Prediction of Housing Location Price by a Multivariate Spatial Method: Cokriging

Author Jorge Chica-Olmo

Abstract

Cokriging is a multivariate spatial method to estimate spatial correlated variables. This method allows spatial estimations to be made and interpolated maps of house price to be created. These maps are interesting for appraisers, real estate companies, and bureaus because they provide an overview of location prices. Kriging uses one variable of interest (house price) to make estimates at unsampled locations, and cokriging uses the variable of interest and auxiliary correlated variables. In this paper, housing location price is estimated using kriging methods, isotopic data cokriging, and heterotopic data cokriging methods. The results of these methods are then compared.

There are two perspectives in the recent literature that consider the spatial autocorrelation of housing prices: spatial econometrics and geostatistics (Chica-Olmo, 1995; Pace, Barry, and Sirmans, 1998; Dubin, Pace, and Thibodeau, 1999; Tse, 2002; and Case, Clapp, Dubin, and Rodriguez, 2004). The regression method is the most widely used to obtain econometric models, while the kriging and cokriging methods are used the most in geostatistics. It is well known that there are many works in which the hedonic regression model has been applied to real estate appraisal. Structural characteristics, neighborhood characteristics, and accessibility are the explanatory variables that can be used (Can, 1990). Structural characteristics are the individual characteristics of the house itself (age, size, bathrooms, etc.), which can be impacted by the property's location. Neighborhood characteristics and accessibility depend on the location of the property. The spatial autocorrelation or spatial dependence of house price is caused by the characteristics that depend on the location. Usually, housing sale price will be directly related to the sale price of other neighboring houses. Location is probably the most important variable used to explain house price. Spatial autocorrelation is present when location is very important to housing price (Dubin, 1998).

It is desirable for researches to map surfaces (e.g., rent or price surfaces) (Clapp, 1997). The kriging and cokriging methods are characterized by the fact that they use the spatial structure of correlation to explain the housing price. In real estate analysis, the kriging method is used to create interpolated maps or continuous

maps (Anselin, 1998). It is important for the appraisal companies, bureaus, investment banking, and administration to speed up mass appraisals and to draw up continuous price maps. These maps reflect patterns in the spatial distribution of location price within a city. Valuers use large databases of houses and the Geographic Information Systems (GIS) provide the means to carry out analyses of these data and create useful models in the mass valuation process. The kriging method is used in GIS as a stand-alone analytical tool to predict residential property values (Deddis, 2002).

If housing price is spatially correlated, the kriging regression technique can be used to estimate unsampled location data. In addition, if there are auxiliary variables that are spatially correlated, along with the variable of interest, then cokriging will increase the estimation accuracy. The cokriging method estimates the value of the variable of interest at an unsampled location from data on said variable and from auxiliary variables in the neighborhood. The spatial correlation is described by a variogram. This variogram expresses the spatial dependence between housing prices at different distances. The cross-dependence between two variables is described by the cross-variogram.

The cokriging interpolation technique uses data defined on different characteristics. The house prices may be estimated from a combination of house prices and structural characteristics. The cokriging method is an extension of kriging when multivariate data are available (Wackernagel, 1995). This method considers the simple and crossed spatial correlation of the housing price and of the auxiliary variables. Moreover, this method is used when two or more explanatory variables are correlated and spatially intercorrelated.

The value of the explanatory variables of the house must be appraised in order to carry out predictions with the regression method; however, this is unnecessary when either the kriging or cokriging method is applied directly to house prices. Hence, it is possible to obtain interpolated maps of the estimated price by applying kriging or cokriging.

It is well known that the classic regression model can be used to assign prices based on a house's characteristics even if the price of the specific house is not observed. The price and characteristics of all the houses included in the sample must be used to estimate the model's parameters; that is, the data must be isotopic. For this reason, another more important characteristic of cokriging in the field of real estate evaluation is that it can be applied when the house price and the explanatory variables have not been previously sampled (heterotopic data). For example, when there are two samples: a sold housing sample (for which its price and characteristics are known) and another not-for-sale housing sample (for which the characteristics are known but not its price). This is typical of most databases obtained from tax assessors, where the sale price is known for only some of the houses. The classical regression model does not have this characteristic and ignores potentially valuable information. This available information on the

characteristics and location of unsold properties has not received much attention in the literature on hedonic price models (LeSage and Pace, 2004). For this reason, these authors discuss an alternative spatial regression model where dependent observations include observed and missing values. The approach in the current study follows the geostatisticical techniques of kriging and cokriging.

The fundamental objectives of this paper are to estimate housing location prices according to location, obtain interpolated maps by applying kriging and cokriging methods, and then to compare the results. The main contribution with regard to other studies is that of modeling house location price bearing in mind not only the spatial dependence of house price, but also the co-regionalization between price and the other characteristics of the house, such as age, size, etc., which also present a spatial dependence. For example, say that house price and size are two co-regionalized variables, since the association between price and surface area is a function of the location: houses that are too large for their neighborhoods do not typically fetch the same price as the same-sized house in a neighborhood of similarly large houses.

The next section presents a brief summary of the kriging and cokriging methods. Then, the results of the application in the case of housing price in Granada, Spain are discussed. The last section gives the conclusions.

Kriging and Cokriging Methods

Assume that the housing price data $Z(S_1), \ldots Z(S_n)$ are a particular realization of a non-stationary process that satisfies the model (Cressie, 1991):

$$Z(s) = m(s) + u(s), \qquad (1)$$

where $m(s)$ is the deterministic mean structure (variable-mean), called the trend or drift and $u(s)$ is the spatially autocorrelated error, intrinsically stationary. The drift represents a surface of the housing location price with ''a large-scale varying mean'' and fluctuations on the surface are due to $u(s)$ ''small-scale variation.'' In this case, different methods are used to estimate the drift: universal kriging, generalized covariance, and residual kriging (Neuman and Jacobson, 1984). When Z is stationary, $m(s) = m$ is constant-mean and unknown, the kriging method is applied to the data.

The residual kriging (cokriging) or detrend method consists of carrying out a polynomial least squares regression of the data to estimate $m(s)$. Kriging (cokriging) of resulting residuals is applied to estimate $u(s)$. In practice, $m(s)$ is represented by the means of polynomial drift terms of first or second degree:

$$\text{linear drift } m(s) = b_0 + b_1x + b_2y$$
$$\text{quadratic drift } m(s) = b_0 + b_1x + b_2y + b_3x^2 + b_4y^2 + b_5xy, \quad (2)$$

where b_j are the regression coefficients and (x,y) are the longitude and latitude of the houses.

Clapp, Kim, and Gelfand (2002) use local polynomial regression instead of the linear or quadratic forms for latitude and longitude and kriging for the residuals in order to capture small-scale variation. One advantage of local polynomial regression is that collinearity can be avoided, which may be problematic when estimating the quadratic form.

Kriging

In matrix form, model (1) can be expressed:

$$\mathbf{z} = \mathbf{Fb} + \mathbf{u}, \quad (3)$$

where: $\mathbf{z}$ is the vector, $n \times 1$, of $Z(s)$; F is the matrix, $n \times k$, what includes the polynomial drift terms; $\mathbf{b}$ is the vector, $k \times 1$, of unknown parameters; $\mathbf{u}$ is the vector, $n \times 1$, of the disturbance term.

Assume that $\mathbf{u}$ is characterized by the variogram $\gamma_u(h)$ defined below.

The ordinary least squares estimator of the regression coefficients in presence of spatial autocorrelation is inefficient. In this case, the generalized least squares estimator (EGLS) can be used, which is BLUE (best linear unbiased estimator).

The EGLS of $\mathbf{b}$ is:

$$\hat{\mathbf{b}}_k = (F' \hat{V}_k^{-1} F)^{-1} F' \hat{V}_k^{-1} \mathbf{z}, \quad (4)$$

where $\hat{V}_k$ is the variance-covariance matrix of the disturbances.

If $Z(s)$ is stationary, the elements of $\hat{V}_k$ are obtained[1] by its relation to the variogram:

$$\hat{v}_{ij} = C_{\hat{u}}(h) = C_{\hat{u}}(0) - \gamma_{\hat{u}}(h), \quad (5)$$

where $\hat{u}$ is least squares residuals; $C_{\hat{u}}(h)$ is the value of the covariance between the pairs of residuals that are at a distance h apart; $C_{\hat{u}}(0)$ is the sill of the variogram of residuals; $\gamma_{\hat{u}}(h)$ represents the variogram of the residuals.

The spatial dependence between $\hat{u}(s_i)$ and the separation vector distance (h) and the direction (θ) is expressed by a variogram (semi-variogram or direct-variogram) $\gamma_{\hat{u}}(h)$. An unbiased estimator of the variogram is (Matheron, 1965):

$$\hat{\gamma}_{\hat{u}}(h_\theta) = \frac{1}{2N(h_\theta)} \sum_{i=1}^{N(h_\theta)} [\hat{u}(s_i + h_\theta) - \hat{u}(s_i)]^2, \tag{6}$$

where $(s_i + h_\theta)$ and (s_i) are locations and $N(h_\theta)$ is the number of h_θ distant point-pairs. The empiric variogram computed on different directions on the map are checked to find directional influences (anisotropy). It is necessary to adjust the model[2] to the empiric variogram to carry out estimations with the kriging method. The following exponential model is used:

$$\gamma(h) = \begin{cases} C_0 + C\left[1 - \exp\left(-\dfrac{h}{a}\right)\right] & h > 0 \\ 0 & h = 0 \end{cases} \tag{7}$$

The model fitted depends on three parameters: nugget effect (C_0), range (a), and sill ($C_0 + C$), where (C_0) is a partial sill. The nugget effect is a measure of spatial continuity; range is the distance where the model levels out, and sill is the value the variogram model attains at the range. The exponential model (Exhibit 1) reaches the sill asymptotically, with the practical range (a′) defined as that distance at which the variogram value is 95% of the sill, a′= 3a (Isaaks and Srivastava, 1989).

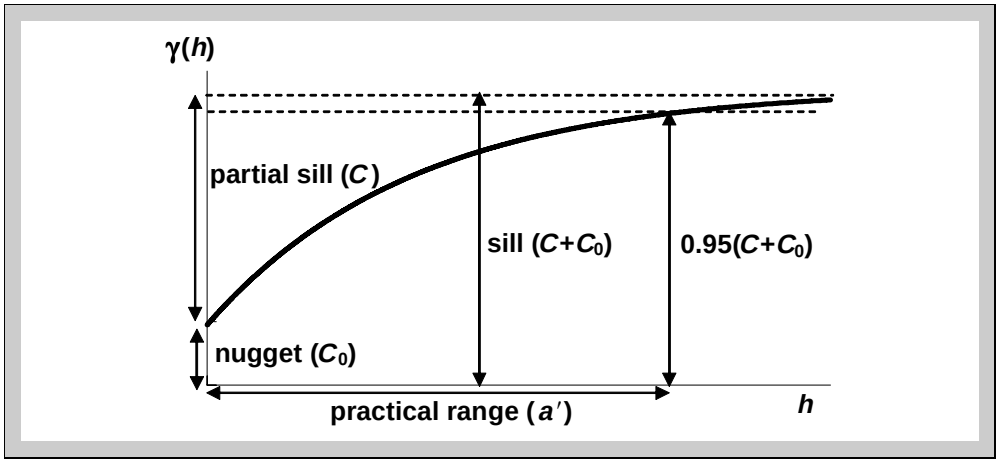
The kriging estimator of the least squares residuals is:

$$\hat{u}_k(s_o) = \sum_{i=1}^{n} \lambda_i \hat{u}(s_i), \tag{8}$$

where λ_i are the kriging weights. In order that the kriging estimator be unbiased, it has to be true that:

$$\sum_{i=1}^{n} \lambda_i = 1. \tag{9}$$

Exhibit 1 | Parameters of Exponential Model



The weights are selected to minimize the variance of error:

$$Var[\hat{u}_k(s_o) - \hat{u}(s_o)], \tag{10}$$

by resolving the kriging system:

$$\begin{bmatrix} \Gamma & \mathbf{1} \\ \mathbf{1}' & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda} \\ \mu \end{bmatrix} = \begin{bmatrix} \gamma_0 \\ 1 \end{bmatrix}, \tag{11}$$

where Γ is a symmetric matrix formed by $\gamma_{\hat{u}}(s_i - s_j)$ for $i,j = 1...n$; $\boldsymbol{\lambda}$ is the weights vector; $\mathbf{1}$ is the vector of ones; $\boldsymbol{\gamma}_0$ is the vector formed by $\gamma_{\hat{u}}(s_0 - s_i)$ for $i = 1...n$ and μ is the Lagrange multiplier.

The best linear unbiased predictor (BLUP) is used to estimate the housing location price located at s_0, $\hat{Z}_k(s_0)$:

$$\hat{Z}_k(s_0) = \mathbf{f}_0'\hat{\mathbf{b}}_k + \hat{u}_k(s_0) = \hat{m}_k(s) + \hat{u}_k(s_0), \tag{12}$$

where $\mathbf{f}_0$ is the vector $k \times 1$ of the known values of the polynomial drift terms of the house located at s_0; n_k represents the number of houses located close to the house s_0; λ_i are kriging weights; $\hat{u}(s_i)$ are GLS residuals; and $\hat{u}_k(s_0)$ is estimation by kriging of $\hat{u}(s_0)$.

Cokriging

When dealing with different variables, each variable is measured at different points on the map. The location of points can be equal for all the variables (isotopy), equal for some variables (partial heterotopy), or different for all the variables (complete heterotopy). In partial heterotopy, cokriging is interesting when the auxiliary variables are available at more points than the main variable (Wackernagel, 1995).

The objective of cokriging is to predict the value of housing location price at an unsampled site employing the auxiliary variables such as surface area, age, etc. In the current study, it is assumed that the auxiliary variables are spatially correlated, as well as being co-regionalized regarding price.

Consider for simplicity[3] only two variables, the variable of interest Z_1 (housing price), a variable with drift, and an auxiliary variable Z_2 (surface area), a variable without drift, with number of samples n_1 and n_2, respectively, not necessarily equal.

The model can then be written in matrix notation as:

$$\begin{pmatrix}\mathbf{z}_1\\ \mathbf{z}_2\end{pmatrix} = \begin{pmatrix}F & 0\\ 0 & \mathbf{1}\end{pmatrix}\begin{pmatrix}\mathbf{b}_1\\ b_2\end{pmatrix} + \begin{pmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{pmatrix} \Rightarrow \mathbf{z}_{ck} = F_{ck}\mathbf{b}_{ck} + \mathbf{u}_{ck}, \qquad (13)$$

the EGLS of $\mathbf{b}_{ck}$ is:

$$\hat{\mathbf{b}}_{ck} = (F'_{ck}V_{ck}^{-1}F_{ck})^{-1}F'_{ck}V_{ck}^{-1}\mathbf{z}_{ck} \text{ with } V_{ck} = \begin{pmatrix}V_{11} & V_{12}\\ V_{21} & V_{22}\end{pmatrix}, \qquad (14)$$

where V_{11} and V_{22} are variance-covariance matrix of the disturbances u_1 and u_2, respectively; and V_{12} and V_{21} are the cross-covariance matrix of the disturbances.

The variogram or direct-variogram measures spatial dependence for one variable. The cross-dependence between two variables is measured with the cross-variogram. In this case, the cross-variogram estimator of least squares residuals is:

$$\hat{\gamma}_{\hat{u}_1\hat{u}_2}(h_\theta) = \frac{1}{2N(h_\theta)}\sum_{i=1}^{N(h_\theta)} [\hat{u}_1(s_{1i} + h_\theta) - \hat{u}_1(s_{1i})][\hat{u}_2(s_{2i} + h_\theta) - \hat{u}_2(s_{2i})], \qquad (15)$$

where $N(h_\theta)$ represents the number of h distant point pairs, where variables $\hat{u}_1$ and $\hat{u}_2$ are measured. The cross-variogram can only be calculated when variables are measured in the same locations (isotopy and partial heterotopy). The cross-variogram can be negative, which indicates a negative correlation between the variables (Journel and Huijbregts, 1978).

The following equation is used to estimate the housing location price located at s_0, $\hat{Z}_{ck}(s_0)$:

$$\hat{Z}_{ck}(s_0) = \mathbf{f}'_{c0}\hat{\mathbf{b}}_{ck} + \hat{u}_{ck}(s_0) = \hat{m}_{ck}(s) + \hat{u}_{ck}(s_0), \tag{16}$$

where $\mathbf{f}'_{c0} = (\mathbf{f}'_0 \quad 0)$ and $\hat{u}_{ck}(s_0)$ is estimation by cokriging of $\hat{u}_1(s_0)$.

The cokriging estimator is a weighted average of observed values of the variables:

$$\hat{u}_{ck}(s_0) = \sum_{i=1}^{n_1} \lambda_{1i}\hat{u}_1(s_{1i}) + \sum_{j=1}^{n_2} \lambda_{2j}\hat{u}_2(s_{2j}), \tag{17}$$

where n_1 and n_2 are the nearest houses to the housing s_0; λ_{1i} and λ_{2j} are the weights associated to each sampling point.

In order that the cokriging estimator be unbiased, two restrictions must be true:

$$\sum_{i=1}^{n_1}\lambda_{1i} = 1 \text{ and } \sum_{j=1}^{n_2}\lambda_{2j} = 0. \tag{18}$$

The cokriging system is:

$$\underbrace{\begin{bmatrix} \Gamma_{11} & \Gamma_{12} & \mathbf{1} & \mathbf{0} \\ \Gamma_{21} & \Gamma_{22} & \mathbf{0} & \mathbf{1} \\ \mathbf{1}' & \mathbf{0} & 0 & 0 \\ \mathbf{0} & \mathbf{1}' & 0 & 0 \end{bmatrix}}_{\Gamma} \underbrace{\begin{bmatrix} \boldsymbol{\lambda}_1 \\ \boldsymbol{\lambda}_2 \\ \mu_1 \\ \mu_2 \end{bmatrix}}_{\lambda} = \underbrace{\begin{bmatrix} \boldsymbol{\gamma}_{10} \\ \boldsymbol{\gamma}_{20} \\ 1 \\ 0 \end{bmatrix}}_{\gamma}, \tag{19}$$

where Γ_{11} and Γ_{22} are direct-variogram matrixes formed by $\gamma_{\hat{u}_1}(s_{1i} - s_{1j})$ for $i,j = 1...n_1$ and $\gamma_{\hat{u}_2}(s_{2i} - s_{2j})$ for $i,j = 1...n_2$; Γ_{12} and Γ_{21} are the cross-variogram matrixes formed by $\gamma_{\hat{u}_1\hat{u}_2}(s_{1i} - s_{2j})$ for $i = 1...n_1$, $j = 1...n_2$ and $\gamma_{\hat{u}_2\hat{u}_1}(s_{2i} - s_{1j})$

for $i = 1...n_2$, $j = 1...n_1$; $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ are weights vectors; $\mathbf{1}$ is the vector of ones; $\boldsymbol{\gamma}_{10}$ and $\boldsymbol{\gamma}_{20}$ are vectors formed by $\gamma_{\hat{u}_1}(s_{10} - s_{1i})$ for $i = 1...n_1$ and $\gamma_{\hat{u}_1\hat{u}_2}(s_{10} - s_{2j})$ for $j = 1...n_2$ and μ_1 and μ_2 are Lagrange multipliers.

When the primary and secondary variables exist at all data locations (isotopic data) and the direct-variograms and cross-variograms are alike, cokriging is similar to kriging (Isaaks and Srivastava, 1989). This also occurs if all the variables are spatially uncorrelated.

To resolve the cokriging system of equations, the condition that the matrix Γ be positive-definite must be met. This condition comes true if any possible linear combination of the variables is always positive. The linear model of coregionalization method to ensures this. The linear model of coregionalization is made up of linear combinations of N structures of variation, typically a nugget effect and one or more structures, characterized by their own basic model of variogram (Chilès and Delfiner, 1999):

$$\Gamma(h) = \sum_{s=1}^{N} B_s \Gamma_s(h), \tag{20}$$

where $\Gamma(h) = [\gamma_{jk}(h)]$ is a matrix of the direct and cross variograms; $\Gamma_s = [\gamma_s(h)]$ is a diagonal matrix of basic variogram models and $B_s = [b^s_{jk}]$ is a symmetric matrix of coefficients (sills of the direct and cross variograms). A sufficient condition for the model to be valid is that all matrixes B_s are positive-definite (Chilès and Delfiner, 1999). For example, consider two variables $\hat{u}_1$ and $\hat{u}_2$, and two basic structures, nugget (n) and exponential (e):

$$\begin{bmatrix} \gamma_{\hat{u}_1}(h) & \gamma_{\hat{u}_1\hat{u}_2}(h) \\ \gamma_{\hat{u}_2\hat{u}_1}(h) & \gamma_{\hat{u}_2}(h) \end{bmatrix} = \begin{bmatrix} b^n_{\hat{u}_1} & b^n_{\hat{u}_1\hat{u}_2} \\ b^n_{\hat{u}_2\hat{u}_1} & b^n_{\hat{u}_2} \end{bmatrix} \cdot \begin{bmatrix} \gamma_n(h) & 0 \\ 0 & \gamma_n(h) \end{bmatrix} + \begin{bmatrix} b^e_{\hat{u}_1} & b^e_{\hat{u}_1\hat{u}_2} \\ b^e_{\hat{u}_2\hat{u}_1} & b^e_{\hat{u}_2} \end{bmatrix} \cdot \begin{bmatrix} \gamma_e(h) & 0 \\ 0 & \gamma_e(h) \end{bmatrix}, \tag{21}$$

where $b^n_{\blacksquare}$ and $b^e_{\blacksquare}$ are, respectively, nuggets and sills of variograms.

The verification, in this case, of the positive-definite condition is:

$$|b^s_{\hat{u}_1\hat{u}_2}| = |b^s_{\hat{u}_2\hat{u}_1}| \leq \sqrt{b^s_{\hat{u}_1}\, b^s_{\hat{u}_2}} \text{ for } s = n, e. \tag{22}$$

In practice, one criterion for considering a linear model is that all variograms have the same ranges (Chilès and Delfiner, 1999).

Cross-Validation

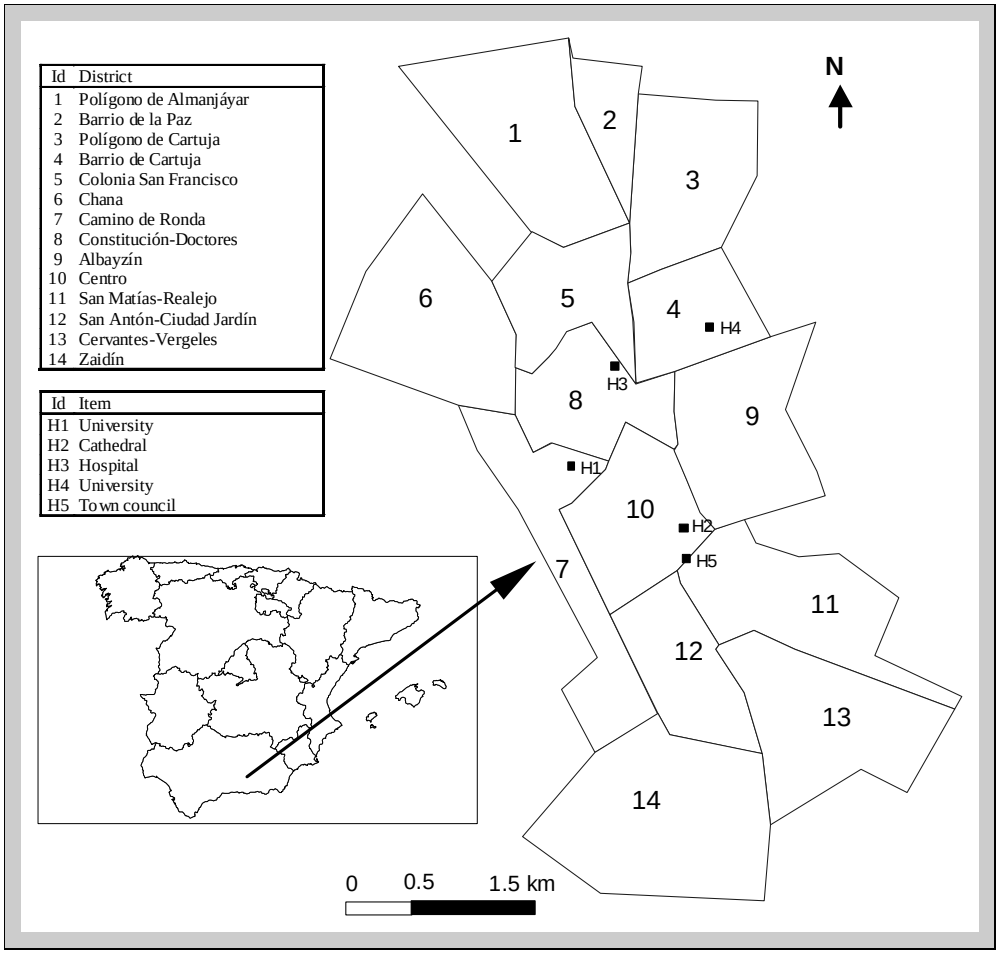
Cross-validation allows selection between different models or methods. Cross-validation removes, one at a time, each data location and predicts data value in this location. This procedure is repeated for all experimental points. In this way, the predicted value is compared with the observed value. Using the simple regression between predicted and observed values and, subsequently, their scatter plot and the coefficient of determination (R^2), gives a measure to compare. Other methods used for comparison are the summary statistics of predicted errors (observed − predicted): Mean Errors (ME ≈ 0), Root Mean Square Errors (RMSE ≈ min.), Mean Kriging/cokriging Standard Error (MKStE ≈ min.), and Root Mean Square Standardized Errors (RMSStE ≈ 1). If the errors are unbiased, ME should be near zero; if the R^2 and regression coefficient are near one, and RMSE is small, then predictions will be close to the observed values; if MKStE is minimal, the uncertainty of predictions will be small; and RMSStE is near to one if the observed error is close, on mean, to the error predicted.

Application

The study area is the city of Granada, Spain. Granada is an ancient city located in the southern part of the country (Exhibit 2). This exhibit shows the city divided into fourteen districts. The current city center is in district 10 and the Central Business District (CBD) covers district 10 and part of districts 11 and 12. District 9 is the oldest district, dating back to before the twelfth century, so many of the buildings are historical. From the sixteenth century through to the beginning of the twentieth century, this historic quarter expanded towards districts 8, 10, and 11, when good quality low-rise buildings of large flats were constructed. In the mid-twentieth century, district 7 was established, containing medium quality high-rise blocks. In the 1970s, districts 6 and 14 appeared: working-class areas with poor quality cheap housing. Around the same time, district 12 was designed, in answer to the demand for better quality modern housing that was closer to the city center. In the 1980s, public funds were used to construct the districts to the north of the city, which were occupied by the lower middle class. The higher crime districts (1, 2, and 3) are in the northern part of the city (the most conflictive being district 2). District 13 sits atop a small hill, at the foot of which there were, originally, small groups of poorly-constructed houses. However, the housing boom of the 1990s meant that this district was converted, and small blocks of well-built houses now cover the entire hill.

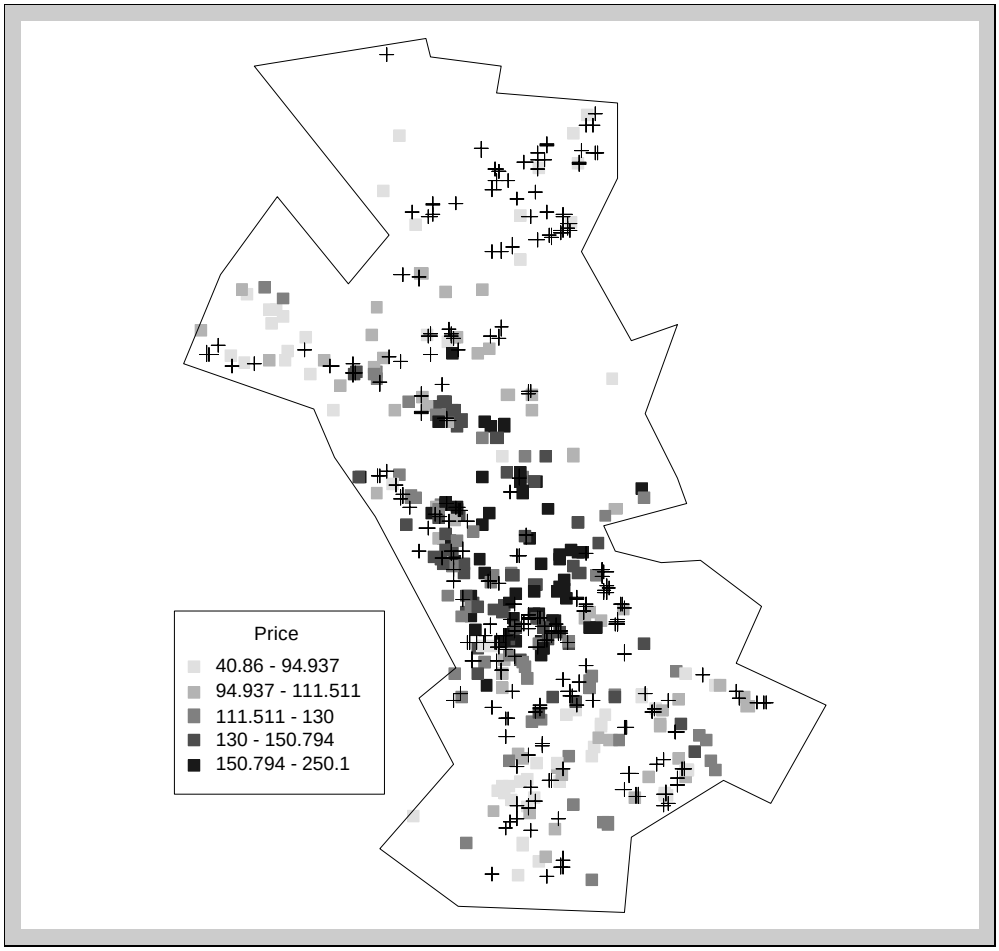
The application has been carried out on a sample of 287 apartments (Exhibit 3). The data comes from market research carried out by the Centro de Gestión Catastral (Official Cadastre Agency) of Granada and completed mainly during the fourth quarter[4] of 1995. This sample represents 23.5% of the total number of second-hand apartment sales registered during this period. Houses in the northern area of the city (districts 1, 3, and 6) and the area to the south (district 14) are,

Exhibit 2 | Study Area, District, and Item



logically, cheaper, since they are poor quality constructions. But they also have low typical deviations, since the building quality in all four districts is similar (see Exhibit 4). However, district 11, which contains housing both in the CBD and on the outskirts, presents the highest mean prices, as well as the greatest variation in price. The lowest typical deviation is found in district 5, which is a transitional area between the northern districts and the city center, with housing having very similar characteristics. In district 7, the mean price is approximately the same as the mean price for the city as a whole, despite covering an elongated area that runs along the southwestern part of the city. Its standard deviation is not high, since it contains buildings constructed during the same period. In the central area of the city, prices and standard deviations are similar in districts 8 and 9, while, in district 10, the mean prices are, logically, higher.

Exhibit 3 | Location of Sample Apartments



Notes: The square symbol represents the price of the sample housing as of 1995. The cross symbol represents the sample as of 1991.

The variable of interest, *PRICE*, represents the apartment price per square meter. The auxiliary variables are: *AGE* (apartment age, adjusted for major rehabilitation, in years); *AMP* (amplitude,[5] quotient between the surface area of the apartment and the number of rooms); and *HEAT* (central heating, binary variable that takes the value 1 if the apartment has central heating and 0 if it does not).

The Exhibit 5 shows the statistics of the variables. The average price is 124.654 thousand pesetas (749.19 euros) and the mean age is 15.84 years. Exhibit 6 shows the simple correlation coefficients of the variables examined. All coefficients of linear correlation are significant (p-value = 0.01). Negative correlation between

Exhibit 4 | Statistics of Error Using the Cokriging Method with Heterotopic Data

		Error		Observed		Predicted	
District	Count	Average		Average	Std. Dev.	Average	Std. Dev.
1	3	15.343	17.856	86.790	6.493	71.446	13.751
3	7	−1.564	11.598	61.893	12.427	63.457	15.403
5	10	−3.787	9.659	103.230	6.161	107.017	11.308
6	20	1.517	12.890	97.532	20.115	96.015	12.380
7	18	−1.209	27.221	130.368	24.758	131.578	12.570
8	29	1.183	19.230	125.231	26.354	124.048	17.784
9	9	−8.364	28.442	124.345	30.089	132.709	17.867
10	53	0.296	17.787	146.307	21.373	146.012	9.176
11	21	−1.257	28.807	160.859	41.237	162.116	23.027
12	56	−0.612	16.456	138.047	30.440	138.658	23.239
13	30	0.097	10.937	107.591	16.222	107.494	13.971
14	31	1.573	11.714	93.908	16.417	92.335	12.488
Total	287	−0.099	17.948	124.654	33.597	124.754	28.756

Note: In districts 2 and 4 there are no homes for sale.

Exhibit 5 | Statistics of Variables

	PRICE	*AGE*	*AMP*	*HEAT*
Average	124.654	15.840	37.047	0.560
Std. Dev.	33.597	9.986	8.935	0.497
Skewness	0.439	0.017	1.106	−0.261
Kurtosis	0.231	−1.109	1.349	−1.945
Minimum	40.860	1.000	23.750	—
25%tile	100.000	6.000	30.000	—
Median	122.951	16.000	35.750	—
75%tile	145.312	23.000	42.000	—
Maximum	250.000	36.000	73.000	—

Exhibit 6 | Correlation Coefficients

	PRICE	*AGE*	*AMP*	*HEAT*
PRICE	1	−0.389	0.439	0.504
AGE	−0.389	1	−0.370	−0.321
AMP	0.439	−0.370	1	0.283
HEAT	0.504	−0.321	0.283	1

age and other variables is as expected, as is the positive correlation between price, amplitude, and central heating.

Using heterotopic data, the Cokriging method has been applied to another sample of 259 houses (the crosses in Exhibit 3 represent their locations). For these houses, the variables *AGE* and *AMP* agree with the data.

Exhibit 3 shows that house prices in Granada are convexly distributed; that is, high prices in the CBD (district 10), which drop in the outskirts. This indicates the presence of quadratic drift in the variable *PRICE*. For this reason, EGLS has been used to estimate the following models:

$$\hat{m}_k(s) = \underset{(-2.017)}{-575.614} + \underset{(2.124)}{0.143x} + \underset{(3.452)}{0.103y} - \underset{(-2.293)}{9.089^{-6}x^2} - \underset{(-6.011)}{7.351^{-6}y_2} - \underset{(-1.347)}{5.034^{-6}xy} \quad (23)$$

$$\hat{m}_{ck}(s) = \underset{(-2.525)}{-564.917} + \underset{(2.768)}{0.146x} + \underset{(3.905)}{0.091y} - \underset{(-2.996)}{9.318^{-6}x^2} - \underset{(-6.486)}{6.219^{-6}y^2} - \underset{(-1.685)}{6.218^{-6}xy} \quad (24)$$

where $\hat{m}_k(s)$ and $\hat{m}_{ck}(s)$ are the estimations of housing location price on a ''large-scale variation'' x and y are UTM coordinates (meters) of houses. The t-values are in parentheses. The t-values of the model estimated by cokriging are larger than those estimated by kriging, which indicates larger efficiency.

Analysis of Spatial Structure of Variability

The variogram is the tool most commonly used in geostatistics to analyze the spatial correlation and continuity (or variability). The empirical direct-variograms

and cross-variograms for the four variables are shown in Exhibit 7. Four direct-variograms and six cross-variograms are displayed.

The calculated variograms are omnidirectional[6] and have been calculated for a maximum distance equal to 1,500 meters (fifteen intervals of width 100 meters each). The examined direct-variograms of the variables show an increasing behavior, which suggests that variables are correlated spatially. Thus, the greater the distance between houses, the greater the variability and, hence, the lower the correlation. The variogram of the price variable, *PRICE*, presents a more continuous behavior in the space than the rest of the variables, since its nugget effect value is lower than that of the other variables.

Likewise, the cross-variograms represent how the degree of association among variables varies with distance. In general, this degree of association would be expected to drop with distance and, thus, the correlation between two variables (for example, price and surface area) of houses that are located close together should be greater than that for homes situated further apart. Thus, if the variables are positively correlated, the variogram is increasing in form. However, when the variables are negatively correlated, the variogram is decreasing in form, but this does not mean that the degree of association increases with distance, but, rather, it decreases inversely. Thus, the cross-variogram between *AGE* and other variables (Exhibit 7) is negative because these variables have a negative correlation. Hence the cross-variogram among *AGE* and other variables has a decreasing tendency as the distance becomes larger, which means that spatial variability increases in absolute terms and therefore decreases the spatial correlation. The rest of the cross-variograms show a positive correlation. The rising trend of these cross-variograms suggests that the variables are less correlated as the distance increases.

The presence of anisotropy in the studied variables has also been examined. A variable is anisotropic if the spatial correlation structure depends on the direction. All variables were assumed to be isotropic because no significant anisotropy was detected for a maximum distance of 1,500 meters. Directional variograms (directions 0°, 45°, 90°, and 135°) were used for the examination of anisotropy. Exhibit 8 shows the directional variograms of the residual of the price variable, *PRICE*. This exhibit demonstrates that directional variograms are similar; thus, isotropy can be observed.

The exponential model has been used to fit the empirical direct-variograms and cross-variograms. In kriging, the method used to estimate the parameters of the variogram is by weighted non-linear least squares (Cressie 1991), the parameters are: $C = 377.274$, $a = 224.126$, and $C_0 = 283.831$. Otherwise, in cokriging, the general procedure for fitting the models of direct-variograms and cross-variograms, in a linear model of coregionalization, is to postulate the range and attempt a fit of the sills by trial and error or other method (Chilès and Delfiner, 1999). Hence, the range to 224.126 has been postulated, and the iterative algorithm proposed by Goulard and Voltz (1992) has been applied. The values of the estimated parameters out of every direct-variogram and cross-variogram are shown

Exhibit 7 | Direct-Variograms and Cross-Variograms of Variables' Residuals
(empirical: dots; model: solid line)

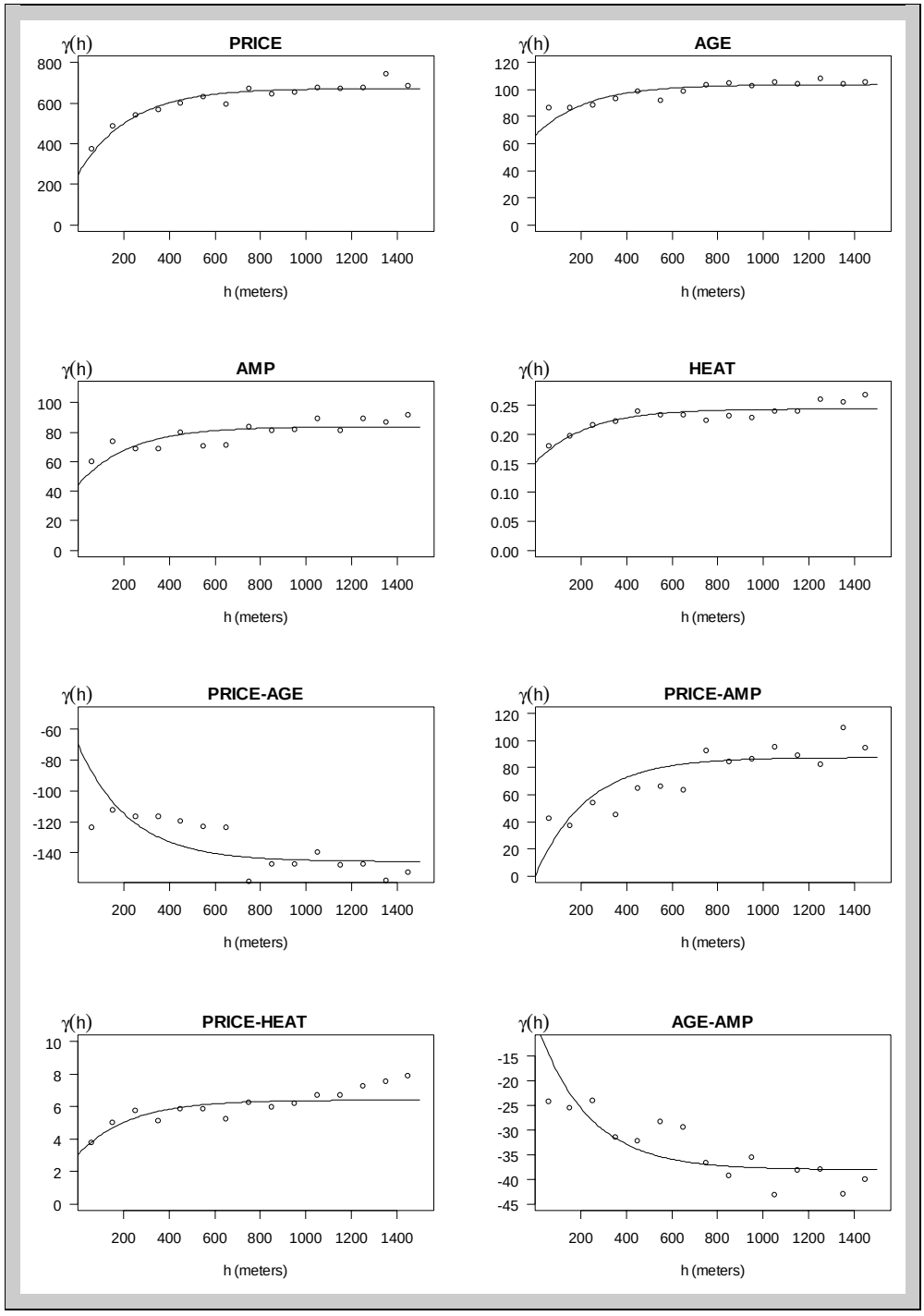


Exhibit 7 | (continued)
Direct-Variograms and Cross-Variograms of Variables' Residuals
(empirical: dots; model: solid line)

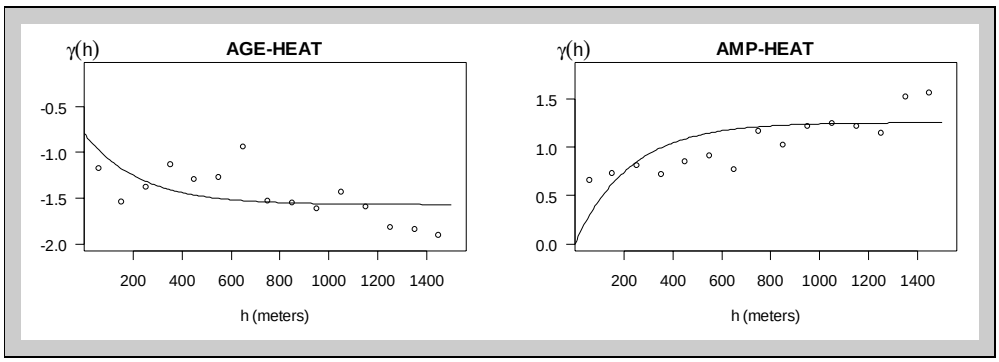
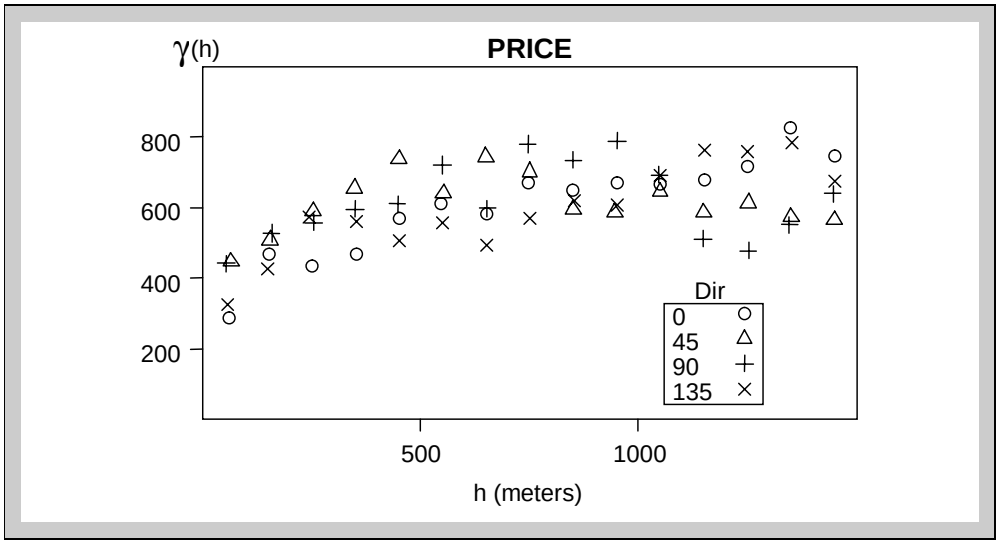


Exhibit 8 | Directional Variograms of Price Variable's Residuals



in Exhibit 9. Column 4 of Exhibit 9 shows the spatial discontinuity (Nugget) of every variable or variables in terms of the total variability (Sill = $C + C_0$). Thus, the variable with the main relative discontinuity is *AGE* (63.82%) while the variable with the least relative discontinuity is *PRICE* (36.88%). The practical range used for spatial correlation is 672.38 meters ($a' = 3a$), which represents the distance as from which the variables stop being spatially correlated.

Exhibit 9 | Parameters of Direct-Variogram and Cross-Variogram Models

Variable	Nugget (C_0)	Partial Sill (C)	%Nugget
PRICE	248.348	425.059	36.88
AGE	66.202	37.525	63.82
AMP	44.220	39.623	52.74
HEAT	0.151	0.093	61.89
PRICE-AGE	−69.706	−75.976	47.85
PRICE-AMP	0.000	87.618	0.00
PRICE-HEAT	2.296	4.367	34.46
AGE-AMP	−7.223	−30.825	18.99
AGE-HEAT	−0.792	−0.779	50.41
AMP-HEAT	0.000	1.254	0.00

Notes: Residuals of variables: apartment price (*PRICE*), apartment age (*AGE*), amplitude of rooms (*AMP*), and central heating (*HEAT*). Where %Nuget = $(C_0/(C_0 + C))*100$ shows spatial discontinuity (Nugget) out of every variable in terms of the total variability (Sill = $C_0 + C$).

Spatial Estimation: Kriging and Cokriging

Kriging (cokriging) can be used to carry out estimations for any point or house on the map. Thus, in order to carry out estimations at a certain point, the estimation system has to be geometrically defined, which means that the neighborhood of the data has to be established that will intervene in the estimation. All the houses can be used in the estimation (global neighborhood), or a set of data (local neighborhood). In practice, only the houses that are found inside a circumference or ellipse centered on the point are used in the estimation. In this application, a minimum of three houses and a maximum of twenty-four have been used, located in a circumference with a radius equal to the range of the variogram. Obviously, the proximity of the data for each point to be estimated will be different, which makes it necessary to solve a different kriging (cokriging) equation system for each of the points on the map. The estimators of housing location price are:

a. Kriging:

$$\hat{Z}_k(s_0) = \hat{m}_k(s_0) + \hat{u}_{1k}(s_0) = \hat{m}_k(s_0) + \sum_{i=1}^{24} \lambda_i \hat{u}_1(s_i), \qquad (25)$$

b. Cokriging:

$$\begin{aligned}\hat{Z}_{ck}(s_0) &= \hat{m}_{ck}(s_0) + \hat{u}_{1ck}(s_0) + \hat{u}_{2ck}(s_0) + \hat{u}_{3ck}(s_0) \\ &\quad + \hat{u}_{4ck}(s_0) \\ &= \hat{m}_{ck}(s_0) + \sum_{i=1}^{24}\lambda_{1i}\hat{u}_1(s_i) + \sum_{i=1}^{24}\lambda_{2i}\hat{u}_2(s_i) + \sum_{i=1}^{24}\lambda_{3i}\hat{u}_3(s_i) \\ &\quad + \sum_{i=1}^{24}\lambda_{4i}\hat{u}_4(s_i), \end{aligned} \tag{26}$$

where $\hat{u}_1$, $\hat{u}_2$, $\hat{u}_3$, and $\hat{u}_4$ are residuals of *PRICE*, *AGE*, *AMP*, and *HEAT*, respectively.

On the other hand, these methods allow the representation of the estimations of housing location prices in the form of isoline maps. These maps are obtained from the estimations carried out on the nodes of a regular mesh. In this paper, the location price has been estimated at each of the nodes of a 50-meter sided regular mesh.

Exhibit 10 shows the prices map obtained applying kriging; Exhibit 11a shows the pricings map using cokriging with isotopic data (original sample of 287

Exhibit 10 | Map of Price Estimated by Kriging

Exhibit 11 | Cokriging with a) Isotopic Data and b) Heterotopic Data

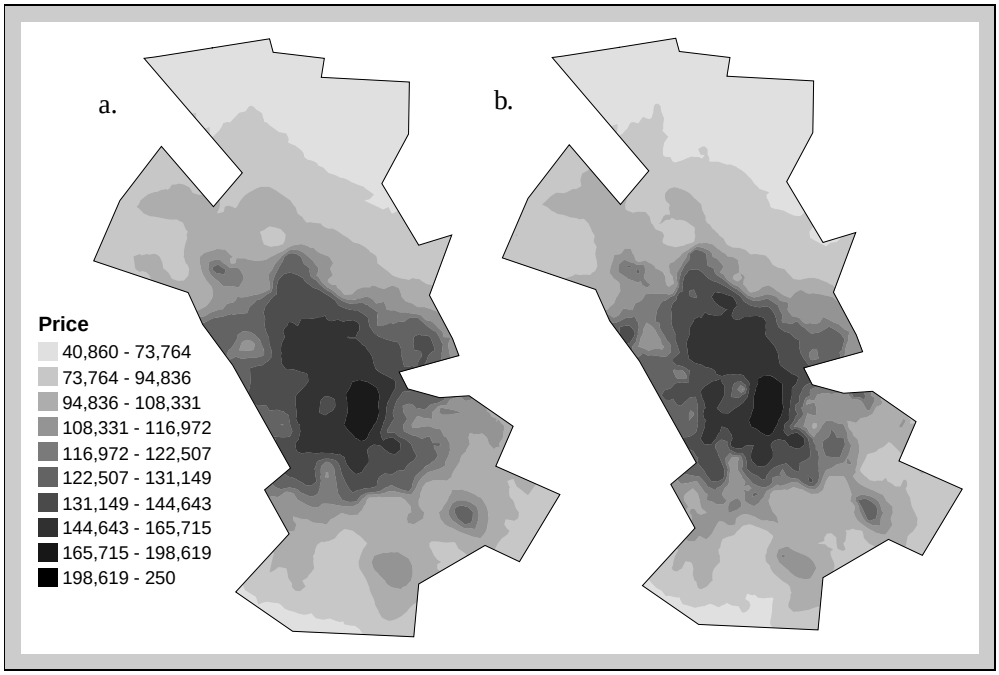
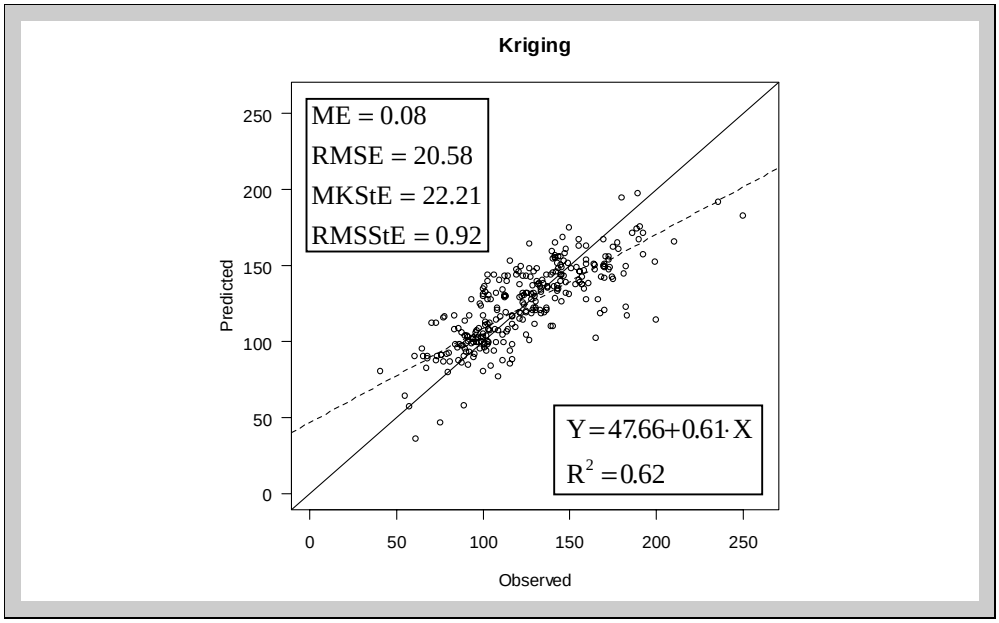
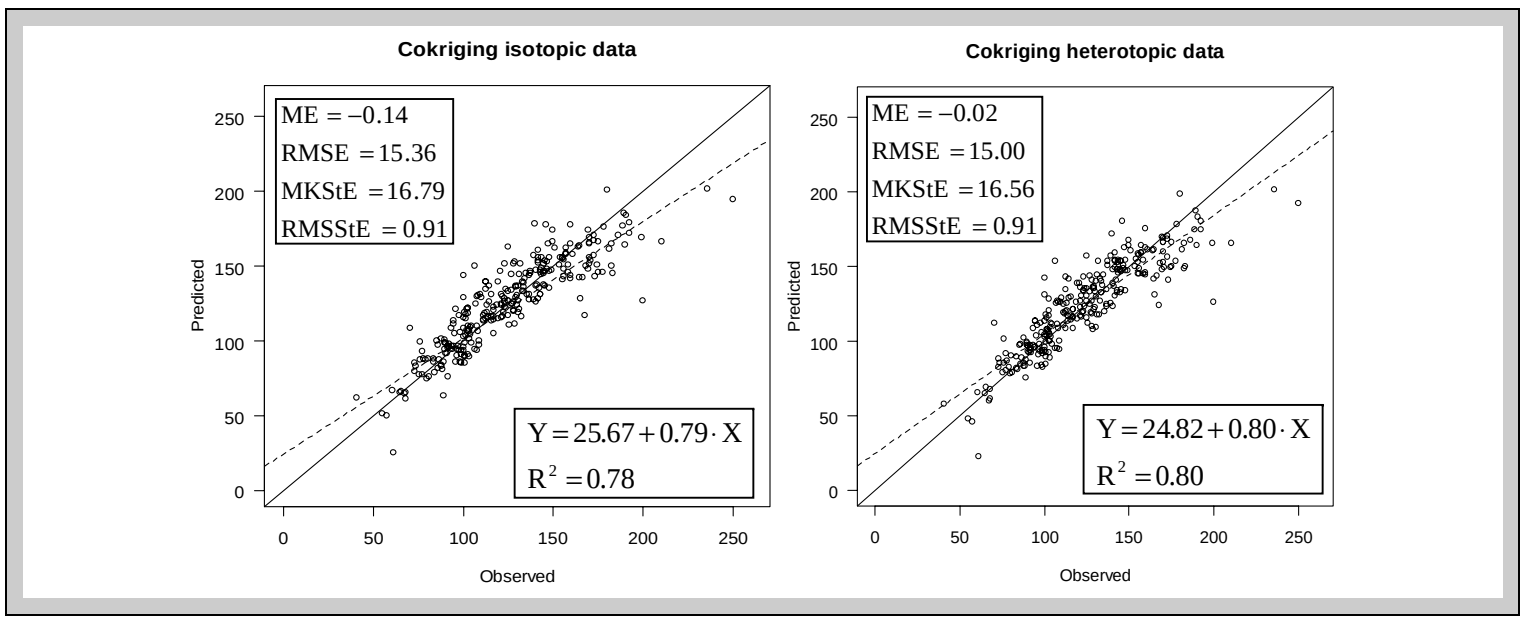


Exhibit 12 | Cross-Validation Kriging



Note: Dashed line represents regression line and solid line represents line 1:1.

Exhibit 13 | Cross-Validation Cokriging with Isotopic and Heterotopic Data



houses). Exhibit 11b shows the pricings map using cokriging with heterotopic data (original sample of 287 houses plus a second sample of 259 houses). On these maps it can be seen how the more highly valued areas coincide with the areas closest to the CBD—districts 10, 8, and part of districts 12, 11, and 9. In these districts, social, hospital, educational, and commercial services predominate. The least valued areas are, principally, in districts 1, 2, and 3, which are the most conflictive areas of the city. Greater spatial heterogeneity is observed in Exhibit 11b than in Exhibits 10 and Exhibit 11a. In Exhibits 11a and b, it can be seen how, in the center of district 13, there is an area with high prices. This is a newly constructed residential area with good views.

The cross-validation results are shown in Exhibits 12 and 13. In this application, the cross-validation shows little differences[7] between cokriging with isotopic and heterotopic data, although there are important differences between kriging and cokriging. Thus, the bias is almost nil for all three methods, although somewhat higher for the cokriging method with heterotopic data. The cokriging method has demonstrated a lower standard error than the kriging method. Furthermore, the cokriging method also shows a better fit and, therefore, the predicted values are closer to the observed values.

Exhibit 4 shows the errors obtained, using cokriging with heterotopic data, in each administrative district. It can be seen that districts 1 and 9 have a small number of homes and, furthermore, present greater average errors. However, districts 10, 12, and 13 contain a high concentration of houses and low average error values. In turn, districts 7, 9, and 11 have the highest standard deviation errors and, furthermore, show high values in the standard deviation of house price. Therefore, they are districts with heterogeneous prices. This latter aspect also means that the prices are heterogeneous, although the average error is only high in district 9, given the low number of houses.

Conclusion

This study suggests that using the cokriging method can be of interest for carrying out mass appraisal. Using this multivariate spatial method, continuous maps can be obtained of location price, which provide appraisers with an overall view of pricing. In the results obtained in the application presented, it is observed that this method provides better results than the kriging method.

Location price has been estimated using the kriging method, shown by the presence of spatial auto-correlation of house price, by means of $\hat{m}_k(s_0)$ (large-scale variation) and $\hat{u}_{1k}(s_0)$ (small-scale variation). Meanwhile, the cokriging method adds to this the location price of auxiliary variables such as age, amplitude and central heating, by means of $\hat{u}_{2ck}(s_0)$, $\hat{u}_{3ck}(s_0)$ and $\hat{u}_{4ck}(s_0)$, which are spatially correlated and co-regionalized to housing price.

An important characteristic of cokriging is that it can be applied when the house price and the auxiliary variables have not been sampled in the same housing

(heterotopic data). This is typical of most databases obtained from tax assessors, where only some of the houses have sales prices. Furthermore, this multivariate method enables house price appraisals to be made when only the location is known (i.e., no individual property characteristics are available).

Also, it is known that the houses that do not sell may be very different to those that do sell. The traditional hedonic regression, using only information of sold houses, can provoke a substantial bias to estimate the value of the entire housing stock (Gatzlaff and Haurin, 1997). Cokriging should be considered in future research to provide a way to deal with this problem.

Endnotes

[1] An iterative method to estimate $\hat{V}_k$ can be seen in Chica-Olmo (1995).

[2] See, for example, Cressie (1991) for the models of variograms and methods used to carry out the fit.

[3] For a generalization, see, for example, Wackernagel (1995).

[4] Significant changes in the house prices were not observed in these months.

[5] This variable is used instead of apartment size, since the correlation coefficient was greater.

[6] The dots of the omni-directional variogram (average-variogram) average out the estimates of $\gamma(h)$ at each distance.

[7] If the second sample would have been bigger, it is possible that the differences would be more significant.

References

Anselin, L. GIS Research Infrastructure for Spatial Analysis of Real Estate Markets. *Journal of Housing Research*, 1998, 9, 113–33.

Can, A. The Measurement of Neighborhood Dynamics in Urban House Prices. *Economic Geography*, 1990, 66, 254–72.

Case, B., J.M. Clapp, R.A. Dubin, and M. Rodriguez. Modeling Spatial and Temporal House Price Patterns: A Comparison of Four Models. *Journal of Real Estate Finance and Economics*, 2004, 29:2, 167–91.

Chica-Olmo, J. Spatial Estimation of Housing Prices and Locational Rents. *Urban Studies*, 1995, 32:8, 1331–44.

Chilès, J.P. and P. Delfiner. *Geostatistics. Modeling Spatial Uncertainty*. Wiley, 1999.

Clapp, J.M. How GIS Can Put Urban Economic Analysis on the Map. *Journal of Housing Economics*, 1997, 6, 368–86.

Clapp, J.M., A. Gelfand, and H. Kim. Predicting Spatial Patterns of House Prices Using LPR and Bayesian Smoothing. *Real Estate Economics*, 2002, 30:4, 505–32.

Cressie, N. *Statistics for Spatial Data*. John Wiley & Sons, 1991.

Deddis, W. Development of a Geographic Information System for Mass Appraisal of Residential Property. RICS Education Trust, 2002.

Dubin, R.A. Spatial Autocorrelation: A Primer. *Journal of Housing Economics*, 1998, 7, 304–27.

Dubin, R.A., J.K. Pace, and T.G. Thibodeau. Spatial Autoregression Techniques for Real Estate Data. *Journal of Real Estate Literature*, 1999, 7, 79–95.

Gatzlaff, D.H. and D.R. Haurin. Sample Selection Bias and Repeat-Sales Index Estimates. *Journal of Real Estate Finance and Economics*, 1997, 14:1/2, 33–50.

Goulard, M. and Voltz, M. Linear Coregionalization Model: Tools for Estimation and Choice of Cross-Variogram Matrix. *Mathematical Geology*, 1992, 24:3, 269–86.

Isaaks, E.H. and R.M. Srivastava. *An Introduction to Applied Geostatistics*. New York: Oxford University Press, 1989.

Journel, A.G. and C.J. Huijbregts. *Mining Geostatistics*. London: Academic Press, 1978.

LeSage, J.P. and J.K. Pace. Models for Spatially Dependent Missing Data. *Journal of Real Estate Finance and Economics*, 2004, 29:2, 233–54.

Matheron, G. La Théorie des Variables Regionalisées et ses Applications. Centre de Géostatistique et de Morphologie Mathematique. Fas. 1. 1965.

Neuman, S.P. and E.A. Jacobson. Analysis of Non-intrinsic Spatial Variability by Residual Kriging with Application to Regional Groundwater Levels. *Mathematical Geology*, 1984, 16, 499–521.

Pace, R.K., R. Barry, and C.F. Sirmans. Spatial Statistics and Real Estate. *Journal of Real Estate Finance and Economics*, 1998, 17:1, 5–13.

Tse, R.Y.C. Estimating Neighbourhood Effects in House Prices: Towards a New Hedonic Model Approach. *Urban Studies*, 2002, 39:7, 1165–80.

Wackernagel, H. *Multivariate Geostatistics*. Germany: Springer-Verlag, 1995.

The author would like to acknowledge the referees' comments, some of which have been included in the paper. In any case, all possible errors can be attributed solely to the author.

Jorge Chica-Olmo, Universidad de Granada, Granada, Spain or jchica@ugr.es.