

# **Geostatistical Air Pollution Indexes in Spatial Hedonic Models:**

## **The case of Madrid, Spain**

Gema Fernández-Avilés, Román Mínguez and José-María Montero

University of Castilla-La Mancha

Gema Fernández-Avilés  
University of Castilla-La Mancha  
Cobertizo S. Pedro Mártir, S/N. 45071 (Toledo) Spain  
Phone: +34 925 26 88 00; Fax: +34 925 26 88 01  
[gema.faviles@uclm.es](mailto:gema.faviles@uclm.es)

Román Mínguez  
University of Castilla-La Mancha  
Avenida de los Alfares, 44, 16002 (Cuenca) Spain  
Phone: +34 969 179 100  
Fax. +34 969 179 107  
[roman.minguez@uclm.es](mailto:roman.minguez@uclm.es)

José-María Montero (University of Castilla-La Mancha)  
University of Castilla-La Mancha  
Cobertizo S. Pedro Mártir, S/N. 45071 (Toledo) Spain  
Phone: +34 925 26 88 00; Fax: +34 925 26 88 01  
[jose.mlorenzo@uclm.es](mailto:jose.mlorenzo@uclm.es)

# **Geostatistical Air Pollution Indexes in Spatial Hedonic Models:**

## **The case of Madrid, Spain**

### **Abstract**

Recent work has shown how spatial econometrics can be applied to a housing-value hedonic equation that includes air pollution variables. In this paper we propose a Spatial Durbin Model that incorporates an Air Pollution Index, instead of one or two pollution variables, and considers spatial and non-spatial endogeneity jointly. We do, however, assume both possibilities for the index: exogenous and endogenous. Unlike most empirical studies in the literature, we have focused our empirical analysis on a European city, Madrid. In order to do so, we constructed a massive data base in 2009 that includes almost 12,000 dwellings.

**Key words:** air pollution, kriging, endogeneity, spatial econometrics, Geostatistics.

**JEL code:** Q51, Q53, R31, C21.

## **I. Introduction**

According to literature on environmental economics, it is reasonable to assume that air pollution enters into the utility function of potential house buyers. Without the intention to be exhaustive, Chay and Greenstone (2005), Anselin and Le Gallo (2006), and Anselin and Lozano-Gracia (2008) are three of the most cited recent examples. It is therefore no surprise that hedonic house price models that incorporate environmental variables among the set of explanatory variables are becoming increasingly popular.

In the context of the framework of hedonic price theory, the traditional approach to this problem has been to use the housing market to infer the implicit prices of these non-market goods. Under standard assumptions of perfect competition, information and mobility, and the maximisation of well-behaved preferences, hedonic theory unambiguously predicts that the implicit price function relating housing prices to an environmental amenity will be positively sloped, all else equal.

But empirical research concerning the effects of pollution on property values does not confirm such an unambiguous positive relation between housing prices and air quality. On the contrary, past results are far from conclusive and there are serious doubts that air pollution significantly affects the price of properties. These non conclusive results could be attributed to the non consideration of the spatial argument in the hedonic model (that is, the implicit prices of housing attributes were considered as stationary). But recent spatial hedonic models continue to provide non conclusive results.

Most of the articles that include pollution in spatial hedonic specifications only consider one or two pollutants —usually  $PM_{10}$  and/or tropospheric ozone, because they are the most visible in the form of “smog”. However, while this is true, it is a fact that “smog” is derived from all kinds of pollutants. This is why we aim to contribute to the

literature by proposing a general spatial specification, the spatial Durbin model (SDM), which includes not only one or two pollutants, as usual, but an air pollution index (API).

Since there is a mismatch between the sites where we observe the prices of houses and the locations of the monitoring stations, geostatistical techniques are required. In the literature, the usual solution to the above mentioned problem is to interpolate the environmental variables to obtain their interpolated values at the locations where house prices are available. Several interpolative alternatives have been considered in recent research: Thiessen polygons, the inverse distance method, splines, and kriging and cokriging. But kriging (when dealing with one environmental variable) and cokriging procedures (when dealing with more than one) have important advantages (Anselin and Le Gallo, 2006, Chica-Olmo, 2007, and Montero and Larraz, 2011). Since we gather all pollution information in an only index, kriging is used as interpolative method in this article. However, we firstly interpolate the environmental variables considered and, subsequently, elaborate the environmental index because the variance of the estimation errors is less than when the index is first constructed and then interpolated. As far as we are aware, no articles in the literature include an API in a spatial hedonic model in the way we propose.

However, this interaction between Spatial Econometrics and the Geostatistical method for constructing APIs that we propose could lead to an “errors in variables” problem (Anselin and Lozano-Gracia, 2008), which would render usual estimation techniques useless and would require instrumental variable estimation. As stated in Anselin and Lozano-Gracia (2008), the use of spatially interpolated values of pollution results in a prediction error which may be correlated with the overall model disturbance term. And this would lead to simultaneity bias in an ordinary least squares (OLS) regression.

In this article we use the same methodology as in the previous literature, that is, a hedonic specification, for comparative reasons. But, (i) we include the spatial argument in the analysis; (ii) as Anselin and Lozano-Gracia (2008) and unlike previous studies, we consider the air pollution interpolated variable not only as an exogenous regressor, but also as an endogenous explanatory variable; and (iii) we plan the same specification as in Osland (2010), which does not include air pollution in the set of explanatory variables. Proceeding in this way we extend Osland's (2010) research by incorporating not only specific air pollutants, but an API. We also propose a new method for constructing APIs and are able to compare our results with the seminal paper of Anselin and Lozano-Gracia (2008).

We apply the above-mentioned novelties to Madrid, Spain. There are several important reasons for choosing Madrid for our case study: (i) First and foremost, the majority of the empirical research in the literature refers to American cities; (ii) population is highly concerned with the environment in general and air quality in particular (iii) construction, and specifically the residential sector, is of great importance in the global economy; and (iii) it can be said that in Madrid there is almost perfect information about air quality all over the city, which solves the problems related to how much is known about air quality variables. According to Clark and Allison (1999), the impact of air quality variables on the hedonic price function depends on how much is known about them.

Our results are in line with previous results in the empirical literature and this is why we suspect that, taking for granted that hedonic models are still the best strategy for measuring the impact of air quality on the price of real estate properties, objective measures of pollution are not the right variable to be included when measuring the impact of air pollution on housing prices.

After this introduction, Section II is devoted to the literature review. Section III includes our proposal for constructing APIs using kriged procedures in such a way that variance of the prediction errors is minimised. In Section IV we briefly describe the SDM, which is the specification that we use in this article. Section V is devoted to the case study: Madrid. First we give some details about air pollution and the housing sector in the study area. Second, we comment on both the air quality and housing data sets used in this research. Third, we present the main results obtained from including the API in the spatial hedonic specification proposed in Section IV. Finally, some concluding remarks are reported in Section VI.

## **II. Literature review**

### *A. Literature on traditional hedonic models for valuating pollution*

There have been two well-known meta-analyses and two comprehensive literature reviews on the effect of contamination on real estate values. But more than 95% of the studies were conducted in US cities. One can also find specific studies for some cities in Europe, Asia and Latin America, Straszheim (1974), Anselin, (1988), Anselin and Bera (1998), but research on the topic in non-US cities is very scarce.

Smith and Huang (1995) conducted a meta-analysis of 37 air pollution studies providing 86 estimates of marginal willingness to pay (MWTP) for a reduction in PM<sub>10</sub> (air pollution particulate of ten microns in diameter) in the period 1982–1984. Simons and Saginor (2006) addressed how proximity to source influences environmental contamination effects on residential property values. Their research includes a summary of 58 peer-reviewed technical journal articles and selected case studies from over 100 articles. Boyle and Kiel (2001) reviewed over 30 exclusively hedonic price studies and their effect on residential property. Jackson (2001) considered about 45 articles that dealt with the effects of environmental contamination on real estate, covering real estate

appraisal theory and sales price analysis. His articles included hedonic regression analysis, case studies and reported appraisal outcomes. The residential studies were published from 1982 onwards. Simons and Saginor (2006) reviewed over 100 peer-reviewed articles on proximity influence (both positive and negative) for residential and commercial property. Hedonic regression was the predominant methodology, but sale-resale analysis, conjoint analysis and similar techniques were also used.

However, Smith and Huang's (1995) meta-analysis suggested that a one-unit reduction in  $PM_{10}$  ( $mg/m^3$ ) results in a 0.05-0.10 percent increase in property values. Boyle and Kiel (2001) found that air studies produced mixed results and posited that measurement factors are not generally known to homebuyers. Jackson (2001) offered no final observations on the consistency of the findings and called for a more systematic study. Simons and Saginor (2006), who included air pollution together with concentrated animal feeding operations, obtained a different sign in the coefficient of the variable depending on whether the model included positive amenities or not.

In summary, as stated in the introductory section, the literature concerning the effects of contamination on property values reveals that the effect of air pollution on property value is far from conclusive. What is more, there are serious doubts that air pollution significantly affects the price of properties. Additionally, the type of study may also generate different results. As stated in Chay and Greenstone (1998), exogenous differences in air quality are extremely difficult to isolate, because the "true" relationship between air pollution and the price of properties may be obscured in cross-sectional analysis by unobserved determinants of housing prices that co-vary with air pollution. For example, areas with high levels of pollution tend to have well educated populations with higher per capita income and population densities. Economic activity is also a driving force in the determination of property values, but differences across

space in the level of activity may also underlie changes in the level of pollution. Of course, the above circumstances lead to a spurious positive relationship between pollution and property values. When Chay and Greenstone use the conventional cross-sectional estimates of the relationship between property values and  $PM_{10}$ , they conclude that the relationship is weak, unstable and indeterminate.

### *B. Literature on spatial hedonic models for valuating pollution*

Recently a successful line of research has emerged that includes the spatial argument in the hedonic specification. As Straszheim (1988) stated many years ago, it may not be appropriate to assume that the implicit prices of housing attributes are stationary across geographic space, even within a big city. On the supply side, homes near each other tend to be similar and, on the demand side, homebuyers regularly emulate one another's behaviour. The result is a process of spatial interaction among market participants, which at least suggests that the first-stage hedonic price function should be modified to include a spatial lag of its dependent variable (Anselin, 1988). This spatial lag can be interpreted as a flexible fixed effect that absorbs the existing and unobserved spatial correlation of supply and/or demand. Recommended literature that considers the spatial argument in the specification of the hedonic model includes the pioneer works: Can (1990) and Can (1992), as well as Kim et al. (2003), Theebe (2004), Brasington and Hite (2005), Anselin and Le Gallo (2006), Anselin and Lozano-Gracia (2008), Osland (2010), Bourassa, et al, 2010, and Ready, 2010, among many others.

### **III. Constructing APIs: an alternative kriged procedure**

When researching air quality in a particular city in a real case, it is impossible to obtain extensive (or even complete) data at every desired point because of practical constraints.

Thus, interpolation is crucial for graphing, analysing and understanding environmental results. Among all the existing interpolation methods, Geostatistics uses kriging to account for spatial dependence. Kriging is a univariate procedure which interpolates the values of the target random function at unobserved locations using the available observations of the same random function. This interpolation procedure produces the best unbiased linear estimator and uses the covariance or variogram function (the spatial equivalent of the autocorrelation function in time series analysis) to account for the correlation in making interpolative estimates.

Kriging can be viewed as a strategy equivalent to time series, but in space. It is based on the idea of stochastic processes or random functions over space, taking into account the multidirectional feature of the space at a specific moment in time. This approach applies to a wide range of phenomena and implies dealing with an infinite family of random variables  $X(\mathbf{s})$  constructed at all points  $\mathbf{s}$  in a region. The variables take different values depending on location and correlation and each set of observed data is supposed to be a realisation of the random function under study. For an extensive review see Cressie (1993).

Next, we focus on the geostatistical contribution of the paper: why kriging the environmental variables and then elaborating an environmental index is a better option than the usual procedure in the literature that consists of building an environmental index to be eventually interpolated (kriged). We use cokriging notations, which are more general than kriging notations, but the simplicity criterion leads us to use kriging, as cokriging obtains a hardly noticeable benefit in relation to kriging in the isotropic case.

Let  $X_1, X_2, \dots, X_K$  be the level of  $k$  different pollutants measured at site  $\mathbf{s}_i$ , assumed to be intrinsic stationary random functions of order zero, and consider an API at such a site given by the weighted mean:

$$API(\mathbf{s}_i) = \sum_{k=1}^K a_k X_k(\mathbf{s}_i) = \mathbf{A}'\mathbf{X} \quad (1)$$

where  $\mathbf{A}' = (a_1, \dots, a_K)$  is the vector that includes the loadings of the  $K$  pollutants in the API, and is usually obtained by principal component analysis (PCA), and  $\mathbf{X}' = (X_1(\mathbf{s}_i), \dots, X_K(\mathbf{s}_i))$  contains the values of the  $K$  pollutants at the  $n$  sites provided with a monitoring station.

Assuming that information about pollution is measured at monitoring stations located at sites  $\mathbf{s}_i, i = 1, \dots, n$ , the two options to linearly estimate the value of API at a location  $\mathbf{s}_j$  where housing prices are known but information on pollution is unknown are:

- (i) Elaborate an API at every site  $\mathbf{s}_i, i = 1, \dots, n$ , where there is a monitoring station using the environmental information provided by the stations, and then obtain kriged estimates of API at locations  $\mathbf{s}_j, j = 1, \dots, m$ , where housing prices have been observed but information on pollution is unknown, that is,

$$API^*(\mathbf{s}_j) = \sum_{i=1}^n \lambda_i API(\mathbf{s}_i) = \sum_{i=1}^n \lambda_i \sum_{k=1}^K a_k X_k(\mathbf{s}_i), \quad j = 1, \dots, m. \quad (2)$$

where the  $\lambda_i, i = 1, \dots, n$ , are the weights of the values of the API at locations provided with a monitoring station in the value of the API at the non observed site  $\mathbf{s}_j$ .

(ii) Cokrige  $X_1(\mathbf{s}), \dots, X_K(\mathbf{s})$  at locations  $\mathbf{s}_j, j=1, \dots, m$ , where housing prices have been observed and there are no monitoring stations, and then form  $API(\mathbf{s}_j)$  as:

$$\square API(\mathbf{s}_j) = \mathbf{A}' \mathbf{X}_j^* = \sum_{k=1}^K a_k X_k^*(\mathbf{s}_j) = \sum_{k=1}^K a_k \sum_{i=1}^{n_j} \lambda_i^k X_k(s_i^k), \quad j=1, \dots, m. \quad (3)$$

That is to say, in a first stage we cokrige separately the value of every pollutant at locations  $\mathbf{s}_j, j=1, \dots, m$ , obtaining  $X_k^*(\mathbf{s}_j)$ . At this stage, the value of the specific pollutant we are interpolating,  $X_k^*(\mathbf{s}_j)$ , is obtained as a weighted mean of the values of the  $k$  pollutants at the  $n$  monitoring stations where they are measured, rather than as a weighted mean of the values of the same pollutant measured at the stations; the  $\lambda_i^k$  represent the weight of the value of the  $k$ -th pollutant at the  $i$ -th monitoring station in the value of the specific pollutant we are interpolating at a non observed site. In a second stage, the interpolated values of the considered pollutants at the non observed sites are combined in an API, the loadings,  $a_k$ , being obtained by PCA.

Following Myers (1983), in general

$$Var[API^*(\mathbf{s}) - API(\mathbf{s})] > Var[\square API(\mathbf{s}) - API(\mathbf{s})] \quad (4)$$

that is, the variance of the prediction errors is less in the case of option (ii) or, in other words, replacing the vector of contaminant values at a given location by a weighted linear combination and then kriging such a linear combination, the usual procedure, is not the best option.

#### **IV. Specification and estimation of spatial hedonic models that include interpolative variables.**

The potential “errors in variables” aspect of interpolated air pollution measures has been practically ignored by the literature. Some worth-mentioning exceptions are van de Kastele and Velders (2006), Anselin and Lozano-Gracia (2008), and Lopiano et al. (2011). But, in our opinion, it is a core aspect when dealing with interpolated environmental measures as regressors. The reason is that the use of spatially interpolated values of pollution results in a prediction error which may be correlated with the overall model disturbance term. And this could lead to simultaneity bias in an OLS regression and also in non-spatial hedonic models (as usual, the spatially lagged dependent variable in spatial models is an endogenous regressor and, as a result, there is a simultaneity bias apart from the interpolated air pollution).

Taking this into account, we propose the specification and estimation of two types of models: a classical non-spatial hedonic model and a SDM defined by the equation:

$$\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \alpha \mathbf{i}_n + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n), \quad (5)$$

where  $\mathbf{y}$  is a  $(n \times 1)$  vector including the observations of the logarithms of the house prices,  $\mathbf{X}$  is a  $(n \times k)$  matrix comprising the observations of the individual and areal characteristics associated to each dwelling and other spatial variables such as the API, surface, condition, mean mortgage in the neighbourhood, etc.,  $\mathbf{i}_n$  is a  $(n \times 1)$  unit vector for the intercept and  $\mathbf{W}$  is the  $(n \times n)$  spatial weights matrix. On the other hand,  $\rho$  is a spatial parameter that measures the existing spatial dependence of the dependent variable,  $\alpha$  is the intercept parameter,  $\sigma^2$  is the variance of the noise under homoskedasticity and  $\boldsymbol{\beta}$  and  $\boldsymbol{\theta}$  are  $(k \times 1)$  vectors of parameters associated to the independent variables and their lags, respectively. While in the SDM we impose the restrictions  $\rho = 0, \boldsymbol{\theta} = \mathbf{0}$ , the non-spatial hedonic model is obtained as a particular case.

It is important to note that in SDM spatial spillovers (effects of changes in the independent variables on the dependent variable) are not given by any vector of parameters directly. The SDM can be rewritten as (LeSage and Page, 2009, pp. 34-35):

$$\mathbf{y} = \sum_{r=1}^k (\mathbf{I}_n - \rho \mathbf{W})^{-1} (\mathbf{I}_n \boldsymbol{\beta}_r + \mathbf{W} \boldsymbol{\theta}_r) \mathbf{x}_r + (\mathbf{I}_n - \rho \mathbf{W})^{-1} \alpha \mathbf{i}_n + (\mathbf{I}_n - \rho \mathbf{W})^{-1} \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n) \quad (6)$$

where  $\mathbf{x}_r$  is a  $(n \times 1)$  vector of observations of the correspondent independent variable.

Now, we can compute the effect of a change in the  $j$ th observation of  $\mathbf{x}_r$  on the  $i$ th observation in  $\mathbf{y}$  as:

$$\frac{\partial y_i}{\partial x_{jr}} = \left( (\mathbf{I}_n - \rho \mathbf{W})^{-1} (\mathbf{I}_n \boldsymbol{\beta}_r + \mathbf{W} \boldsymbol{\theta}_r) \right)_{ij} = S_r(\mathbf{W})_{ij} \quad (7)$$

Therefore, the  $(n \times n)$   $S_r(\mathbf{W})$  matrix includes all the spatial spillovers or effects of the independent variables on the dependent variable. Of course, these spillovers are a complex function of all the parameters of the model and, as a consequence, we should be cautious when it comes to interpreting the parameters separately. The principal diagonal of  $S_r(\mathbf{W})$  measures the direct effects, that is, the impact of a change in  $x_{ir}$  on the observation  $i$  of the dependent variable, and the off-diagonal elements represent indirect impacts, i.e., changes in the observation  $i$  derived from a change in  $x_{jr}$ .

The sum across the  $i$ th row of  $S_r(\mathbf{W})$  represents the total impact on the individual observation  $y_i$  resulting from changing the  $r$ th explanatory variable by the same amount across all  $n$  observations. On the other hand, the sum across the  $i$ th column reflects the total impact on all  $y_i$  arising from changing the  $r$ th explanatory variable by the amount in the  $j$ th observation.

One of the main advantages of SDM is that if we set some restrictions in this model, it is possible to obtain other well-known spatial models (see Osland, 2010, for

details). As the SDM framework nests those models, it is robust under different specifications.

In the estimation process we have considered the potential endogeneity of the API as it acts as a regressor. Thus, we have estimated the model using both the classical Maximum Likelihood (ML) procedure (assuming that the only endogenous regressor is the spatially lagged dependent variable) and the spatial Two-Stage Least Squares (TSLS) method (assuming that the spatially lagged dependent variable is not the only endogenous regressor, as the API variable is also endogenous). This method allows us to compare classical spatial ML estimates (see Le Sage and Pace, 2009, for details) to TSLS spatial estimates (Kelejian and Prucha, 1998, and Anselin, 2008, are two excellent references). In order to implement the TSLS estimation procedure, we have used spatial coordinates and lagged APIs (first, second, third and fourth quarter of 2008) as instrumental variables for the errors-in-variables problem. As in Anselin and Lozano-Gracia (2008), we opted for latitude and longitude, which are able to proxy the overall spatial pattern of pollution as a global spatial trend. As a consequence, they are unlikely to be correlated with hedonic error terms, which reflect local spatial patterns of omitted variables. As stated in Anselin and Lozano-Gracia (2008), these instruments may also help to correct endogeneity due to other factors. In addition, lagged APIs –which are strongly correlated with the API and are not supposed to be correlated with the regression error– substantially improve the efficiency of the estimators.

We have tested whether the spatial coordinates and lagged APIs are weak instruments or not. For this purpose we have implemented the Staiger and Stock (1997) test and followed the Stock and Watson rule-of thumb (Stock and Watson, 2007). Removing the spatial coordinates from the set of regressors, the F-statistic is 363.76. When considering the lagged APIs as instruments, the F-statistic is 13,447. Taking into

account that in the Staiger-Stock test the null is “weak instruments” and that the critical value proposed by Stock and Watson is  $F\text{-statistic} = 10$ , we can conclude that both the spatial coordinates and the lagged APIs are not weak instruments.

Finally, as far as we know there is no software for carrying out TSLS estimation of spillovers. For this purpose we used a proprietary code based on the following steps:

1. Estimate the spatial models (Spatial Autoregressive Model, or SAR, and SDM) by TSLS using the coordinates and the lagged APIs as instruments.
2. Define an ML object as the estimated spatial model (in this model spillovers can be computed by simulation, as usual).
3. Replace the coefficients, standard deviations and the remainder of results in the ML object with the TSLS estimates.
4. Compute the spillovers in the ML object with the new estimates.

## **V. Case study: Incorporating an API in a spatial hedonic housing price model for Madrid**

Madrid (the capital of Spain) is the third most populous city in the European Union (6,271,638 inhabitants in 2009, 3,213,271 of whom live in the city). In the last decade, the number of vehicles in Madrid has increased by 5.6%, amounting in 2009 to a total of 1,917,382. This implies 1,202.5 vehicles per km. and 683.5 vehicles per 1,000 inhabitants. Two million drivers enter and leave the city on a daily basis.

**[Figure 1 here]**

So, car pressure is increasing as well as its negative environmental impacts. Nevertheless, air pollution in Madrid may also be attributed to other factors, such as manufacturing and heating systems during winter, among others.

Following EU directives, in this article we have considered the following six pollutants: sulphur dioxide (SO<sub>2</sub>), nitrogen oxides (NO<sub>x</sub>) —which is a generic term for mono-nitrogen oxides (nitric oxide (NO) and nitrogen dioxide (NO<sub>2</sub>))—, carbon monoxide (CO), particulate matter (PM<sub>10</sub>) and ground-level ozone (O<sub>3</sub>). According to the Department of Environmental Assessment, Control and Quality of Madrid, the mean values of those pollutants in the city in 2009 were as follows: SO<sub>2</sub> (11 micrograms/m<sup>3</sup>), suspended particulate matter (27 micrograms/dry standard m<sup>3</sup>), CO (0.4 milligrams/m<sup>3</sup>), NO<sub>x</sub> (96 micrograms/m<sup>3</sup>), NO<sub>2</sub> (55 micrograms/m<sup>3</sup>) and O<sub>3</sub> (39 micrograms/m<sup>3</sup>).

Construction, particularly residential, is an extremely important part of Madrid's global economy. According to the Spanish Regional Accounts 2009, this sector contributes 8.6% of total GDP. Madrid is the city with the largest housing stock in Spain (11.5% of the total) and is also the main housing market: in 2009 some 53,513 housing transactions were carried out in Madrid (Spanish Housing Office). On the other hand, in 2008 the percentage of home ownership was 78.7% (2,275,188 out of 2,890,229).

The most central districts of the city are well established areas where few new houses are built and large projects are uncommon due to the lack of available land. The second-hand home market is witnessing a slight upturn due to many homeowners putting their dwellings up for sale as a result of not being able to meet their mortgage payments; however, this circumstance does not invalidate the use of a hedonic pricing model. In addition, there is only a token presence of State-subsidised housing. Current supply focuses mainly on second-hand homes, the price of which, due to the characteristics of the area itself, has remained unchanged due to their quality and advantageous location. Prices will however tend to decrease as these two arguments lose strength.

It is worth highlighting the districts of: a) Salamanca (prices ranging from 3,750 to 11,549 euros/m<sup>2</sup>) which currently has a highly variable number of empty dwellings, most of which belong to the second-hand home market. The majority of sales made have affected dwellings in the lower echelon of prices, although the high level of purchasing power required to buy a house in the area does not significantly affect prices, which can rise in some cases; b) Chamartín (prices ranging from 4,225 to 11,183 euros/m<sup>2</sup>) has a very small housing stock, generally open blocks with few individual houses, communal gardens and swimming pools. New house development is minimal due to a lack of building sites, the majority of dwellings being second-hand, some 30 years old and medium to high quality; c) Hortaleza (prices ranging from 2,667 to 6,458 euros/m<sup>2</sup>), where there are mainly multi-family dwellings, although semi-detached and luxury stand-alone houses are abundant in some areas. These districts are practically new and have grown markedly in recent years. In the vicinity of the M-40 urban motorway, luxury houses are currently being built in closed housing developments with large communal areas.

Peripheral districts are characterised by the balance between public and private housing development. At present, new house prices are decreasing, while this downward trend is much more pronounced in the case of second-hand homes, with very few transactions being made. Building activity has stagnated, focusing on a very small number of developments to replace houses and with little sign of the market picking up. As in some central districts, and for the same reason, the second-hand home market is witnessing a slight upturn.

### *A. Data sets*

The issue of housing prices remains unresolved in Spain. This is the reason why we have constructed our own database for Madrid. The final database we have created contains the price and characteristics of 11,796 owner-occupied single family homes (see Figure 1). This database was created from the sales that took place in Madrid in the last quarter of 2009. It is important to note that the sample accounts for 90% of the sales in that quarter. The list of variables we have used mirrors the usual set used in the literature (see Table A and B in Appendix). Most of them have been codified as categorical to allow for more flexibility in the specification of the model. This allows for nonlinearities between the different levels of each variable.

**[Figure 2 here]**

As for the data relative to pollution, they were provided by the Atmosphere Pollution Monitoring System of Madrid. As said above, we deal with six pollutants: SO<sub>2</sub>, NO<sub>x</sub>, NO<sub>2</sub>, CO, PM<sub>10</sub> and O<sub>3</sub> (see Table C in Appendix). Note that in the specialised literature, hedonic specifications typically include only O<sub>3</sub> (Banzhaf, 2005, Hendrix et al., 2005, and Anselin and Le Gallo, 2006, among others), PM<sub>10</sub> (Chay and Greenstone, 2005 and Murthy et al. 2009), or both O<sub>3</sub> and PM<sub>10</sub> (Anselin and Lozano-Gracia, 2008 is a recent example), as these two are the most visible in the form of “smog” and are thought to have the greatest health impact. But a workable approach to environmental data should consider multiple contaminants. Obviously, including six variables in a spatial hedonic house price model is not an easy task, so we decided to incorporate an air pollution index that gathers the information contained in such variables.

The location of the air quality monitoring stations is shown in Figure 2. Most of them are located in the urban centre and relatively few in peripheral areas. Therefore,

the monitoring stations cover the area under study reasonably well, as most of Madrid's population is concentrated in the urban centre. Pollution measurements were taken in February 2009 at 10 a.m, and we have used the monthly average. There were two reasons for this decision: (i) February is the month of the year that records the most pollution; (ii) following the "population affected criterion", 10 a.m. is a critical time. We have also tried with the simultaneous API (fourth quarter of 2009) and results are practically the same.

Another possibility (that used by Anselin and Lozano-Gracia, 2008) is to average the daily maxima during the worst quarter of a particular year. We rejected this option because the spatial structure of dependencies is not the same every hour, and the averaging process could lead to compensate such different structures. In any case, the question of when to measure pollution is a core aspect. The mismatch between the location of monitoring stations and the sites where the houses have been transacted must be overcome by kriging and the structure of the spatial dependencies of the level of a particular pollutant depends on time and is highly sensitive to temporal aggregation.

The values of the pollutants measured at the monitoring stations have been interpolated to the locations of every house in the sample. In light of the empirical data, we found ordinary kriging (OK) to be the best strategy for interpolating the above mentioned values. We used the ML method to select the valid semivariograms. These semivariograms and their corresponding estimated parameters are displayed in Table 1. We used the GeoR package (Diggle and Ribeiro, 2007) for that purpose as well as carrying out the cross validation procedure and obtaining kriged estimates.

**[Table 1 here]**

It is a well-known fact that interpolated values become less accurate for locations far from monitoring stations. Thus, following Anselin and Lozano-Gracia

(2008), we removed the house locations within the upper 5% of the prediction error distribution of either pollutant included in the API from the sample. Proceeding in this way, we adjust for the possible bias stemming from high error interpolated values. This resulted in a final set of 11,796 house locations.

Table 2 reports the weights of the pollutants included in the API (we used PCA for this purpose). The variables have been previously standardised. When estimating the API for Madrid, our alternative reduces the variance of the prediction errors estimated with the usual procedure by more than 7%. Figure 4 reports the interpolated values of the API at house sites. The reported values are arranged in a quartile map so they can be compared to the spatial quartile map of house prices shown in Figure 3.

**[Table 2 here]**

**[Figure 3 here]**

**[Figure 4 here]**

## *B. Estimation results*

### *B.1. The standard hedonic model*

We first obtain OLS estimates (assuming that all regressors are exogenous) and TSLS estimates (assuming that the API is an endogenous regressor) for a standard hedonic model (Table 3), and test for the presence of spatial autocorrelation in the residuals using the usual Lagrange Multiplier test statistics for error and lag dependence (Table 4). Computation for this and the rest of the econometric models has been carried out using the Spatial Econometrics Toolbox written in Matlab by LeSage (1999) and the `spdep` library written in R by Bivand (2010).

In order to perform the TSLS estimation, we use the longitude and latitude coordinates (available at all spatial points and observed without measurement error) and the values of the API in the first, second, third and fourth quarters of 2008 as instrumental variables for the errors-in-variables problem. Initially, it seems feasible

that due the kriging method itself, the API variable will be correlated to the chosen instruments, and the instruments should not be correlated to the disturbances of the hedonic model.

**[Table 3 here]**

**[Table 4 here]**

As regards the influence of the API on the logarithm of the price per square metre of housing (Table 3), the OLS estimation is positive but not significant, while the TSLS estimation is, unexpectedly, positive and significant. The rest of the coefficients of the model generally display the signs initially expected and most of them are significant both in the OLS and also TSLS estimations. It is important to highlight that, in the case of an endogenous API, only the TSLS estimates are consistent (assuming that the instruments satisfy the usual orthogonality conditions). Moreover, the Durbin-Wu-Hausman statistic records a value of 56.84 with a p-value of 0.00. Therefore, this test rejects the null hypothesis that the API is exogenous and suggests that only TSLS estimations would be consistent.

However, from Table 4 it can be deduced that there is a strong evidence of spatial autocorrelation, which suggests the specification of alternative spatial hedonic models.

### *B.2. The spatial hedonic specifications*

SDM is the best option to capture, as generally as possible, the detected spatial dependence because this specification includes spatial lags both of the dependent variable and also the regressors (equation 5). In the specification of this spatial alternative we have used a spatial weights matrix that takes into account the six closest neighbours. As usual, the weights matrix is used in row-standardised form.

Nevertheless, we have checked that results do not vary significantly when other weights matrices are used (matrices with a different number of neighbours, Delaunay triangles from a Voronoi tessellation, etc.).

The SDM was estimated using ML, assuming that the API is exogenous, and TSLS, under the assumption that the API is endogenous. In this last case, only the TSLS method provides consistent estimations although, if the API were exogenous, ML estimations would be consistent and efficient.

When estimating the SDM with TSLS, after several trials we included  $\mathbf{W}^2\mathbf{X}$ ,  $\mathbf{W}^3\mathbf{X}$ ,  $\mathbf{W}^4\mathbf{X}$  and  $\mathbf{W}^5\mathbf{X}$  as instruments for  $\mathbf{W}\mathbf{y}$  in addition to the instruments for the errors-in-variables problem. The new instruments have been chosen in line with Anselin and Lozano-Gracia (2008), Kelejian and Prucha (1998) and Anselin (2007). Note that the TSLS estimation method must use the same set of instruments for both  $\mathbf{W}\mathbf{y}$  and API (see Verbeek, 2008). As for potential multicollinearity problems derived from the inclusion of the new instruments for the spatial lag, the results of the SDM could be affected by some numerical instability.

Tables 5 and 6 display total, direct and indirect spillovers calculated on the basis of equation (7) using ML and TSLS, respectively. As such spillovers are generally different for each observation  $i = 1, \dots, n$ , Tables 5 and 6 include the average values of the spillovers for all observations. In order to take into account the uncertainty regarding the parameters estimated when calculating the spillovers, 1,000 simulations are performed using different values for parameters each time. These values are obtained from the asymptotic distribution of the estimators, that is, in each simulation the values  $\boldsymbol{\gamma} = (\rho, \boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T$  are obtained by extracting a value from the distribution  $N(\bar{\boldsymbol{\gamma}}, \bar{VAR}(\hat{\boldsymbol{\gamma}}))$  where  $\hat{\boldsymbol{\gamma}}$  represents the vector of parameters estimated in the SDM and the matrix

$\hat{VAR}(\hat{\gamma})$  is the correspondent estimated variance-covariance matrix (LeSage and Pace, 2009).

The following findings can be extracted from those Tables:

(i) The  $\rho$  parameter, which measures spatial dependence in SDM, is significant and positive. The absolute value of  $\rho$  with ML estimation is approximately 0.25, in line with recent research (between Osland, 2010, and Anselin and Lozano-Gracia, 2008, estimates). The value of  $\rho$  when estimating the SDM with TSLS is 0.86, although the API spillovers are still insignificant. This result could be due to multicollinearity problems derived from the inclusion of the new instruments for the spatial lag, which produces numerical instability.

(ii) The total spillover of the API is positive and not significant in both the ML and TSLS estimation. What is more, neither direct nor indirect spillovers were found to be significant in either of the estimations, which bearing in mind the possible endogeneity of the API (in which case the TSLS estimation would be more suitable), shows that, when the model takes spatial dependence into account, there is no evidence that the API has any influence on housing prices.

(iii) The rest of total spillovers estimated display the expected signs and are significant in the ML estimation. This is not the case with the TSLS estimation due to the numerical instability that stems from the inclusion of  $\mathbf{W}^2\mathbf{X}$ ,  $\mathbf{W}^3\mathbf{X}$ ,  $\mathbf{W}^4\mathbf{X}$  and  $\mathbf{W}^5\mathbf{X}$  as instruments for  $\mathbf{W}\mathbf{y}$ . Note that the above instruments are strongly correlated with the term  $\mathbf{W}\mathbf{X}$  in the SDM and, as a result, the estimate of  $\rho$  is inflated and indirect spillovers almost (and in some cases more than) offset direct spillovers. In any case, direct spillovers have the expected sign and, in general, are significant.

**[Table 5 here]**

**[Table 6 here]**

As is well known, SAR and the Spatial Error Model (SEM) are more parsimonious spatial alternatives to the standard hedonic model than SDM, as they are nested in it. Of course, in case that there is no significant difference between SDM and one of those models when it comes to fit the data, the parsimonious option would be preferred.

When deciding if we can use one of the above parsimonious spatial alternatives instead of the SDM, we distinguish two cases: the API is considered endogenous and the API is assumed to be exogenous. As for the first case, Table 7 summarizes the results of the two statistical measures we have used to take the decision: the Akaike's Information Criterion (AIC) and the standard deviation of the error. As can be appreciated, the SDM (estimated using both ML and TSLS) minimises the value of the AIC and therefore appears to be a more suitable specification than the other spatial alternatives.

Note that Table 7 also includes the results for the standard hedonic model, to compare it with the spatial strategies. As was to be expected, the non-spatial model (OLS and TSLS) performed worse than the spatial specifications.

In addition, we have performed an ANOVA test to compare the SAR nested model vs. the SDM. The test is based on the sum of squared TSLS residuals and clearly rejects the null hypothesis (SAR) and favours the alternative (SDM).

**[Table 7 here]**

As for computing the AIC in the case of estimating the SDM with TSLS, note that for linear models the AIC can be computed using the ML of  $\sigma^2$ , given

by  $\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n}$ , where  $e_i$  are the ML residuals. In this case, the AIC can be written as

(Verbeek, 2008, p. 61):

$$AIC(M) = \log \hat{\sigma}^2 + 2 \frac{p(M)}{n}$$

where  $M$  indicates the model,  $p(M)$  represents the number of parameters in such a model and  $n$  is the number of observations.

If we compute  $\hat{\sigma}^2$  from the TSLS residuals in the above expression, it is possible to obtain the AIC for the TSLS estimation method. Of course, as we use TSLS residuals instead of ML residuals, the value of the AIC should be taken only as an approximate value.

In case that the API is assumed to be exogenous, as the SAR and SEM spatial models are individual cases of the general SDM, we can proceed performing likelihood-ratio (LR) tests, the null hypothesis being the suitability of the restricted model (SAR or SEM) in comparison to the general SDM. Table 8 shows the result of these tests, which reject the null hypothesis in all cases, indicating the preference for the SDM model ahead of the rest.

**[Table 8 here]**

Summing up, irrespective of whether the API is considered exogenous or endogenous, SDM is preferred to SAR and SEM, and the spatial specifications are preferred to the standard hedonic model.

### *B.3. Comments on API spillovers*

Although SDM is preferred to SAR and SEM specifications and, of course, to the standard hedonic model, in Table 9 we have summarized the API spillovers for all the above specifications. It can be observed that, irrespective of the specification, the API total spillover is positive and, in general, not significant. The only exception is the non-spatial model estimated with TSLS (where spillovers are clearly significant). The

spillovers obtained by the SAR model estimated with TSLS are on the verge of being significant.

**[Table 9 here]**

The positive sign (albeit not significant) displayed by API total spillover constitutes a counterintuitive result and does not empirically confirm the hedonic theory. This “unexpected” result could be explained as follows: (i) it is not true that empirical evidence unambiguously confirms the hedonic theory. In fact, after more than thirty years of research, the relationship between pollutants and housing prices appears, in the best of cases, relatively weak; (ii) it could be more than possible that the subjective environmental perception of citizens does not match objective measures; and (iii) we are afraid to say that the American and European cases are not equal. Thus, it is not surprising that the influence of clean air (or pollution) on house prices is quite different. American and European cities are organised differently (and the people who live in them are entirely different in terms of culture). The decision to buy a house in a European city like Madrid does not depend on air quality in the area, but on the commercial services, communications, level of income, economic activity, etc. of the neighbourhood. And it just so happens that the areas with the highest levels of the above-mentioned variables are the most polluted areas in the city. If our finding is confirmed, it would lead to an open question: Is the concept of “willingness to pay for clean air” the same in American and European cities? In other words, does both the different concept of city in the United States and Europe and different environmental culture (especially in Mediterranean countries) lead to air quality having different effects on house prices?

However, the other possible reason behind this “unexpected” finding is that our results could be questionable from the point of view of data, instruments and

methodological issues. From the perspective of the instruments, as indicated, the question of how to specify the right instruments remains unsolved for many economic problems and this could perfectly be one of them. Therefore, more theoretical and empirical research is needed to confirm our findings. In any case, the use of coordinates as instruments seems to be an appropriate option as errors-in-variables problems that arise from the use of interpolative procedures are avoided. The use of lagged APIs as instruments improves the efficiency of estimators due to the strong positive correlation with the API. As for the data relative to air pollution, although kriging, and especially OK, is the recommended interpolative procedure (Anselin and LeGallo, 2006, and Anselin and Lozano-Gracia, 2008) for this type of problems, we would like to outline some drawbacks of the procedure in the environmental field, including:

- (i) First, the amount of actual data on pollution is enormously scarce with respect to the number of dwellings in the sample because there used to be between 20 and 30 monitoring stations in the best of cases in large cities. This means that estimates are needed for practically all locations where dwelling prices are sampled.
- (ii) Second, it is an interpolative method and, as a consequence, provides a weighted mean of the values of neighbouring locations as estimates. That means that "peak-estimates" cannot be obtained, the value of which exceeds the values of neighbouring locations. Obviously this feature makes the OK strategy strongly dependent on where monitoring stations are located. And we have to bear in mind that the location of monitoring stations in a large city does not obey statistical criteria (design of an optimal monitoring network) but other criteria that include municipal regulations.

(iii) And third, in large cities most of the monitoring stations are located in the city centre, only a few being on the periphery. This implies that in peripheral neighbourhoods, OLS estimates tend to be near to the mean pollution value and their variance is higher than desirable.

These three drawbacks could explain why empirical results contradict or, in the best of cases, only slightly confirm, the hedonic theory.

And finally, we should not forget that, as stated in Chay and Greenstone (1998), exogenous differences in air quality are extremely difficult to isolate, because the “true” relationship between air pollution and the price of properties may be obscured in cross-section analysis by unobserved determinants of housing prices that co-vary with air pollution. And this is still an unsolved question.

## **VI. Conclusion and discussion**

After more than thirty years of research, the relationship between pollutants and housing prices appears, in the best of cases, relatively weak. Additionally, research has been focused on American cities. Since American cities are quite different to European cities, we have applied the newest and most sophisticated procedures provided by Geostatistics and Spatial Econometrics to a European case: Madrid. We have also proposed a new way of constructing APIs. But results do not confirm the hedonic theory: the implicit price function relating housing prices to air quality will be positively sloped, assuming all else equal; and we conclude that, at least in this European case, housing prices could be related to the commercial services, communications, level of income, economic activity, etc. of the neighbourhood, but not to air pollution. Note that in some of the central districts of Madrid during the sample period, the second-hand

home market is witnessing a slight upturn, but this circumstance does not invalidate the use of a hedonic pricing model.

In particular, the sign of the API coefficient is positive (albeit not significant) and contradicts the results in Anselin and Lozano-Gracia (2008). Initially, this finding could be considered surprising, but there is some rationale behind it, as we will see next.

The two areas under study are quite different and represent quite different frameworks and even cultures. The first great difference is that in Anselin and Lozano-Gracia (2008), the total surface area of the region under study (Orange County and the non-desert portions of Los Angeles, San Bernardino, and Riverside Counties) encompasses a population of almost 17 million and an area of over 10,000 square miles. However the population in Madrid City was 3,213,271 in 2009 and its surface area only 375.9 square miles.

The second important difference is that in the South Coast Air Quality Management District (AQMD) of South California, although air quality has significantly improved over the past few decades due to the efforts of the AQMD, the region still has the worst air quality in the nation in regard to smog (Ozone) and fine particulate matter ( $PM_{2.5}$ ). However, in Madrid City, as in other important European cities, troposphere ozone, as well as particulate matter, is not the main problem, but  $NO_2$  due to traffic emissions. Troposphere ozone is considered a problem in the Madrid Mountains and particulate matter is a problem in the municipalities located in the southern part of Madrid.

Additionally, in our modest opinion, Anselin and Lozano Gracia (2008) do not have a sufficient number of monitoring stations to obtain credible interpolations for suspended particulate matters.

In light of our results—in line with previous results in the empirical literature—and taking for granted that hedonic models are still the best strategy for measuring the impact of air quality on the price of real estate properties, we have the suspicion that objective measures of pollution are not the right variable to be included when measuring the impact of air pollution on housing prices.

API total spillovers in the selected model, SDM, are found not to be significant (irrespective of the estimation method). It is true that, in this specification, TSLS estimation could be affected by multicollinearity. Nonetheless, this problem does not affect direct spillovers to a great extent, but indirect spillovers and, as a consequence, total spillovers. In this case, that is, when the API is assumed to be endogenous, the SAR specification, which is not affected by multicollinearity when estimating with TSLS, could be a more reliable specification to estimate indirect and total effects. In any case, the indirect and total spillovers derived from the TSLS-SAR model are in line with the effects obtained by ML-SAR and ML-SDM.

Of course, our results could be questionable from the point of view of data, instruments and methodological issues (see section V). But in any case, after three decades of inconclusive research, some data and methodological issues should be addressed and new hypotheses considered. More research in the field of instrumental variables is required, new non linear specifications which include alternatives to  $\mathbf{W}$  for taking into account spatial dependencies should be developed (for example, including a term for collecting the large scale spatial dependencies along with  $\mathbf{W}$  representing short and medium scale dependencies) and innovation in interpolation procedures should be tried, at least while there are so few monitoring stations in comparison to the number of dwellings in the model. On the other hand, studies including perceived pollution (using

annoyance surveys or census-based measures) are needed and we should consider the hypothesis that individuals may not really care that much about air pollution.

Hence, three open questions arise from the above lines: Do citizens identify living in the city with pollution and, as a consequence, does pollution not enter into their utility function? Should the willingness to pay for clear air only be assessed for secondary residences? And, should the objective measures of pollution be replaced by house buyers' subjective perception of pollution? These are undoubtedly three promising avenues for future research. Other interesting lines of research, in the context of data and methods, refer to extending this work to the field of spatial time series—this approach might make it possible to test whether variation in pollution has a significant impact on housing prices—to the design of optimal observation networks for pollution and to search for optimal instruments that are able to proxy the overall spatial pattern of pollution as a global spatial trend.

**Table 1. Valid Semivariograms for Pollutants**

Variable	Semivariogram model	Nugget effect	Partial sill	Range	$\phi_A$	$\phi_B$
CO	Spherical	0.15	0.82	4000.00	0.00	1.00
PM <sub>10</sub>	Spherical	0.77	0.24	3999.97	0.00	11.30
O <sub>3</sub>	Wave	1.04	1.00	1940.40	0.00	1.00
NO <sub>x</sub>	Pure Nugget	1.00	0.00	3301.91	6.23	3345.90
NO <sub>2</sub>	Pure Nugget	1.00	0.00	1184.26	0.00	1.01
SO <sub>2</sub>	Exponential	0.08	1.09	3999.99	0.00	1.07

$\phi_A$ : value (in radians) for the anisotropy angle.

$\phi_B$ : value for the anisotropy ratio (always greater than 1).

**Table 2. Weights of Pollutants in the API (PCA)**

<b>SO<sub>2</sub></b>	<b>CO</b>	<b>NO<sub>2</sub></b>	<b>PM<sub>10</sub></b>	<b>NOx</b>	<b>O<sub>3</sub></b>	<b>Percentage of variance extracted</b>
0.64	0.86	0.74	0.84	0.96	-0.74	61.80

**Table 3. Coefficient estimates for Non-Spatial Models**

VARIABLE	OLS		TSLS	
	Coeff.	t-st.	Coeff.	t-stat.
API	0.0039	1.46	0.0129	4.40
Relative altitude	0.0191	9.01	0.0186	8.78
Noise	-0.0007	-1.72	-0.0007	-1.68
M.30	0.1125	13.21	0.1000	11.55
M.30.2	0.0360	6.11	0.0342	5.80
Shopping area	-0.0027	-0.27	-0.0025	-0.25
Historical quarter	0.0971	8.64	0.0883	7.81
Built up area	-0.0005	-13.35	-0.0005	-13.47
Mortgage reference area	0.0001	36.56	0.0001	36.43
Monthly mortgage	0.0001	24.86	0.0001	24.89
Age	-0.0016	-7.96	-0.0016	-8.03
Density pop. distr.	-0.0001	-1.51	-0.0002	-2.47
Retired (% distr.)	0.0011	0.95	0.0005	0.40
Children (% distr.)	-0.0191	-8.02	-0.0209	-8.74
Immigrants (% distr.)	-0.0044	-6.37	-0.0043	-6.25
Trees (% Ha. distr.)	-0.0002	-0.22	0.0006	0.75
Baths	0.0142	3.57	0.0144	3.61
House	0.0272	1.75	0.0294	1.89
Studio-apartment	0.0580	3.78	0.0574	3.75
Flat	-0.0589	-6.99	-0.0588	-6.99
Floor.1st	0.0450	3.86	0.0453	3.89
Floor.2nd-3rd	0.0326	2.83	0.0333	2.89
Floor.4th-5th	0.0437	3.89	0.0442	3.94
Floor.6th or more	0.0590	4.68	0.0592	4.70
Elevator	0.0609	11.85	0.0607	11.82
Air conditioning	0.0506	11.06	0.0508	11.12
Good condition	0.0620	10.61	0.0623	10.67

Swimming pool	0.0613	9.16	0.0620	9.27
Garage	0.0437	7.68	0.0440	7.73
<hr/>				
Durbin-Wu-Hausman	56.8463 (0.00)			

\* The  $t$ -statistics test, as null hypothesis, the non significance of the corresponding variable. The asymptotic distribution is one  $N(0,1)$ , with critical values at 5% and 1% are 1.96 and 2.58, respectively.

\* In the Durbin-Wu-Hausman test, the null hypothesis is that the API is exogenous. The value in parentheses is the p-value associated to the value of the statistic.

**Table 4: Lagrange multiplier diagnostics for spatial dependence**

	<b>LM-Lag</b>	<b>LM-Lag Rob.</b>	<b>LM-Err</b>	<b>LM-Err Rob.</b>
<b>OLS</b>	555.045 (0.00)	58.048 (0.00)	605.872 (0.00)	108.875 (0.00)
<b>TSLS</b>	543.262 (0.00)	50.651 (0.00)	610.841 (0.00)	118.230 (0.00)

\* LM-lags test a non-spatial hedonic model (null hypothesis) against a SAR model (alternative hypothesis). LM-Err tests contrast a non-spatial hedonic model (null hypothesis) against a SEM (alternative hypothesis). In both cases, we display the test statistic, the asymptotic distribution under  $H_0$  being a  $\chi^2_{(1)}$ , together with the associated  $p$ -value.

**Table 5. Spillovers and Spatial Dependence ( $\rho$ ) in SDM Models**

**(ML estimation)**

Variables	Total		Direct		Indirect	
	Coeff.	t-stat.	Coeff.	t-stat.	Coeff.	t-stat.
$\rho$	0.2670	29.78				
API	0.0057	1.52	-0.0164	-0.56	0.0222	0.74
Relative altitude	-0.0129	-3.67	0.0806	26.19	-0.0935	-21.80
Noise	-0.0016	-3.08	0.0046	0.80	-0.0063	-1.08
M.30	0.0760	6.29	-0.0095	-0.27	0.0855	2.27
M.30.2	0.0253	2.96	-0.0275	-1.20	0.0527	2.17
Shopping area	0.0085	0.55	0.0364	2.20	-0.0280	-1.27
Historical quarter	0.0603	3.54	0.1207	6.55	-0.0604	-2.47
Built up area	-0.0005	-3.63	-0.0006	-14.47	0.0001	0.67
Mortgage refer. area	0.0002	22.62	0.0001	23.22	0.0001	7.49
Monthly mortgage	0.0001	10.00	0.0001	23.42	0.0001	2.59
Age	-0.0004	-0.83	-0.0018	-8.50	0.0014	2.85
Density pop. distr.	-0.0002	-1.64	-0.0001	-0.51	-0.0001	-0.69
Retired (% distr.)	0.0006	0.33	0.0068	3.25	-0.0062	-2.30
Children (% distr.)	-0.0213	-5.71	-0.0106	-2.72	-0.0106	-2.14
Immigrants (% distr.)	0.0016	1.44	-0.0043	-3.87	0.0059	4.10
Trees (% Ha. distr.)	0.0010	0.78	-0.0015	-0.86	0.0025	1.18
Baths	0.0202	1.69	0.0150	4.05	0.0052	0.48
House	-0.1030	-2.40	0.0256	1.67	-0.1286	-3.15
Studio-apartment	0.1229	2.63	0.0514	3.52	0.0715	1.71
Flat	-0.0700	-2.74	-0.0565	-7.12	-0.0135	-0.59
Floor. 1 <sup>st</sup>	0.0315	0.90	0.0434	4.10	-0.0119	-0.36
Floor. 2nd-3rd	0.0008	0.02	0.0294	2.72	-0.0285	-0.87
Floor.6th or more	0.0117	0.35	0.0440	4.28	-0.0324	-1.03
Floor.4th-5 <sup>th</sup>	0.0265	0.72	0.0582	5.03	-0.0317	-0.92

Elevator	0.0650	5.13	0.0566	10.90	0.0084	0.69
Air conditioning	0.0886	6.00	0.0495	11.47	0.0391	2.89
Good condition	0.0595	3.16	0.0543	9.96	0.0053	0.30
Swimming pool	0.0684	4.22	0.0486	6.94	0.0199	1.27
Garage	0.0493	3.07	0.0412	7.59	0.0081	0.54

---

**Table 6. Spillovers and Spatial Dependence ( $\rho$ ) in SDM Models**

**(TSLS estimation)**

Variables	Total		Direct		Indirect	
	Coeff.	t-stat.	Coeff.	t-stat.	Coeff.	t-stat.
$\rho$	0.8584	15.82				
API	0.0197	0.18	-0.0379	-0.37	0.0576	0.37
Relative altitude	-0.0149	-0.41	0.0862	23.23	-0.1011	-2.95
Noise	0.0003	0.01	0.0040	0.60	-0.0037	-0.10
M.30	0.0583	0.27	-0.0121	-0.32	0.0703	0.33
M.30.2	0.0582	0.18	-0.0164	-0.70	0.0745	0.24
Shopping area	0.1081	0.10	0.0428	1.39	0.0653	0.06
Historical quarter	0.0035	0.02	0.1226	6.09	-0.1192	-0.79
Built up area	0.0009	0.12	-0.0005	-2.62	0.0014	0.19
Mortgage refer. area	0.0002	0.50	0.0001	10.79	0.0001	0.18
Monthly mortgage	0.0001	0.25	0.0001	10.55	0.0001	-0.16
Age	0.0001	0.02	-0.0018	-5.43	0.0019	0.28
Density pop. distr.	-0.0005	-0.32	-0.0001	-0.79	-0.0003	-0.24
Retired (% distr.)	0.0059	0.09	0.0076	2.85	-0.0017	-0.03
Children (% distr.)	-0.0263	-0.23	-0.0116	-2.27	-0.0147	-0.13
Immigrants (% distr.)	0.0069	0.09	-0.0045	-2.13	0.0114	0.16
Parkland (% Ha. distr.)	0.0022	0.22	-0.0014	-0.75	0.0036	0.37
Baths	-0.0332	-0.10	0.0123	1.26	-0.0455	-0.14
House	-0.2316	-0.28	0.0256	0.83	-0.2571	-0.32
Studio-apartment	0.3205	0.14	0.0564	0.95	0.2641	0.12
Flat	0.0251	0.03	-0.0512	-2.28	0.0763	0.10
Floor. 1 <sup>st</sup>	-0.0949	-0.06	0.0423	1.04	-0.1372	-0.09
Floor.2nd-3rd	-0.0914	-0.06	0.0309	0.74	-0.1223	-0.08
Floor.4th-5 <sup>th</sup>	-0.1547	-0.08	0.0430	0.88	-0.1977	-0.10
Floor.6th or more	-0.1364	-0.07	0.0564	1.12	-0.1929	-0.10

Elevator	0.0781	0.38	0.0569	6.79	0.0212	0.11
Air conditioning	0.1363	0.95	0.0500	6.36	0.0863	0.63
Good condition	0.0225	0.14	0.0515	4.99	-0.0290	-0.19
Swimming pool	0.0444	0.29	0.0452	4.38	-0.0008	-0.01
Garage	0.1110	0.18	0.0442	2.69	0.0668	0.11

---

**Table 7: The Endogenous Case:**

**Measures of the Relative Goodness of Fit of the Standard and Spatial Hedonic Models; ANOVA TSLS test: SAR vs. SDM**

	NON-SPATIAL MODELS		SPATIAL MODELS				
	OLS	TSLS	SDM		SAR		SEM
			ML	TSLS	ML	TSLS	ML
<i>n</i>	11796	11796	11796	11796	11796	11796	11796
$\sigma$	21.25%	21.26%	19.97%	20.29%	20.74%	20.76%	20.64%
<i>p</i> ( <i>M</i> )	37	37	73	73	37	37	37
AIC	-1.34	-1.34	-1.38	-1.37	-1.36	-1.35	-1.36
Log-Likelihood			6280.29		5849.07		5875.38

**ANOVA TSLS TEST**

SAR( $H_0$ ) vs. SDM( $H_1$ ) F = 15.02 (0.00)

*p*(*M*) represents the number of parameters in the model.

ANOVA TSLS test: The nested SAR model vs. the SDM. This test is based on comparing the sum of squared residuals between the restricted and unrestricted models, both estimated by TSLS. The asymptotic distribution of the test statistic is  $F_{n_1, n_2}$  with  $n_1$  the number of restrictions imposed by the restricted model (SAR) and  $n_2$  the degrees of freedom of the unrestricted model (SDM). The values in parentheses are the p-values associated to each test statistic.

**Table 8. The Exogenous Case: LR Test for Selecting Spatial Models**

LR TESTS (ML estimation)		
SAR( $H_0$ ) – SDM( $H_1$ )	862.45	(0.00)
SEM( $H_0$ ) – SDM( $H_1$ )	809.83	(0.00)

\* Likelihood ratio tests: The nested (SAR or SEM) model vs. the more general model (SDM). The asymptotic distribution of the test statistic is a  $\chi^2$  with degrees of freedom equal to the number of restrictions imposed by the corresponding nested model. The values in parentheses are the p-values associated to each test statistic.

**Table 9: Summary of API Spillovers and Spatial Dependence ( $\rho$ )**

	NON-SPATIAL MODELS		SPATIAL MODELS				
	OLS	TSLS	SDM		SAR		SEM
			ML	TSLS	ML	TSLS	ML
$\rho$			0.267	0.858	0.234	0.216	0.277
			(29.78)	(15.82)	(65.44)	(11.88)	(46.43)
<b>API spillovers</b>							
Total	0.0039	0.0129	0.006	0.019	1.1E-4	0.0074	0.0034
	(1.462)	(4.404)	(1.52)	(0.18)	(0.33)	(2.014)	(0.993)
Direct	0.0039	0.0129	-0.016	-0.038	8.8E-4	0.0059	0.0034
	(1.462)	(4.404)	(-0.56)	(-0.37)	(0.33)	(2.010)	(0.993)
Indirect			0.022	0.057	2.6E-3	0.0015	
			(0.74)	(0.37)	(0.33)	(1.990)	

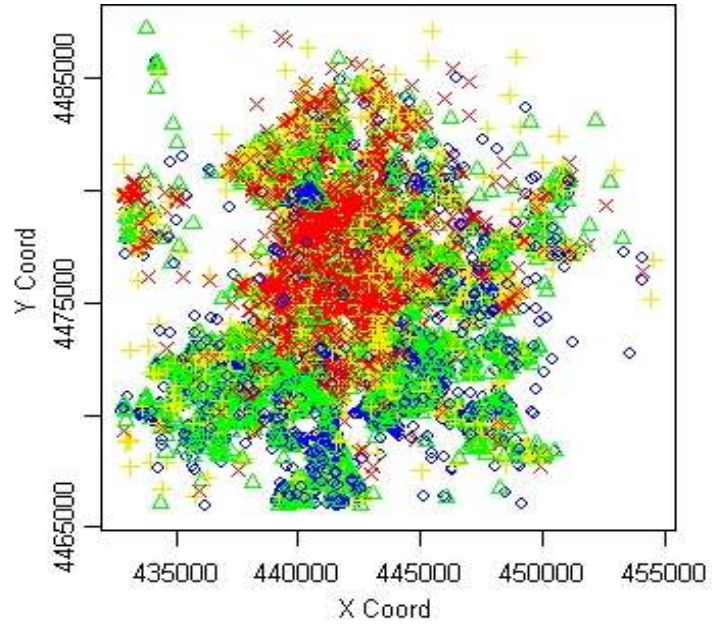
\*The direct and total spillovers for the non-spatial and SEM models coincide with the  $\beta_i$  coefficients of the corresponding models. There are no indirect spillovers in these models. For the SDM and SAR models, the spillovers (direct, indirect and total) are computed using equation (7).



**Figure 1. Location of Madrid**

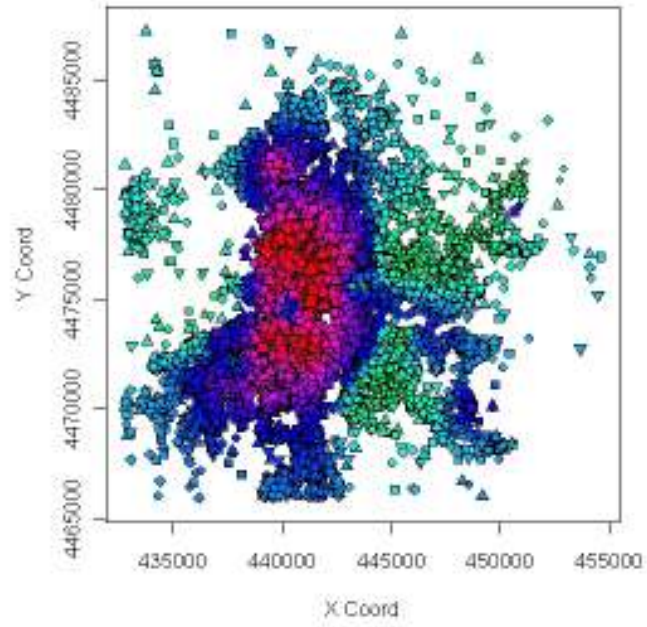


**Figure 2. Location of monitoring stations**



(x): First quartile; (+) Second quartile; ( $\Delta$ ) Third quartile; ( $\bullet$ ) Fourth quartile;

**Figure 3. Quartile map of house prices**



Red: First quartile; Dark blue: Second quartile; Light blue: Third quartile; Light green: Fourth quartile

**Figure 4. API: interpolated (kriged) values at houses sites**

## References

- Anselin, L., *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers: Boston, MA, 1988.
- Anselin, L., Spatial Econometrics. In Mills, T.C. and Patterson, K (Eds.) *Palgrave Handbook of Econometrics*. Vol I. Palgrave MacMillan: New York, NY, 2007, 901–970.
- Anselin, L. and Le Gallo, J., Interpolation of Air Quality Measures in Hedonic House Price Models: Spatial Aspects. *Spatial Economic Analysis*, 2006, 1(1), 31–52.
- Anselin, L. and Lozano-Gracia, N., Errors in variables and spatial effects in hedonic house price models of ambient air quality, *Empirical Economics*, 2008, 34, 5–34.
- Banzhaf, H.S., Green Price Indices, *Journal of Environmental Economics and Management*, 2005, 49(2), 262–280.
- Bivand, R., *Spdep: Spatial dependence: weighting schemes, statistics and models*. R package version 0.5-21, 2010. Available at <http://CRAN.R-project.org/package=spdep>.
- Boyle, M.A., Kiel, K.A., A Survey of House Price Hedonic Studies of the Impact of Environmental Externalities, *Journal of Real Estate Literature*, 2001, 9, 117–144.
- Bourassa, S.C., E. Cantoni, and M. Hoesli. Predicting House Prices with Spatial Dependence: A Comparison of Alternative Methods. *Journal of Real Estate Research*, 2010, 32:2, 139–60.
- Brasington, D.M. and Hite, D., Demand for Environmental Quality: A Spatial Hedonic Analysis. *Regional Science and Urban Economics*, 2005, 35, 57 – 82.
- Can, A., The Measurement of Neighborhood Dynamics in Urban House Prices. *Economic Geography*, 1990, 66, 254 – 272.

- Can, A., Specification and Estimation of Hedonic Housing Price Models. *Regional Science and Urban Economics*, 1992, 22, 453 – 474.
- Chay, K.Y. and Greenstone, M., Does Air Quality Matter? Evidence from the Housing Market, 1998. NBER Working Paper No. 6826.
- Chay, K.Y. and Greenstone, M., Does air quality matter? Evidence from the housing market. *Journal of Political Economy*, 2005, 113: 2, 376 – 424.
- Chica-Olmo, J. Prediction of Housing Location Price by a Multivariate Spatial Method: Cokriging. *Journal of Real Estate Research*, 2007, 29:1, 91–114.
- Clark, D.E. and Allison, T., Spent Nuclear Fuel and Residential Property Values: The Influence of Proximity, Visual Cues and Public Information. *Papers in Regional Science*, 1999, 78, 403 – 421.
- Cressie, N.A.C., *Statistics for Spatial Data*. John Wiley & Sons: New York, NY. 1993.
- Diggle, P.J., Ribeiro, P.J., *Model-Based Geostatistics*. Springer: New York. 2007.
- Hendrix, M.E., Hartley, P.R. and Osherson, D., Real Estate Values and Air Pollution: Measured Levels and Subjective Expectations; Discussion Paper, Rice University, 2005.
- Jackson, T., The Effects of Environmental Contamination on Real Estate: A Literature Review, *Journal of Real Estate Literature*, 2001, 9(2), 93–116.
- Kelejian, H.H. and Prucha, I., A Generalized Spatial Two Stages Least Squares Procedures for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances. *Journal of Real Estate Finance and Economics*, 1998, 17, 29 – 121.
- Kim, C-W., Phipps, T.T. and Anselin, L., Measuring the benefits of air quality improvement: A spatial hedonic approach. *Journal of Environmental Economics and Management*, 2003, 45, 24 – 39.

- LeSage, J., *The Theory and Practice of Spatial Econometrics*, 1999. Available at <http://www.spatial-econometrics.com/html/sbook.pdf>.
- LeSage, J. and Pace, R.K., *Introduction to Spatial Econometrics*. Chapman & Hall/CRC: Boca Raton, Florida, 2009.
- Lopiano, K.A., Young, L. and Gotway, C.A., A comparison of errors in variables methods for use in regression models with spatially misaligned data. *Statistical Methods in Medical Research*, 2011, 20(1), 129-147.
- Montero, J.M. and Larraz, B. Interpolation Methods for Geographical Data: Housing and Commercial Establishment Markets. *Journal of Real Estate Research*, 2011, 33 21, 233–244.
- Murthy, M.N., Gulati, S.C. and Banerjee, A., Hedonic property prices and valuation of benefits from reducing urban air pollution in India. Delhi Discussion Papers, 2009, 61, Institute of Economic Growth: Delhi, India.
- Myers, D.E., Estimation of linear combinations and cokriging. *Mathematical Geology*, 1983, 15, 633 – 637.
- Osland, L., An Application of Spatial Econometrics in Relation to Hedonic House Price Modeling. *Journal of Real Estate Research*, 2010, 32(3), 289 – 320.
- Ready, R.C. Do Landfills Always Depress Nearby Property Values? *Journal of Real Estate Research*, 2010, 32:3, 233–244.
- Simons, R. and Saginor, J., A meta-analysis of the effect of environmental contamination and positive amenities on residential real estate, *Journal of Real Estate Research*, 2006, 28(1), 71-104.
- Smith, V.K. and Huang, J.C. Can markets value air quality? A meta-analysis of hedonic property value models. *Journal of Political Economy*, 1995, 103, 209-227.

- Staiger, D. and Stock, J.H., Instrumental Variables Regression with Weak Instruments. *Econometrica*, 1997, 65, 557-586.
- Stock, J.H. and Watson, M.W., *Introduction to Econometrics*, International 2<sup>nd</sup> edition, Addison-Wesley: Boston, MA. 2007.
- Straszheim, M., Hedonic Estimation of Housing Market Prices: A Further Comment, *Review of Economics and Statistics*, 1974, 56, 404 – 406.
- Theebe, M.A.J., Planes, Trains, and Automobiles: The Impact of Traffic Noise on House Prices. *Journal of Real Estate Finance and Economics*, 2004, 28, 209 – 235.
- van de Kastele, J. and Velders, G.J.M., Uncertainty assessment of local NO<sub>2</sub> concentrations derived from error-in-variable external drift kriging and its relationship to the 2010 air quality standard, *Atmospheric Environment*, 2006, 40(14), 2583-2595.
- Verbeek, M., *A Guide to Modern Econometrics*. John Wiley & Sons: Chichester, 2008.

## Appendix:

**Table A. Variable names and description**

Variable name	Description
<b>Dependent variable</b>	
Price	House price
<b>Variable of interest</b>	
API	Air Pollution Indicator
<b>Coordinates</b>	
Coordinate x	Longitude
Coordinate y	Latitude
<b>House characteristics</b>	
Relative altitude	Indicator for relative altitude (radius: 800m.)
Noise	Indicator for noise dB(A)
Good condition	Indicator variable for good condition
Flat	Indicator variable for flats
Studio-apartment	Indicator variable for studios
Top-floor flat	Indicator variable for top-floor flats
House	Indicator variable for houses
Age	Age of houses
Ground level	Indicator variable for ground level
Floor.1st	Indicator variable for floor 1st
Floor.2nd - 3rd	Indicator variable for floor 2nd and floor 3rd
Floor.4th - 5th	Indicator variable for floor 4th - 5th
Floor.6th or more	Indicator variable for floor 6th or more
Baths	Number of bathrooms
Garage	Indicator variable for parking space

Elevator	Indicator variable for lift
Air conditioning	Indicator variable for central air-conditioning
Swimming pool	Indicator variable for swimming pool
Monthly mortgage	Monthly mortgage

**Areal characteristics**

M.30	Indicator for housing inside the M-30
M.30.2	Indicator for housing close to the M-30
Shopping area	Indicator for houses in the shopping area
Historical quarter	Indicator for houses in the historical quarter
Built up area	Number of square metres of built-up area
Density pop. distr.	Population density in the district
Retired (% distr.)	Percentage of retired people in the district
Children (% distr.)	Percentage of children under 14
Immigrants (% distr.)	Percentage of immigrants in the district
Trees (% Ha. distr.)	Trees per Ha. in the district
Mortgage reference area	Mean mortgage in the area

---

**Table B. Main descriptive statistics****Table B.1. Quantitative Variables**

<b>QUANTITATIVE VARIABLES: MAIN DESCRIPTIVES</b>					
	<b>Range</b>	<b>1st Qu.</b>	<b>Median</b>	<b>Mean</b>	<b>3rd Qu.</b>
<b>Price (Euro)</b>	1,113	2,827	3,562	3,738	4,429
<b>API</b>	10.45	9.16	9.64	9.86	10.38
<b>Relative altitude</b>	1.84	1.96	2.04	2.04	2.12
<b>Noise</b>	55.41	58.94	61.28	61.34	63.55
<b>Monthly mortgage</b>	12,701	762	1067	1,435	1,639
<b>Mortgage reference area</b>	8,346	2,851	3,337	3,456	3,950
<b>Age</b>	100	15	28.3	23.7	30
<b>Built up area</b>	2,14.1	62.2	81	100.2	110
<b>Density pop. distr.</b>	302.9	73.63	180	168.1	272.0
<b>Retired (% distr.)</b>	11.17	16.56	18.68	18.50	20.78
<b>Children (% distr.)</b>	9.47	11.18	12.78	12.94	14.24
<b>Immigrants (% distr.)</b>	14.47	8.04	12.07	12.55	17.56
<b>Trees (%Ha distr.)</b>	0.44	0.07	0.11	0.14	0.20
<b>Baths</b>	6	1	1	1.5	2

**Table B.2. Qualitative Variables**

QUALITATIVE VARIABLES: MAIN CHARACTERISTICS					
<b>Type</b>	Flat	House	Studio- apartment	Top-floor flat	
	87.30%	3.92%	2.41%	6.37%	
<b>Floor</b>	Ground floor	Floor.1st	Floor.2nd-3rd	Floor.4th-5th	Floor.6th
	4.39%	23.09%	37.51%	22.42%	12.59%
<b>Condition</b>	Good condition	Alterations needed			
	77.71%	22.29%			
<b>Elevator</b>	Yes	No			
	67.91%	32.09%			
<b>Air Conditioning</b>	Yes	No			
	51.16%	48.84%			
<b>Swimming Pool</b>	Yes	No			
	23.72%	76.28%			
<b>Garage</b>	Yes	No			
	39.60%	60.40%			
<b>M.30</b>	Yes	No			
	86.70%	13.30%			
<b>M.30.2</b>	Yes	No			
	75.08%	24.92%			
<b>Shopping area</b>	Yes	No			
	58.91%	41.09%			
<b>Historical quarter</b>	Yes	No			
	65.95%	34.05%			

**Table C. Pollution variables: Main descriptive statistics (\*)**

	SO <sub>2</sub> <sup>(a)</sup>	CO <sup>(b)</sup>	NO <sub>2</sub> <sup>(a)</sup>	PM <sub>10</sub> <sup>(a)</sup>	NO <sub>x</sub> <sup>(a)</sup>	O <sub>3</sub> <sup>(a)</sup>
<b>Mean</b>	16.516	0.748	73.768	42.992	197.332	10.601
<b>S.D.</b>	5.455	0.242	17.319	10.425	52.091	2.453
<b>Max.</b>	27.237	1.267	99.645	71.547	298.324	14.866
<b>Min.</b>	7.337	0.226	41.149	22.438	82.870	6.133
<b>Skewness</b>	0.330	-0.389	-0.173	0.741	-0.083	-0.059
<b>Kurtosis excess</b>	-0.753	0.388	-0.767	0.760	-0.381	-0.666
<b>Legal standard</b>	125(c)	10(e)	210(d)	50 (c)	210 (d)	120(e)

(a):  $\mu/m^3$ ; (b):  $mg/m^3$ ;

(c): daily mean; (d): in an hour; (e) maximum of eight hours mean

(\*) The pollution measures are published in the ‘Atmosphere Pollution Monitoring System’, which can be downloaded from the Municipality of Madrid’s web page ([www.munimadrid.es](http://www.munimadrid.es)). The six air pollution variables are measured on an hourly basis at 25 fixed and operative monitoring stations.

### Acknowledgments

This work has been partially funded by Junta de Comunidades de Castilla-La Mancha, under FEDER research project POII10-0250-6975. We would like to thank the anonymous reviewers for their useful, constructive and valuable comments, which have undoubtedly improved the original version of the manuscript.