

Kriging Methodology for Regional Economic Analysis: Estimating the Housing Price in Albacete

MATÍAS GÁMEZ MARTÍNEZ, JOSÉ MARÍA MONTERO LORENZO,
AND NOELIA GARCÍA RUBIO*

Because of the socioeconomic importance of the housing subsector in the local, regional, and national economy and its implications for housing policy, this paper attempts to analyze the spatial behavior of the free housing price in the city of Albacete. To achieve this aim, the authors have used the models and estimators imported from geology called kriging. To do this, it is necessary to know the spatial dependence structure of the process, which is shown in the variogram. (JEL O47; Int'l. Advances in Econ. Res., 6(3): pp. 438-450, Aug. 2000. ©All Rights Reserved.)

Introduction

The importance of geographic space and its incorporation to economic analysis, both in theoretical and empirical studies, is based on two main reasons: It is the natural support upon which many regional economic variables are measured and influenced and whose values show a spatial pattern of behavior. Traditionally, spatial distribution has not been taken into account in regional economic analysis. However, economic variables in time have been studied more. Spatial study was removed perhaps due to its greater complexity (countless ways in space as opposed only one way in time) and the lack of specific computer programs.

Today, both obstacles are being partially mitigated, first, by publication of studies that are more rigorous and homogeneous, accomplished by Matheron [1971], Cliff and Ord [1981], Anselin [1988, 1995], and Cressie [1993], who have carried spatial statistic-taking entity as a knowledge area. Second, these obstacles are being studied by using statistics computer programs (such as S-Plus, Variowin, and the like), which make working in spatial statistics easier. Third, the development of geographical information systems creates powerful tools for analyzing georeferenced variables.

The aim of this paper is to model the free housing price in the city of Albacete from a spatial perspective because it seems reasonable to consider that specific location of housing is influenced, among other factors, by adjacent housing prices (spatial diffusion phenomenon). For this purpose, this paper will use spatial linear models and the best linear unbiased estimators, called kriging estimators, on a geostatistic field [Krige, 1951; Goldberger, 1962; Matheron, 1962, 1963].

The theoretical methodology is precise since it estimates a spatial process in a point or region in space as a linear combination of a part of the spatial observations or as all of them.

* Universidad de Castilla La Mancha—Spain.

To do this, it is necessary to know the the spatial dependence structure of the process, which is shown in the variogram. Assuming the process verifies second order or intrinsic stationarity, then the ordinary or the universal kriging estimator can be calculated respectively. However, at this point, some important problems arise in practice because the real spatial dependence structure of the process is difficult to discover. These two essential problems are, first, how to estimate the parameters of the theoretical variogram model from the empirical variogram in order to optimize the fit, since the usual procedures are not valid, and second, how to forecast in the presence of spatial trend.

Theoretically, in this case, universal kriging could be applied. Nevertheless, in the presence of trend, it is impossible to estimate the theoretical variogram. Two solutions are proposed for the latter problem. The first solution deletes the trend previously estimated by median polish or a generalized additive model (GAM) and then applies ordinary kriging on the residuals. The other solution applies universal kriging iteratively until the theoretical variogram has been appropriately estimated.

Before applying this procedure to study the housing price in Albacete, the factors that decide the price are analyzed (that is, size, age, quality, and with or without garage). By deleting such effects and by reducing all the observations to a class of equivalent housing, four methods of modelization and estimation are used: universal kriging and ordinary kriging on the residuals of a GAM for the point estimation and universal kriging and median polish kriging for the block estimation. This last one is chosen as the best to model the spatial behavior of the housing price in Albacete.

Modelization of the Housing Price in Albacete

The importance of the housing sector in the Spanish economy, and particularly in Castilla La Mancha, comes from its share in the gross domestic product (more than 5 percent). The economic and physical qualities of Albacete make it a suitable case to apply a spatial study. It is worth pointing out that the geometric form of the city, practically circular, and the absence of relevant topographical unevenness allow the city to be considered as an ideal environment for spreading the spatial dependency equally in all directions (isotropy).

The empirical study starts with the description of the sampling. Then the results obtained from four different kriging methods are compared and analyzed.

The Sampling and Treatment of the Information

Available information was obtained by a sampling procedure from data supplied by real estate agencies due to the lack of official information about housing prices related to the spatial aspect. The sample contains very rich, wide-range information which includes houses of every type regarding age, quality, surface, and so forth. In other words, the sample is representative of the housing market in Albacete in 1997.

The initial database had 505 records with the following fields: street and number, zone, age, surface, quality, parking facilities, and total price of sale in current pesetas. Once the information has been collected, the next sequence was followed:

- 1) each house was exactly located on the city plan of Albacete;

- 2) the city expansion belt was delimited for predicting the expansion zones of the city contemplated in future general town planning;
- 3) age was recodified in a new ordinal variable (cod.age) with five modalities (in years): less than 1 (new), from 1 to 5, from 6 to 10, from 11 to 20, and more than 20;
- 4) surface was recodified in an ordinal variable (cod.surf) with five categories (in square meters): very small (surface below 50), small (50 to 70), medium (70 to 90), large (90 to 125), and very large (more than 125);
- 5) quality was studied by using five categories of quality: luxury, very good, good (intermediate), bad (for houses built with bad materials and in bad habitability and conservation conditions), and substandard (for new or semi-new houses built below good standards or formerly rated as bad but were renovated and improved); and
- 6) the square meter price was determined as the ratio between the total housing price and its useful surface.

Construction of the Equivalent Housing Class

Once the dataset has been selected, there are two options. The first option consists of filtering the available information, that is, to consider only those cases that fulfill some given characteristic relating to age, surface, and quality [Chica Olmo, 1994]. In doing so, the database used for the spatial modeling will be the most homogeneous possible. The principal disadvantage of this alternative is that much information is lost. This is not advisable if the size of the sample is not excessively large.

The other solution adopted in this investigation consists of referring all the house prices to a sole support in order to use all the available information in the sample. For this purpose, this paper accomplishes an analysis of the variance of the price by square meter, using as factors the surface, age, parking facilities availability, and quality of the housing. By eliminating these effects from the data, an equivalent housing class is obtained.

Analysis of Variance

The aim here is to estimate the effects of age, surface, quality, and parking availability on the housing sale price in Albacete. A multiplicative model has been used, assuming that the effect of any one of the previous factors is proportional to housing price. The model used is:

$$\log(\text{price}) \sim \text{cod.surf} + \text{cod.age} + \text{quality} + \text{parking} + \text{error} \quad .$$

Once the effects have been estimated and the appropriate transformations have been applied, the following expression is obtained:

$$\text{price.m}^2 \sim k \cdot e^{\beta_1 \text{cod.surf}} \cdot e^{\beta_2 \text{cod.age}} \cdot e^{\beta_3 \text{quality}} \cdot e^{\beta_4 \text{parking}} \cdot e^{\beta_5 \text{error}} \quad ,$$

where k is the mean price of the square meter filtered of all effects, and the remaining terms are the error and the antilogarithms of the estimated effects expressed as indices.

The previous model does not include different order interactions among the factors because the estimated effects were not significantly different from 0. Table 1 shows the estimated effects for all factors as well as their antilogarithms. The reference price, to which all the effects refer, is 72.078 pesetas and corresponds to a very small house (up to 50 square meters) more than 20 years old, of bad quality, and without parking. From this table it can be observed that:

- 1) The effect of the surface on the price decreases the size of housing increases. For example, for a very large house, this would be 79.36 percent of the price of a very small house.
- 2) The price of housing decreases as the age of housing increases. The maximum price is reached in the group of houses from 1 to 5 years old, with a very similar effect in the group of newest houses.
- 3) Quality is the most important factor. It reflects 58.6 percent of the difference in prices in the poor quality group of houses. To the luxury quality group, it reflects 130 percent difference in prices. Surprisingly, this factor is omitted by ministry and valuation companies when they do valuations.
- 4) The scarcity of public parking in some city areas means that, with the same characteristic of surface, age, and quality, a house with parking space is 16.4 percent more expensive, on average, than one with no parking.

Once the previous effects have been estimated, they are eliminated from the observations, transforming them into a house of reference, that is, a very small house more than 20 years old, of poor quality, and without parking space. This is called the equivalent house.

TABLE 1
Housing Price Factors: Analysis of Variance

Factors	Levels	Estimated Effects	Exp(Effects)
Constant		11.18551	72078.10000
Surface	Very Small	0	1
	Small	-0.07724	0.92567
	Medium	-0.16508	0.84782
	Large	-0.15951	0.85256
	Very Large	-0.23120	0.79358

TABLE 1 (CONT.)

Factors	Levels	Estimated Effects	Exp(Effects)
Age	> 20 Years	0	1
	11 to 20 Years	0.12353	1.13148
	6 to 10 Years	0.13373	1.14309
	1 to 5 Years	0.29364	1.34130
	< 1 Year	0.28124	1.32477
Quality	Bad	0	1
	Substandard	0.30054	1.35055
	Good (Standard)	0.46133	1.58618
	Very Good	0.67931	1.97252
	Luxury	0.83255	2.29918
Parking	No	0	1
	Yes	0.15157	1.16365

Only 355 records of initial data had information about quality. In order to use all the records for the spatial analysis, a classification tree¹ was applied to estimate the quality for the remaining cases [Breiman et al., 1984; Clark and Pregibon, 1992; Venables and Ripley, 1996]. The predictors are location, age, and surface. The classification tree is shown as:

```
tree(formula = quality ~ x + y + cod.age + cod.surf; data = datos;
na.action = na.omit; mincut = 5; minsize = 10; mindev = 0.01) ,
```

where the number of terminal nodes is 41; the residual mean deviance is $0.8392 = 234.1/279$, and the misclassification error rate is 0.1812. The proportion of misclassified cases is 18.12 percent, which seems acceptable.

Equivalent Information

Only 30 houses were rejected after the location process and were removed from the analysis, leaving 475 valid cases remaining. All the effects were filtered in this group. Thus, an estimation of the logarithm of price and the residue of each case were calculated. By manipulating them conveniently, the following expression was obtained:

$$price.m^{\square} = base.price * surface * age * quality * parking * resid \quad ,$$

where the reference price (the constant in the multiplicative model is 72.078 pesetas per square meters) and the residuals are known. The class of equivalent houses is obtained by multiplying the base price by the residuals:

$$\text{equivalent.price.m}^2 = \text{base.price} * \text{resid} .$$

The equivalent house database is now ready for applying kriging techniques.

Kriging on the Housing Price in Albacete

The objective here is to model the spatial behavior of the housing price in Albacete. First, point kriging will be carried out on the observations irregularly distributed on the plane. Alternatively, block kriging will be applied where observations are taken as the mean price calculated in different areas or blocks. Generally, these areas have an irregular form, but this paper uses squares from a regular grid. In particular, the following four methods have been applied: universal kriging and ordinary kriging on the residuals of a GAM [Hastie and Tibshirany, 1990; Venables and Ripley, 1996] for the point estimation and universal kriging and median polish kriging [Cressie, 1993] for the block estimation. The usual procedure can be summarized as follows:

- 1) estimate and eliminate the trend if the procedure requires it;
- 2) estimate the empirical variogram, removing anisotropies if they exist, from the original data or from the residuals with respect to the trend;
- 3) fit a theoretical variogram model from the empirical variogram;
- 4) estimate the corresponding kriging model (ordinary or universal);
- 5) predict the price and calculate the prediction error over 1,330 nodes of a regular grid which covers the city; and
- 6) analyze by cross-validation the goodness of fit for each model.

Model Selection and Results

Here, the relative advantages of the four methods will be compared briefly. Criteria will be the means of the prediction results, cross-validation, and measure errors.

Selection of Modeling and Prediction Method of the Housing Price

Regarding the spatial dependency structure of housing prices, the four procedures give similar results. This structure is spherical but shows different range, sill, and nugget effect. Thus, the price in a particular location mainly depends on the prices of the nearest houses. This influence decreases as the distance increases and vanishes when the distance range is reached. This kind of spatial dependence is called the neighborhood effect.

The four methods show similar price trends in Albacete. The highest prices are reached in the downtown area (such as Paseo de la Libertad, Altozano, Catedral, Plaza de la Mancha, and Calle Ancha), decreasing gradually from there to peripheral areas (such as Barriada de la Seiscientas, Carretera de Murcia, Alto de los Molinos, Hoya de San Ginés, Mortero de Pertusa, Carretera de Jaén, Barrio de San Pablo, Campollano, and the end of Paseo de La Cuba). A quadratic function was used to model the structure of the trend, as

previously explained, when applying universal kriging methods (for point and block data). The same form (quadratic function) is revealed for trend by kriging methods that eliminate this component, modeling it through a local regression surface, and median polish when point and block kriging are respectively applied.

When the prediction error is considered, the apparent similarity between methods disappears (shown in Table 2). When comparing prediction error from different methods, note that $\varepsilon(u)$, can be broken down into two second order stationary and incorrelated error processes:

$$\varepsilon(u) = e(u) + e_{\square}(u) \quad ,$$

where $e_{\square}(u)$ is a measure error process (nugget effect) and $e(u)$ is the prediction error process due exclusively to the model. Furthermore, the process covariance function verifies that:

$$\sigma_{\varepsilon}(u) = \sigma_{\square}(u) + \sigma_e(u) \quad ,$$

for whatever u and w locations. According to this breakdown and using all of the above-mentioned methods, prediction of the square meter price for equivalent housing class and the error (total and free of measure error) have been obtained.

TABLE 2
Total Prediction Error (Including Measure Error)

Kriging Methods	Minimum	Q_{\square}	Median	Mean	Q_{\square}	Maximum
Universal Point	10940	14760	16180	16390	18090	19250
GAM	11030	14460	15350	15280	16170	16230
Universal Block	8315	12560	13310	14030	15010	22730
Median Polish	6073	10290	11350	11270	12310	12320

The estimation of the measure error variance is different for each method and is equal to the nugget effect corresponding to each one of the fitted theoretical models of variogram. These effects, (expressed in $10^{\square x}$ pesetas) are shown in Table 3.

TABLE 3
Estimation of Measure Error Variance (Nugget Effect)

Kriging Methods	Estimate
Universal Point	0.016075
GAM	0.016896
Universal Block	0.012798
Median Polish	0.006380

It is observed that all are very similar except the median polish kriging model. Eliminating these effects, the prediction errors can be obtained due to the estimation method. A summary of the statistics is shown in Table 4.

TABLE 4
Prediction Error Due to the Estimation Model (Measure Error Free)

Kriging Methods	Minimum	Q_0	Median	Mean	Q_1	Maximum
Universal Point	5475	7572	10070	10170	12900	14490
GAM	4740	6338	8164	7922	9618	9726
Universal Block	4945	5450	7017	7981	9863	19720
Median Polish	5411	6494	8069	7906	9370	9387

Comparing the four models through cross-validation statistics (mean error (ME), mean quadratic error (MQE), adimensional mean quadratic error (AMQE), and theoretical mean error (TME), as shown in Table 5) or any of the previous prediction errors, the conclusions are as follows:

- 1) In point kriging as well as in block kriging, the best result is provided when the trend has been previously eliminated.
- 2) If the two models are compared for deciding whether or not to group the observations in blocks, block methods are chosen because they provide better MQE and AMQE statistics.²
- 3) Comparing kriging on residuals from GAM with median polish kriging, the errors due exclusively to the model (as shown in Table 4) are practically equal for both models,

but the cross-validation statistics favor the latter. The only disadvantage of this method is that its empirical residuals tails are not well fitted to the normality hypothesis.

From this discussion, it can be concluded that we prefer blocks as support for the observations, designing a regular grid that covers the city and averaging the observations in such grid blocks. At the same time, median polish kriging is used to predict the square-meter price of housing in Albacete because this method provides the best results in terms of cross-validation and total error of prediction (as shown in Table 5).

TABLE 5
Cross-Validation Statistics for All Estimation Methods

Kriging Methods	ME	MQE	AMQE	TME
Universal Point	-0.00027	0.02117	0.999462	0.146492
GAM	-0.00010	0.02092	1.010452	0.143707
Universal Block	-0.00074	0.01507	0.957974	0.128710
Median Polish	4.21874E-13	0.01364	0.951739	0.122727

Results

The results obtained by this method are graphically shown.³ Figure 1 shows a nonstationary variogram with a "cyclical in space" structure, similar to a quadratic trend. If trend is eliminated by the median polish method, the experimental variogram of the median polish residuals can be calculated (shown in Figure 2). This variogram is stationary and follows a spherical variogram model: range is 500 meters, sill is 0.009, and nugget effect is 0.007.

FIGURE 1
Omnidirectional Variogram of Housing Prices

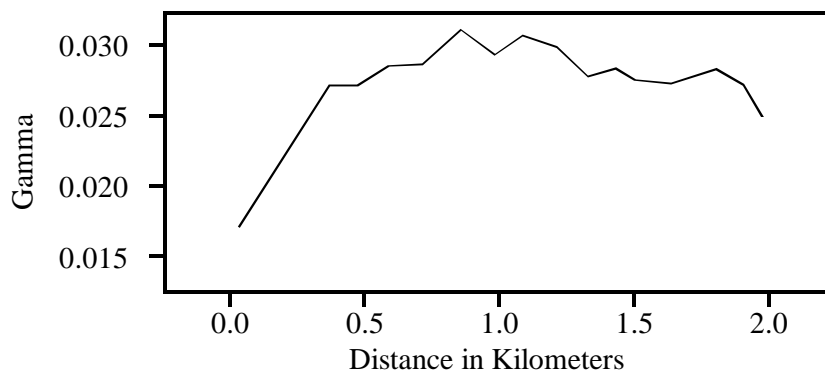
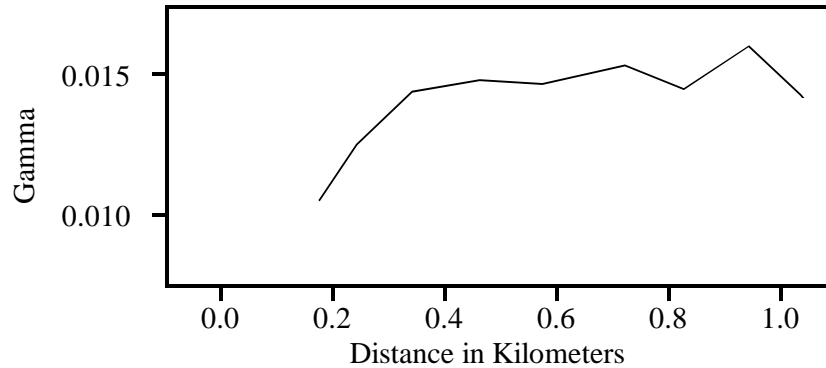
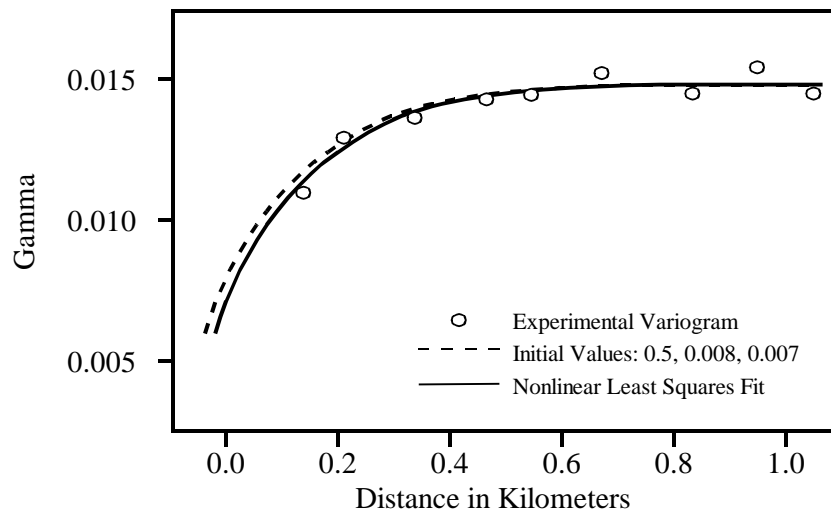


FIGURE 2
Omnidirectional Variogram of Median Polish Residuals



Taking these values as initial parameters to fit the spheric model by nonlinear least squares [Cressie, 1985], the fitted values of the variogram parameters are: range is 500.42 meters, sill is 0.008594, and nugget effect is 0.006377. Figure 3 shows the median polish experimental residuals and the initial and fitted theoretical variograms.

FIGURE 3
Fitting a Spherical Variogram on the Median Polish Residuals



Once the theoretical variogram model has been estimated, median polish kriging can be applied [Cressie, 1993]. The housing price prediction will be obtained as the sum of the trend, estimated by the Akima [1978] method, from the nodes in which the median polish trend has been obtained and from the ordinary kriging prediction of the residuals, in the 1,330 nodes of a regular grid whose squares are 100 meters by 100 meters. Finally, the

goodness of fit of the median polish kriging model is analyzed. To do this, its position and cross-validation statistics are summarized in Tables 6 and 7.

TABLE 6
Statistics of the Observed and Predicted Values (Cross-Validation)

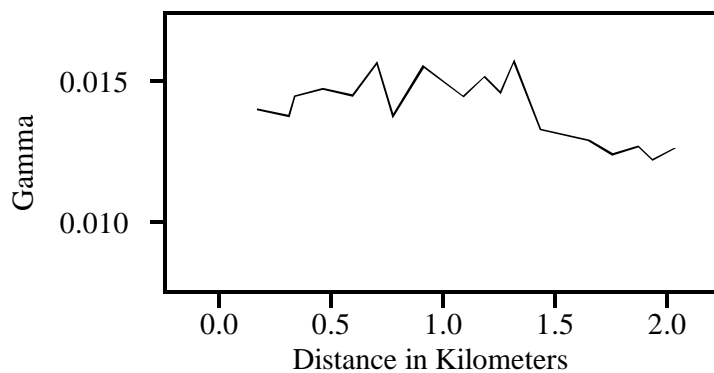
Values	Minimum	Q_1	Median	Mean	Q_3	Maximum
Observed	0.35100	0.61700	0.67930	0.70060	0.78080	1.19900
Predicted	0.42830	0.63890	0.70190	0.70060	0.75820	0.87790

TABLE 7
Cross-Validation Statistics from Ordinary Kriging on Median Polish Residuals

ME	MQE	AMQE	TME
4.21874E-013	0.013643235	0.9517394	0.1227271

Figure 4 shows the omnidirectional variogram of experimental errors. It is observed that there is no spatial dependence among residuals. It is concluded that the theoretical model has successfully explained the spatial dependence structure.

FIGURE 4
Variogram of Experimental Errors from Median Polish Kriging (Cross-Variation)



Footnotes

1. The learning method is applied in expert systems.
2. Cross-validation statistics are:

$$ME = \frac{1}{n} \sum e_{\bar{j}}; \quad TME = \frac{1}{n} \sum \sigma_{\bar{j}};$$

$$MQE = \frac{1}{n} \sum e_{\bar{j}}^2; \quad AMQE = \sqrt{\frac{1}{n} \sum \left(\frac{e_{\bar{j}}}{\sigma_{\bar{j}}} \right)^2},$$

where $e_{\bar{j}}$ is the experimental error and $\sigma_{\bar{j}}$ is the theoretic error due to the model.

3. The estimated values and the estimated errors can also be seen in Gámez [1998, Appendix C].

References

- Akima, H. "A Method of Bivariate Interpolation and Smooth Surface Fitting for Irregularly Distributed Data Points," *Association for Computing Machinery Transactions on Mathematical Software*, 4, 1978, pp. 148-59.
- Anselin, L. *Spatial Econometrics: Methods and Models*, Dordrecht, Netherlands: Kluwer Academic Publishers, 1988.
- Anselin, L.; Florax, R. J. G. M. *New Directions in Spatial Econometrics*, Heidelberg, Germany: Springer-Verlag, 1995.
- Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*, Monterey, CA: Wadsworth and Brooks, 1984.
- Chica Olmo, J. M. *Teoría de las variables regionalizadas: Aplicación en economía espacial y valoración inmobiliaria*, Granada, Spain: Cuadernos de la Universidad de Granada, 1994.
- Clark, L. A.; Pregibon, D. "Tree-Based Models," in J. M. Chambers; T. J. Hastie, eds., *Statistical Models in S*, 1992, Ch. 9.
- Cliff, A.; Ord, J. K. *Spatial Processes, Models and Applications*, London, United Kingdom: Pion, 1981.
- Cressie, N. A. C. "Fitting Variogram Models by Weighted Least Squares," *Journal of the International Association for Mathematical Geology*, 17, 1985, pp. 563-86.
- _____. *Statistics for Spatial Data*, revised, New York: Wiley, 1993.
- Gámez, M. *Nuevas técnicas de Estadística Espacial para la Economía: Modelización del precio de la vivienda libre en la ciudad de Albacete*, doctoral thesis, Universidad de Castilla La Mancha, 1998.
- Goldberger, A. S. "Best Linear Unbiased Prediction in the Generalized Linear Regression Model," *Journal of the American Statistical Association*, 57, 1962, pp. 369-75.
- Hastie, T. J.; Tibshirany, R. J. *Generalized Additive Models*, London, United Kingdom: Chapman and Hall, 1990.
- Krige, D. G. "A Statistical Approach to Some Basic Mine Valuation Problems on the Witwatersrand," *Journal of the Chemical, Metallurgical, and Mining Society of South Africa*, 52, 1951, pp. 119-39.
- Matheron, G. *Traité de Géostatistique Appliquée Tome I*, memoires, 14, Bureau de Recherches Géologiques et Minières, 1962.

- ___ . *Traité de Géostatistique Appliquée Tome II: Le Krigeage*, memoires, 24, Bureau du Recherches Geologiques et Minières, 1963.
- ___ . *The Theory of Regionalized Variables*, Paris, France: Cahiers Centre de Morphologie Mathématique, 1971.
- Venables, W. N.; Ripley, B. D. *Modern Applied Statistics with S-Plus*, 3th ed., New York, NY: Springer-Verlag, 1996.