SPATIAL EFFECTS ON HOUSING PRICE PREDICTIONS FOR GUARDA CITY

MªJosé Andrade P. Valente
ESTG - Instituto Politécnico Guarda
mjvalente@ipg.pt

Rui Nuno Baleiras
Escola de Economia e Gestão
Universidade do Minho

Housing prices are influenced by a variety of physical, tax, proximity, neighbourhood and local attributes. Hedonic specifications are able to measure the influence of dwelling size and age (and other physical characteristics) on housing prices.

This paper examines the influence that local housing market characteristics have on price prediction accuracy using hedonic housing price equations with single-family transactions for Guarda city.

The regression analysis employs cross-section data; in this instance specialized methods of spatial econometrics are used to avoid potentially biased results and faulty inference.

An alternative to a spatial autoregressive process to avoid the complexity of definition of the spatial weight matrix can be expressed in the form of autocorrelation tests. So we propose the serial correlation *LM* test – specifically the Breusch-Godfrey Lagrange Multiplier test – as a *proxy* of spatial correlation tests.

JEL: C51; C21.

Keywords: Housing Methods
Spatial Econometrics

I. Introduction

Existing efforts toward political and economic integration have in some cases led to a surprising reconsideration of a role for taxes that do not cross borders—taxes on immovable property being one of the few that meet this description. The simplicity with which capital, technology, and information can be transferred across geographic boundaries poses a fundamental challenge to tax systems developed in response to an entirely different model of economic activity. At the same time the growth in the economic importance of intangible property prescriptions some contemporary policy to a reduction or elimination of the tax on real property. So, current property tax reforms face the challenge of identifying the appropriate role for a tax on immobile physical assets in an economy ever more reliant on mobile and intangible factors.

This is not necessarily the case. In fact, three closely related considerations argue for an expanded role for the real property tax in an era of globalization. First, at a time when markets and economic activity cross borders with ease, immovable property offers one of the few tax bases that cannot be readily shifted to another jurisdiction. Second, in an era of tax harmonization, only a purely local revenue source permits local fiscal autonomy. Third, globalization has drawn renewed attention to the goals of devolution and subsidiarity, which require stable local revenue sources to be effective. Each of these factors influences and reinforces the others. The mobility of capital, sales and labour encourages harmonization or convergence in response to tax competition. A period of globalization also places a great premium on preserving accountable, independent local governments to deal with those tasks for which they are best suited.

The property tax can offer a stable revenue source particularly well suited for local government. At the same time, it requires administrative capability, legislative support and political acceptance that are often lacking in highly developed and long established systems as well as in transition economies. Technological advances offer potential efficiency gains in assessment, administration and collection, but they can also consume vast sums for glamorous but inappropriate projects that yield little additional revenue.

International economic and political integration, far from diminishing the role of local governments, has heightened attention to the importance of independent, accountable and responsive subnational governments. The greatest political drawback to a centralized intergovernmental apparatus such as the European Commission may be a lack of accountability. So the efficiency of centralization must be tempered by political representation and responsiveness. The property tax offers an independent revenue source particularly suited for local governments, permitting the option of decentralized rate-setting, administration, and collection. The highly visible nature of the property tax is its greatest political burden, yet also its most important contribution to transparent, politically accountable taxing and spending decisions.

The property tax offers an independent revenue source particularly suited for local governments, permitting the option of decentralized rate-setting, administration, and collection. The highly visible nature of the property tax is its greatest political burden, yet also its most important contribution to transparent, politically accountable taxing and spending decisions.

Assuring equitable assessment values among properties for taxation purposes is an important problem. One cannot, of course, ignore the abundant evidence that tax appraisal are often arbitrary and inequitable – in the Portuguese case, the situation with property tax, in 2002, demonstrated substantial vertical and horizontal inequities in tax assessment practices.

The assessor's primary responsibility is, after all, to provide for equity in assessed valuations so that the property tax burden is fairly distributed which requires that all properties in each property class be valued at the same ratio. Value-based taxes on immovable property are a natural target for taxpayer dissatisfaction. They are highly visible, generally not withheld from income, and not necessarily accompanied by cash earnings with which to make payment; so the public perception of the tax as regressive charge.

Regression analysis can be used to provide a hedonic price as a function of the characteristics of a property and this seems a good approach to make available the attributes of both equity and efficiency to process of assessment. The application of regression analysis to the housing's attribute data would assure consistency and could facilitate an analysis of the relationship of assessment values to selling prices of the properties if the parameter estimates were divided from a model based on sales data. The objectives of consistency, equity and market value thereby would be simultaneously served in the process of determining the assessed values.

The method of predicting house values using ordinary least squares (*OLS*) as the statistical technique to estimate a hedonic regression, sometimes, ignores a potentially large source of information regarding house prices—the correlations existing between the prices of neighbouring houses. From the theoretical point of view spatial autocorrelation seems to be a common phenomenon on real estate markets, although, it is not very often included in classical models constructed in purpose of describing the behaving of particular market.

This paper investigates the sensitivity of hedonic models of house prices to the spatial autocorrelation to address them to typical hedonic real estate data, if necessarily. One can improve efficiency by assigning spatial effects, since we gain insight into the size of bias that can occur in parameters. So, the purpose of this paper is to explore some of the issues involved in estimating housing models with spatially autocorrelated error terms. Section II introduces the housing market and discusses the price of a property as a function of values describing its characteristics, known as the hedonic price function,[1] Section III estimates the resulting hedonic regression after discusses the functional form, while Section IV concludes with the basic issues involved in modelling the autocorrelation structure and compare by difference the most commonly used techniques.

II. Housing Markets

One of the most familiar models in economics is that of price determination in the market. The market mechanism works to reconcile the needs of consumers and firms by establishing the price at which aggregate demand is equal to aggregate supply and the market clears. For many goods, however, this simple model is inadequate. In a market such as that for housing we observe different properties commanding different prices. Indeed, housing is an example of what is called a *differentiated good*. Such goods consist of a diversity of products that, while differing in a variety of characteristics, are so closely related in consumers' minds that they are considered as being one commodity.

Though the simple model does not adequately explain the workings of markets in differentiated goods, it would appear market forces determine that different varieties of the product command different prices and that these prices depend on the individual products' exact characteristics. For example, properties that have more bedrooms will tend to command a higher price in the market than properties that have fewer bedrooms. Furthermore, the set of prices in the market would appear to define a competitive

[1] "Hedonic" because it is determined by the different qualities of the *differentiated good* and the utility derived from their consumption.

equilibrium. That is, in general, the market will settle on a set of prices for the numerous varieties of the differentiated good that reconcile supply with demand and clear the market.

Clearly the set of properties in the market represent a differentiated good. We could describe any particular property by the qualities or characteristics of its structure, environs and location. A succinct means of denoting this is as a vector of values; effectively a list of the different quantities of each characteristic of the property. In general, therefore, the vector could describe any house,

$$\mathbf{z} = (\boldsymbol{z}_{\mathbf{1}}, \boldsymbol{z}_{\mathbf{2}}, \ \ldots, \ \boldsymbol{z}_{\boldsymbol{k}})$$

where $\boldsymbol{z}_{\boldsymbol{i}}, \quad \boldsymbol{i} = \mathbf{1}, \ldots, \boldsymbol{k}$, is the level or amount of any one of the many characteristics describing a property. Indeed, the vector ***z*** completely describes the services provided by the property to a household. When households select a particular property in a particular location they are selecting a particular set of values for each of the $\boldsymbol{z}_{\boldsymbol{i}}$. We can imagine this market for properties as being one in which the consumers consider a variety of somewhat dissimilar products which differ from each other in a number of characteristics including, amongst many characteristics, number of rooms, size of garden, distance to shops and environmental characteristics such as levels of pollution or noise.

The price of any one of these properties will be determined by the particular combination of characteristics it displays. Naturally we would expect properties possessing larger quantities of good qualities to command higher prices and those with larger quantities of bad qualities to command lower prices. Again we can use a succinct piece of notation to illustrate this point;

$$\boldsymbol{P} = \mathrm{P}(\mathbf{z})$$

Which can be read as; the price of a property, *P*, is a function of the vector of values, ***z***, describing its characteristics. This function is known as the *hedonic price function*; ‘hedonic’ because it is determined by the different qualities of the differentiated good and the ‘pleasure’ (in economic terms utility) these would bring to the purchaser.

A number of researchers (e.g. Rosen, 1974; Epple, 1987) have attempted to analytically model equilibrium in hedonic markets. To do this it is necessary to make specific assumptions concerning the various behavioural functions that determine household and landlord behaviour. Specifically, researchers must assert a particular functional form for the utility function of households, $U(.)$, and the cost function of landlords, $C(.)$. Further they have to make assumptions concerning the distribution in the population of household characteristics, ***s***, and landlord characteristics, ***r***. Given specific forms for each of these different functions it should be possible to solve for an expression that gives the equilibrium hedonic price function. This expression will be a function of the arguments in the underlying functions. Hence, using such models it is possible to investigate how changes in the underlying arguments influence the hedonic price schedule.

Unfortunately, the complexity of the hedonic market is such that one must make very restrictive assumptions concerning the various functions in order to end up with an expression for the hedonic price schedule that is reasonably tractable. In general, therefore, research has concentrated on empirical analyses of hedonic markets.

In many ways, the problem for empirical analysis is the reverse of analytical research. Rather than assuming the functional forms of the underlying behavioural functions and working through the problem to solve for the hedonic price schedule, researchers estimate the hedonic price schedule from real world market data and work back through the problem to discover the form of the underlying behavioural functions. We shall return to how this might be achieved in Section III.

III. Functional Form

Many functional forms of the variables and parameters lead to pricing functions that agree with the information amassed by the substantial theoretical and empirical work in hedonic pricing and mass assessment. Consequently, the exact specification to adopt remains one of the central uncertainties of empirical work, especially since the ‘‘wrong’’ functional form leads to all sorts of disastrous consequences for traditional estimators. In response to this problem, many nonparametric estimators have been proposed which adapt to the data and do not require an *a priori* functional specification. However, nonparametric estimator performance typically declines as the dimensionality of the problem increases. As a compromise, various semiparametric estimators have arisen that possess the adaptive traits of nonparametric regression while retaining the estimation efficiency of parametric estimators.

Many functional forms have been proposed and used for hedonic property models including linear, quadratic, log-log, semi-log, inverse semi-log, exponential, and Box-Cox transformation. Theory only suggests that the first derivative of the hedonic price function with respect to the characteristic in question be positive (negative) if the characteristic is desirable (undesirable). Properties of the second derivative cannot be deduced from the general features of the model (Freeman, 1993). Popular methods to select the functional form include using a linear relationship and altering any variables which are believed *a priori* to be nonlinear and using flexible forms (e.g. Box-Cox Transformations) to determine the best fit (Kulshreshtha and Gillies, 1993).

A good way to establish a model is to use cross-validation.[2] To determine the best model we use a goodness of fit index, the average mean average error, average-MAE. This is our option. Despite complex functional forms and sophisticated regression techniques we are used 400 observations on house sales, in 2002, from Guarda city, to illustrate applicable techniques for hedonic house price estimation. By going from a benchmark linear specification to logarithmic functional forms and Box-Cox transformations, the mean average error, *MAE*,[3] in 1000 regressions can be reduced significantly. The best results however can be achieved using a semi-log model. The average *MAE* indicates a 14,2% reduction in the prediction error: from 12552,243 on the benchmark linear model to 10793,232 in the log-linear specification.

[2] Cross-validation refers to the process of removing one of the n observation points and using the remaining $n-1$ points to predict its value. This process is repeated at each data point; for each estimate, $n-1$ points are used. The interpolation error at each data point is the difference between its observed and predicted values.

[3] $MAE=\frac{1}{T}\sum_{t=1}^{T}\left|\hat{P}_t - P_t\right|$, where $\hat{P}_t$ is the expected price, P_t is the observed price and T is the number of estimations.

	Linear Model	Log-linear Model	Log-log Model	Quadratic Model	BoxCox linear Model	BoxCox Quadratic Model
Average-MAE	12552,243	10793,232	12323,039	11294,027	10983,189	11110,866

The formal econometric model is stated as:[4]

$$
\begin{aligned}
LNVMERCADO = {} & \underset{(86{,}743)}{10{,}733} - \underset{(-5{,}654)}{0{,}0140} * ANOS \;+\; \underset{(0{,}123)}{0{,}0029} * AQCENTRAL + \underset{(3{,}859)}{0{,}0035} * AREA \\
& + \underset{(2{,}467)}{0{,}068} * QUARTOS + \underset{(0{,}923)}{0{,}0361} * CBANHO + \underset{(0{,}608)}{0{,}0200} * ARRECA \text{ -} \\
& - \underset{(-0{,}916)}{0{,}0167} * ELEV + \underset{(0{,}058)}{0{,}0040} * EXPSOL + \underset{(2{,}398)}{0{,}0618} * GCANAL \;+ \\
& + \underset{(2{,}515)}{0{,}1246} * GSDOS \;+\; \underset{(2{,}531)}{0{,}0623} * GSUM \;-\; \underset{(-0{,}436)}{0{,}0093} * JANDUPLAS + \\
& + \underset{(1{,}841)}{0{,}0646} * LAR \;-\; \underset{(-1{,}048)}{0{,}0578} * LCNTRO + \underset{(0{,}468)}{0{,}0317} * LCVLHOS \text{ -} \\
& \text{-}\underset{(-2{,}087)}{0{,}1091} * LGGARE \;-\; \underset{(-0{,}220)}{0{,}0138} * LLAMNH + \underset{(0{,}038)}{0{,}0020} * LLUZ \text{ -} \\
& \text{-}\underset{(-1{,}474)}{0{,}0827} * LPMILEU \;-\; \underset{(-1{,}334)}{0{,}0777} * LPNHR \;+ \underset{(0{,}197)}{0{,}0113} * LRDIZ \text{ -} \\
& \text{-}\underset{(-0{,}426)}{0{,}0233} * LSREMD - \underset{(-0{,}611)}{0{,}0407} * CV \;-\; \underset{(-0{,}753)}{0{,}0281} * RCH \;+\; \underset{(1{,}206)}{0{,}0329} * PPR \\
& \text{-}\underset{(-0{,}734)}{0{,}0399} * PQRT - \underset{(-1{,}136)}{0{,}0455} * PSEG + \underset{(0{,}726)}{0{,}0233} * PTERC
\end{aligned}
$$

R^2 = 0,625821 R^2 Adjusted= 0,597581

IV - Spatial Dependence

Spatial dependence among hedonic regression residuals was iniatially revealed by Brigham (1965) who carried out topographic error projections. Sibert built a model of residential values based on spatial autocorrelation but did not adopt the hedonic conceptual

[4] The variables are identified in Annex I

framework. Anas and Eum (1984) assumed the absence of spatial autocorrelation but implicitly incorporated a spatial autoregressive term; they used the most recent nearby sale as a temporal proxy. Dubin (1988) carried out a formal verification of the existence of spatial dependence among the hedonic regression error terms. Since the end of 80's there has been a marked increase in studies highlighting concerns about spatial autocorrelation.

In fact the hedonic model remains a reflection of the price formation mechanism, since housing price translates the marginal utility of the housing's characteristics over all characteristics; so neighbouring house prices correspond to the combined effect of the individual preferences of the other consumers. According to this assumption, the price of neighbouring houses can be interpreted as a local quality factor that interprets both location characteristics and physical factors shared by neighbouring houses.

The consequences of spatial autocorrelation are the same as those of time series autocorrelation: the *OLS* estimators are unbiased but inefficient, and the estimates of the variance of the estimators are biased.[5] Thus the precision of the estimates as well as the reliability of hypotheses testing can be improved by making a correction for autocorrelation. Once the structure of the autocorrelation has been estimated, this information can be incorporated into any predictions, thereby improving their accuracy.[6]Just as with time series autocorrelation, maximum likelihood (*ML*) techniques are commonly used to estimate the autocorrelation parameters and the regression coefficients.[7]

[5] Unbiased and consistent estimation by *OLS* requires that error term and regressors are uncorrelated. This assumption is violated when,

• The 'spatially lagged' or 'average neighbouring' dependent variable (housing price) $\mathbf{WP}$ is correlated with the unobserved error term:

$$\mathbf{WP} = \mathbf{W}(\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{Z}\beta + \mathbf{W}(\mathbf{I} - \rho\mathbf{W})^{-1}\varepsilon$$

• The matrix, $\mathbf{W}(\mathbf{I} - \rho\mathbf{W})^{-1}$, does not have all zeros on leading diagonal, hence

$$E\left[\varepsilon'\mathbf{W}(\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{P}\right] \neq 0$$

Efficiency of *OLS*, and correct standard errors requires that error term is homoscedastic and has no autocorrelation, i.e.

$$E(\varepsilon, \varepsilon') = \Omega = \sigma_v^2\mathbf{I}$$

[6] This technique is known as kriging in the geostatistics literature and best linear unbiased prediction (BLUP) in the econometrics literature. Dubin (1992) and Basu and Thibodeau (1998) use this technique to predict housing prices. Also, Dubin (1998) and Dubin *et al.* (1998) discuss the issues involved in kriging.

[7] With the log-likelihood function,

Despite the similarities, spatial autocorrelation is conceptually more difficult to model than time series autocorrelation, because of the ordering issue. In a time series context, the researcher typically assumes that earlier observations can influence later ones, but not the reverse. In the spatial context, an ordering assumption such as this is not possible: if A affects B, it is likely that the reverse is also true. Also, the direction of influence is not limited to one dimension as in time series, but can occur in any direction (although we generally restrict the problem, at least in the case of housing, to two dimensions).

A method for modelling the autocorrelation structure is to depict the process itself. This approach is based on the work of geographers (Cliff and Ord, 1981) and requires the use of a weight matrix. This approach is probably the more common of the two in the real estate literature (see Can (1992) and Pace and Gilley (1998) for examples).[8] The first task is quantification of the location aspects of our sample data. Given that we can always map a set of spatial data observations, we have two sources of information on which to draw: One source of information is the location in Cartesian space represented by latitude and longitude. This information would also allow us to calculate distances from any point in space, or the distance of observations located at distinct points in space to observations at other locations. If the relationship we are modelling varies over space, observations that are near should exhibit similar relationships and those that are more distant may exhibit dissimilar relationships. The second source of location information is contiguity, reflecting

$$\boldsymbol{L}(\rho,\beta,\sigma^{2}) = \tfrac{1}{2}\ln|\mathbf{V}| - \tfrac{1}{2}\left[\boldsymbol{n}\ \ln(2\pi\sigma^{2}) + \right]\sigma^{-2}(\mathbf{P}-\mathbf{Z}\beta)'\boldsymbol{V}\,(\mathbf{P}-\mathbf{Z}\beta)$$ where $\mathbf{V}$ equals $(\mathbf{I}-\rho\mathbf{W})'(\mathbf{I}-\rho\mathbf{W})$. The maximum likelihood method efficiently estimates the model asymptotically (given the assumptions hold).

[8] The usual prediction of the dependent variable, $\mathbf{P} = \mathbf{Z}\beta + \varepsilon$ is correct by a weighted average of the errors on nearby properties as in, $\mathbf{P} = \mathbf{Z}\beta + \rho\mathbf{W}(\mathbf{P}-\mathbf{Z}\beta) + \varepsilon$, where $\mathbf{W}$ represents an *400* by *400* comparable weighting matrix with 0s on the diagonal (the observation cannot predict itself). The rows of $\mathbf{W}$ sum to 1 as implied by below. The non-zero entries on the *i*th row of $\mathbf{W}$ represent the observations whose errors interact with the error on the *i*th observation. We assume independent, 0 mean errors from a normal distribution. These assumptions appear in,

(a) $\underset{(400*400)}{\boldsymbol{W}}\ \underset{(400*1)}{[\mathbf{1}]} = \underset{(400*1)}{[\mathbf{1}]}$

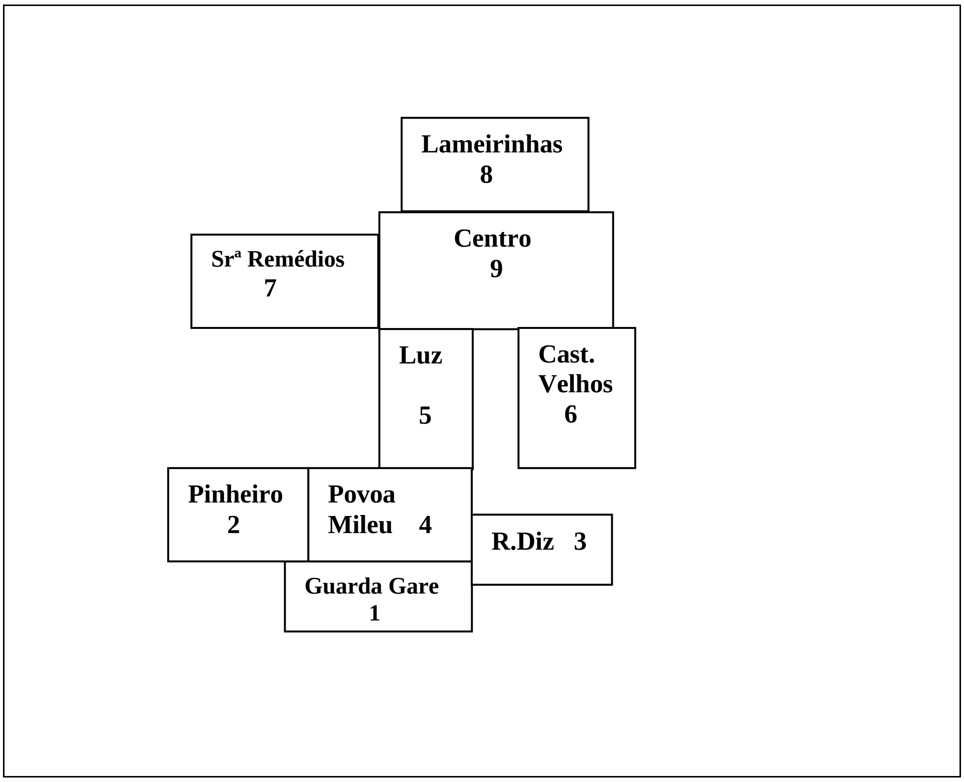
(b) $\boldsymbol{diag}(\boldsymbol{W}\ ) = \underset{(400*1)}{[\mathbf{0}]}$

(c) $\mathbf{0} \le \rho < \mathbf{1}$

(d) $\varepsilon \ \square\ N(0,\sigma_{\varepsilon}^{2}\mathbf{I})$

the relative position in space of one regional unit of observation to other such units. From this we can determine which units are neighbours (have borders that touch) or represent observational units in reasonable proximity to each other.

In our case we wish to construct a *9* by *9* binary matrix $\mathbf{W}$ containing 81 elements taking values of 0 or 1 that captures the notion of "connectiveness" between the nine entities (different neighbourhoods in Guarda city) depicted in the map configuration below,



We record the contiguity relations for each area in the row of the matrix $\mathbf{W}$. For example the matrix element in row 1, column 2 would record the presence (represented by a 1) or absence (denoted by 0) of a contiguity relationship between places 9 and 8. As another example, the row 3, column 4, element would reflect the presence or absence of contiguity between places 7 and 5. Of course, a matrix constructed in such fashion must be symmetric — if regions 7 and 5 are contiguous, so are regions 5 and 7.

It turns out there are a large number of ways to construct a matrix that contains contiguity information regarding the regions. Below, we define a binary matrix $\mathbf{W}$ that reflects the “linear contiguity” relationships between the nine entities in figure above. So we define $W_{ij} = \mathbf{1}$ for entities that share a common edge to the immediate right or left of the region of interest and $W_{ij} = \mathbf{0}$ otherwise.

Then the matrix $\mathbf{W}$ shows the first order linear’s contiguity relations for the nine zones,

$$\mathbf{W} = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

Note that $\mathbf{W}$ is symmetric, and by convention the matrix always has zeros on the main diagonal. In applied work a transformation often used converts the matrix $\mathbf{W}$ to have row-sums of unity; so we can obtain a standardized version of $\mathbf{W}$

$$\mathbf{W} = \begin{bmatrix} 0 & 1/4 & 1/4 & 1/4 & 1/4 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/4 & 0 & 1/4 & 1/4 & 1/4 \\ 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 1/3 & 1/3 & 1/3 & 0 \end{bmatrix}$$

The motivation for the standardization can be the possibility to obtain a new variable,

$$m_i(P) = \sum_j w_{ij} P_j$$

equal to the mean of observations from contiguous regions as shown in,

$$m_9(P) = \tfrac{1}{4}P_8 + \tfrac{1}{4}P_7 + \tfrac{1}{4}P_6 + \tfrac{1}{4}P_5,$$

$$m_8(P) = m_7(P) = m_6(P) = P_9$$

$$m_5(P) = \tfrac{1}{2}P_9 + \tfrac{1}{2}P_4$$

$$m_4(P) = \tfrac{1}{4}P_5 + \tfrac{1}{4}P_3 + \tfrac{1}{4}P_2 + \tfrac{1}{4}P_1$$

$$m_3(P) = \tfrac{1}{2}P_1 + \tfrac{1}{2}P_4$$

$$m_2(P) = \tfrac{1}{2}P_1 + \tfrac{1}{2}P_4$$

$$m_1(P) = \tfrac{1}{3}P_3 + \tfrac{1}{3}P_4 + \tfrac{1}{3}P_2$$

This new variable is important to calculate Moran's *I*, which allow us to test, in a statistical sense, for unevenness in the spatial distribution of some characteristic z ,

$$I = \frac{Cov(P_i, m_i(P))}{Var(P_i)}$$ [9]

In space, the error variances are also heteroskedatic; wich is not the case in the time domain (Anselin and Bera 1998). The heteroskedasticity is induced by the spatial process and will complicate specification testing (i.e. distinguishing "true" heteroskedasticity from that induced by a spatial process).

In our case the problem is not present. To capture heteroskedasticity we can look at the residuals of our model and test the hypothesis null of no heteroskedasticity, then the

[9] Moran's I can take any value between –1 and +1. A value close to zero indicates no spatial association: there is no systematic relationship between the value of x in any region with that in its neighbours. +1 indicates a perfect correlation between the value of x in any region and that in its neighbours. –1 indicates perfect negative correlation. In small samples, the expected value of I when there is data is randomly distributed across space is 1 . If n is large, this is clearly zero.

result of *White's test* is carried out by obtaining, $[(n\ * R^2) = 26{,}6789] : \ c^2_{[32\ ]}$, which is highly significant; so we don't reject the null of no heteroskedasticity. To increase the power of this result we can obtain a more powerful test, the *Goldfeld-Quandt test*, by ordering all observations on the basis of the independent variable suspected of causing heteroscedasticity,[10] so the test statistic is,

Variables	Test G-Q	Resultado
age	$F_{[203,71]} = 1,2507$	Don't reject the null of no heteroskedasticity
area	$F_{[218,55]} = 2,53869$	Reject the null of no heteroskedasticity
rooms	$F_{[197,77]} = 2,47512$	Reject the null of no heteroskedasticity
bathrooms	$F_{[197,77]} = 2,401797$	Reject the null of no heteroskedasticity

The results are heterogeneous; they recommend the introduction of *Glesjer's Test*. In each case, the test for the specific context of variables, "*old*", "*rooms*", "*bathroom*" and "*area*", suspected of causing heteroscedasticity provide the results,

(i) $F\ [4,395\ ]= 1,4723$, when the functional formulation is linear;

(ii) $F\ [4,395\ ]= 0,89183$, when the functional formulation is quadratic;

(iii) $F\ [4,395\ ]= 1,3153$, when the functional formulation is logarithmic.

Finally, we can reject the hypothesis of heteroscedasticity. This is a good consequence since we don't need concerned to distinguish true heteroscedasticity from that induced by the spatial process, so the tests against spatial dependence will have more power.

The most straightforward testing approach is to use *Lagrange Multiplier Tests* that are based on the residuals of an *OLS* regression. Separate tests are available for a spatial lag and a spatial error alternative. Other tests with high power are based on the application of Moran's *I* to regression residuals, which is a valid misspecification test against a wide

[10] For the Goldfeld-Quandt test we assume that the observations can be divided into two groups in such a way that under the hypothesis of homoscedasticity, the disturbance variances would be the same in two groups, and under this hypothesis the test has an *F* distribution.

range of alternatives and applicable in various econometric specifications (Kelejian and Prucha 1999).

However spatial econometric methods are not routinely incorporated in commercial software packages. Hence, several authors have developed “tricks” to carry out estimation and specification testing using macro or script facilities in statistical computing software. Examples are routines in Limdep, Gauss and S-Plus in Anselin and Hudak (1992) and maximum likelihood estimation in SAS (Griffith 1993) or Matlab (Pace and Barry 1998).

In this case we need expand the weighting matrix, $\mathbf{W}$, to the 400 observations, by the construction of a block matrix, and this is a great limitation without specific software; so an adaptation is required to achieve the aim without spatial computing software, in way to allow the application of the standard autocorrelation tests bid by Eviews 4.

In order to apply the principles of the standard autocorrelation tests we have to reduce the two-dimensional space in which the regions are located to a one-dimensional space like the time dimension.[11] Lining up all the regions according to their location does this; so that we start from one corner of the Guarda city and then take the nearest neighbouring area, and the next-nearest, etc. There is scope for ambiguity here, of course, but fortunately we didn’t have to really do this ordering, because the municipalities were originally numbered according to location, so we just had to sort all data according to municipality number postal. The results are,

Centro	**Lameirinhas**	**Sr^{a}Remédios**	**Castelos Velhos**	**Luz**	**Povoa Mileu**	**Rio Diz**	**Pinheiro**	**Guarda Gare**
9	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**

When that is done, we can use the large sample Breusch-Godfrey Serial Correlation *LM* test statistic as a proxy for a test of spatial correlation. We make the test with nine neighbours, which is also the limit we have chosen for the construction of spatial variables. When the null of no spatial correlation were rejected, we tried to include additional spatial variables to reduce the correlation problem. This usually helped and the remaining spatial correlation problems are small.

[11] Starting from an idea developed by Andersen, Lykke and Reis E. (1997)

At the same time we revisit the Moran's *I*, already consider above, which is applied to residuals; so appears as, $I = e'\mathbf{W}e/e'e$. We can observe the formal nearness with Durbin-Watson[12]statistic, which presents also similar properties asymptotically comparable to *LR* e *LM* statistics.

The housing model estimation, presented in Section III, by the *Ordinary Least Squares* (*OLS*), after introducing the alignment of data according the spatial codification present in table, establish the value $1,9981$ for the Durbin Watson (*DW*) statistic. If there is no serial correlation the *DW* statistic will be around 2. In this case *DW* statistic will fall practically 2, so we don't reject the null hypothesis of *no serial correlation*. But there are three main limitations of *DW* test as a test for serial correlations; first the distribution of the *DW* statistic under the null hypothesis depends on the data matrix $\mathbf{Z}$ (the housing characteristics matrix); second if there are lagged dependent variables on the right-hand side of a regression; finally, you may only test the null hypothesis of no serial correlation against the hypothesis of first-order serial correlation.

To overcome these limitations we applied one other test of serial correlation – the Breusch-Godfrey *LM* test. The null hypothesis of this test is no serial correlation in the residuals up to specified order. In our case study, since the highest contiguous spatial relations we can associate are 4 regions, we should introduce a 4 order *lag* to be tested. Eviews4 reports a statistic labelled "*F*-statistic", and an "$Obs.{*}R^2$" (NR^2),[13] $0,1565$ and $0,6811$, respectively. The test doesn't reject the hypothesis of no serial correlation up to order four.[14]

Eviews4 will display the autocorrelation and partial autocorrelation functions of the residuals with the Ljung-Box *Q*-statistics[15] for high-order serial correlation. The application of these testing procedures produces the following view,

[12] $DW = e'\mathbf{A}e/e'e$, **A** is a band matrix, $-1,2,-1$. *DW* is a test for formal first order serial correlation.

[13] O NR^2 statistic has an asymptotic c^2 distribution under the hypothesis null, the distribution of *F*-statistic is not known, but is often used to conduct an informal test o f the null.

[14] If instead a lag 4, defined as function of the maximum value of space contiguous if it had introduced the values of 1, or of 2, or of 3, in test *LM* of Breusch-Godfrey, we would be continued not to reject the hypothesis of inexistence of correlation in series above of order 1, or of order 2, or of order, respectively

[15] The *Q*-statistics at lag *k* is a test for the null hypothesis that there is no autocorrelation up to order *k* and is computed as, $Q_{LB} = T(T+2)\overset{k}{\underset{j=1}{\mathring{a}}} t_j^2/(T-J)$, where, t_j^2, is the squared of *j*-th residuals autocorrelation,

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat.	Prob.
.|. |	.|. |	1	0.037	0.037	0.5414	0.462
.|. |	.|. |	2	0.007	0.006	0.5631	0.755
.|. |	.|. |	3	0.007	0.007	0.5856	0.900
.|. |	.|. |	4	-0.010	-0.010	0.6232	0.960

Notice that all *Q*-statistics are insignificant with large *p*-values, and at all lags the autocorrelations and the partial autocorrelations are nearly zero; so there is no serial correlation in the residuals.

Our conclusion can achieve best performance if we interpret the *Q*-statistics as a χ^2 distribution, so the results of one $\chi^2_{[1]} = 0,5414$, one $\chi^2_{[2]} = 0,5631$, one $\chi^2_{[3]} = 0,5856$ and one $\chi^2_{[4]} = 0,6232$, reveal, for highly significance level, we can accept the hypothesis of no autocorrelation.

Finally some doubts can exist about the accuracy and precision of the artifice used, the transformation of two-dimensional space in a one-dimensional space and the consistency of the results produced. To overcome those we try calculating the Moran's I, although the difficulty to create the weight matrix $\mathbf{W}$, *400* by *400*, as function of contiguous spatial keep going; so we segment the housing sample by the number of rooms to obtain efficiency gains from the time of computing results. We choose the two smaller sub-segments sample, one the houses with one room, denominated T1, with 33 observations, the others with four rooms, denominated T4, with 39 observations, both with the follow spatial distribution,

	Centro	**Lanhas**	**SReméd**	**CVelhos**	**Luz**	**PMileu**	**RDiz**	**Pinheiro**	**GGare**
T1	16	3	3	1	1	0	0	0	9
T4	9	3	4	4	4	1	1	1	12

The oldest and best test known is Moran's *I* for regression residuals, is a locally best invariant test, and moments and estimation details are given in Cliff and Ord (1981), and Anselin (1988). For those sub-segments the Moran's *I* are, $I_{T1} = 0,1786$ and

$\tau_j^2 = \sum_{t=j+1}^{T} e_t e_{t-j} \Big/ \sum_{t=1}^{T} e_t^2$, and T is the number of observations. Under the null hypothesis, Q is asymptotically distributed as a χ^2, with degrees of freedom equal to the number of autocorrelation.

$I_{T4} = -0,0310$, respectively. The Moran coefficient[16] is identical to LM-error tests, and reads as:

$$LM_{\rho} = \frac{1}{T}(\frac{\mathbf{e}^{'}\mathbf{We}}{s^2})^2$$

where s^2 is the maximum likelihood variance $\mathbf{e'e}/n$, T is a scalar computed as the trace of a quadratic expression in the weight matrix, $T = tr(\mathbf{W'W} + \mathbf{W}^2)$, and the test asymptotically follows a χ^2 distribution with one degree of freedom. In these cases we have $\chi^2_{[1]\mathrm{T1}} = 4{,}131$ and $\chi^2_{[1]\mathrm{T4}} = 0{,}173$, then the hypothesis of one specific misspecification with spatial error, under high level of significance, can be rejected.

V. Conclusion

This spatial question appears as a marginal problem when the goal is achieving an efficient appraisal method for housing assessment; so we analysed this question in this context, nevertheless the growing availability of easy–to–use software for spatial analysis makes it tempting to simply ‘try’ all kinds of different spatial models and techniques. The tools of spatial statistics including geostatistics bring new explorative opportunities on real estate markets. From the statistical point of view, a usage of spatial statistics method gives us more accurate estimations enabling more precise inference what means in practice that we have more explicit insight in mechanisms and processes occurring on real estate market than previously.

The cost-efficiency analysis relatively to this research recommend parsimonious but no omission, therefore the present essay try answered to the spatial doubt underlying to hedonic regressions.

[16] Is a scaled coefficient, since it’s calculated for row-standardized weights.

Bibliography

Anas, A.; Eum, S. J.. 1984. “Hedonic Analysis of a Housing Market in Disequilibrium”. *Journal of Urban Economics,* 15 (1) .

Anselin, L., and Bera, A. K. 1998. “Spatial Dependence in Linear Regression Models with an Introduction to Spatial Econometrics”. In *Handbook of Applied Economic Statistics* (A.Ullah, and D. Giles, Eds.). Marcel Dekker, New York.

Basu, S. and Thibodeau, T.G. 1998. “Analysis of Spatial Autocorrelation in House Prices”. *The Journal of Real Estate Finance and Economics*, 17(1), 61 - 85.

Can, Ayse.1992. “Specification and Estimation of Hedonic Housing Price Models”. *Regional Science and Urban Economics*, 22, 453 – 474.

Cliff, A.D., and Ord, J.K. 1981. SPATIAL PROCESSES: MODELS &APPLICATIONS. Pion, London.

Dubin, R. A..1992. “Spatial Autocorrelation and Neighborhood Quality”. *Regional. Science Urban Economics,* 22, 433–452.

Dubin, R. A..1998. “Spatial Autocorrelation: A Primer”.*Journal of Housing Economics,* 7, 304-327.

Epple, D.. 1987. “Hedonic Prices and Implicit Markets: Estimating Demand and Supply Functions for Differential Products”. *Journal of Political Economy*, 95(1), 59 – 80.

Eviews4. 2002. COMMAND AND PROGRAMMING REFERENCE. Quantitative Micro-Software.

Florax, Raymond, and Nijkamp P. 2004. “Misspecification in Linear Spatial Regression Models”. Forthcoming in the *Encyclopedia of Social Measurement*, San Diego: Academic Press.

Freeman, A. Myrick. 1993. “The Measurement of Environmental and Resource Values: Theory and Methods”. in *Resources for The Future*, Washington D.C., 104-105.

Kelejian, H. H., and Prucha, I. R. 2001. “On the Asymptotic Distribution of the Moran I test Statistic with Applications. *Journal of. Econometrics*, 104, 219–257.

Kulshreshtha, S. N., and Gillies J. A.. 1993. “Economic Evaluation of Aesthetic Amenities: A Case Study of River View,” *Water Resources Bulletin*, 29(2), 257-266.

Lykke E. Andersen, Reis E. J.. 1997. “Deforestation, Development, and Government Policy in the Brazilian Amazon: An Econometric Analysis”. *IPEA*-Texto para Discussão Nº 513 - Trabalho desenvolvido no âmbito do NEMESIS.

Pace, R. K. and Gilley O.. 1998. “Optimally Combining OLS and the Grid Estimator”. *Real Estate Economics*, 26 (2),331-347.

Pace, R. Kelley e Gilles, Otis W.. 1997. “Using The Spatial Configuration of the Data to Improve Estimation”. *Journal of Real Estate Finance and Economics*, 14 (3), 333 – 340.

Rosen, S. 1974. “Hedonic prices and implicit markets: Product differentiation in pure competition”. *Journal of Political Economy*, 82: 34 – 55.

Annexe I: Identification and Definition for Variables

Variable	**Definition**
AGE	Age of house (years)
AQCENTRAL	Dummy variable: one if house is heated by central system, else zero.
AREA	Total floor space house (unit for area, m^2)
ARRECA	Dummy variable: one if house has loft, else zero.
BATHROOMS	Number of bathrooms
CV	Dummy variable: one if house has placed on underground floor, else zero.
ELEV	Dummy variable: one if house has lift system, else zero.
EXPSOL	Dummy variable: one if house has a good sunlight, else zero.
GCANAL	Dummy variable: one if house has central gas system, else zero.
GSDOS	Dummy variable: one if house has two garages, else zero.
GSUM	Dummy variable: one if house has one garage, else zero.
JANDUPLAS	Dummy variable: one if house has double windows, else zero.
LAR	Dummy variable: one if house has inglenook, else zero.
LCNTRO	Dummy variable: one if house has situated at Centro, else zero.
LCVLHOS	Dummy variable: one if house has situated at CVelhos, else zero.
LGGARE	Dummy variable: one if house has situated at GuardaGare, else zero.
LLAMNH	Dummy variable: one if house has situated at Lameirinhas, else zero.
LLUZ	Dummy variable: one if house has situated at BairroLuz, else zero.
LNVMERCADO	Transaction price of the *i*-th house sold in 2002.
LPMILEU	Dummy variable: one if house has situated at PovoaMileu, else zero.
LPNHR	Dummy variable: one if house has situated at BPinheiro, else zero.
LRDIZ	Dummy variable: one if house has situated at RioDiz, else zero.
LSREMD	Dummy variable: one if house has situated at SraRemedios, else zero.
PPRIM	Dummy variable: one if house has located on first floor, else zero.
PQRT	Dummy variable: one if house has located on fourth floor, else zero.
PSEG	Dummy variable: one if house has located on second floor, else zero.
PTERC	Dummy variable: one if house has located on third floor, else zero.
ROOMS	Number of rooms
RCH	Dummy variable: one if house has located on ground floor, else zero.