NEW STATISTICAL TECHNOLOGIES APPLIED TO THE ESTIMATION OF FREE HOUSING PRICE: KRIGING.

José María Montero Lorenzo (e-mail: jmlorenzo@jur-to.uclm.es).
Matías Gámez Martínez (e-mail: gamez@ecem-ab.uclm.es).

ABSTRACT.

The aim of this paper is to analyse the spatial behaviour of the free housing price in Albacete. We decided to write it because of the socieconomic importance of the housing subsector in the local, regional and national economy and its implications for Housing Policy.

To achieve this aim, we have used the models and estimators imported from Geology, called kriging. In the geostatistics bibliography, 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 all of them. In order to do this, it is neccesary to know the structure of spatial dependence of the process which is shown in the variogram and, assuming that the process verifies second order or intrinsic stationarity, then the ordinary or the universal kriging estimator can be calculated respectively.

Before applying this procedure to study the housing price in Albacete, the factors which settle their price are analysed (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 generalized additive model for the point estimation, and universal kriging and median polish kriging for the block estimation. We select the last one as the best one to model the spatial behaviour of the housing price in Albacete.

Keywords: Kriging prediction, spatial model, housing price.

1.- 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 a large of regional economic variables are measured.
- Its influence on these variables, whose values show a spatial pattern of behaviour.

Traditionally, spatial distribution has not been taken into account in regional economic analysis. However, economic variables in time have been more studied, perhaps removing spatial study due to its greater complexity (countless ways in space against only one way in time) and the lack of specific computer programs.

Nowadays, both obstacles are being partially mitigated. Firstly by publication of studies more rigorous and homogeneous, the efforts accomplished by Matheron (1970), Cliff and Ord (1981), Anselin (1988, 1995) and Cressie (1993), who have carried Spatial Statistic taking entity as a knowledge area. Second by the appearance of statistics computer programs (S-plus, Variowin, etc.) which make working in Spatial Statistics easier. Finally, by the development of Geographical Information Systems (G.I.S), which constitutes a powerful tool for the analysis of georeferenced variables.

The aim of this paper is to model the free housing price in 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, we will use spatial linear models and the Best Linear Unbiased Estimators (BLUE), called kriging estimators on Geostatistic field (Krige (1951), Goldberger (1962), Matheron (1952,1963)).

2.- SPATIAL DATA MODEL.

A spatial data collection can be considered as a stochastic process or random function realization

$$y(\boldsymbol{u}), \quad \boldsymbol{u} \in D \subset \mathbb{R}^d.$$

where **u** is a location in D domain, contained in a *d* dimension space (generally 1,2 or 3). For each location $\boldsymbol{u}$, $y(\boldsymbol{u})$ is a random variable. To carry out this analysis, we must model the spatial stochastic process $y(\boldsymbol{u})$. Generally, inference is not possible based on only one stochastic process realization $y(\boldsymbol{u})$. Therefore we will set some restrictions in terms of stationarity hypothesis. We will assume a certain regular spatial behaviour for the first moments process or their increments.

We will suppose that, for each $\boldsymbol{u}$, first and second order moment exist and they are finite, then we have that E[$y(\boldsymbol{u})$] and var[$y(\boldsymbol{u})$]exist and $y(\boldsymbol{u})$ can be separated into:

$$y(\boldsymbol{u}) = m(\boldsymbol{u}) + e(\boldsymbol{u})$$

where $m(\boldsymbol{u})$ is a function that represents $y(\boldsymbol{u})$ mean

$$E[y(\boldsymbol{u})] = m(\boldsymbol{u}),$$

and $e(\boldsymbol{u})$ is a stochastic error process with

$$E[e(\boldsymbol{u})] = 0.$$

In particular, $e(\boldsymbol{u}) = y(\boldsymbol{u}) - m(\boldsymbol{u})$. For our purposes, we must model $m(\boldsymbol{u})$ and $e(\boldsymbol{u})$. More formally, if we assume a linear structure for mean m(u), we obtain what is known in Geostatistic as "universal kriging model". The existence of p known functions, $x_1(\boldsymbol{u}),..., x_p(\boldsymbol{u})$ is supposed in such a way that

$$m(\boldsymbol{u}) = \sum_{j=1}^{p} \beta_j x_j(\boldsymbol{u})$$

for certain unknown fixed parameters $\beta_1,...,\beta_p$.

We can distinguish two particular cases:

a) Ordinary kriging in which the mean of the process is constant.
b) Simple kriging in which the mean of the process is constant and known.

We will obtain the most general estimator (Universal kriging). Furthermore, some small changes will provide us with the estimator in the other cases.

For modelling error process, we will be especially interested in second order error properties. The covariance function is defined as:

$$\sigma(\boldsymbol{u},\boldsymbol{w}) = \mathrm{cov}[e(\boldsymbol{u}),e(\boldsymbol{w})].$$

Using the following notation

$$\begin{aligned}
y_i &= y(\boldsymbol{u}_i), \quad i = 0,...,n, \\
\boldsymbol{y} &= (y_1,...,y_n)', \\
x_{ij} &= x_j(\boldsymbol{u}_i); \quad i = 0,...,n; \; j = 1,...,p \\
\boldsymbol{x}_i' &= (x_{i1},...,x_{ip}), \quad i = 0,...,n, \\
\boldsymbol{X} &= \begin{bmatrix} \boldsymbol{x}_1' \\ \vdots \\ \boldsymbol{x}_n' \end{bmatrix}, \\
\beta &= (\beta_1,...,\beta_p)', \\
e_i &= e(\boldsymbol{u}_i)
\end{aligned}$$

and

$$\boldsymbol{e} = (e_1,...,e_n)'$$

the universal kriging model can be written on matrix form:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{e};$$

$$\mathrm{E}[\boldsymbol{e}] = \boldsymbol{0},$$

$$\mathrm{cov}(\boldsymbol{e}) = \boldsymbol{\Sigma}$$

where $\boldsymbol{\Sigma} = (\sigma_{ij})$ y $\sigma_{ij} = \sigma(\boldsymbol{u}_i, \boldsymbol{u}_j)$.

Generally, the matrix $\boldsymbol{\Sigma}$ is not singular and the functions $x_j(\boldsymbol{u})$ and locations $\boldsymbol{u}$ are taken so that $\boldsymbol{X}$ has maximum range, by columns, with $J \in C(X)$. Under maximum range hypothesis, since $\boldsymbol{\beta}$ can be estimated for any location $\boldsymbol{u}_0$, then $\boldsymbol{x}_0'\boldsymbol{\beta}$ can also be estimated. Let

$$\boldsymbol{\Sigma}_{\boldsymbol{Y0}} = \begin{pmatrix} \sigma(\boldsymbol{u}_1, \boldsymbol{u}_0) \\ \vdots \\ \sigma(\boldsymbol{u}_n, \boldsymbol{u}_0) \end{pmatrix}.$$

The best linear unbiased predictor for y_0 is (Christensen, 1991):

$$\hat{y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}} + \delta'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}$ y $\delta = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y0}}$, that can be written as:

$$\hat{y}_0 = \boldsymbol{b}'\boldsymbol{y}$$

where $\boldsymbol{b}' = \boldsymbol{x}_0(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1} + \delta'(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1})$.

The mean of the squared prediction errors (variance of prediction) is:

$$\mathrm{var}(y_0 - \hat{y}_0) = \sigma(\boldsymbol{u}_0, \boldsymbol{u}_0) - 2\boldsymbol{b}'\boldsymbol{\Sigma}_{\boldsymbol{Y0}} + \boldsymbol{b}'\boldsymbol{\Sigma}\boldsymbol{b}\,.$$

Kriging estimator properties.

The most important properties of this estimator are:

1.- The kriging estimator is linear, unbiased and optimum (BLUE).

2.- The kriging estimator is exact. That is to say that if $\boldsymbol{u}_0$ coincides with the location of some experimental points $\boldsymbol{u}_i$ (points which we have information), then the estimated value equals observed value and the estimation error variance is null.

3.- The kriging method as spatial estimation instrument is fundamentally characterized by taking into account the geometry or spatial structure of data.

4.- The cross validation method allows us to evaluate experimental errors of

estimation. It consists of fictionally suppressing each observation y_i and estimating it using kriging. In this way we obtain the estimation of experimental errors $e_i = \hat{y}_i - y_i$ The analysis of these errors through different statistics provides a goodness-of-fit measure.

Covariance function and variogram estimation

All information concerning spatial dependency comes from covariance function or variogram. Since both of them belong to universal kriging estimator expressions, we must estimate these functions by fitting them to useful practical models. These models are spherical, gaussian, linear, exponential and potential, which can be estimated by maximum likelihood, restricted maximum likelihood, minimum quadratic norm or traditional geostatistics method (method of moments). The first two methods can only be applied in the presence of normal residuals.

In this way, we obtain an experimental estimated variogram, but only for a few distances. Therefore, in order to predict we need a theoretical model for covariance function or variogram in such a way that we can evaluate these functions for distances on which we usually have no information.

3.- METHODOLOGICAL CONSIDERATIONS.

The theoretical side of methodology is quite clear, since the aim is to estimate a spatial process in a point or set of points of space as a linear combination of all (or part) of the observations. So it is enough to know the spatial dependency structure of the process integrated in the variogram and, if the process is stationary of second order or intrinsically stationary (second order stationary of increments), then it is possible to calculate the ordinary or universal kriging estimator, respectively, to obtain the prediction. But just at this point the main problems of a practical type arise, due to difficulties in knowing the real structure of the spatial variability of the process.

There are two main problems. First, how to estimate the parameters of a theoretical variogram from empirical variogram, so that the fitting is accomplished in an optimum way and taking into account that the usual fitting procedures are invalid. Second, how to do spatial processes prediction in the presence of spatial drift or trend. In theory, if the process is intrinsically stationary we should apply universal kriging, but it can not be done because it is impossible to estimate directly the theoretical variogram in trend presence.

Theoretical variogram estimation.

The variogram appears in kriging equations whenever we make predictions or calculate their variance.The variogram function is defined for any distance between locations.Since it may be location pairs for which we have no information, we must estimate the theoretical model from the empirical variogram.

When we have calculated the experimental variogram (having amended any possible anisotropy), we decide the kind of theoretical model we are going to use. Then, the fitting can be carried out in two ways:

1) By an “interactive way”. Programs such as Variowin or Splus Spatial Stats allow us an iterative estimation of theoretical variograms starting from some initial values and reducing the error of estimation until it reaches a satisfactory level.

2) Estimating by “non-linear least squares”. We can fit the theoretical model by optimization, usually through non-linear least squares. The usual statistical hypothesis for non-linear regression models are invalid to fit variograms since the different values of variogram function are not independent. Cressie (1985) describes procedures based on generalized least squares and weighted least squares for variogram fitting, that take into account the particular structure of the variogram values. Zimmerman and Zimmerman (1991) compares different estimation methods and concludes that estimation by non-linear or weighted (through some specific function) least squares can be considered valid for most of the more sophisticated and computationally intensive methods.

Cressie (1985) suggests the following weighted non-linear least squares estimator:

$$\sum_{\boldsymbol{h}} |N(\boldsymbol{h})| \cdot \left\{ \frac{\gamma(\boldsymbol{h})}{\gamma(\boldsymbol{h};\cdot)} - 1 \right\}^2$$

where $|N(\boldsymbol{h})|$ is the number of different pairs at distance $\boldsymbol{h}$, and the sum is computed for every distance in which the empirical variogram has been calculated. $\gamma(\boldsymbol{h})$ is the value for the empirical variogram on the distance $\boldsymbol{h}$, and $\gamma(\boldsymbol{h};\cdot)$ is the theoretical variogram with unknown parameters. These weights can be easily included in the non-linear least squares procedure (nls) through a function for the residuals that incorporates the previous weights.

Estimation and prediction under spatial trend.

Under spatial trend, some directional variogram, and the omnidirectional, may be non-stationary . Their analysis will show us the direction or directions in which some type of trend is manifested. Since the process mean is not constant, we can not apply the ordinary kriging procedure. In this case, we can follow two different ways. First, to eliminate the trend by obtaining a stationary process. Second, to integrate the trend in the model and use it to make spatial predictions. We will call these procedures “kriging on the residuals” and “iterated universal kriging”, respectively.

In the first case, there are some procedures to estimate the trend. It is worth pointing out the median-polish procedure and other estimation methods related to polynomic regression and generalized additive models.

Median-polish: This is a non-parametric procedure for estimation of the trend, that is to say, it does not impose any restriction on its functional form. Starting from the spatial process

$$\boldsymbol{y(u) = m(u) + \epsilon(u)},$$

when its mean $m(\cdot)$ is unknown and it is not constant, it can be estimated supposing that it can be broken down as the sum of its directional components. In the most usual case, i.e. in $\mathbb{R}^2$, this is expressed as:

$$m(\boldsymbol{u}) = a + c(x) + r(y), \quad \boldsymbol{u} = (x,y)' \in D \subset \mathbb{R}^2$$

where a is the global effect, $c(\cdot)$ is the column effect, $r(\cdot)$ is the row effect and the points $\{\boldsymbol{u}_i: i=1,...,n\}$ are located on the nodes of a mesh or grid of dimension pxq (p rows and q columns): $\left\{ (x_l, y_k)' : l = 1,...,q; k = 1,...,p \right\}$ in such a way that $\boldsymbol{u}_i = (x_l, y_k)'$ and $m(\boldsymbol{u}_i) = a + r_k + c_l$.

The row effect can be estimated through an analogous procedure by using medians instead of means. The estimated principal effects, under very general conditions, minimize the norm L_1 (absolute value norm)

$$\sum_{k=1}^{p} \sum_{l=1}^{q} | Y_{kl} - a - r_k - c_l |$$

and the procedure converges.

When calculating spatial trend in locations without observations, we will not have any explicit formulation to calculate it, as we have not imposed any restriction on the functional form of the trend.In this case, we can estimate interpolation planes for each four points of the median-polish trend or use other more sophisticated procedure called "Akima method" or "spline surfaces method". This method is valid for regular grids as well as for regularly distributed data in the plane. It consists of fitting polynomic surfaces of small degree, the minimum possible, to the regions obtained through the triangulation of Delaunnais, so that the surfaces will be differentiable on edges (obtained from surfaces intersection). Thus, we get a soft surface from composition of all the small surfaces, called "spline surfaces". Since the form of each spline surface is known, we can interpolate the trend in every point of the spatial domain. Extrapolation by any one of the previous methods can not be done unless we make some hypothesis about the behaviour of the surface at domain limit.

Generalized additive models (G.A.M.): When the data are not regularly distributed on the plane, there are other procedures to estimate the trend.

These methods are: polinomic regression, generalized least squares and local regression surfaces, that are particular cases of "generalized additive models". In all of them it is hypothesized that trend follows the form:

$$\tilde{m}(\boldsymbol{u}) = \tilde{m}\left((x,y)\right) = \sum_i a_i f_i(x,y)$$

where f_i can be any type of function, for example polynomic functions, trigonometrical functions, etc. The main problem of this method is that it relies on the hypothesis of independence of the data, therefore the results must be handled with care.

Using global trend surfaces implies the following difficulties: the variability of extrapolated values of the trend depend critically on how far we have moved from the observational domain. One solution consists of fitting a polynomic surface in each point using only the nearby points. This is known as "local regression", obtained, generally, by weighted least squares, giving more weight to the nearest points.

Ordinary Kriging on the residuals.

Once we have estimated the trend, we can express the spatial process as

$$y(\boldsymbol{u}_i) = \tilde{m}(\boldsymbol{u}_i) + R(\boldsymbol{u}_i); \qquad \boldsymbol{u}_i = (x_k, y_l)'$$

where the residuals $\{R(\boldsymbol{u}_i): i=1,...,n\}$ are a set of spatial data devoid of trend, and they are ready to have ordinary kriging procedure applied to them.

We can fit a theoretical variogram with these residuals, thus applying ordinary kriging equations we get:

$$\hat{R}(\boldsymbol{u}_0) = \sum_{i=1}^{n} \lambda_i R(\boldsymbol{u}_i); \qquad \boldsymbol{u}_0 \in \mathbb{R}^2.$$

This estimation approximates the unknown real errors $\{\epsilon(\boldsymbol{u}_i): i=1,...,n\}$. We can obtain an estimation of the trend $\tilde{m}(\boldsymbol{u}_0)$ using any one of the previous procedures. Thus, now for every point on the plane we have two estimations. One for the trend and the other for the residual $\hat{R}(\boldsymbol{u}_0)$. Combining both, we get the kriging predictor based on the residuals, $y(\boldsymbol{u}_0)$, defined as follows:

$$\tilde{y}(\boldsymbol{u}_0) \equiv \tilde{m}(\boldsymbol{u}_0) + \tilde{R}(\boldsymbol{u}_0); \qquad \boldsymbol{u}_0 \in \mathbb{R}^2.$$

This estimator is called "median-polish kriging" in the event that we have estimated trend through median-polish.

This predictor is also an exact interpolator, i.e., $\tilde{y}(\boldsymbol{u}_i) = y(\boldsymbol{u}_i); \quad i=1,...,n.$

The variability of this estimator $\tilde{y}(\boldsymbol{u}_0)$ comes only from the ordinary kriging predictor applied to the residuals. This variability is measured by the variance of the ordinary kriging estimator and we denote it by $\sigma^2_m(\boldsymbol{u}_0)$.

Iterated Universal Kriging.

The second method consists of applying universal kriging, that estimates both the trend, generally polynomic, and the process in the point or points. The main problem of this method is the impossibility of correctly estimating variogram unless we have eliminated trend previously. And we can not estimate trend by usual procedures if the data are correlated. One solution is to apply universal kriging iteratively following the next algorithm:

Step 1. Apply universal kriging using empirical variogram of the original data as first approximation of the theoretical variogram.

Step 2. Once trend has been eliminated using estimations from step 1, we obtain the residuals.

Step 3. Compute empirical variogram using the residuals (step 2) and estimate the

theoretical variogram.

Step 4. Apply universal kriging using former estimation of theoretical variogram as input.

Step 5. Return to step 2 and repeat the process until the results do not vary substantially.

As this algorithm concerns only the estimation of theoretical variogram, the variance of the error of estimation will be the one given by the universal kriging method.

There are no general results for comparing the precision of the two previous procedures. In cases in which they have been applied (Cressie, 1993) the obtained predictors have been almost identical, the mean of squared error being smaller in median-polish-kriging. The same conclusion has been reached in the application to the case of the real-estate market in the case of Albacete city.

4.- MODELIZATION OF 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 GDP (more than 5 %). The economic and physical qualities of Albacete city make it a suitable case to apply a spatial study. It is worth pointing out the geometric form of the city, practically circular, and the absence of relevant topographical unevenness that allow us to consider the city as an ideal environment for the spread of 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, finally, they are analysed.

SAMPLING AND TREATMENT OF THE INFORMATION.

Available information was obtained by a sampling procedure over the 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 information covering a wide range which includes houses of every type from the point of view of: age, quality, surface, and so forth. In other words, the sample representativeness is in accordance with the housing-market in Albacete city in 1997.

The initial data base had 505 records with the following fields:

- street and number,
- zone,
- age, expressed in years,
- surface, in square meters,
- quality, different levels according to the Agencies,
- parking facilities, if it possesses it or not, and
- total price of sale, in current pesetas.

Once the information has been collected, the next sequence was followed:

a) Each house was exactly located on the city plan of Albacete.

b) The city expansion belt was delimited for predicting in the expansion zones of the city contemplated in the future P.G.O.U. (General Town Planning).

c) Age was recodified in a new ordinal variable (cod.age) with the following modalities:

- new, for the finished and delivered houses in the last year,
- from 1 to 5, for the houses that were delivered more than a year ago and less than five,
- from 6 to 10, for those six to ten years old,
- from 11 to 20, for those eleven to twenty years old, and
- more than 20.

d) Surface recoding. Surface variable was recodified in an ordinal variable, called "cod.surf", with five categories:

- very small, surface below 50 m^2,
- small, 50 - 70 m^2,
- medium, 70 - 90 m^2,
- large, 90 - 125 m^2, and
- very large, more than 125 m^2.

e) Treatment of the quality. We have worked with five categories of quality: intermediate level "good" corresponding to the standard quality; two categories above it, "very good", comparable to the new houses of first quality, and "luxury"; and two lower categories, "bad" for old houses built with bad material and in very bad habitability and conservation conditions, and "substandard" for new or semi-new houses that were built below the standards of quality of the houses classified as good or houses of the previous type that have been renovated and improved.

f) Determination of square meter price, as the ratio between the total housing price and its useful surface.

CONSTRUCTION OF THE EQUIVALENT HOUSING CLASS.

When we select the data set which will be used in the analysis, we have two options. The first consists of filtering the available information, that is to say, to consider only those cases that fulfil some given characteristics relating to age, surface and quality (Chica Olmo (1994)). In doing so, the data base used for the spatial modelling will be the most homogeneous possible. The principal disadvantage of this alternative is that much information is lost, which is not advisable if the size of the sample is not excessively large, as in our case.

The other solution which is 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, we have accomplished an analysis of the variance of the price by square meter using as factors the surface, the age, the possession or not of parking facilities and the quality of the housing. In this way we obtain the corresponding effects at different levels of the previous factors referenced to a housing pattern. Eliminating these effects from the data, we obtained their reduction to an "equivalent housing class".

Analysis of variance. The aim here is to estimate the effects on housing sale price in

Albacete city of age, surface, quality and parking possesion. Therefore, we will be able to eliminate them reducing data to homogeneous support. We used a multiplicative model assuming that the effect of any one of the previous factors is proportional to housing price. The model used was:

$$log(price) \sim cod.surf + cod.age + quality + parking + error$$

Once the effects have been estimated and the appropriate transformations applied, we will obtain the following:

$$price.m2 \sim k \cdot e^{cod.surf} \cdot e^{cod.age} \cdot e^{quality} \cdot e^{parking} \cdot e^{error}$$

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 zero.

The following table shows the estimated effects for all factors as well as their antilogarithms.The "reference price", to which all the effects refer, is 72.078 ptas. and corresponds to a very small house (up to 50 m^2) more than twenty years old, of bad quality and without parking.

Factors	Levels	Estimated efects	exp(efects)
Constant		11.18551	72078.1
Surface	very small	0	1
	small	-0.07724	0.92567
	medium	-0.16508	0.84782
	big	-0.15951	0.85256
	very big	-0.2312	0.79358
Age	more than 20	0	1
	from 11 to 20	0.12353	1.13148
	from 6 to 10	0.13373	1.14309
	from 1 to 5	0.29364	1.3413
	new	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

Table 1: Housing price factors (analysis of variance).

From the table above we observe:

a) The effect of the surface on the price decreases the size of housing increases, being for a very large house 79,36% of the price corresponding to a very small house.

b) 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 ones.

c) The quality is the most important factor. It explains 58,6% of the difference of prices in the poor quality group of houses. For the luxury houses group, differences in prices due to different qualities reach 130%. Surprisingly, this factor is omitted by Ministry and Valuation Companies when they do valuations.

d) 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% more expensive, on average, than the one which does not possess it.

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 of more than 20 years, poor quality and without parking space. We call it "equivalent house".

Only 355 records of initial data had information about quality. In order to use all the records for the spatial analysis we estimated the quality for the remaining cases through a *classification tree* (Breiman et al.(1984), Clark & Pregibon (1992), Venables & Ripley (1996)), a learning method used in expert systems. The predictors are location, age and surface:

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

Number of terminal nodes:  41
Residual mean deviance:  0.8392 = 234.1/279
Misclassification error rate: 0.1812
```

The proportion of misclassified cases is 18,12%, which seems acceptable.

Equivalent Information. Only 30 houses were rejected after the location process and they were removed from the analysis, 475 valid cases remaining. All the efects were filtered in this group. Thus we obtained an estimation of the logarithm of price and the residue of each case. By manipulating them conveniently, we have:

$$price.m2 = base.price * surface * age * quality * parking * resid$$

where we know the reference price (the constant in the multiplicative model: 72.078 ptas/m^2) and the residuals. The class of equivalent houses is obtained by multiplying the base price times the residuals:

$$equivalente.price.m2 = base.price * resid$$

Then, the equivalent houses data base is ready for applying kriging techniques.

KRIGING ON HOUSING PRICE IN ALBACETE.

Our objective is to model the spatial behaviour of housing price in Albacete city. First, we will carry out point kriging on the observations irregularly distributed on the plane. Alternatively we also apply block kriging, where observations are taken as the mean price calculated in different areas or blocks. Generally, these areas have an irregular form but we have used squares from a regular grid. In particular, we will apply the following four methods: universal kriging and ordinary kriging on the residuals of a generalized additive model (Hastie & Tibshirany (1990), Venables & 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 1330 nodes of a regular grid which covers the city.
6) Analyse by cross validation the goodness-of-fit for each model.

MODEL SELECTION AND RESULTS.

In this section, we will compare briefly the relative advantages of the four methods by taking as criteria the means of the prediction results, cross validation and measure errors.

A) Selection of modelling and prediction method of the housing price. With regard to the spatial dependency structure of housing prices, the four procedures give us similar results. This structure is spherical, but show different range, sill and nugget effect. Thus, the price in a particular location depends, mainly, 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 “neighborhood effect”.

The four methods show similar price trends in Albacete. The highest prices are reached in the downtown area (Paseo de la Libertad, Altozano, Catedral, Plaza de la Mancha, Calle Ancha,...) decreasing gradually from there to peripheral areas (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, end of Paseo de La Cuba,...). A quadratic function was used to model the structure of the trend, 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, modelling it through a local regression surface and median polish when we use point and block kriging respectively.

When we consider the prediction error, the apparent similarity between methods disappears (table 2). When comparing prediction error from different methods, we must bear in mind that $\varepsilon(u)$, can be broken down into the sum of two second order stationary and

incorrelated error processes,

$$\varepsilon(\boldsymbol{u}) = e(\boldsymbol{u}) + e_M(\boldsymbol{u})$$

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

$$\sigma_\varepsilon(\boldsymbol{u}) = \sigma_e(\boldsymbol{u}) + \sigma_M(\boldsymbol{u})$$

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

Total prediction error (includes measure error)						
Universal Point Kriging	Minimum	Q_1	Median	Mean	Q_3	Maximum
	10940	14760	16180	16390	18090	19250
G.A.M. Kriging	Minimum	Q_1	Median	Mean	Q_3	Maximum
	11030	14460	15350	15280	16170	16230
Universal Block Kriging	Minimum	Q_1	Median	Mean	Q_3	Maximum
	8315	12560	13310	14030	15010	22730
Median-Polish Kriging	Minimum	Q_1	Median	Mean	Q_3	Maximum
	6073	10290	11350	11270	12310	12320

Table 2: Prediction error for every estimation method.

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^{10} pesetas) are:

Estimation of measure error variance (nugget effect)	
Universal Point Kriging	0.016075
G.A.M. Kriging	0.016896
Universal Block Kriging	0.012798
Median-Polish Kriging	0.00638

Table 3: Nugget effect estimation.

We see that all of them are very similar except from the median-polish kriging model. Eliminating these effects, we can obtain the prediction errors due to the estimation method.A summary of the statistics is shown in table 4.

Prediction error due to the estimation method (measure error free)						
Universal Point Kriging	Minimum	Q_1	Median	Mean	Q_3	Maximum
	5475	7572	10070	10170	12900	14490
G.A.M. Kriging	Minimum	Q_1	Median	Mean	Q_3	Maximum
	4740	6338	8164	7922	9618	9726
Universal Block Kriging	Minimum	Q_1	Median	Mean	Q_3	Maximum
	4945	5450	7017	7981	9863	19720
Median-Polish Kriging	Minimum	Q_1	Median	Mean	Q_3	Maximum
	5411	6494	8069	7906	9370	9387

Table 4: Prediction error (measure error free) for every estimation method.

Comparing the four models through cross validation statistics (Table 5), or any of the previous prediction errors, we can conclude that:

- In point kriging as well as in block kriging, the best result is provided when we previously have eliminated the trend.

- If we compare the two models for deciding whether or not to group the observations in blocks, we choose block methods because they provide better MQE and AMQE statistics.[1]

- Comparing kriging on residuals from generalized additive model with median-polish kriging, the errors due exclusively to the model (table 4) are practically equal for both models, but the cross-validation statistics favour 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, we conclude that we prefer blocks as support to take the observations, designing a regular grid that covers the city and averaging the observations in such grid blocks. At the same time, we choose median-polish kriging to predict the square meter price housing of Albacete because this method provides the best results in terms of cross validation and total error of prediction.

[1] Cross-validation statistic are: $ME = \frac{1}{n}\sum e_i$; $TME = \frac{1}{n}\sum \sigma_i$; $MQE = \frac{1}{n}\sum e_i^2$; $AMQE = \sqrt{\frac{1}{n}\sum \left(\frac{e_i}{\sigma_i}\right)^2}$

where e_i = experimental error, σ_i = theoretic error due to the model.

Comparative table of cross-validation statistics for all estimation methods.				
Method	**ME**	**MQE**	**AMQE**	**TME**
Universal Point Kriging	-0.00027	0.02117	0.999462	0.146492
G.A.M. Kriging	-0.0001	0.02092	1.010452	0.143707
Universal Block Kriging	-0.00074	0.01507	0.957974	0.12871
Median-Polish Kriging	4.21874e-13	0.01364	0.951739	0.122727

Table 5: Cross-validation statistics.

B) Results. The results obtained by this method are graphically shown below. The estimated values and the estimated errors for the four methods can be seen in appendix.

Figure1 shows a non-stationary variogram with a "cyclical in space" structure, similar to a quadratic trend. If we eliminate the trend, estimated by median-polish method, we can obtain the experimental variogram of the median-polish residuals (Figure 2). This variogram is stationary and follows a spherical variogram model with range 500 m., sill 0.009 and nugget effect 0.007. If we take these values as initial parameters to fit the spheric model by non-linear least squares (Cressie, 1985), the fitted values of the variogram parameters are: range 500.42 m., sill 0.008594 and nugget effect 0,006377. Figure 3 shows the median-polish experimental residuals and the initial and fitted theoretical variograms:

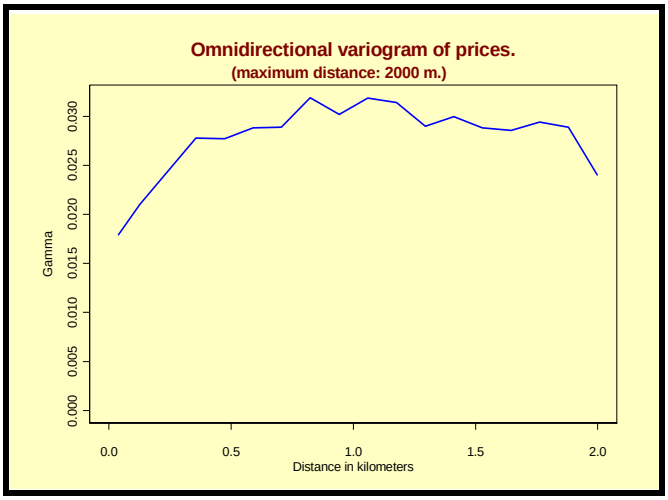

Figure 1: Omnidirectional variogram of housing prices.

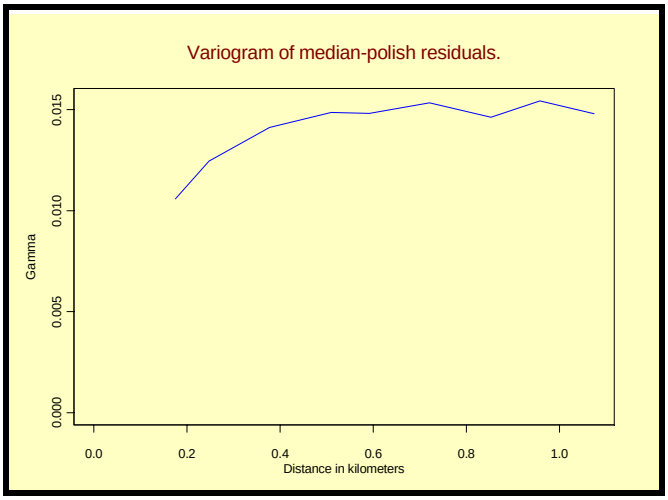

Figure 2: Omnidirectional variogram of median-polish residuals.

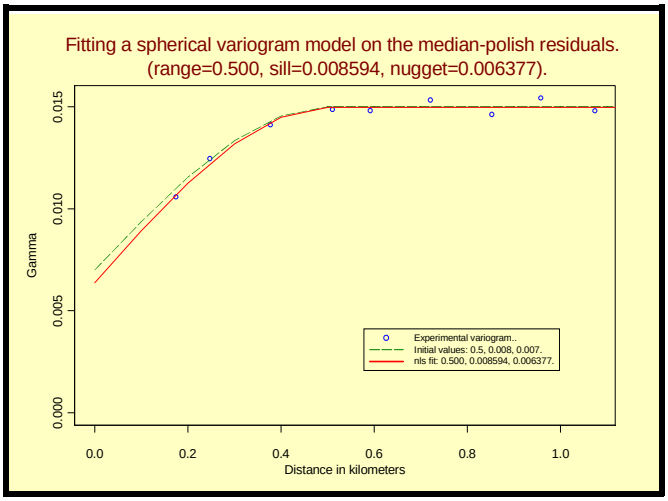

Figure 3: Fitting a spherical variogram model on the median-polish residuals.

Once we have estimated the theoretical variogram model, we can apply median-polish kriging (Cressie, 1993). The housing price prediction will be obtained as the sum of trend, estimated by Akima method (Akima, 1978), from the nodes in which the median-polish trend has been obtained, and the ordinary kriging prediction of the residuals, in the 1330 nodes of a regular grid whose squares have 100m. x 100m. Figures 4 to 11 show the aforementioned results.

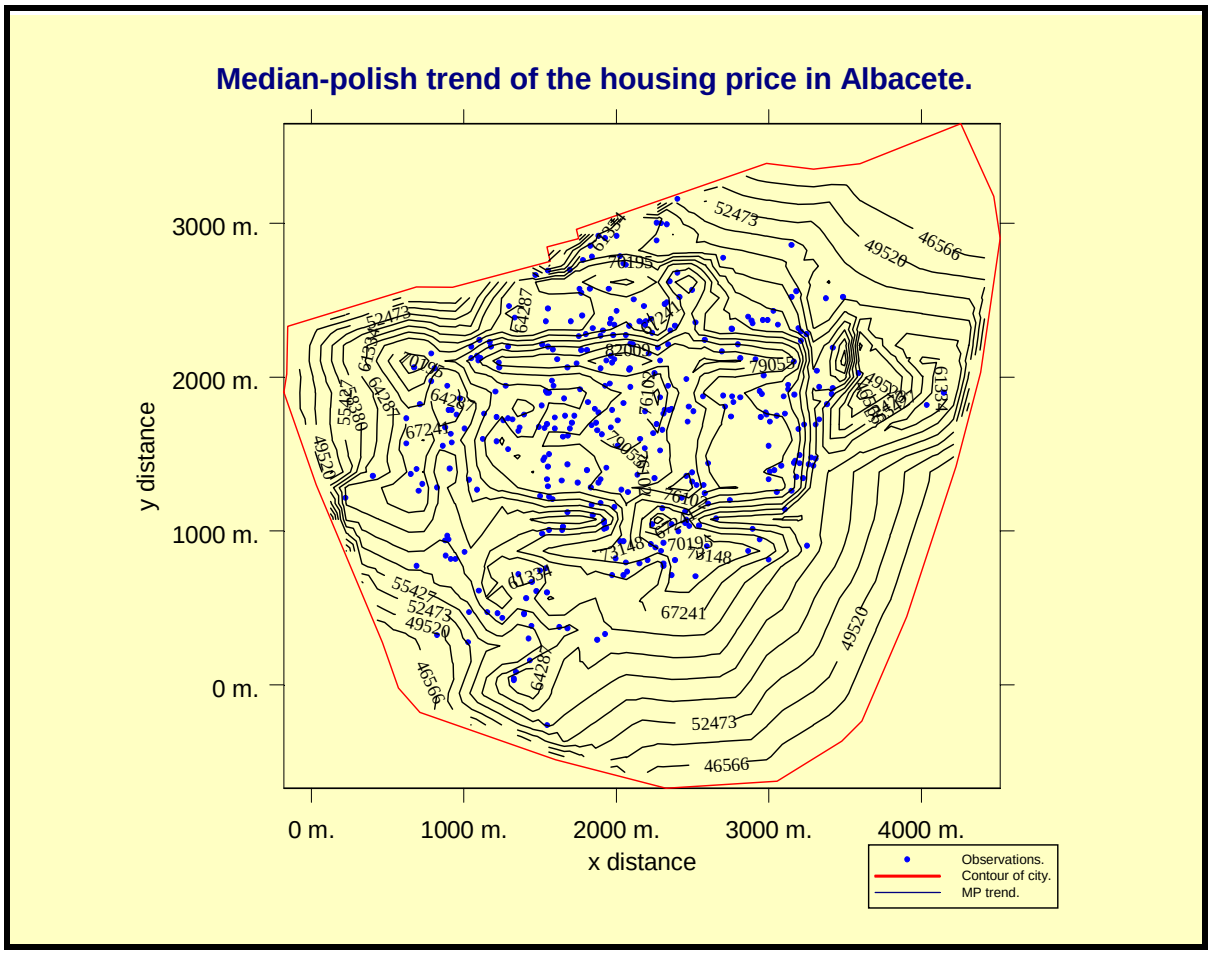

Figure 4: Contour plot of the median-polish trend.

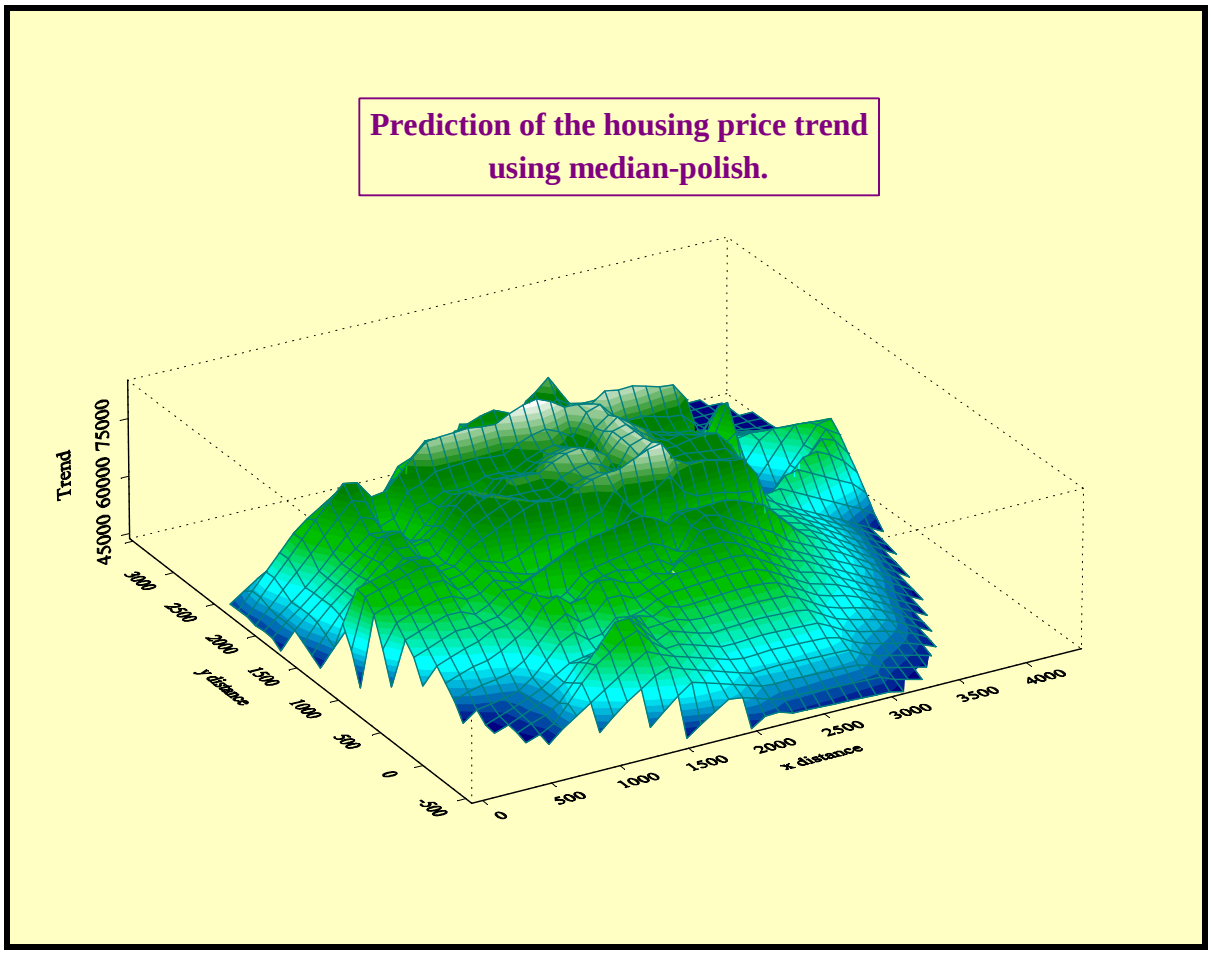

Figure 5: Median-polish trend surface of the housing price.

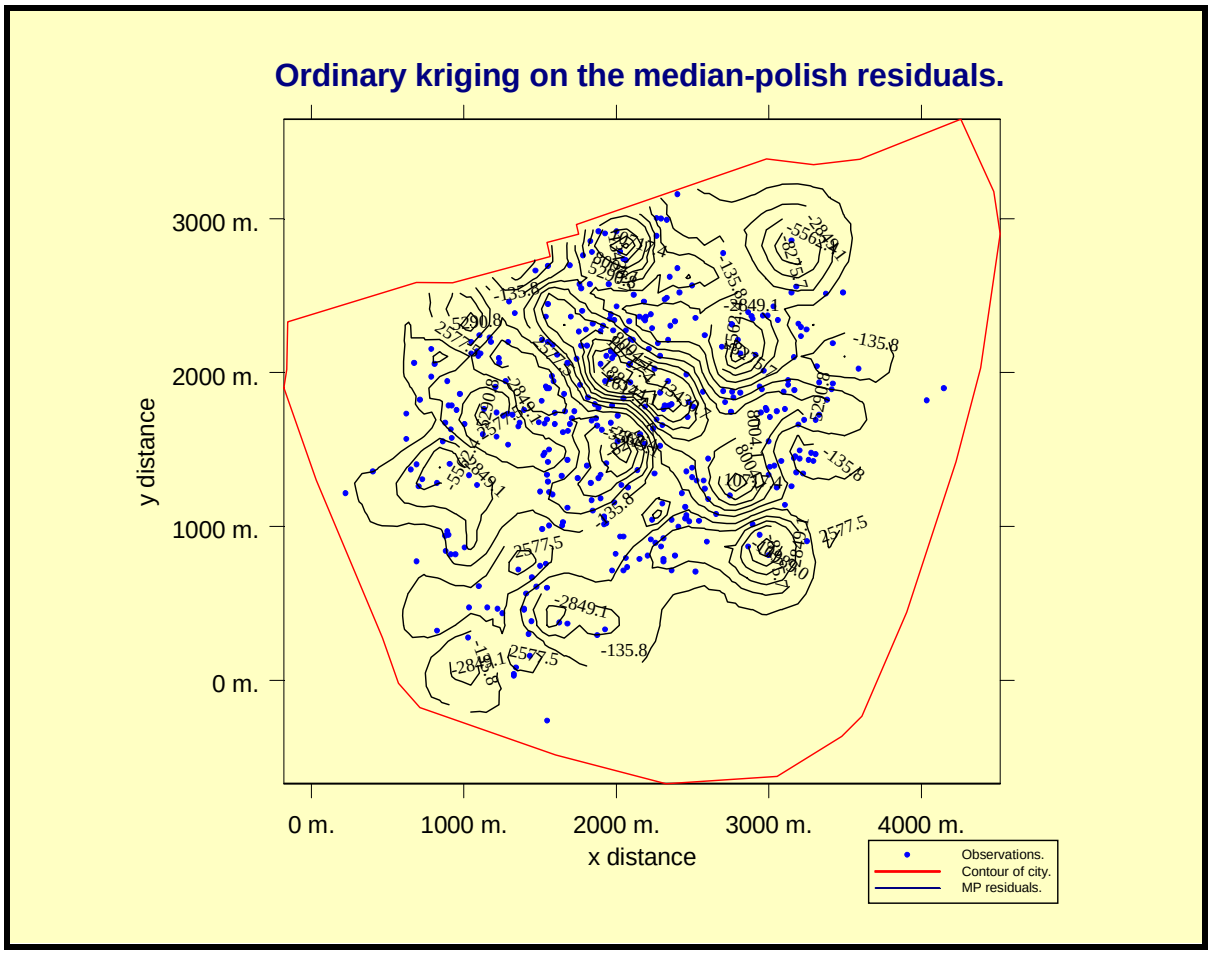

Figure 6: Contour plot of the median-polish residuals. Ordinary kriging prediction.

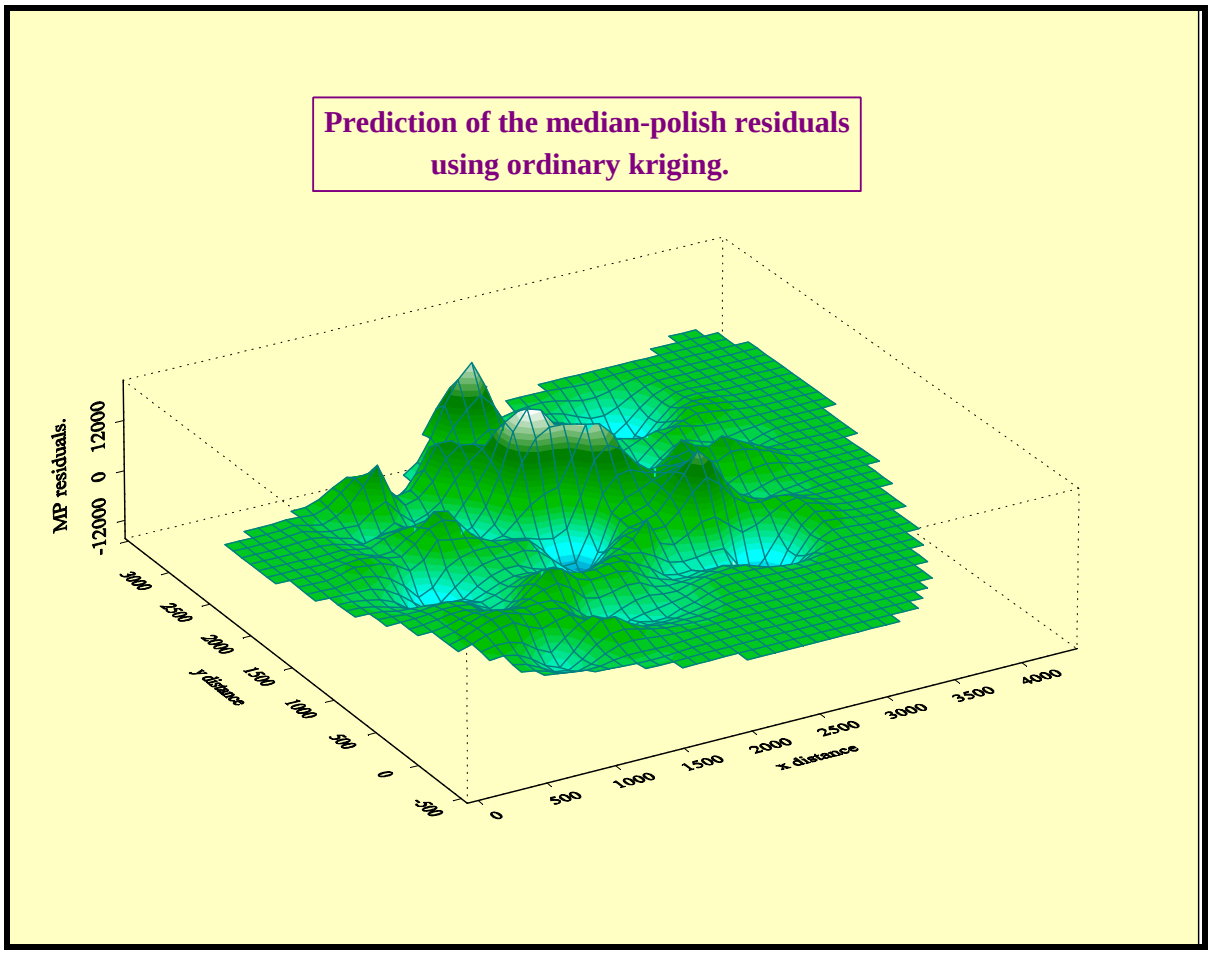

Figure 7: Surface of the median-polish residual. Prediction using ordinary kriging.

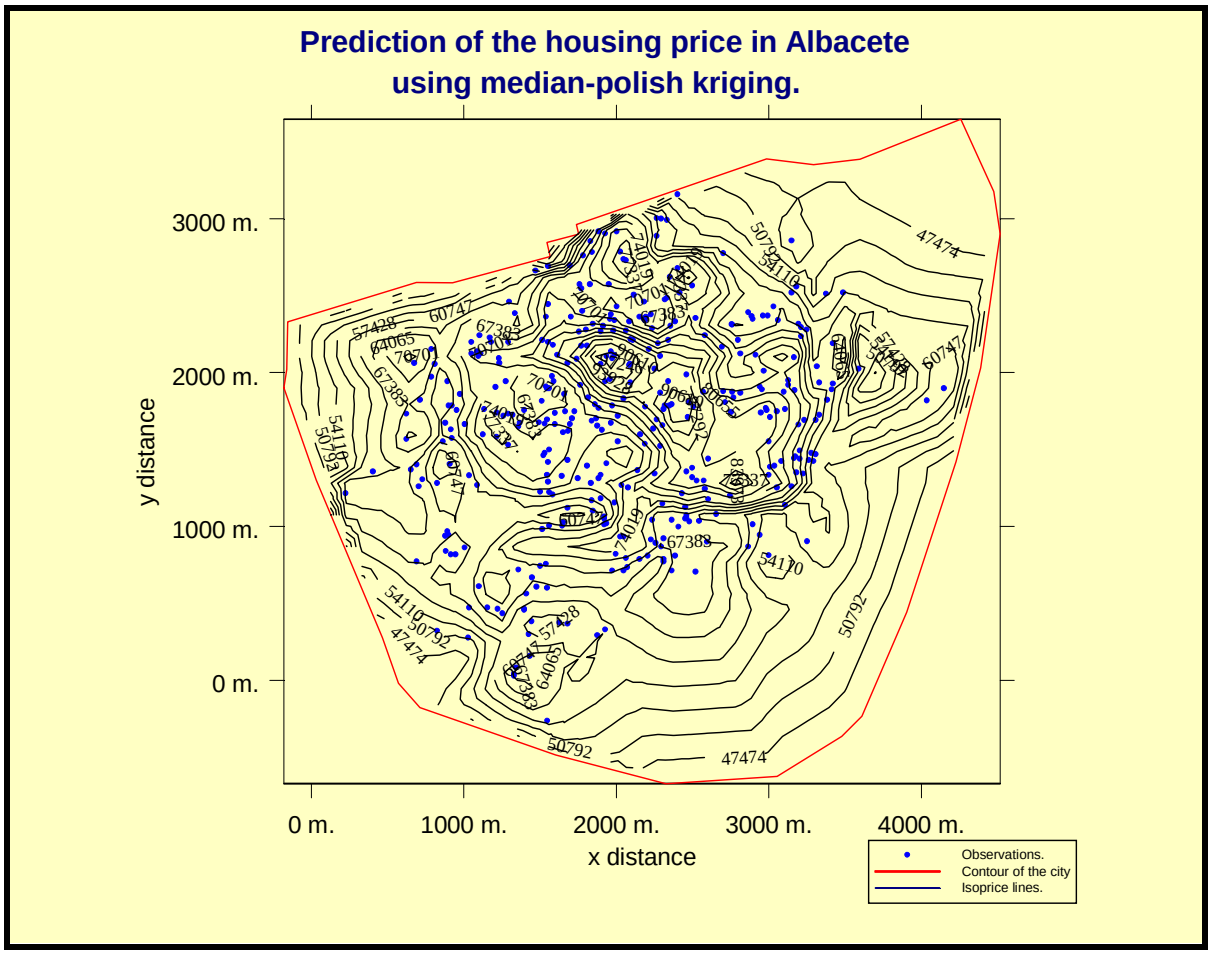

Figure 8: Isoprice lines predicted using median-polish kriging.

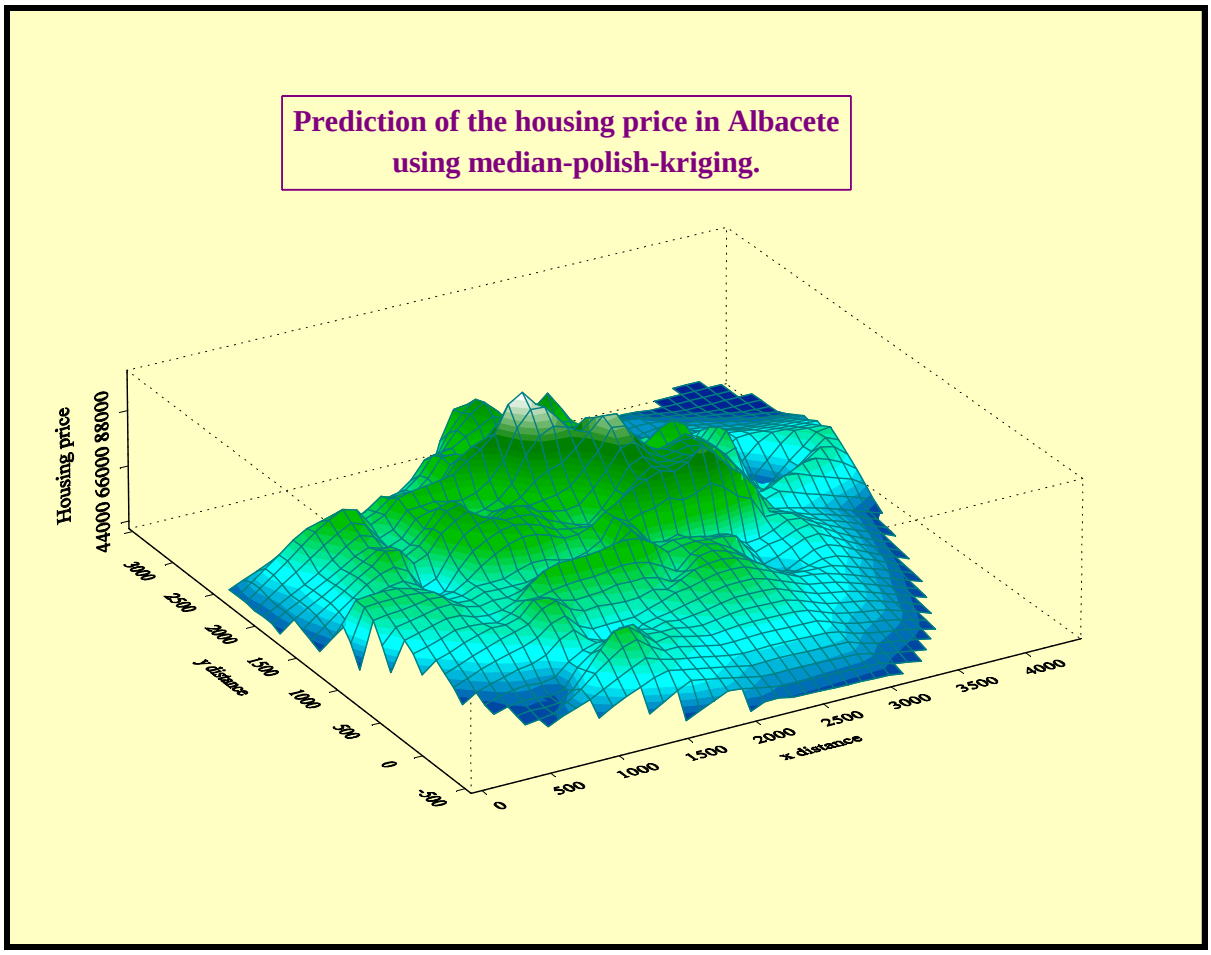

Figure 9: Surface of the housing prices predicted using median-polish kriging.

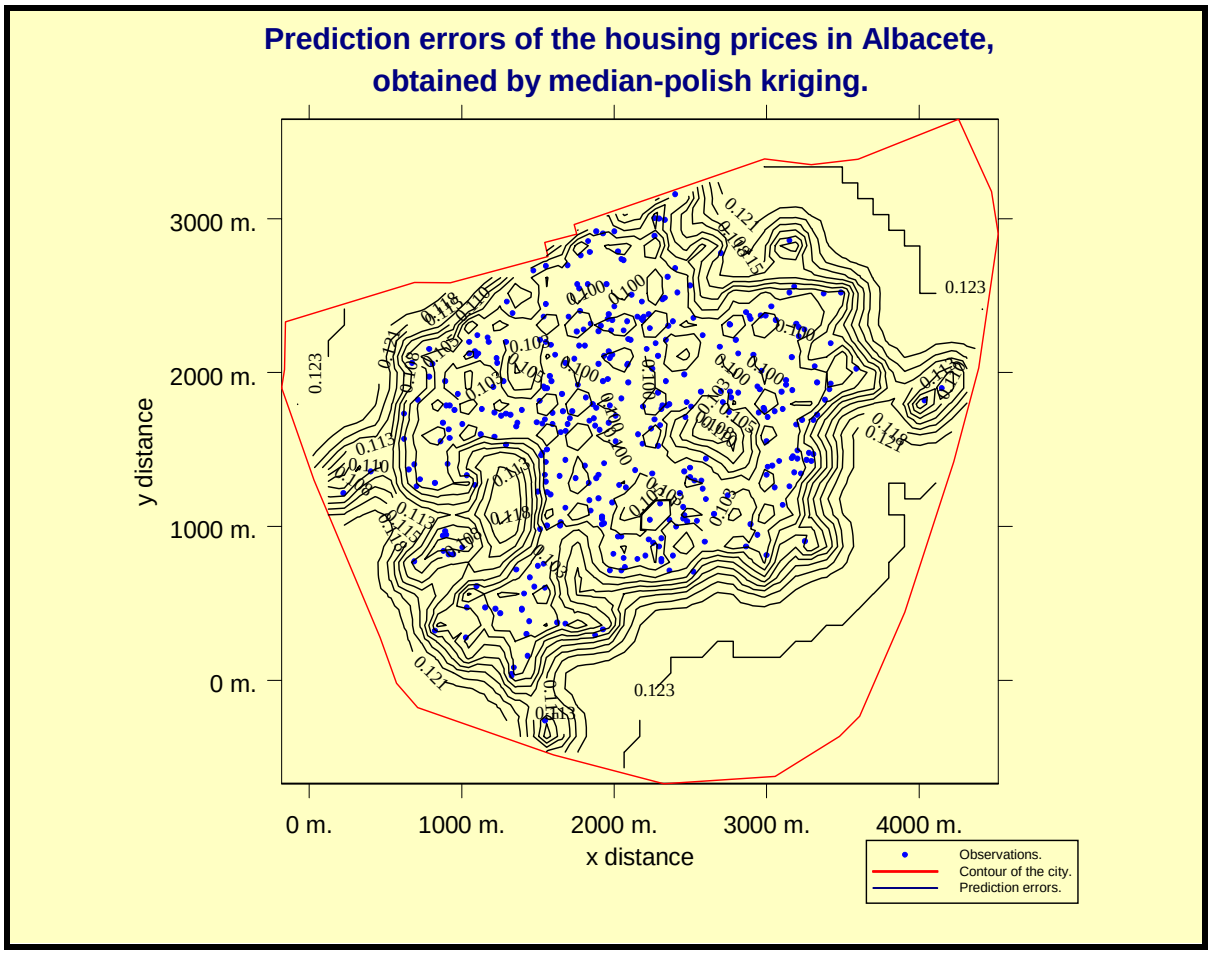

Figure 10: Contour plot of the median-polish-kriging errors.

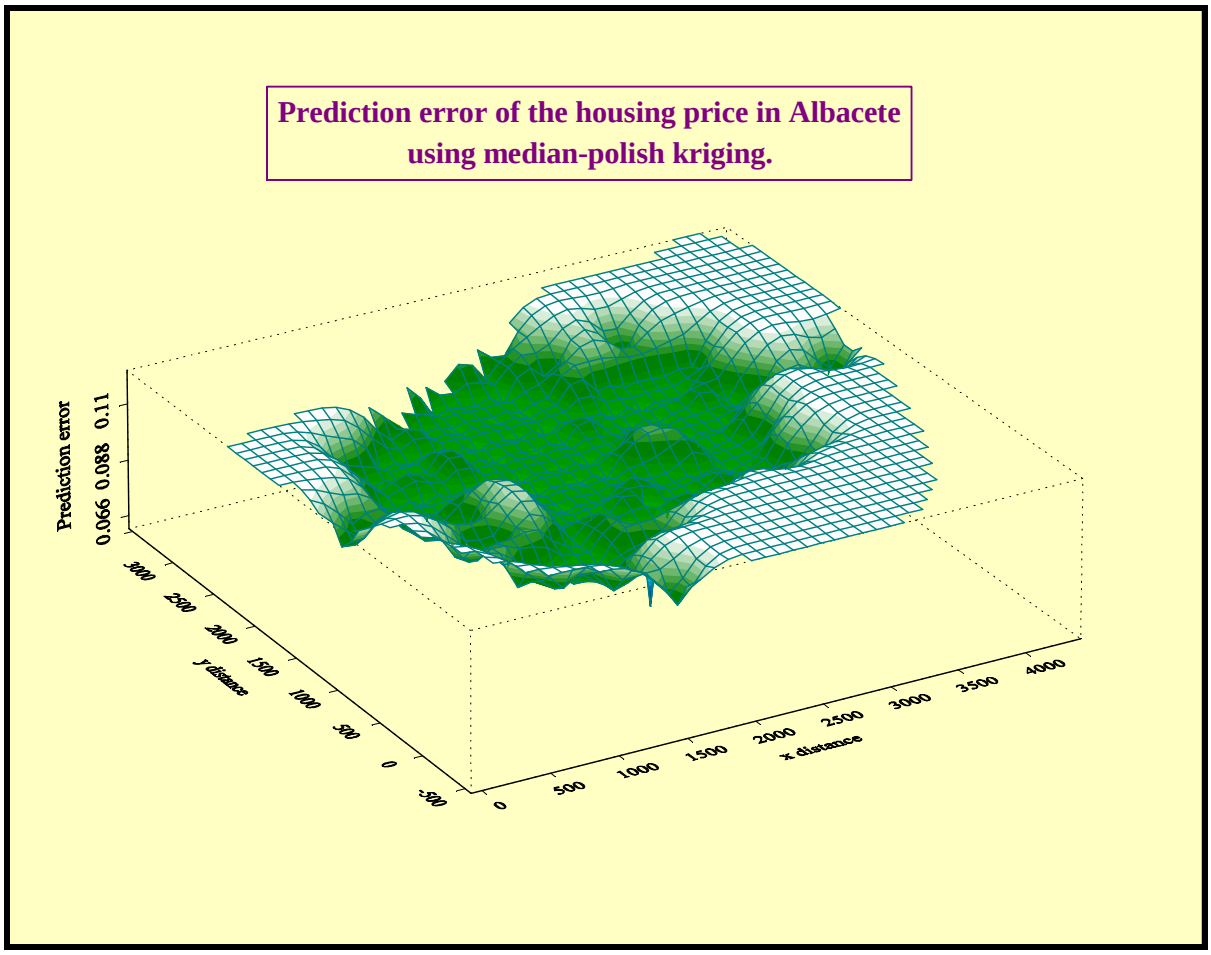

Figure 11: Surface of the median-polish-kriging errors.

Finally we analyse the goodness-of-fit of median-polish kriging model. In order to do that, we summarize its position and cross-validation statistics in the following table.

Observed values		Predicted values	
Minimum	0,35100	Minimum	0,42830
Q_1	0,61700	Q_1	0,63890
Median	0,67930	Median	0,70190
Mean	0,70060	Mean	0,70060
Q_3	0,78080	Q_3	0,75820
Maximum	1,19900	Maximum	0,87790

Table 6: Statistics of the observed and predicted values (cross-validation).

Cross-validation statistics from the ordinary kriging on median-polish residuals.	
ME = 4,21874e-013	AMQE = 0,9517394
MQE = 0,013643235	TME = 0,1227271

Table 7: Cross-validation statistics from the median-polish kriging.

The following figure show the omnidirectional variogram of experimental errors. We can observe that there is no kind of spatial dependence among residuals. So we conclude that the theoretical model has successfully explained the spatial dependence structure.

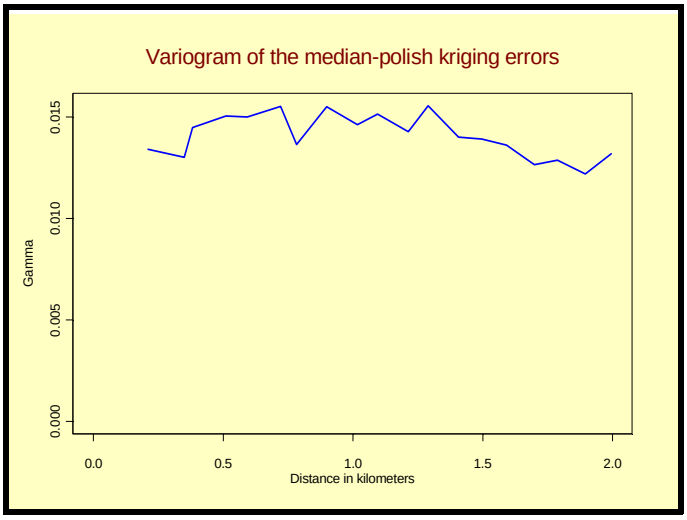

Figure 12: Variogram of the experimental errors from median-polish kriging (cross-validation).

REFERENCES.

Akima, H. (1978): *"A method of bivariate interpolation and smooth surface fitting for irregularly distributed data points"*. ACM Transactions on Mathematical Software, 4, 148-159.

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

Anselin, L. & Florax, R.J.G.M. (1995): *"New Directions in Spatial Econometrics"*. Heidelberg: Springer-Verlag Berlin.

Arlinghaus, S.L. (1996) (ed.): *"Practical Handbook of Spatial Statistics"*. New York: CRC Press.

Bailly, A.S.; Coffey, W.J.; Paelinck, J.H.P. & Polèse, M. (1992): *"Spatial Econometrics of Services"*. Aldershot: Avebury.

Breiman, L.; Friedman, J.H.; Olshen, R.A. & Stone, C.J. (1984): *"Classification and Regression Trees"*. Monterey: Wadsworth and Brooks/Cole.

Brooker, P. (1986): *"A parametric study of robustness of kriging variance as a function of range and relative nugget effect for a spherical variogram"*. Mathematical Geology,18, 477-488.

Chambers, J.M. & Hastie, T.J. (1992) eds.: *"Statistical Models in S"*. New York: Chapman and Hall.

Chica Olmo, J.M. (1994). *"Teoría de las variables regionalizadas. Aplicación en economía espacial y valoración inmobiliaria"*. Granada: Cuadernos de la Universidad de Granada.

Christensen, R. (1991). *"Linear models for multivariate, time series and spatial data"*. New York: Springer-Verlag.

Clark. L.A. & Pregibon, D. (1992): *"Tree-based models"*. Chap. 9 in Chambers & Hastie (1992).

Cleveland, W.S.; Grosse, E. & Shyu, W.M. (1992): *"Local Regression Models*". Chap. 8 in Chambers & Hastie (1992).

Cliff, A. & Ord, J.K. (1981). *"Spatial processes, models and applications"*. London: Pion.

Cordy, C. & Griffitf, D. (1993): *"Efficiency of least squares estimators in the presence of spatial autocorrelation"*. Communications in Statistics-Simulation and Computation, 22, 1161-1179.

Cressie, N.A.C. (1985). *"Fitting variogram models by weighted least squares"*. Journal of the International Association for Mathemetical Geology, 17, 563-586.

Cressie, N.A.C. (1986). *"Kriging nonstationary data"*. Journal of the American Statistical Association, 81, 625-634.

Cressie, N.A.C. (1993). *"Statistics for spatial data"*. Revised ed. New York: Wiley.

Gámez, M. (1998). "Nuevas técnicas de Estadística Espacial para la Economía. Modelización del precio de la vivienda libre en la ciudad de Albacete". Tesis Doctoral de la Universidad de Castilla-La Mancha.

Goldberger, A.S. (1962). *"Best linear unbiased prediction in the generalized linear regression model"*. Journal of the American Statistical Association, 57, 369-375.

Granelle, J.J. (1970): *"Espacio urbano y precio del suelo"*. Paris: Sirey.

Haining, R. (1990). *"Spatial data analysis in the social and environmental sciences"*. Cambridge: C.U.P.

Hastie, T.J. & Tibshirany, R.J. (1990): *"Generalized Additive Models"*. London: Chapman and Hall.

Hawkins, D.M. & Cressie, N.A.C. (1984): *"Robust Kriging. A proposal"*. Journal of the International Association for Mathematical Geology, 16, 3-18.

Hüijbregts, C.J. & Matheron, G. (1971): *"Universal Kriging (An optimal method for*

estimating and contouring in trend surface analysis)". In Proceedings of Ninth International Symposium of Techniques for Decision-Making in the Mineral Industry, McGerrigle, J.I. (ed), The Canadian Institute of Mining and Metallurgy. Special Volume 12, 159-169.
Journel, A.G. & Hüijbregts, CH.J. (1991) *"Mining Geostatistics"* (2ª edición revisada). New York: Academic Press.
Kramer, W. & Donninger, C. (1987). *"Spatial autocorrelation among errors and the relative efficiency of OLS in the linear regression model"*. Journal of the American Statistical Association, 82, 577-579.
Krige, D.G. (1951). *"A statistical approach to some basic mine valuation problems on the Witwatersrand"*. Journal of the Chemical, Metallurgical and Mining Society of South Africa, 52, 119-139.
Matheron, G. (1962). *"Traité de Géostatistique Apliquée. Tome I."* Memoires du Bureau de Recherches Geologiques et Minieres, Nº. 14. Paris: Ed. Technip.
Matheron, G. (1963): *"Traité de Géostatistique Apliquée. Tome II: Le krigeage"*. Memoires du Bureau du Recherches Geologiques et Minieres. Nº 24. Paris: Ed. Bureau du Recherches Geologiques et Minieres.
Matheron, G. (1971): *"The Theory of Regionalized Variables"*. Paris: CCMM de Fontainebleau.
Matheron, G. (1978): *"Estimer et choisir"*. Paris: Les Cahiers du Centre de Morphologie Mathématique. Fasc. 7, Ecole Nationale Supérieure des Mines de Paris.
Ripley, B.D. (1988). *"Statistical inference for spatial processes"*. Cambridge: C.U.P.
Romero Colunga, M. (1991): *"La Valoración Inmobiliaria"*. 2ª ed. Madrid: Aranzadi.
Samper Calvete, F.J. & Carrera Ramírez, J. (1996): *"Geoestadística. Aplicaciones a la hidrología subterránea"*. Barcelona: CIMNE.
Sposito, V. A. (1987): *"On median polish and L_1 estimators"*. Computational Statistics and Data Analysis, 5, 155-162.
Starks, T.H. & Fang, J.H. (1982): *"The effect of drift on the experimental variogram"*. Journal of the International Association for Mathematical Geology, 14, 309-320.
Stein, M.L. & Handcock, M.S. (1989): *"Some asymptotic properties of kriging when the covariance function is misspecified"*. Mathematical Geology, 21, 171-190.
Venables, W.N. & Ripley, B.D. (1996): *"Modern Applied Statistics with S-Plus*" (3th edition). New York: Springer-Verlag.
Wakernagel, H. (1995): *"Multivariate Geostatistics"*. New York: Springer Verlag.
Watson, G.S. (1984): *"Smoothing and interpolation by kriging with splines"*. Journal of the International Association for Mathematical Geology, 16, 601-615.
Zimmerman, D.L. & Zimmerman, M.B. (1991). *"A Monte Carlo comparison of spatial variogram estimators and kriging predictors"*. Technometrics.