structure of the spatial dependence existing when dealing with prices of properties. Section 3 presents these two strategies.

3. Kriging and cokriging as as alternative

Kriging is a method of interpolation which estimates unknown values from data observed at known locations (univariate approach) while cokriging estimates unknown values using also additional information about other phenomena – usually more extensively sampled – correlated with the main one (multivariate approach).

3.1. Kriging: The univariate approach

Kriging is a spatial smoothing/interpolation strategy based on the assumption that the data stem from the realization of a set of random function or stochastic process over the space. This implies dealing with an infinite family of random variables constructed at all points of the given region $-X(\mathbf{s})$, being **s** the different locations where the data are or could been observed – and means that each variable takes different values depending on its spatial location, being the sample a particular realization $x(\mathbf{s})$ of $X(\mathbf{s})$. These variables are usually spatially correlated and, in this case, kriging is considered as a suitable estimation method because it incorporates the structure of that spatial correlation providing the best linear unbiased estimator. Depending on the kind of stochastic process we deal with, three different types of punctual kriging can be distinguished: simple kriging, ordinary kriging (OK) and universal kriging. In this paper OK has been used, given that the random function we deal with is intrinsically stationary with unknown mean.

In particular, the unobserved value of the variable is estimated as a weighted average of the known values in the sample locations $\{\, \mathbf{s}_i, i=1,\ldots,n \,\}$ through $X^*(\mathbf{s}_0)=\sum_{i=1}^{n}\lambda_i X(\mathbf{s}_i)$, where the weights λ_i are obtained solving the following system of $n+1$ linear equations in $n+1$ unknowns (the n weights and the Lagrange parameter, α)

$$\begin{cases}\sum_{j=1}^{n}\lambda_j\gamma(\mathbf{s}_i-\mathbf{s}_j)+\alpha=\gamma(\mathbf{s}_i-\mathbf{s}_0), \forall i=1,\ \ldots,\ n \\ \sum_{i=1}^{n}\lambda_i=1\end{cases}. \tag{1}$$

This system has been obtained after imposing the classical conditions of unbiasedness $E[X^*(\mathbf{s}_0)-X(\mathbf{s}_0)]=0 \Leftrightarrow \sum_{i=1}^{n}\lambda_i=1$, and minimum error variance

$\min V[X^*(\mathbf{s}_0)-X(\mathbf{s}_0)]=\min\left[2\sum_{i=1}^{n}\lambda_i\gamma(\mathbf{s}_i-\mathbf{s}_0)-\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma(\mathbf{s}_i-\mathbf{s}_j)\right]$, as we wish the estimator to be the BLUE one. The variogram $\gamma(\mathbf{h})$ necessary to solve this ordinary kriging system is defined as the expected squared increment of the values between a pair of locations at distance $\mathbf{h}$: $\gamma(\mathbf{h})=\frac{1}{2}V[X(\mathbf{s}+\mathbf{h})-X(\mathbf{s})]=\frac{1}{2}E\left[(X(\mathbf{s}+\mathbf{h})-X(\mathbf{s}))^2\right]$. It has to be estimated (from the observed sample values) fitting to the experimental variogram, usually calculated by the classic moments estimator [Matheron 1965], one or some theoretical variogram models following the linear model of regionalization (see, e.g. [Chilès and Delfiner 1999; Wackernagel 2003]). The usual variogram models can be seen in [Emery 2000].

3.2. Cokriging: The multivariate approach

Cokriging is a multivariate interpolation technique that allows one to better estimate map values if the distribution of a secondary process sampled more intensely than the primary process is known. Cokriging allows estimating a variable of interest – the so-called main process – at a specific location from data about itself and about auxiliary processes in the neighbourhood.

To estimate the unknown value of a particular variable X_i at the point $\mathbf{s}_0$, cokriging estimator is based on a weighted linear combination of the data values from the variables X_j ($j=1,\ ...,\ m$) located at sampled points in the neighbourhood of $\mathbf{s}_0$: $X_i^*(\mathbf{s}_0)=\sum_{j=1}^{m}\sum_{\alpha=1}^{n_j}\lambda_\alpha^j X_j(\mathbf{s}_\alpha^j)$, with $\{\mathbf{s}_\alpha^j,\ \alpha=1,\ \dots,\ n_j\}$ being the set of locations where the variable X_j has been sampled.

As in the univariate case, this section reviews the theoretical considerations for ordinary cokriging (OCK), in the joint intrinsic hypothesis, in the partial heterotopy case – some variables share some sample locations. The full details can be found in, e.g. [Wackernagel 2003]. In the same way as in OK approach, the weights λ_α^j, $\alpha=1,\ \dots,\ n_j$, $j=1,\ \dots,\ m$, are calculated to ensure that the estimator is optimal, i.e., it is unbiased and minimum error-variance. The OCK system is obtained as

$$\begin{cases}\sum_{k=1}^{m}\sum_{\beta=1}^{n_k}\lambda_\beta^k\gamma_{jk}(\mathbf{s}_\alpha^j-\mathbf{s}_\beta^k)+\omega_j=\gamma_{ji}(\mathbf{s}_\alpha^j-\mathbf{s}_0) & \forall j=1,\ ...,\ m;\quad \forall\alpha=1,\ ...,\ n_j\\ \sum_{\alpha=1}^{n_j}\lambda_\alpha^j=\delta_{ij}=\begin{cases}1 & si\ i=j\\ 0 & si\ i\neq j\end{cases}\end{cases}, \quad (2)$$

being the unbiasedness warranted by choosing weights which sum up to one for the variable of interest and which have zero sums for auxiliary variables and having minimized the variance expression

$$V\left[X_i^{*}(\mathbf{s}_0)-X_i(\mathbf{s}_0)\right]=2\sum_{j=1}^{m}\sum_{\alpha=1}^{n_j}\lambda_\alpha^j\gamma_{ji}(\mathbf{s}_\alpha^j-\mathbf{s}_0)-\sum_{j=1}^{m}\sum_{k=1}^{m}\sum_{\alpha=1}^{n_j}\sum_{\beta=1}^{n_k}\lambda_\alpha^j\lambda_\beta^k\gamma_{jk}(\mathbf{s}_\alpha^j-\mathbf{s}_\beta^k). \quad (3)$$

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 corregionalization (see, e.g. [Goovaerts 1997, p. 108-115]).

4. Applications in the real estate market

In this section, OK and OCK are used to map the price of houses and premises, respectively, in the historical area of Toledo city (Spain). The database for the kriging application was collected during September 2003 and contains information about 121 houses. The database for the cokriging case study includes information about 223 houses and 123 premises in the third quarter of 2004. The information has been provided by the Real Estate Agencies operating in Toledo. Taking these two databases as a starting point, "equivalent classes" of houses and premises have been constructed to isolate the spatial component of the properties. The factors taken into account for such a procedure have been condition, edge, surface and parking space for houses and condition, edge, surface and basement, in the case of premises. The analysis has been developed in price per square meter terms.

4.1. Mapping house prices

The structure of the positive autocorrelation that the house prices present, was estimated and fitted following the linear model of regionalization, as can be seen in Table 1 and Figure 1.

Table 1. Linear Model of Regionalization

Model	Sill	Range
Nugget effect	20000	
Spheric	50000	120 meters
Exponential	20002	136 meters

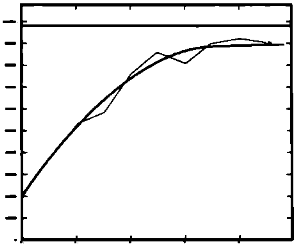


Fig. 1. Fitted variogram γ

The fitted variogram has been checked using a cross-validation procedure, and 118 robust prices – those whose standardized errors belong to the interval $[-2.5; 2.5]$ – have been obtained out of 121 that form the original set. As that variogram stabilizes around the estimated variance of the process, the random function relative to the price of houses can be considered second-order or intrinsically stationary. Therefore, OK is the procedure to map the estimations. In order to carry out the kriging estimations we have drawn a regular grid of 36 meter side over the map, having performed the estimations in the nodes of the grid. As the neighbourhood was a moving one, with a radius of 136 meters, 95 550 estimations were carried out. The estimation map and the estimation surface are shown in Figures 2 and 3 respectively.

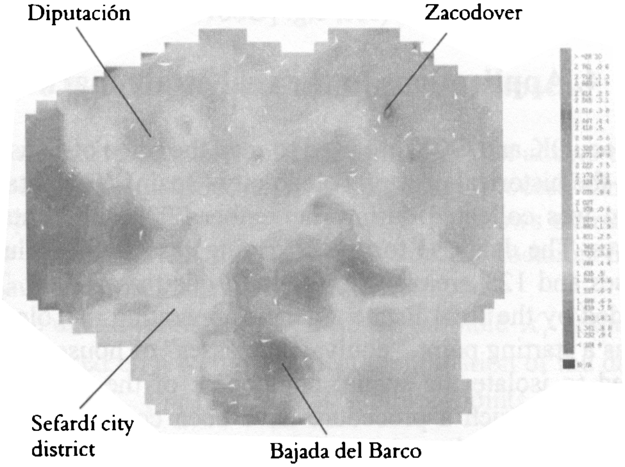


Fig. 2. Kriging house prices. Estimation map

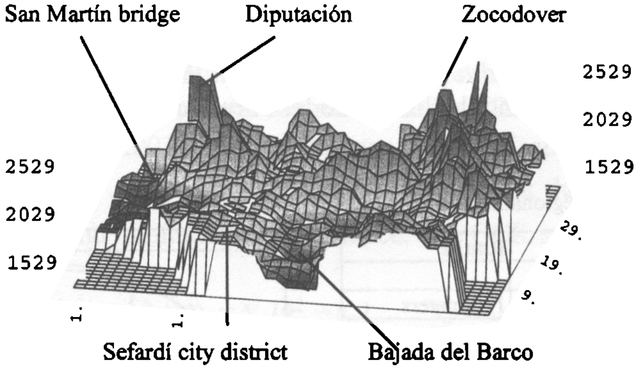


Fig. 3. Kriging house prices. Estimation surface

As it can be appreciated, it is easy to recognize the areas where house prices are cheaper (red zones) and the locations where it has been estimated a higher price (blue zones). Figure 4 shows the map of the standard deviation of the estimation error. When the colour becomes green, the standard deviation decreases.

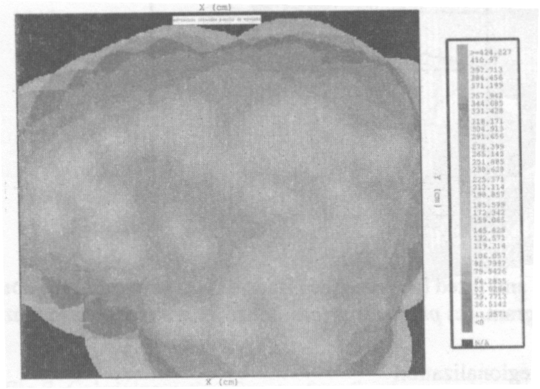
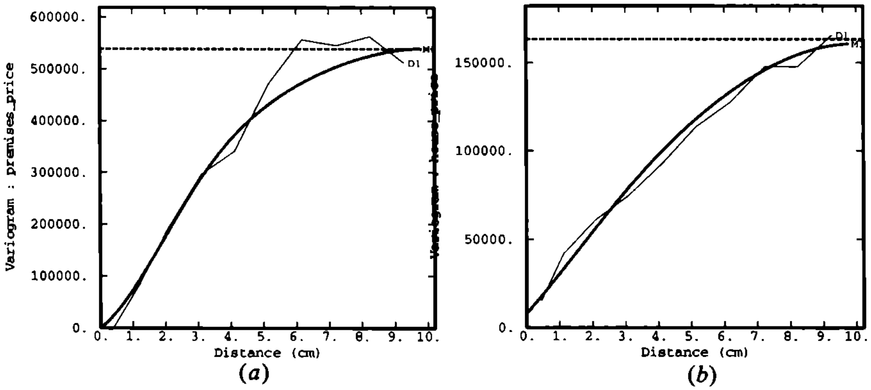


Fig. 4. Standard deviation of estimation error obtained by kriging the house prices

Figures 1 to 4 have been obtained using ISATIS v4.1.1 (2001). Real estimates could be computed by incorporating the factor effects relative to each house.

4.2. Cokriging estimation of premises pricès

We next proceed to show the results obtained from the application of the OCK to the estimation of premises prices in the old part of Toledo city, taking the price of houses in that area as an auxiliary process. First, we have represented the spatial dependence in both houses and premises markets by the appropriate theoretical variogram model, as well as the cross-dependence between both processes selecting the suitable cross variogram. The experimental and fitted cross and direct variograms appear in Figure 5, being the values of their parameters reported in Table 2.



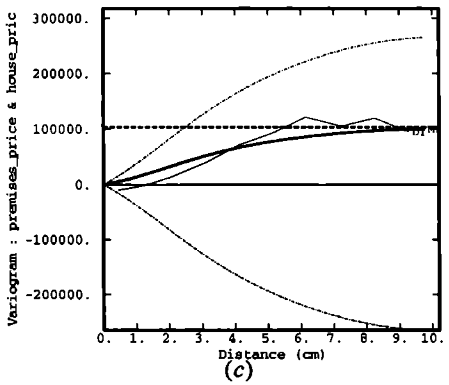


Fig. 5. Experimental and fitted (a) *premises prices* direct variogram, (b) *house prices* direct variogram, (c) *premises prices-house prices* cross-variogram

Table 2. Linear Model of Coregionalization

Model	Sill		
	Premises prices direct variogram	House prices direct variogram	Premises prices-house prices cross variogram
Spherical – 330m. range	340 978.332	142 783.006	70 505.189
Nugget effect	1	8 000	-85
Gaussian – 165m. range	200 000	10 000	30 000

These variograms have been checked for validity using the cross-validation or “leave-one-out” procedure (see, for example [Sinclair and Blackwell 2002, p. 221]). In concrete, models from Table 2 provide 119 robust estimates when estimating premises prices (96.7% from a total of 123) and 214 in the case of the house prices (96% from a total of 223), what permit us to consider models from Table 2 and Figure 5 valid for cokriging estimation.

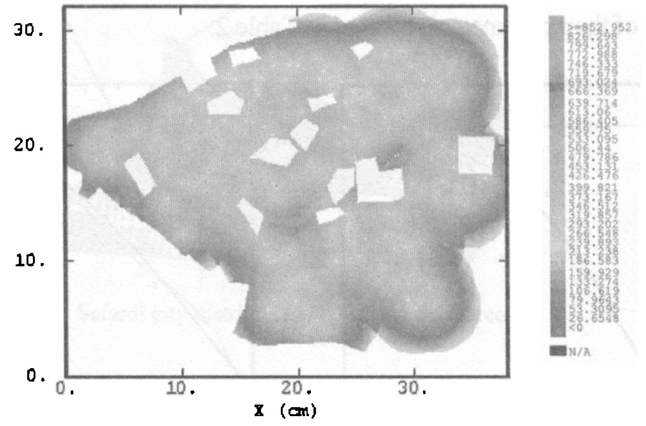


Fig. 6. Cokriging estimation map for the premises prices (€/m^2)

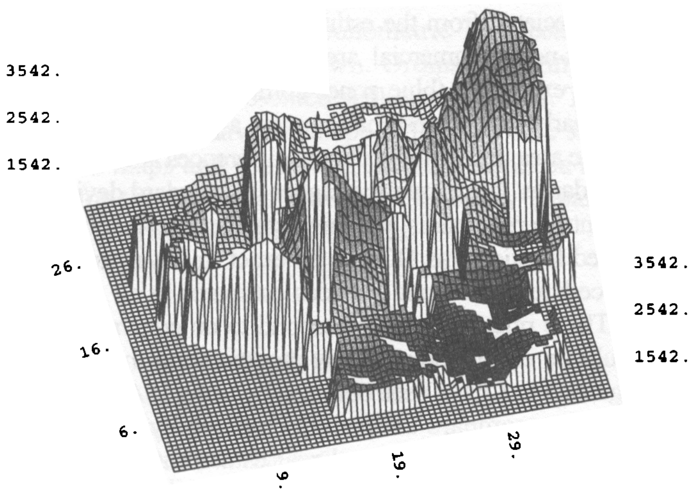


Fig 7. Cokriging estimation surface for the premises prices (€/m^2)

As the fitted variograms stabilize around the variance of the data, the random functions relative to the price of premises and houses can be considered second--order stationary and OCK is used to map the estimates. We have also estimated the price of premises by OK. The aim is to compare both procedures and check, as expected from the theoretical literature, that OCK is more accurate than OK. To perform the OCK estimation we have we have drawn a regular grid of 3.30 meter mesh over the above mentioned polygon, having performed the estimation in the nodes of the grid. As the neighborhood was a moving one with a radius of 132 meters, 68 911 estimations were carried out. These 68 911 estimates are depicted in the OCK estimation map (Figure 6) and the OCK estimation surface (Figure 7).

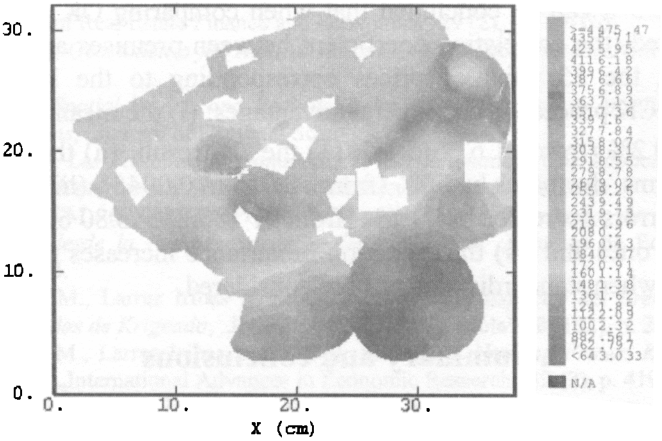


Fig. 8. Standard deviation map obtained by cokriging

As it can be appreciated from the estimation map, the areas where premises prices are cheap (red zones-non commercial areas) are easily distinguished from the areas where they are more expensive (blue zones-commercial areas). The estimation surface shows, even most clearly than the estimation map, a general view of the cokriged prices of premises across the area under study and the differences among them.

In Figure 8, the darker the colour, the higher the standard deviation. It can be clearly appreciated 123 points in red, that is, with null standard deviation; obviously they correspond to the sampled locations, as OCK is an exact multivariate interpolator. The ordinary cokriging procedure carried out provides estimates in all and each location of the area under study. These prices would correspond to an equivalent set of premises, and real estimates would be easily computed by incorporating the factor effects relative to each premises. Figures 5 to 8 have been obtained by using ISATIS v 4.1.1., 2001.

We next proceed to compare the results obtained by OCK and OK procedures. Comparison is based on the same actual real estate data. OK estimates have been computed incorporating in the weighting mechanism the premises prices direct variogram reported in Table 2, in order to compare them with the OCK ones. The comparison criterion is the interpolation accuracy when carrying out a cross validation procedure. In particular, cokriging versus kriging estimation variances are compared. Additionally, OCK and OK standardized errors $\left({}^{(x^*(s_0)-x^*(s_0))}\big/_{\sigma^*(s_0)}\right)$ have been calculated. The comparison results are reported in Table 3.

Table 3. Cross-validation results

Interpolation Method	Error		Standardized error	
	Mean	Variance	Mean	Variance
Kriging	–1.672	91 186.233	–0.0150	1.097
Cokriging	1.501	80 621.447	0.0045	0.922

From Table 3 it can be concluded that when comparing OK versus OCK results, as expected (the correlation coefficient between premises and house prices, computed with the 65 pairs of prices corresponding to the homotopic case, is $\rho = 0.696$), OCK procedure has several advantages: (i) the mean estimation error decreases by 10.2% (from –1.672 to 1.5011) the OK result, (ii) the mean error, in standardized terms, decreases by 70% (from –0.015 to 0.0045), (iii) the variance of the estimation errors decreases by 11.6% (from 91 186.233 to 80 621.447) the variance of the OK ones and (iv) the reduction in variance increases to 15.95% (from 1.097 to 0.922) when standardized errors are considered.

5. Summary and conclusions

This paper has analyzed how the space can be taken into account when estimating a value from two different points of view: a spatial econometric and a spatial statistic approaches. Related to spatial econometric perspective, the models with spatial dependence have been briefly shown. Ordinary kriging and cokriging, the statistical alternative, have been presented with more details. These univariate and multivariate spatial alternatives have been applied in two case studies in the properties market. Estimation maps and estimation surfaces of house and premises prices have been obtained for the historical part of Toledo city (Spain) and it has been shown how cokriging improves kriging estimates.

References

Anderson S.T., West S.E. (2006), *Open Space, Residential Property Values, and Spatial Context*, „Regional Science and Urban Economics" 36(6), p. 773-789.

Anselin L. (1988), *Spatial Econometrics: Methods and Models*, Kluwer Academic Publishers, Boston.

Case B., Clapp J.M., Dubin R.A., Rodríguez M. (2004), *Modeling Spatial and Temporal House Price Patterns: A Comparison of Four Models*, „Journal of Real Estate Finance and Economics" 29 (2), p. 167-191.

Chilès J.P., Delfiner P. (1999), *Geostatistics: Modelling Spatial Uncertainty*, Wiley & Sons, New York.

Clapp J.M., Kim H.J., Gelfand A.E. (2002), *Predicting Spatial Patterns of House Prices Using LPR and Bayesian Smoothing*, „Real Estate Economics" 30(4), p. 505-532.

Cressie N (1993), *Statistics for Spatial Data*, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, Inc., New York.

Emery X. (2000), *Geoestadística Lineal*, Departamento de Ingeniería de Minas, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Chile.

Gámez M., Montero J.M., García N. (2000), *Kriging Methodology for Regional Economic Analysis: Estimating the Housing Price in Albacete*, „International Advances In Economic Research" 6 (3), p. 438-451.

Gelfand A.E., Ecker M.J., Knight J.R., Sirmans C.F. (2004), *The Dynamics of Location in Home Price*, „Journal of Real Estate Finance and Economics" 29 (2), p. 149-166.

Goovaerts P. (1997), *Geostatistics for Natural Resources Evaluation*, Oxford University Press, New York.

Haining R. (1986), *Spatial Models and Regional Science: A Comment on Anselin's Paper and Research Directions*, „Journal of Regional Science" 26, p. 793-798.

Matheron G. (1965), *Les Variables Régionalisées et leur Estimation. Une Application de la Théorie des Functions Aléatories aux Sciences de la Nature*, Masson & Cie, Paris.

Montero Lorenzo J.M. (2004), *El Precio Medio del Metro Cuadrado de la Vivienda: Una Aproximación desde la Perspectiva de la Geoestadística*, „Estudios de Economía Aplicada" 22-3, p. 675-693.

Montero Lorenzo J.M., Larraz Iribas B. (2006a), *Estimación Espacial del Precio de la Vivienda Mediante Métodos de Krigeado*, „Revista Estadística Española" 48 (162), p. 201-240.

Montero Lorenzo J.M., Larraz Iribas B. (2006b), *Estimating Housing Prices: Kriging the Mean – Research Note*, „International Advances in Economic Research" 12 (2), p. 419.

Montero Lorenzo J.M., Larraz Iribas B. (2006c), *Cokriging Versus Univariate Interpolation Methods: An Application to the Premises Market*, „International Workshop on Spatio-Temporal Modelling (METMA3)", p. 223-226.

Sinclair A.J., Blackwell G.H. (2002), *Applied Mineral Inventory Estimation*, Cambridge University Press, Cambridge.

Wackernagel H. (2003), *Multivariate Geostatistics. An Introduction with Applications*, 3ª Ed., Springer-Verlag, Berlin.

ESTYMACJA CEN DOMÓW: *KRIGING* I *COKRIGING* JAKO ALTERNATYWA

Streszczenie

Podejście modelowe jest tradycyjnie wykorzystywane do szacowania cen nieruchomości. Istnieje również podejście alternatywne, tj. wywodzące się z analizy danych, które zawiera m.in. metody oparte na ważeniu odwrotnej odległości, *kriging* i *cokriging*. Ostatnie dwie dają doskonałe wyniki, ponieważ biorą pod uwagę strukturę zależności przestrzennej występującą w danych. W opracowaniu pokazano, jak działają obie wspomniane metody i wyprowadzono równania odpowiednio do analizowanych sytuacji. Przedstawiono dwie różne aplikacje w odniesieniu do rynku nieruchomości w Toledo (Hiszpania).