A scan test for spatial groupwise heteroscedasticity in cross-sectional models with an application on houses prices in Madrid

Coro Chasco
Dpto. Economía Aplicada
Universidad Autónoma de Madrid
Avda. Francisco Tomás y Valiente, 5
Madrid, 28049 Madrid (Spain)
mail: coro.chasco@uam.es
Tel: + 34 914 974 266

Julie Le Gallo
CESAER
Agrosup Dijon
DSHS Longelles
26 Boulevard Petitjean
21079 Dijon Cedex
mail: julie.le-gallo@agrosupdijon.fr
Tel: +33 (0)3 80 77 23 66

Fernando A. López[1]
Dpto. Métodos Cuantitativos e Informáticos
Universidad Politécnica de Cartagena
Facultad de Ciencias de la Empresa
C/ Real, 3
30201 Cartagena (Spain)
mail: fernando.lopez@upct.es
Tel: + 34 968 325 619

Abstract: We propose a scan test for the presence of spatial groupwise heteroskedasticity in cross-sectional data. The scan approach has been used in different fields before, including spatial econometric models, to detect instability in mean values of variables or regression residuals. In this paper, we extend its use to second order moments. Using large Monte Carlo simulations, we check the reliability of the proposed scan procedure to detect instabilities in the variance, the size and power of the test and its accuracy to find spatial clusters of observations

[1] Corresponding author

with similar variances. Finally, we illustrate the usefulness of this test to improve the specification search in a spatial hedonic model, with an empirical application on housing prices in Madrid.

Key words: Spatial scan procedure, spatial groupwise heteroskedasticity, spatial variance clusters, Monte Carlo simulation, House prices, Madrid.

JEL classification: C21, C52, C63, R15

This work was supported by Spanish Ministry of Economics and Competitiveness (ECO2015-65758-P) and the project from Programa de Ayudas a Grupos de Excelencia de la Región de Murcia, Fundación Seneca (#19884-GERM-15). The usual disclaimers apply.

1. Introduction

When analyzing spatial data, one has to deal with its two main specificities, namely spatial autocorrelation and spatial heterogeneity. Spatial autocorrelation, or the coincidence between value similarity and locational (Anselin and Bera, 1998), is now largely documented with multiple possible specifications for cross-sectional, qualitative, spatio-temporal and panel data and extensive testing procedures (see Arbia 2014, 2016; Dubé and Legros 2014; Elhorst 2014; LeSage and Pace 2009 for recent textbooks on the topic). On the other hand, spatial heterogeneity means that the spatial process is not uniform over space. Frequent causes of heterogeneity are instability in (i) the mean, (ii) the variance or (iii) both.

Mean instability implies local clustering of the values of a spatial variable. For instance, in the case of parameter instability in a regression, regression coefficients may follow a number of distinct spatial regimes such as North-South or Center-Periphery patterns, or they can evolve continuously over space (Brunsdon et al. 1999; Páez et al. 2002). In other cases, the variance, which varies over space, is the source of instability in the model. This phenomenon is known as spatial heteroskedasticity. The variance can vary continuously over space or it can take different values between separate parts of the area. It is then called spatial groupwise heteroskedasticity (SGWH from now one).

The consequences of wrongly omitting spatial effects are well known (Anselin, 1988; LeSage and Pace, 2009; Le Gallo, 2014). Omitted spatial autocorrelation in the mean equation or the wrong assumption of a common vector of parameters for the whole sample both leads to biased and inconsistent Ordinary Least Squares (OLS) estimators. Omitted spatial error dependence and/or heteroscedasticity leads to biased inference.

From an empirical point of view, spatial autocorrelation and spatial heterogeneity are often both present. These two effects entertain complex links. First, there may be observational equivalence between these two effects in a cross-section (Anselin and Bera, 1998). Secondly, heteroskedasticity and structural instability tests are not reliable in the presence of spatial autocorrelation (Anselin and Griffith, 1988). Conversely, spatial autocorrelation tests are affected by heteroscedasticity (Kelejian and Robinson, 2004; Zhang and Lin, 2016). Thirdly, spatial autocorrelation is sometimes the result of unmodelled parameter instability (Brunsdon et al., 1999). To tackle these issues, joint tests for spatial error autocorrelation and heteroscedasticity have been proposed by Anselin (1988) and Kelejian and Robinson (1998). Conditional tests of heteroscedasticity or instability in the regression coefficients, under the presence of spatial autocorrelation, can be found in Anselin (1988, 1990) and Páez et al. (2001), who introduce a Chow test for spatially switching regressions in a spatial lag model (see also López et al., 2009). Models allowing for both spatial autocorrelation and spatial heterogeneity in the coefficients have been suggested (see Geniaux and Martinetti, 2017 for a review of these models and a proposal of a new class of models where both the regression parameters and spatial autocorrelation coefficients vary over space).

All these possibilities are based on some parametrization of spatial error dependence and heteroscedasticity. Another possibility is to leave spatial error dependence and heteroscedasticity unmodelled using HAC methods, which allow performing robust inference in the mean equation for different departures of the iid clause. For instance, Kelejian and Prucha (2007) suggest a non-parametric heteroscedasticity and autocorrelation consistent (HAC) estimator of the variance-covariance matrix within a spatial context. This route has been followed by Kim and Sun (2011), who generalize the spatial HAC estimator for nonlinear spatial models, and Dorn and Egger (2014), who analyze the small sample performance of the spatial HAC estimators. On their side, Kelejian and Prucha (2010) and Lin and Lee (2010) keep spatial autocorrelation parametrized and discuss the instrumental variables and generalized method of moment approaches to estimate spatial autoregressive models with unknown heteroskedasticity in the disturbances.

In this paper, we depart from these approaches by arguing that looking for specific spatial patterns in heteroscedasticity can be a useful help in specification search in empirical analysis. Note that hetereroskedasticity in spatial models that do not follow a spatial pattern is not of interest in this paper as it can be treated as usual, using the classical White (1980), the Breusch-Pagan (Breusch and Pagan, 1980) or Koenker and Basset (Koenker and Basset, 1982) tests. Instead, we consider the perspective initiated by Ord and Getis (2012). They study the problem of local instability in the variance introducing the so-called LOSH (Local spatial heteroscedasticity) statistic, whose aim is to identify the limits of the area where the variance changes. The authors draw the attention to the lack of papers directed at examining the spatial structure of the variance (p. 530): '*Spatial statistics' cluster identification is now common to many fields. (...however) these studies have focused attention upon local means, to the extent that variability is considered at all it is typically assumed that the process has a constant* *variance (i.e., that it is homoscedastic). A moment's thought indicates that such an assumption could overlook important information*'.

This paper aims at filling such a gap. Indeed, we introduce a formal test for SGWH for the residuals of a multiple regression analysis with a null hypothesis of constant variance in the residuals. Then, when the null is rejected, the procedure is able to detect the locations of the clusters of observations, for which the residual variance is significantly high or low. In order to implement such a test, we follow the approach originally suggested by Openshaw et al. (1987) in the so-called Geographical Analysis Machine, GAM, and later improved by Kulldorff et al. (1995, 2009) in the nowadays popular scan algorithms. This paper therefore complements that of López et al. (2015) who explore scan methods to test for spatial structure in mean values. Cucala (2016) has also suggested a scan statistic to detect high-variance clusters. We extend his work by providing a formal proof of the consistency of the test, detailing its relations with spatial autocorrelation, performing an extensive Monte-Carlo simulations to assess its small-sample properties in various configurations and presenting a detailed empirical analysis.

With respect with the latter aspect, we show that the detection of spatial clusters of observations with similar, high or low, variance of residuals is a useful guide to specification search in applied econometrics. Consider for instance hedonic models for house prices. The empirical challenge with these models is important as the estimation of hedonic prices should tackle, *inter alia*, spatial autocorrelation due to shared local amenities and disamenities, spatial heterogeneity stemming from the existence of housing submarkets and omitted variables, as house prices usually depend on unobserved microgeographic characteristics. All these problems interact in a complex way, as mentioned above, so that there is a need to provide the applied econometrician with procedures helping to deal with them. In the application, we show that by determining the location of clusters of high and low variance, the SGWH test is a step in that direction. In addition, we show that after having removed the effects of conditional heteroscedasticity and spatial autocorrelation, clusters of high and low variance of residuals persist, that correspond to areas with particular characteristics of the housing market that would have been difficult to identify without the proposed test.

The paper is organized as follows. Section 2 introduces some basic results from the scan methodology, including our proposal to detect SGWH. The design of a Monte Carlo experiment is presented in Section 3, together with the main results related to estimated size and power. Section 4 presents a discussion about the links between spatial dependence and SGWH. Section 5 illustrates the use of this methodology with an empirical application on housing prices in Madrid and shows the SGWH scan test can help with the search for the best empirical specification. Main conclusions appear in Section 6.

2. A scan test for spatial groupwise heteroskedasticity

Spatial scan statistics are widely used in epidemiology, criminology or ecology. Their purpose is to analyze the spatial distribution of points or geographical regions by testing the hypothesis of spatial randomness of this distribution on the basis of different distributions (e.g. Bernoulli, Poisson or Normal distributions). Using the popular scan methodology, we introduce a test aimed at detecting the presence of SGWH in the residuals of a regression model. Our proposal has two objectives: (i) check for the null of homoskedasticity and, in case of rejection of the null, (ii) identify the points, spatially linked, for which the residuals share the same variance. We first present the test for a variable with a Gaussian distribution, then show how it can be used to detect secondary clusters and finally, establish its consistency.

2.1 Presentation of the $Scan_\sigma$ test

Formally, suppose that $\{x_i\}$ is a spatial process with $i = 1,.., n$ being set of spatial coordinates. We are interested in testing the null hypothesis that all the variances of this variable are equal (note that the null also implicitly assumes normality and spatial independence):

$$H_0 : \quad x_i \sim \text{i.i.d. } N(\mu;\sigma) \tag{1}$$

The alternative hypothesis states that there is a single group of spatially connected observations, Z, for which the variance is different than that for the rest of the observations:

$$H_A : \begin{matrix} x_i \sim \text{i.i.}d.N(\mu;\sigma_Z) \text{ for } i \in Z \\ x_i \sim \text{i.i.}d.N(\mu;\sigma_{\bar{Z}}) \text{ for } i \notin Z \end{matrix}; \text{ with } s_Z \text{ ¹ } s_{\bar{Z}} \tag{2}$$

In order to proceed with the scan methodology, it is necessary to derive the likelihood function under the null and alternative hypotheses, respectively. The log-likelihood function under the null hypothesis is:

$$l(H_0) = ln\left(L_0\left(x, m, s\right)\right) = -n\,ln\sqrt{2p} - n\,ln\,s - \sum_{i=1}^{n}\!\!\text{å}\ \frac{\left(x_i - m\right)^2}{2s^2} \tag{3}$$

The maximum likelihood estimates of the mean and variance are:

$$\hat{\mu}_{H_0}=\sum_{i=1}^{n}\frac{x_i}{n}\qquad \hat{\sigma}^2_{H_0}=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\hat{\mu}_{H_0}\right)^2 \tag{4}$$

which produce a value in the corresponding concentrated log-likelihood function of:

$$l(H_0)=-\frac{n}{2}\left[ln\left(2\pi\times\hat{\sigma}^2_{H_0}\right)+1\right] \tag{5}$$

Under the alternative hypothesis, the log-likelihood is:

$$\begin{aligned} l(H_A)&=ln\left(L_A\left(x,\mu,\sigma_Z,\sigma_{\bar{Z}}\right)\right)=\\ &=-n\,ln\sqrt{2\pi}-n_Z\,ln\,\sigma_Z-\left(n-n_{\bar{Z}}\right)ln\,\sigma_{\bar{Z}}-\frac{1}{2\sigma^2_Z}\sum_{i\in Z}\left(x_i-\mu\right)^2-\frac{1}{2\sigma^2_{\bar{Z}}}\sum_{i\notin Z}\left(x_i-\mu\right)^2\end{aligned} \tag{6}$$

where n_Z is the number of observations in set Z. The maximum likelihood estimates of the mean and variance for this case are:

$$\hat{\mu}_{H_A}=\sum_{i=1}^{n}\frac{x_i}{n}\qquad \hat{\sigma}^2_{H_A}(Z)=\sum_{i\in Z}\frac{\left(x_i-\hat{\mu}_{H_A}\right)^2}{n_Z}\qquad \hat{\sigma}^2_{H_A}(\bar{Z})=\sum_{i\notin Z}\frac{\left(x_i-\hat{\mu}_{H_A}\right)^2}{n-n_Z} \tag{7}$$

The value of the log-likelihood function in this point is:

$$l^I(H_A)=-\frac{n}{2}\left[ln\,2\pi+1\right]-\frac{n_Z}{2}ln\,\hat{\sigma}^2_{H_A}(Z)-\frac{n-n_Z}{2}ln\,\hat{\sigma}^2_{H_A}(\bar{Z}) \tag{8}$$

Consequently, the scan statistic for the assumption of equal variances can be written as:

$$Scan_{\sigma}=\max_{Z\in\Theta}\left[l^I(H_A)-l(H_0)\right]=\max_{Z\in\Theta}\left[2n\left(ln\left(\frac{\hat{\sigma}^2_{H_0}}{\hat{\sigma}^2_{H_A}(\bar{Z})}\right)+\frac{n_Z}{n}ln\left(\frac{\hat{\sigma}^2_{H_A}(\bar{Z})}{\hat{\sigma}^2_{H_A}(Z)}\right)\right)\right] \tag{9}$$

where Θ is a set of connected regions Z, called *windows*, for which the *Scan*$_\sigma$ statistic is computed. The size and shape of the windows must be defined *a priori* by the researcher with the aim of getting a good balance between cost and effectiveness. For example, the evaluation of elliptical windows is more time consuming but it provides greater flexibility than circular windows. The window *Z* moves across the entire map, changing its size and possibly shape while looking for the maximum likelihood ratio. It is usually recommended that the maximum number of cases entering any given window does not exceed 50 per cent of all available cases. Once the window with maximum differential is detected, it is evaluated with the test to check whether the difference is statistically significant. The set Z where the scan test attains its maximum value is usually called the *Most Likelihood Cluster*, MLC.

Inference for the *Scan*$_\sigma$ test is based on a permutational framework, which is more robust –since it avoids data mining and the assumption of normality– but also more computationally demanding. More precisely, a *p*-value is obtained through a Monte Carlo testing procedure, by comparing the value of the scan statistics for the real data set with a large sequence of values corresponding to purely random data sets, according to the null hypothesis of the test. The procedure is as follows:

1. Compute the *Scan*$_\sigma$ statistic for the original sample $\left\{x_i\right\}_{i\,\hat{I}S}$, where S is a set of spatial coordinates, $S=\{cx_i;cy_i\}$; $i=1,2,...,n$.
2. Relabel the set of locations by randomly drawing, without replacement, the spatial coordinates; $\left\{x_i^r\right\}_{i\,\hat{I}S}$ is the new, permuted, series, where *r* is the permutation index.
3. Compute the *Scan*$_\sigma^r$ statistic for each permuted sample $\left\{x_i^r\right\}_{i\in S}$.
4. Repeat steps 2 and 3 (*B*–1) times to obtain *B*–1 realizations of the $\left\{\text{Scan}^r\right\}_{r=1}^{B-1}$ permuted statistic.
5. Compute the pseudo-probability as:

$$p_p\text{-}value = \frac{1}{B-1}\overset{B-1}{\underset{r=1}{\mathring{a}}} t(Scan_s^r - Scan_s) \tag{10}$$

   where $\tau(\bullet)$ is an indicator function which assigns a value of 1 to a true statement and 0 otherwise.
6. Reject the null hypothesis if p$_p$-value<α for a nominal size *α*.

The MLC is the window Z where the difference in likelihood is the maximum. In case of this difference is greater than percentile (1–α)% of the empirical distribution obtained by permutation bootstrap the MCL will be significant at α% level.

2.2 Secondary clusters

The *Scan*$_\sigma$ is a maximum likelihood ratio statistic that under the alternative hypothesis, establishes that there is a single cluster of unknown localization, shape and size. If the *Scan*$_\sigma$ test identifies a cluster (the MLC), a natural question arises: do there additional clusters, which do not overlap with the MLC, have significantly large likelihood ratio? If yes, these are the so-called *secondary clusters*. There are different possibilities to assign *p*-values to secondary clusters in order to evaluate the significance of those clusters (Zhang et al., 2010). The standard approach for assigning *p*-values is to compare the likelihood ratio of secondary clusters with the empirical distribution of the statistic obtained by permutational bootstrapping. Then, the first secondary cluster is considered as significant (e.g. p-values$<\alpha$) if the likelihood ratio of this cluster is over (1-α)% of the values of the empirical distribution obtained under the null provided that it is non-overlapping with the MLC. The second secondary cluster is significant (*p*-values$<\alpha$) if the likelihood ratio of this cluster is over (1-α)% of the values of the empirical distribution obtained under the null provided that it is non-overlapping with the MLC and the first secondary cluster. The procedure continues until no clusters have a likelihood ratio over (1-α)% of values in the empirical distribution. This method gives conservative p-values (Kulldorff, 1997) since they are calculated ignoring the existence of the MLC. Some authors (Zhang et al., 2010) have suggested the use of an iterative method by deleting from the data the observations included in the MLC (getting a 'hole' in the sample). The standard scan statistic is then newly computed for the reduced dataset. The procedure continues until no significant clusters are found. This process proposed by Zhang et al. (2010) shows higher power for secondary clusters that the standard one. For this reason, we will use this approach in this paper.

2.3 Consistency of the *Scan*$_\sigma$ test

To the best of our knowledge, only few results on the asymptotic properties of scan tests are available in the literature. Cressie (1980) derives the asymptotic distribution of the one-dimensional scan statistic for testing uniformity on [0,1] against a clustering alternative. More recently, Zhang and Lin (2017) provide the non-trivial asymptotic behaviour (consistency and local efficiency) of two-dimensional scan test in case of a Bernoulli distribution. In this subsection, we derive the consistency of the *Scan*$_\sigma$ test. The main results of this subsection are synthesized in the following theorem:

Theorem 1. Let a Spatial Gaussian process of i.i.d. $\{x_i\}_{i=1}^{n}$ with finite E[x$_i$]=μ and assume that there is a subset $Z\in\Theta$ where the variance of x$_i$ is different if $i\in Z$ that the variance of x$_i$ if $i\notin Z$. Then,

$$\lim_{\substack{n\to\infty \\ n_Z\to\infty}} \Pr(\text{Scan}_\sigma > C) = 1 \text{ for all real number } C > 0.$$

if 1/n is an infinitesimal of higher order than $1/n_Z$. The proof is in Appendix 1.

A consequence of this result is that the power of $Scan_\sigma$ test is directly related to the size of true cluster (n_z) and the ratio between inside and outside Z variance (δ) under the specific alternative H_A.

3. Evaluating Size and Power of *Scan*$_\sigma$ test

In this section, we evaluate the performance of the permuted scan test introduced in Section 2, applied on the least square residuals of a linear model without spatial effects. The set-up of our Monte-Carlo study is as follows:

i. A linear equation is specified, including one regressor ($u_i \equiv U(0,1)$) plus a constant term as:

$$y_i = a + bu_i + e_i \tag{11}$$

The values of the parameters guarantee an expected $R^2=b^2/(b^2+12)$ coefficient close to 0.4.[2]

ii. Regarding the error term, we consider seven different situations to evaluate the impact of a departure from non-normality:

● DGP1: the error terms are distributed as a N(0,1);

● DGP2: the error terms are distributed as a $\chi^2(2)$;

● DGP3: the error terms are distributed as a Beta(0.5,0.5);

● DGP4: the error terms are distributed as a Lognormal(0,1);

● DGP5: the error terms are distributed as a Binomial distribution, B(n,0.1);

● DGP6a: the error terms are distributed as a weighted average of a $\chi^2(2)$ and a Student's *t* distribution with 2 degrees of freedom, *t*(2); the unit weights are obtained randomly in the interval (0;1)[3];

- DGP6b: the error terms are distributed as a weighted average of three distributions ($\chi^2(2)$; *t*(2); U(0,1));

● DGP7: the error terms are heteroskedastic, with a random spatial structure in the variance.

[2] Results for the case *a* = 2 and *b* = 7, which guarantees an expected $R^2=b^2/(b^2+12)$ coefficient of close to 0.8 available on request present similar results.

[3] The resulting variable is a stochastic mixture of two non-normal distributions, which generates a random variable with an unknown distribution. We call it a "*mixture error structure*" (see Lin et al., 2010).

iii. Hexagonal regular lattices of orders (6×6), (7×7), (10×10), (15×15) and (20×20) have been used as the spatial support for the data, implying sample sizes (n) of 36, 49, 100, 225 and 400 observations, respectively;

iv. Irregular lattices with the same sample size n=36, 49, 100, 225 and 400. The coordinates of each observation (cx_i,cy_i) were obtained from a uniform distribution U(0,1);

v. Q is the set of all possible elliptical windows whose center corresponds to the centroid (cx_i,cy_i) of each observation and have different parameters: an eccentricity of e = 1, 2, 3, 4 and a rotation angle of $\theta=\pi(2k+1)/18$; k=0,1,…,8. Moreover, the number of location entering in any given window should not exceed 50% of all locations;

vi. Each combination has been repeated 1,000 times. The number of permutations for each simulated dataset, in order to compute the p_P-values, has been 999, so that B=1,000.

3.1. Size of the $Scan_\sigma$ test

Table 1 shows the estimated size for the $Scan_\sigma$ test under the 8 DGPs. The results show that under DGP1 (ε~iidN(0,1)) and DGP7 (random heteroscedasticity), the estimated size for the $Scan_\sigma$ tests is very close to the nominal value of 0.05, even for very small sample sizes. For the other DGPs, the $Scan_\sigma$ test is a bit oversized for small samples but for larger sample sizes, it behaves properly. The test is then quite robust to departures of normality.

---------- Table 1 is about here ----------

3.2. Power of the $Scan_\sigma$ test

Two types of spatially varying variances may be considered: discrete or continuous. The first situation means that there are blocks of locations that share the same variance, which differs from block to block. This corresponds to the traditional interpretation of SGWH (see Ertur et al., 2006 or Ramajo et al., 2008 for examples). The second situation is continuous instability that should be interpreted in analogy with the concept of '*parameter surface*' in Geographically Weighted Regression models, where the parameters associated to each location is space are the image of the corresponding location in the surface (Fotheringham et al., 1999). Yan (2007) introduces the term of

Spatial Stochastic Volatility in reference to a process, i.e. $\{\zeta_i; i = 1,...,n\}$, whose variance is changing smoothly over space:

$$V(\zeta_i) = h(cx_i, cy_i) \tag{12}$$

where (cx_i,cy_i) are the spatial coordinates of location i.

In what follows, we investigate the behavior of the *Scan*$_\sigma$ test for these two categories of variance instability.

For the case of SGWH, we consider six different patterns. In the first two cases, the heteroskedastic cluster is formed by 7 locations and has a circular shape (SGWH1 and SGWH2 in Figure 1 below). The number of locations included in this heteroskedastic cluster (that is, the size of the cluster) remains the same for the different sample sizes used in the experiment; this means that the symptoms of instability are weaker as sample size increases. Two other cases exhibit a North-South dichotomy with different values of variances in each regime (SGWH3, SGWH4 in Figure 1). The size of the cluster in the last two cases is proportional to the size of the sample. For the SGWH5 and SGWH6 the heteroskedastic cluster and has an elliptic form with growing number of localizations for different sample sizes (SGWH5, SGWH6 in Figure 1).

For continuous spatial patterns in the variance, three cases have been considered. The first, SGWH7, is inspired by Casetti and Can (1999) where the variance of the error terms is expanded into a monotonic function of the distance of each location to the geographical central point of the system. The second case, SGWH8, reflects a continuous North-South variation while the third, SGWH9, extends the Casetti and Can (1999) example by considering two central foci.

---------- Figure 1 is about here ----------

Table 2 shows the results obtained for the *Scan*$_\sigma$ tests. We also include the results obtained for well-known classical test of Breusch-Pagan test of heteroskedasticity, computed with the explanatory variable of the model (Breusch and Pagan, 1979). Additionally, we include the BP test using the coordinates of localization (latitude, longitude) of each observation as heteroskedasticity source, namely BP_c. It is worth reminding that, in all the cases, these tests have been applied to the least square residuals of an estimated equation, similar to that of (11). We also include the Moran's *I* statistic (MI) to obtain information about the sensibility of this test of spatial dependence to the presence of SGWH[4].

---------- Table 2 is about here ----------

The results highlight the following results:

(i) Overall, the estimated power for the *Scan*$_\sigma$ test is higher than that obtained for the classical heteroskedasticity tests.

(ii) The BP$_c$ test shows similar power to that of the *Scan*$_\sigma$ for some DGPs although it cannot give information about the spatial structure of heteroskedasticity.

(iii) The power of the *Scan*$_\sigma$ test improves as the sample size increases, also when there is a great difference between the variances of the two regimes.

(iv) The power of the BP test is very poor independently of the heteroskedastic pattern and/or sample size. It is not designed to deal with the spatial nature of the data and, consequently, they have low power.

(v) In case of complex pattern of SGWH (continuous heteroscedastic pattern SGWH7-SGWH9 and elliptic cluster of high variance SGWH5-SGWH6) the *Scan*$_\sigma$ test has the highest power in almost all cases.

(vi) The MI test is sensitive to SGWH with sample size over 400 observations. Kelejian and Robinson (2004) provide a theoretical explication for this result.

3.3. Spatial precision in cluster identification

Subsection 3.2 on the estimated size and power of the *Scan*$_\sigma$ test under different heteroskedasticity patterns lead to encouraging results. However, in case of rejecting the null hypothesis of homoskedasticity, another important question emerges: the necessity of identifying, as accurately as possible, the heteroskedasticity pattern present in the data. It is clear that, in order to improve the specification, the researcher needs to know where and how are produced the clusters in the variance. Hence, we evaluate in this section the ability of the *Scan*$_\sigma$ to precisely identify the location of these spatial clusters of observations with similar variances.

To that purpose, we define *Local Sensibility (LS)* as the percentage of times, in the Monte Carlo simulation, that each cell is selected as a member of the significant cluster (MLC) over the times that the test rejects the null, that is:

[4] The results for the White test and the Lagrange multiplier LM-err test are available upon request from the authors. The performances of these tests are similar than those of BP test and MI respectively.

$$\mathrm{LS(i)} = \frac{\text{Number of times that localization i is assign to the MLC}}{\text{Number of times that the test rejects the null hypothesis}} \qquad (13)$$

Indeed, the maximum number of times that a location can be selected as pertaining to the MLC is the number of times that the test rejects the null. A procedure such as the scan technique will be useful for the researcher if its *LS* is close to 1 for the cells that really pertain to the true variance cluster, and 0 for the cells that do not belong to the cluster.

Figure 2 shows the estimated *LS* corresponding to the six discrete SGWH introduced in Subsection 3.2 (SGWH1 to 6) using the $Scan_\sigma$ test. This figure displays the percentiles of the decisions taken with respect to each cell in the spatial lattice.

---------- Figure 2 is about here ----------

Figure 2 demonstrates the utility of $Scan_\sigma$ test to identify the area with a different variance when the null is rejected. In particular, it shows that (i) the size of the windows with different variance is not important. The precision needed to identify the windows with 7 observations is similar for n=100 and n=225; (ii) in case of low differences in variance inside and outside of windows (SGWH2, SGWH4, SGWH6) the ability to identify the true area is low but still provides an acceptable information about the localization of the cluster. This information is vital to improve the model specification in a regression exercise as we show in the next section.

4. Spatial dependence and spatial groupwise heteroscedasticity

As underlined in the introduction, spatial dependence and SGWH entertain complex links. To shed more light on this issue, we analyze in this section the size of the $Scan_\sigma$ test in presence of spatial patterns in the mean of the process.

It is well-known that the diagonal elements of the variance-covariance matrix of a spatial autoregressive (SAR) model or a spatial error model (SEM) are not equal, *i.e.,* that SAR or SEM processes imply a specific pattern of heteroscedasticity. However, given our objective to evaluate the size of the $Scan_\sigma$ test only in presence of spatial autocorrelation but not of SGWH, we need to define a homoscedastic spatial process.

Definition 1. We say that a spatial process $\{\mathrm{x}_\mathrm{i}\}_{\mathrm{i}\,\hat{\imath}\mathrm{S}}$, where *S* is the set of spatial coordinates, is homoscedastic if the covariance matrix of $\{\mathrm{x}_\mathrm{i}\}_{\mathrm{i}\,\hat{\imath}\mathrm{S}}$ is diagonal constant.

It is simple to prove that any spatial process can be transformed into a homoscedastic spatial process. For instance, in the case of a SAR model, we have the next results:

Theorem 2. If the SAR process *y* is defined by $y = (I - rW)^{-1}(Xb + e);\ e \equiv N(0, s^2W)$ where $W = (w_{ij}) = (I - rW)^{-1}(I - rW')^{-1}$, then the spatial process H defined as $H = j^{-1/2}y$ with $j = \mathrm{diag}(w_{11},...,w_{nn})$ is a homoscedastic SAR process (HSAR) and $H \equiv N(\tilde{X}\beta, \sigma^2(I-\rho\tilde{W})^{-1}\varphi^{-1}(I-\rho\tilde{W}')^{-1})$ with $\tilde{X} = \varphi^{-1/2}X$ and $\tilde{W} = \varphi^{-1/2}W\varphi^{1/2}$ (proof in Appendix 2). This result can be extended to the other well-known usual spatial processes.

To obtain the empirical performance of $Scan_\sigma$ test in presence of a spatial pattern in the mean we develop a Monte Carlo exercise with the following characteristics, which depends of the type of spatial autocorrelation:

- DGP8: a SAR process (DGP8a) or a HSAR process (DGP8b), with coefficients of ρ=0.2; 0.5; 0.8.
- DGP9: a SEM process (DGP9a) or a HSEM process (DGP9b), with coefficients of λ=0.2; 0.5; 0.8.

Secondly, we also specify instability in the trend of the spatial process. The DGP is then defined as:

$$y_i = 2 + b_i\, u_i + \varepsilon_i$$

where $u_i \equiv U(0,1)$ and we consider two (DGP10a) or four (DGP10b) spatial regimes:

- DGP10a Low: b_i=2.5 if $cy_i \leq Me_y$ and b_i=3.5 if $cy_i > Me_y$, where Me_y is the median of the latitude coordinates.
- DGP10a High: b_i=2 if $cy_i \leq Me_y$ and b_i=4 if $cy_i > Me_y$
- DGP10b Low: b_i=1.5 if $cy_i \leq P^y_{33}$ and $cx_i \leq P^x_{33}$; b_i=5 if $cy_i \leq P^y_{33}$ and $cx_i > P^x_{33}$; b_i=2.5 if $cy_i > P^y_{33}$ and $cx_i \leq P^x_{33}$; b_i=3 if $cy_i \geq P^y_{33}$ and $cx_i \geq P^x_{33}$, where P^x_{33} is the 33-percentile of $\{cx_i\}_{i=1}^n$ and P^y_{33} is the 33-percentile of $\{cy_i\}_{i=1}^n$.
- DGP10b High: b_i=1 if $cy_i \leq P^y_{33}$ and $cx_i \leq P^x_{33}$; b_i=6 if $cy_i \leq P^y_{33}$ and $cx_i > P^x_{33}$; b_i=2 if $cy_i > P^y_{33}$ and $cx_i \leq P^x_{33}$; b_i=3 if $cy_i \geq P^y_{33}$ and $cx_i \geq P^x_{33}$

In all cases, the model is estimated with constant coefficients. Table 3 summarizes the results.

---------- Table 3 is about here ----------

The $Scan_\sigma$ test is clearly affected by the presence of a spatial pattern in the mean. The $Scan_\sigma$ test is oversized for cases where the error terms exhibit strong patterns of spatial autocorrelation. In case of homoscedastic spatial processes, the test also shows high rates of rejection of the null hypothesis, though they are lower than for SAR or

SEM processes. The over-reaction of the test increases with sample size reflecting the well-known problem of observational equivalence between spatial autocorrelation and spatial heterogeneity in cross-section models. Table 3 also indicates that if the model is misspecified, in the sense that the spatial trend is not stable over space, the OLS residuals will produce symptoms of both spatial dependence (Lopez et al., 2015) and spatial clustering in the variance. Overall, these results confirm the difficulty of isolating the symptoms of spatial dependence and heteroskedasticity, as shown previously by Kelejian and Robinson (1995) and Mur and Angulo (2009).

5. Application on housing prices

In this section, the behavior of the $Scan_\sigma$ test as a useful guide to specification search is illustrated with an application on a hedonic model for house prices in the city of Madrid. Our study focuses on downtown Madrid, 'Central Almond', which is an area administratively formed by seven districts, these being subdivided into 43 neighborhoods. Our records refer to January 2015 and were drawn from an on-line real estate database, 'idealista.com' since, due to confidentiality constraints, it is almost impossible to obtain housing prices microdata from Spanish official institutions. The asking price has then been used as a proxy for the selling price as usual in many other cases (Cheshire and Sheppard 1998, Orford 2000, Chasco and Le Gallo, 2013). In total, 5,541 housing prices were finally recorded after the corresponding consolidation and geocoding processes.

As a benchmark model, we first specify a standard hedonic house price model with a broad set of explanatory variables: twenty-two are attribute variables and fourteen are accessibility measures, since they are frequently advertised by real estate agents and often capitalized in housing prices. Additionally, in order to proxy all the micro-geographic determinants that buyers and sellers can observe, but are hidden for the econometricians, we also include the Earth coordinates of latitude and longitude plus fifty dummy variables corresponding to the seven central districts and forty-three neighborhoods, which can be considered as contextual variables (e.g. Anselin and Lozano-Gracia 2008).[5] Table 4 contains a listing of all variables together with their definitions.

[5] Districts are official administrative units defined by the Spanish National Statistics Office (INE) and neighborhoods, which are nested in the districts, are officious divisions recognized by the city council

The standard hedonic house models are expressed in semi-log form:

$$lpri_i = b_0 + \sum_{s=1}^{S} a_s x_{si} + \sum_{c=1}^{C} g_c x_{ci} + \sum_{g=1}^{G} d_g x_{gi} + u_i \tag{14}$$

where *lpri*$_i$ is the log of price of transaction *i*; *S* is the number of property structural attributes, x_{si}; *C* is the number of accessibility variables x_{ci}; *G* is the number of geographic contextual variables x_{gi} and u_i is a well-behaved error term.

---------- Table 4 is about here ----------

In the benchmark **Model 1**, we retain 33 significant regressors: 17 structural characteristics, 12 accessibility indicators and 4 geographical variables, 14 of which are expressed as orthogonal splines to capture their nonlinear impact on house price. An ordinary least squares (OLS) estimation of this first model captures almost 90% of housing price variance (adjusted R^2=0.8653), as shown in Table 5.[6] Table 5 also contains several diagnostics and specification tests. It appears that model (1) is affected by multicollinearity problems, as shown by the high value of the condition number test (4,753), above the acceptable limit of 30-40 (Belsley 1991). The Jarque-Bera statistic, which takes on a very significant value (27.43), is an indicator of clear non-normality in the error terms. The Lagrange Multiplier (LM) for spatial autocorrelation have been computed for an inverse squared distance matrix for a radius of 375 meters, which is the minimum distance for which every dwelling has at least one neighbor, and different sets of (2, 5, 6 and 30) nearest neighbor matrices[7]. In every case, either the robust LM test against a spatial error model or its counterpart against a spatial lag model, are very significant, though the first is always higher. However, because of non-normality, these LM results must be taken with caution.

---------- Table 5 is about here ----------

Finally, we test the homoskedasticity assumption considering different forms of variability in the error terms. A significant Koenker-Basset test[8] (212.04) shows the existence of heteroskedasticity as a linear function of the independent variables (conditional heteroskedasticity). In order to investigate the existence of a SGWH form

(http://www.munimadrid.es). Neighborhoods are characterized by certain homogeneity in terms of population density, infrastructure, historical and socioeconomic features.

[6] The complete results for the six models are available upon request from the authors.

[7] The 2 and 30 nearest neighbor matrices represent very narrow and sparser neighborhood specifications, respectively. As for 5 and 6 nearest neighbor matrices, they are the commonest connectivity structure after creating Voronoi or Thiessen polygons from the point-data set of dwellings in downtown Madrid.

[8] The Koenker-Basset test is a Breusch and Pagan (1979)'s studentized version suggested by Koenker and Bassett (1982), which is reported when the errors are non-normal.

in the errors, we have also computed the $Scan_\sigma$ test,[9] which is quite robust to departures of normality (section 3.1). This test is significant at 5% for six clusters. Particularly, they are low-variance residual clusters, one of which is the MLC with 1,324 observations (Figure 3).

At this point, we do not have enough information to identify the causes for which the LM and $Scan_\sigma$ tests reject the null: spatial dependence, SGWH or instability in the trend of the spatial process or combination of all these issues. We first deal with instability of spatial trend and conditional heteroscedasticity. In **Model 2**, dummies for districts and neighborhoods have been included in order to check whether the existence of spatial real-estate submarkets. However, as Table 5 shows, misspecification problems are still present with all LM tests remaining high values. The $Scan_\sigma$ test identifies 5 low-variance significant spatial clusters broadly extended across the eastern side of downtown Madrid (Figure 3). **Model 3** operates a change of strategy focused on reducing conditional heteroscedasticity with a more parsimonious model. Interaction variables are also allowed, heteroskedasticity is carefully evaluated with the Koenker-Basset test so that the significant –but heteroskedastic– regressors are excluded (column 1, Table 6). This model captures practically the same housing price variance than the previous ones (adjusted R^2=0.8491). Though still high, the condition number test reveals a significant reduction in multicollinearity (64).

---------- Figure 3 is about here ----------

Although Model 3 eradicates conditional heteroskedasticity, the $Scan_\sigma$ test still rejects the null assumption of a common variance. It identifies three significant low-variance spatial clusters and two high-variance ones, though they are smaller in size. Therefore, while this strategy allows considerably better results than the previous specifications, the model residuals are still affected by spatial autocorrelation and/or groupwise heteroskedasticity, which could be interacting with each other and it is not possible to know the source of instability. Thus, in order to avoid possible oversizing of the $Scan_\sigma$ test in presence of spatial autocorrelation (section 4), we propose to remove this effect first and isolate SGWH in the residuals.

Due to the presence of splines and interaction variables, it is not meaningful to estimate a spatial Durbin model. We therefore estimate firstly a SEM (**Model 4**) and secondly the SAR model (**Model 5**), both with a 5-nearest neighbor spatial weight

[9] Running on a desktop with Inter(R) Core i7 with 2.93 GHz, and 12Gb of Ram, the elapsed CPU times for performing the $Scan_\sigma$ test -in this case, for 5,541 observations- was 1,373 seconds.

matrix, which represents the most common connectivity structure of this sample (see **Table 6**, columns 2 and 3).[10] Since ML is not appropriate when the error terms are not normally distributed, we use general method of moments (GMM) and spatial two-stage least squares (STSLS), respectively. Either the Anselin-Kelejian or the Moran's *I* test on the errors do not allow to accept the null of no spatial autocorrelation, and the $Scan_\sigma$ test detects error variance clusters, particularly for the SEM.[11] Consequently, we estimate a SARAR model (**Model 6**) with a spatial lag dependent variable and a spatial autoregressive process in the error terms. The SARAR model is estimated by a three-step procedure (see **Table 6**, column 4), which combines spatially weighted least squares with GMM (Anselin and Rey 2014). This is the only model capable of absorbing spatial autocorrelation in the residuals whatever spatial weights matrix is being used (in **Table 6**, column 4, we have shown the results of the Moran's *I* test).

---------- Table 6 is about here ----------

Nevertheless, the $Scan_\sigma$ test cannot accept the null, since there are still three significant spatial clusters. Therefore, once conditional heteroskedasticity and spatial autocorrelation are controlled, a certain degree of spatial groupwise heteroskedasticity persists. There is a persistent big cluster with 816 observations located in the eastern side of Central Almond, which contains residuals with a lower inside dispersion and higher outside dispersion. That is, in this area the model fits significantly well house prices, compared to the rest of the sample, probably because it is located in a compact cohesive zone, the 'Retiro' district, which is one of the last urban developments of downtown Madrid built in the middle of the last century (Martínez 2014). Conversely, there are two smaller high-variance clusters at the northwestern neighborhood of 'Valdeacederas' (182 observations) and part of the central neighborhoods of 'Embajadores', 'Cortes' and 'Universidad' (23 observations), where the model tends to both overestimate and underestimate the observed prices due to the existence of unobserved local variables and housing micro-markets. In effect, these are older city developments placed in well-communicated quarters close to cultural amenities, in which modern and degraded buildings coexist, leading to gentrification processes and some kind of property speculation (Leković 2013, García 2014, Muñoz 2014, Camacho

[10] In fact, the most significant specifications of the spatial weights matrix are achieved by sparser neighborhood structures (2, 5 and 6-nearest neighbors), with respect to broader ones (30-nearest neighbors and inverse squared distance). Consequently, house price spatial spillovers in downtown Madrid are operating –ceteris paribus– at a reduced scale.

[11] The $scan_\sigma$ test has a similar performance when computed for GMM estimations than in ML. Complete results are available from the authors under request.

et al. 2015, von Breymann 2017). These effects would have been very difficult to identify without the *Scan*$_\sigma$ test.

Because we do not have additional variables able to control for this remaining heteroscedasticity, we perform the KP-HET inference of the coefficient covariance matrix, proposed by Kelejian and Prucha (2010), which is robust to the presence of spatial heteroskedasticity in the error terms.

6. Conclusion

As pointed by Anselin and Bera (1998, p. 238), spatial autocorrelation and heteroskedasticity may be observationally equivalent in cross-sections: '*For example, a spatial cluster (...) of extreme residuals may be interpreted as due to spatial heterogeneity (e.g., groupwise heteroskedasticity) or to spatial autocorrelation*'. In the same vein, Mur and Angulo (2009) show that most patterns of SGWH are indistinguishable from the cases of spatial dependence or heterogeneous mean values. Hence, there is a high risk of misinterpreting the symptoms, especially if the variance follows some regular spatial pattern. Obviously, if the symptoms are misinterpreted, decisions will be erroneous and the inference probably wrong. Therefore, it is of great practical importance to develop tests capable of detecting different forms of spatially structured heteroskedasticity. Our impression is that the *Scan*$_\sigma$ test is a first step towards tackling this problem.

In this paper, we have shown that the *Scan*$_\sigma$ test is a simple and powerful method to identify SGWH in the residuals of a regression model. The principal advantage of this test is that it is not necessary to provide information about the pattern of instability in the variance. Moreover, an output of this test is to indicate the localizations of the spatial clusters of observations with higher (or lower) variances. This information supplied by the test is vital information to improve the model specification. We think that both tests (spatial dependence and SGWH) must be used in a correct exercise of specification for spatial regression model. The way these tests can be combined in a specification search is left for future research.

References

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

Anselin, L., Bera, A., Florax, R. and Yoon, M., (1996): Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics* 26: 77-104.

Anselin, L., Rey, S.J. (2014): *Modern Spatial Econometrics in Practice*. Geoda Press, LLC, Chicago (IL, USA).

Anselin, L. (1990): Spatial dependence and spatial structural instability in applied regression analysis. *Journal of Regional Science*, 30: 185–207.

Anselin, L. (2010): Thirty years of spatial econometrics. *Papers in Regional Science* 89(1): 3-25.

Anselin, L. and Bera A. (1998): Spatial dependence in linear regression models with an introduction to spatial econometrics. In D. Giles and A. Ullaah (eds). *Handbook of Applied Economic Statistics* pp.237-289. Dekker: New York.

Anselin, L. and Griffith, D. (1988): Do spatial effects really matter in regression analysis? *Papers of the Regional Science Association*. 65: 11-34.

Arbia, G. (2014): *A Primer for Spatial Econometrics. With Applications in R*. Palgrave MacMillan.

Arbia, G. (2016): Spatial Econometrics: A Broad View. *Foundations and Trends in Econometrics* vol.8 (3-4).

Belsley, D. (1991): *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. New York: Wiley. ISBN 0-471-52889-7.

von Breymann, H (2017) Residential satisfaction: the resident´s experience as an urban planning tool. The case of the Embajadores neighborhood, Madrid, *Revista Urbano* 35, 88-101.

Breusch, T. and Pagan A. (1979): A simple test for heteroskedasticity and random coefficient variation, *Econometrica* 47, 1287-94.

Brunsdon, C., Fotheringham, A.S. and Charlton, M. (1999): Some notes on parametric significance tests for geographically weighted regression. *Journal of Regional Science* 39: 497-52.

Camacho J., F. Díaz, E. Gadea, X Ginés and M.L. Lourés (2015): Madrid: the end of an urban model and the construction of new proposals for a changing city, *Quid 16* 5, 5-45.

Casetti, E. and Can, A. (1999): The econometric estimation and testing of DARP models. *Journal of Geographical Systems*, 1: 91-106.

Chasco, C. and Le Gallo, J. (2013): The impact of objective and subjective measures of air quality and noise on house prices: a multilevel approach for downtown Madrid. *Economic Geography* 89: 127-148.

Cheshire, P. and Sheppard S. (1998): Estimating the demand for housing, land, and neighborhood characteristics. *Oxford Bulletin of Economics and Statistics* 60: 357–82.

Cressie, N. (1980): The asymptotic distribution of the scan statistic under uniformity. *The Annals of Probability* 8(4): 828-840.

Cucala, L. (2016): Scan statistics for detecting high-variance clusters, *Journal of Probability and Statistics*, Article ID 7591680, 8 pages.

Dorn, S. and Egger, P.H. (2014): Small-sample inference with spatial HAC estimators. *Economics Letters* 2: 236-239.

Dubé, J. and Legros, D. (2014): *Spatial Econometrics using Microdata*. Wiley-ISTE.

Elhorst, J.P. (2010): Applied spatial econometrics: Raising the bar, *Spatial Economic Analysis* 5(1), 9-28.

Elhorst, J.P. (2014): *Spatial Econometrics. From Cross-Sectional Data to Spatial Panels*. Springer-Verlag, Berlin.

Ertur, C. and Le Gallo, J. and Baumont, C. (2006): The regional convergence process, 1980-1995: Do spatial regimes and spatial dependence matter? *International Regional Science Review* 29 3-34.

Fotheringham, S and Zhan, B. (1996): A comparison of three exploratory methods for cluster detection in spatial point patterns. *Geographical Analysis* 28: 200-218.

Fotheringham, A., Charlton, M. and Brunsdon, C. (1999): Geographically weighted regression. A natural evolution of the expansion method for spatial data analysis. *Environment and Planning A* 30: 1905-1927.

García, E. (2014) Gentrificación en Madrid: de la burbuja a la crisis, *Revista de Geografía Norte Grande* 58, 71-91.

Geniaux G., Martinetti D. (2017): A new method for dealing simultaneously with spatial autocorrelation and spatial heterogeneity in regression models, *Regional Science and Urban Economics*, forthcoming.

Godfrey, L.G. (1996): Some results on the Glejser and Koenker tests for heteroscedasticity. *Journal of Econometrics* 72: 275-299.

Goldfeld, S.,R. and Quandt (1965): Some Tests for Homoscedasticity. *Journal of the American Statistical Association* 60: 539–547.

Griffith, D. (2003): *Spatial Autocorrelation and Spatial Filtering*. Berlin: Springer.

Huang, L., Pickle, W., Das, B. (2008): Evaluating spatial methods for investigating global clustering and cluster detection of cancer cases. *Statistics in Medicine* 27: 5111–5142.

Judge, G., Hill, C., Griffiths, W, Lee, T. and Lutkepol H. (1985): *The Theory and Practice of Econometrics*. New York: John Willey and Sons.

Kelejian, H.H., Prucha, I. (1999): HAC estimation in a spatial framework. *Journal of Econometrics* 140: 131-154.

Kelejian, H.H. and Prucha I.R. (2010): Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances, *Journal of Econometrics* 157: 53-67.

Kelejian, H.H. and Robinson, D.P. (1995): The influence of spatially correlated heteroskedasticity on test for spatial correlation. In Anselin L., Florax R.G.J.M. (Eds.), *New Directions in Spatial Econometrics* pp. 79-97, Springer-Verlag, Berlin.

Kelejian, H.H. and Robinson, D.P. (1998): A suggested test for spatial autocorrelation and/or heteroskedasticity and corresponding Monte Carlo results. *Regional Science and Urban Economics* 28: 389–417.

Kelejian, H.H. and Robinson, D.P. (2004): The influence of spatially correlated heteroskedasticity on tests for spatial correlation. In Anselin L., Florax RJGM, Rey S.J. (Eds.), *Advances in Spatial Econometrics* pp. 79-97, Springer Berlin Heidelberg.

Kim, M.S. and Sun, Y. (2011): Spatial heteroskedasticity and autocorrelation consistent estimation of covariance matrix. *Journal of Econometrics* 2 349-371.

Koenker, R. and Bassett G. (1982): Robust tests for heteroskedasticity based on regression quantiles, *Econometrica* 50: 43-61.

Kulldorff, M., Huang, L. and Konty, K. (2009): A scan statistic for continuous data based on the normal probability model, *International Journal of Health Geographics* 8: 58-73.

Kulldorff, M. and Nagarwalla, N. (1995): Spatial disease clusters: detection and inference, *Statistics in Medicine* 14: 799-810.

Lesage, J. and Pace, K. (2009): *Introduction to Spatial Econometrics*. Boca Raton: Chapman & Hall/CRC.

Le Gallo, J. (2014): Cross-section spatial regression models. In: Fischer M.M., Nijkamp P. (Eds.), *Handbook of Regional Science* pp. 1511-1533, Springer-Verlag, Berlin.

Leković, M. (2013) *Barrio de Valdeacederas: entre abandono, remodelacion y gentrificación*, Departament d'Urbanisme i Ordenació del Territori. Universitat Politècnica de Catalunya, Barcelona.

Lin, X. and Lee, L.F. (2010): GMM estimation of spatial autoregressive models with unknown heteroskedasticity, *Journal of Econometrics* 157: 34-52.

Lin, Hartz, P.S., Zhang, Z., Saccone, S., Wang, J., Tischfield, J., Edenberg, H., Kramer, J., Goate, A., Bierut, L. and Rice, J. (2010): A new statistic to evaluate imputation reliability. *PloS ONE* 5 e9697.

López, F.A., Mur, J. and Angulo, A. (2009): Maps of continuous spatial dependence. *Region et Development* 30: 3-20.

López, F.A., Chasco, C. and Le Gallo, J. (2015): Exploring scan methods to test spatial structure with an application to housing prices in Madrid. *Papers in Regional Science* 94: 317-346.

Martínez B (2014), Cityscape and environmental issues: the case of district Retiro (Madrid), *Espacio, tiempo y forma* 6–7, 119–160.

Muñoz, O. (2014) Gentrification, segregation and social restructuring in Madrid, *Revista de Direito da Cidade* 06(01), 180-207.

Mur, J. and Angulo, A. (2009): Model selection strategies in a spatial setting: Some additional results. *Regional Science and Urban Economics* 39: 200-213.

Mur, J., López F.A. and Angulo, A. (2009): Testing the hypothesis of stability in spatial econometric models. *Papers in Regional Science* 88: 409-444.

Openshaw, S., Charlton, E., Wymer, C. and Craft, A. (1987): A Mark 1 geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems* 1: 359-377.

Ord, J.K. and Getis, A. (2012): Local spatial heteroscedasticity (LOSH) *Annals of Regional Science* 48: 529–539.

Orford, S. (2000): Modelling spatial structures in local housing market dynamics: A multilevel perspective. *Urban Studies* 37: 1643–1671.

Páez, A., Uchida, T. and Miyamoto, K. (2001): Spatial association and heterogeneity issues in land price models. *Urban Studies* 38: 1493-1508.

Páez, A., Uchida, T. and Miyamoto, K. (2002): A general framework for estimation and inference of geographically weighted regression models: 1. Location-specific kernel bandwidths and a test for locational heterogeneity. *Environment and Planning A* 34: 733-754

Palm, R. (1978): Spatial segmentation of the urban housing market, *Economic Geography* 54: 210–21.

Straszheim, M. (1974): Hedonic estimation of housing market prices: A further comment. *Review of Economics and Statistics* 56: 404–06.

White, H. (1980): A Heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48: 817–838.

Yan, J. (2007): Spatial stochastic volatility for lattice data *Journal of Agricultural, Biological, and Environmental Statistics* 12: 25-40.

Zhang, Z., Assunção, R., and Kulldorff, M. (2010): Spatial scan statistics adjusted for multiple clusters. *Journal of Probability and Statistics*, 2010.

Zhang, T. and Lin, G. (2016): On Moran's image coefficient under heterogeneity. *Computational Statistics and Data Analysis* 95: 83-94.

Zhang, T. and Lin, G. (2017): Asymptotic properties of spatial scan statistics under the alternative hypothesis, *Bernoulli* 23(1): 89-109.

TABLES

Table 1: Estimated size for the $Scan_{\sigma}$ test. Percentage p_p-value < 0.05

		n	DGP1	DGP2	DGP3	DGP4	DGP5	DGP6a	DGP6b	DGP7
Elliptic Windows***	Regular Lattice*	36	0.062	0.051	0.077	0.073	0.062	0.052	0.047	0.053
		49	0.053	0.050	0.069	0.063	0.055	0.055	0.051	0.046
		100	0.056	0.051	0.054	0.049	0.054	0.046	0.052	0.066
		225	0.043	0.065	0.046	0.042	0.050	0.064	0.050	0.055
		400	0.045	0.057	0.050	0.055	0.050	0.050	0.049	0.043
			DGP1	DGP2	DGP3	DGP4	DGP5	DGP6a	DGP6b	DGP7
	Irregular Lattice**	36	0.054	0.054	0.064	0.081	0.056	0.045	0.047	0.051
		49	0.054	0.057	0.067	0.062	0.050	0.058	0.055	0.041
		100	0.054	0.043	0.046	0.051	0.053	0.048	0.060	0.058
		225	0.057	0.044	0.059	0.057	0.067	0.039	0.051	0.056
		400	0.051	0.048	0.052	0.055	0.054	0.048	0.053	0.055

1000 iterations with 999 boots

* Hexagonal lattice

** Coordinates X, Y=U(0,1)

*** Q is a set of elliptic windows with eccentricity of e = 1, 2, 3 and 4 and rotation angles $\theta=\pi(2k+1)/18$; k=0,1,…,8

Table 2: Power of the $Scan_{\sigma}$ tests with elliptic windows*. Percentage p-value < 0.05

		Regular Lattice†									Irregular Lattice‡								
		Discrete SGWH						Continuous SGWH			Discrete SGWH						Continuous SGWH		
	n	SGWH1	SGWH2	SGWH3	SGWH4	SGWH5	SGWH6	SGWH7	SGWH8	SGWH9	SGWH1	SGWH2	SGWH3	SGWH4	SGWH5	SGWH6	SGWH7	SGWH8	SGWH9
$Scan_{\sigma}$	36	0.378	0.082	0.523	0.097	0.140	0.047	0.183	0.320	0.078	0.384	0.082	0.505	0.094	0.147	0.055	0.092	0.256	0.311
	49	0.449	0.058	0.677	0.113	0.205	0.047	0.450	0.490	0.267	0.478	0.069	0.755	0.121	0.213	0.046	0.107	0.374	0.352
	100	0.609	0.064	0.999	0.264	0.672	0.081	0.997	0.893	0.987	0.599	0.068	0.997	0.292	0.740	0.093	0.192	0.800	0.808
	225	0.606	0.065	1.000	0.659	0.992	0.187	1.000	1.000	1.000	0.645	0.071	1.000	0.721	0.991	0.216	0.454	0.998	0.999
	400	0.627	0.065	1.000	0.962	1.000	0.393	1.000	1.000	1.000	0.678	0.064	1.000	0.974	1.000	0.446	0.725	1.000	1.000
BP_{coor}	36	0.139	0.035	0.680	0.123	0.126	0.055	0.023	0.554	0.122	0.543	0.103	0.614	0.119	0.139	0.036	0.059	0.407	0.231
	49	0.275	0.063	0.816	0.193	0.217	0.061	0.109	0.653	0.121	0.613	0.104	0.862	0.193	0.161	0.040	0.061	0.573	0.216
	100	0.513	0.067	1.000	0.497	0.300	0.039	0.202	0.957	0.202	0.614	0.091	0.999	0.465	0.284	0.046	0.077	0.900	0.287
	225	0.600	0.081	1.000	0.872	0.475	0.061	0.971	0.999	0.939	0.526	0.085	1.000	0.888	0.376	0.056	0.114	0.998	0.409
	400	0.561	0.069	1.000	0.990	0.744	0.063	1.000	1.000	0.997	0.469	0.054	1.000	0.990	0.557	0.061	0.153	1.000	0.489
BP_{us}	36	0.134	0.053	0.092	0.049	0.170	0.055	0.076	0.161	0.169	0.173	0.065	0.105	0.054	0.176	0.054	0.069	0.090	0.089
	49	0.266	0.090	0.105	0.059	0.140	0.052	0.119	0.057	0.092	0.223	0.056	0.105	0.049	0.173	0.048	0.057	0.089	0.091
	100	0.231	0.073	0.139	0.057	0.270	0.063	0.165	0.111	0.148	0.238	0.059	0.135	0.078	0.281	0.061	0.065	0.098	0.100
	225	0.261	0.050	0.206	0.108	0.296	0.077	0.257	0.107	0.280	0.197	0.060	0.170	0.066	0.264	0.071	0.080	0.094	0.115
	400	0.240	0.049	0.173	0.086	0.184	0.047	0.545	0.164	0.492	0.165	0.040	0.162	0.079	0.335	0.065	0.081	0.120	0.108
MI	36	0.043	0.040	0.079	0.059	0.046	0.039	0.049	0.089	0.068	0.104	0.048	0.075	0.064	0.068	0.057	0.078	0.069	0.067
	49	0.070	0.046	0.082	0.063	0.049	0.050	0.061	0.091	0.060	0.105	0.059	0.090	0.055	0.071	0.056	0.063	0.079	0.056
	100	0.099	0.054	0.120	0.061	0.086	0.048	0.108	0.097	0.091	0.124	0.053	0.107	0.066	0.127	0.046	0.062	0.082	0.078
	2	0.0	0.0	0.1	0.0	0.1	0.0	0.3	0.0	0.2	0.1	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.0

25	93	51	08	59	36	46	04	81	37	25	61	05	60	63	57	72	78	84
400	0.089	0.054	0.113	0.073	0.141	0.058	0.419	0.106	0.368	0.107	0.043	0.126	0.074	0.196	0.056	0.055	0.087	0.096

In shadow, cases where the test has more power than the other tests; 1000 iterations with 999 boots;
[†] Hexagonal lattice; [‡] Coordinates X, Y=U(0,1); [*] Q is a set of elliptic windows with eccentricity of e = 1, 2, 3 and 4 and rotation angles θ=π(2k+1)/18; k=0,1,…,8

Table 3: Performance *Scan*$_\sigma$ test in presence of spatial pattern in mean (elliptic windows). Percentage p_p-value<0.05

		DGP8a: SAR			DGP9a: SEM			DGP8b: HSAR			DGP9b: HSEM			DGP10a		DGP10b	
	n	ρ=0.2	ρ=0.5	ρ=0.8	λ=0.2	λ=0.5	λ=0.8	ρ=0.2	ρ=0.5	ρ=0.8	λ=0.2	λ=0.5	λ=0.8	Low	Strong	Low	Strong
Regular lattice	36	0.054	0.084	0.205	0.062	0.071	0.152	0.068	0.050	0.116	0.056	0.048	0.100	0.053	0.114	0.064	0.071
	49	0.051	0.073	0.258	0.050	0.079	0.170	0.038	0.070	0.150	0.040	0.078	0.134	0.052	0.092	0.049	0.102
	100	0.040	0.090	0.405	0.051	0.095	0.357	0.046	0.086	0.328	0.054	0.088	0.266	0.045	0.176	0.044	0.590
	225	0.053	0.114	0.629	0.049	0.103	0.624	0.048	0.104	0.534	0.054	0.108	0.462	0.048	0.421	0.041	0.965
	400	0.061	0.131	0.789	0.046	0.121	0.740	0.045	0.120	0.682	0.066	0.112	0.612	0.052	0.572	0.048	1.000
Irregular lattice	36	0.053	0.086	0.356	0.055	0.073	0.332	0.076	0.084	0.210	0.066	0.068	0.122	0.060	0.061	0.047	0.114
	49	0.058	0.097	0.584	0.050	0.103	0.493	0.068	0.086	0.340	0.072	0.068	0.242	0.059	0.098	0.054	0.227
	100	0.063	0.129	0.740	0.055	0.118	0.714	0.060	0.114	0.366	0.064	0.096	0.424	0.059	0.126	0.049	0.361
	225	0.052	0.179	0.918	0.060	0.162	0.896	0.060	0.148	0.550	0.050	0.132	0.668	0.053	0.599	0.040	0.995
	400	0.068	0.216	0.964	0.060	0.178	0.964	0.060	0.200	0.746	0.074	0.162	0.836	0.052	0.635	0.048	1.000

Table 4: Variables used in the model

Var.	*Description*	*Source*	*Units*	*Var.*	*Description*	*Source*	*Units*
lpri	Sale price	Idealista	log €				
Structural characteristics:				*Variables of accessibility, distance to:*			
flo1	Third and higher floors plus houses	Idealista	0-1	*dcbd*	The business center (CBD)	Self-elabor., C. Madrid, SABI	Km.
flbs	Basement and semi-basement		0-1	*dsol*	Historical center		Km.
flb0	Basement, semi-basement and ground floor		0-1	*dm30*	M-30 road-belt		Km.
atti	Attic		0-1	*dm40*	M-40 road-belt		Km.
hous	House (detached, semi-detached, row)		0-1	*dmet*	Closest metro station		Km.
deta	Detached house		0-1	*dtra*	Closest train station		Km.
sdet	Semi-detached house		0-1	*dhub*	Intermodal transport hubs		Km.
hflo	House floors		numb.	*dair*	International airport		Km.
hplo	House on a plot		0-1	*dpar*	Closest green area (over 1 Ha)		Km.
dupl	Duplex		0-1	*dhip*	Closest hypermarket		Km.
beds	Bedsit		0-1	*dsho*	Closest shopping centers		Km.
lsqm	Surface		log m^2	*dkil*	Closest 'category killer' center[12]		Km.
bedr	Bedrooms		numb.	*dser*	Service providers' outlets: retailing, hotels and restaurants		Km.
refb	Refurbished		0-1	*hot5*	Closest 5-star hotel		Km.
refo	Needs renovation		0-1	*Geographical characteristics:*			
new	New		0-1	*D*	District 1 to 7	Self-elabor., GIS, C. Madrid	Km.
nele	Building without elevator		0-1	*N*	Neighborhood 1 to 43		Km.
inne	All the rooms are facing an inner courtyard		0-1	*xcoo*	Longitude coordinate		Km.
outd	All the rooms are facing outdoor public areas		0-1	*ycoo*	Latitude coordinate		Km.
gara	Garage space		0-1	*xyco*	Longitude × Latitude		Km.
terr	Dwelling with terrace		0-1				
view	Nice views		0-1				

[12] A 'category killer' is marketing industry jargon for big-box retail chains, such as Leroy Merlin that has such an advantage over other firms in its market that competing firms find it almost impossible to operate profitably and have to leave the industry, thereby increasing the dominant firm's concentration ratio.

Table 5: Regression diagnostics for Models 1 and 2

Model	Model 1 Basic model	Model 2 Spatial submarkets	Model	Model 1 Basic model	Model 2 Spatial submarkets
Estimation	**OLS**	**OLS**	**Estimation**	**OLS**	**OLS**
Adjusted R^2	0.8653	0.8669			
Condition #	4753	22829			
Jarque-Bera	27.43^{***}	34.12^{***}			
Koenker-B.	212.04^{***}	240.35^{***}			
RLMe, 2nn	107.90^{***}	76.43^{***}	MLC, cases	Low, 1324	Low, 727
RLMl, 2nn	79.41^{***}	67.81^{***}	SC1, cases	Low, 145	Low, 126
RLMe, 5nn	240.73^{***}	141.10^{***}	SC2, cases	Low, 95	Low, 65
RLMl, 5nn	113.96^{***}	99.49^{***}	SC3, cases	Low, 1048	Low, 1791
RLMe, 6nn	198.08^{***}	107.49^{***}	SC4, cases	Low, 72	Low, 190
RLMl, 6nn	118.64^{***}	89.90^{***}	SC5, cases	Low, 120	
RLMe, 30nn	538.00^{***}	196.46^{***}			
RLMl, 30nn	118.64^{***}	98.88^{***}			
RLMe, dinv	154.38^{***}	94.39^{***}			
RLMl, dinv	88.81^{***}	88.38^{***}			

Note: *** significant at 1%, ** significant at 5%, * significant at 10%, *Koenker-B* is Koenker-Basset test, RLMe and RLMl are the robust LM tests for spatial error and spatial lag models, respectively, 2nn, 5nn, 6nn, 30nn are the 2, 5, 6, 30 nearest neighbor W matrix, respectively, dinv is an inverse distance W matrix for a distance band of 375 meters, MLC is most likely cluster, SC is secondary cluster.

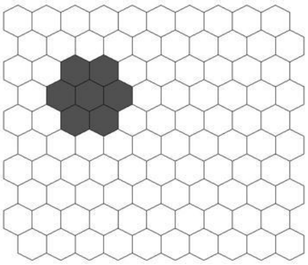
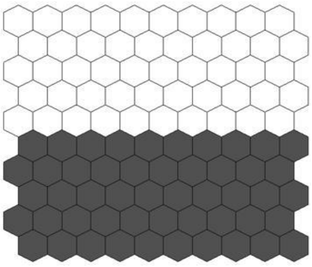
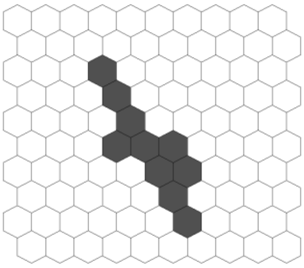
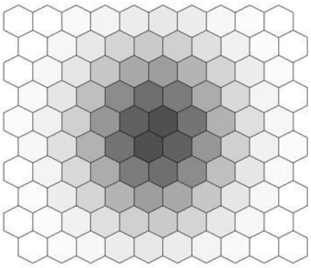
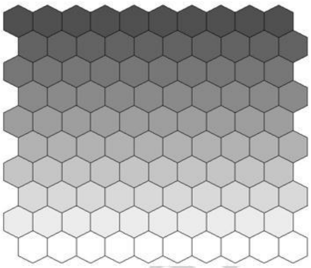
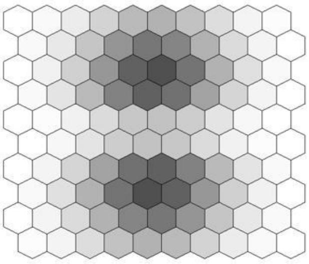
Table 6: Main regression results for Models 3, 4, 5 and 6

Models	Model 3 Submarkets & interactions	Model 4 SEM model (W = 5 nn)	Model 5 SAR model (W = 5 nn)	Model 6 SARAR model (W = 5 nn)
Estimation	**OLS**	**GMM**	**S2SLS**	**GMM-KPHET**
Spatial variable coefficients:				
Rho	-	-	0.1652^{***}	0.1618^{***}
Lambda	-	0.4056^{***}	-	0.2684^{***}
Goodness-of-fit and diagnostics:				
Adjusted -R^2	0.8491	-	-	-
Jarque-Bera	38.66^{***}	-	-	-
Koenker-B.	12.108	-	-	-
Quasi-White	-	59.047^{**}	59.180^{**}	48.302^{**}
RLMe, 5nn	357.56^{***}	-	-	-
RLMl, 5nn	202.10^{***}	-	-	-
Moran's I	-	0.2612^{***}	0.1236^{***}	-0.0043
Anselin-Kel.	-	-	158.281^{***}	-
MLC, cases	Low, 1253	Low, 27	Low, 1231	Low, 816
SC1, cases	Low, 19	Low, 1307	High, 182	High, 182
SC2, cases	Low, 541	High, 53	Low, 19	High, 122
SC3, cases	High, 65	Low, 19	High, 163	-
SC4, cases	High, 57	Low, 541	-	-
SC5, cases		High, 31		
		High, 389		

Note: *** significant at 1%, ** significant at 5%, * significant at 10%, *Koenker-B* is Koenker-Basset test, RLMe and RLMl are the robust LM tests for spatial error and spatial lag models, respectively, *Anselin-Kel.* is the Anselin-Kelejian spatial test on the STSLS residuals, W is spatial weight matrix, 5nn is the 5 nearest neighbor W, MLC is most likely cluster, SC is secondary cluster.

FIGURES:

Figure 1: Spatial heteroskedastic patterns: discrete and continuous processes (n=100).

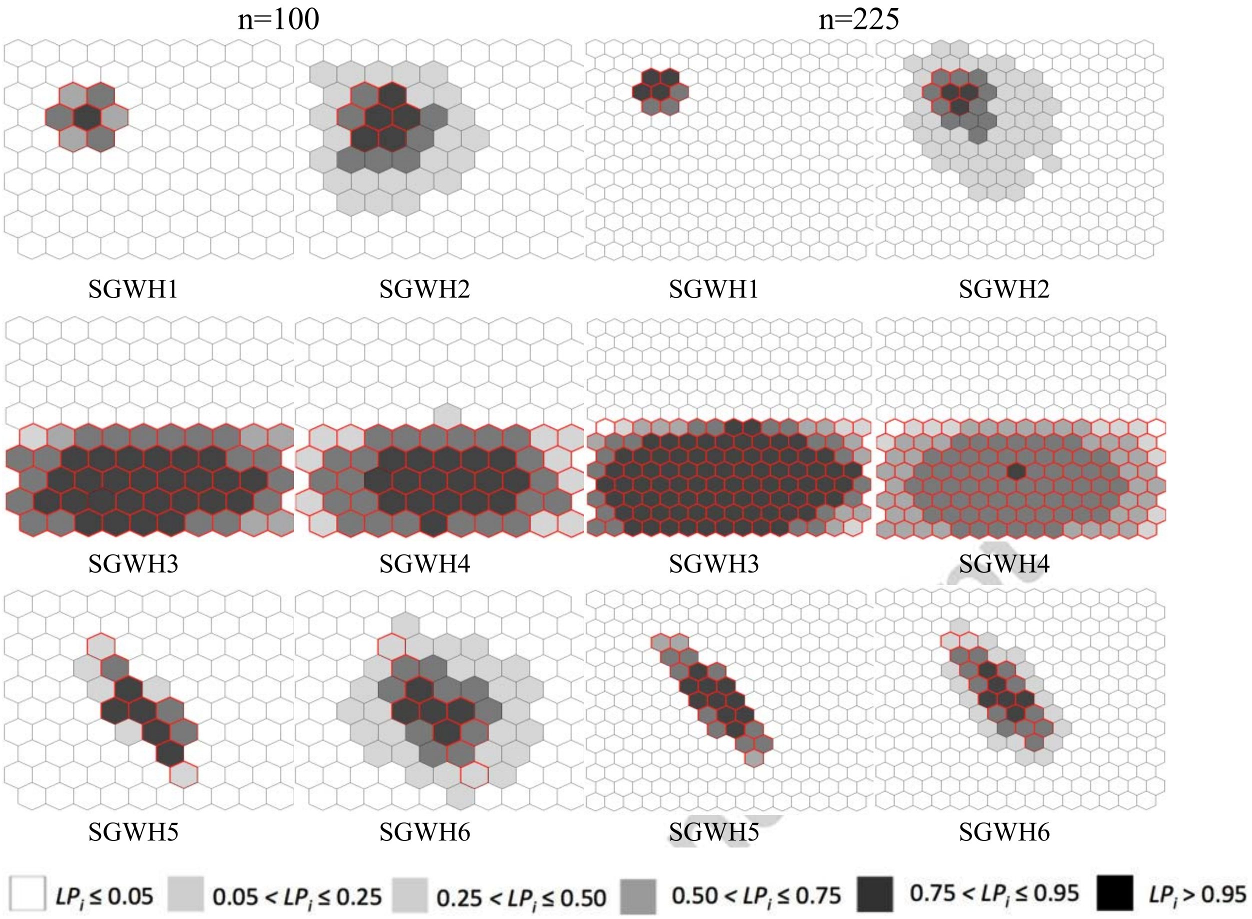
SGWH1 – SGWH2	SGWH3 – SGWH4	SGWH5 – SGWH6
		
$h(cx_i,cy_i)=\begin{cases} a \text{ if } (cx_i,cy_i)\in Z^{\dagger} \\ b \text{ if } (cx_i,cy_i)\notin Z^{\dagger} \end{cases}$ in SGWH1 $a=0.9; b=0.1$ in SGWH2 $a=0.7; b=0.3$	$h(cx_i,cy_i)=\begin{cases} a \text{ if } cy_i < Me_Y \\ b \text{ if } cy_i \geq Me_Y \end{cases}$ in SGWH3 $a=0.9; b=0.1$ in SGWH4 $a=0.7; b=0.3$	$h(cx_i,cy_i)=\begin{cases} a \text{ if } (cx_i,cy_i)\in Z^{\ddagger} \\ b \text{ if } (cx_i,cy_i)\notin Z^{\ddagger} \end{cases}$ in SGWH5 $a=0.9; b=0.1$ in SGWH6 $a=0.7; b=0.3$
SGWH7	**SGWH8**	**SGWH9**
		
$h(cx_i,cy_i)=$ $=\exp\left[-0.05(cx_i - Me_X)^2+(cy_i - Me_Y)^2\right]$	$h(cx_i,cy_i)=\dfrac{yc_i}{\max(yc_i)}$	$h(cx_i,cy_i)=0.5\lvert cy_i - Me_Y\rvert$ $\times\exp\left[-0.05(cx_i - Me_X)^2+(cy_i - Me_Y)^2\right]$

Me_x is the median of the cx coordinates. *Me_y* is the median of the cy coordinates.

[†] the circle Z has the same size, 7 cells, for all the cases

[‡] the size of the ellipse Z is variable: [n/10] cells

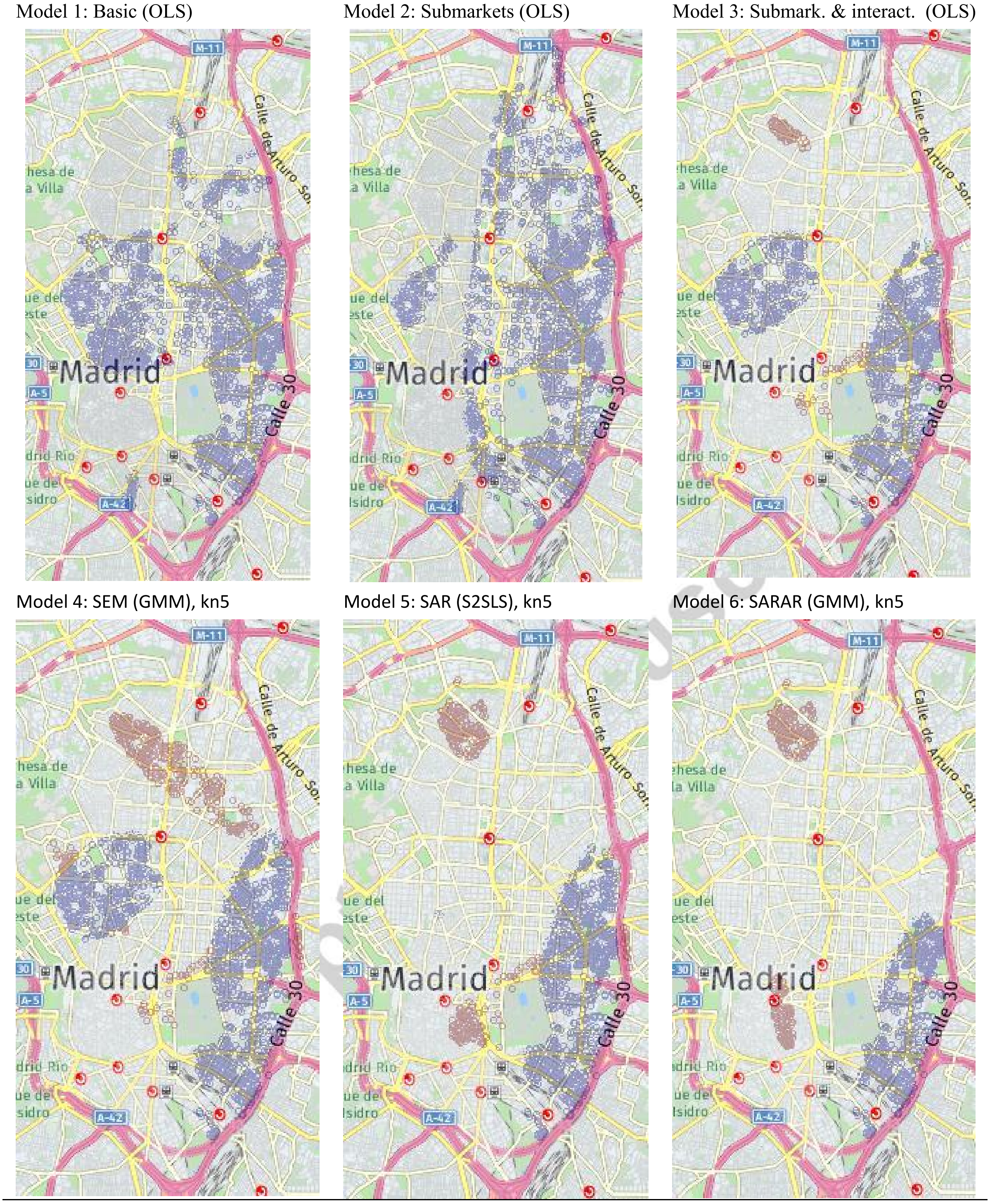
Figure 2: Local Sensibility of $Scan_{\sigma}$ for discrete heteroskedasticity patterns*



* Border in red for the true cluster

Results for n=36, 49 and 400 non report to save space (avalaible under request).

Figure 3: Significant spatial groupwise heteroskedasticity clusters of the error terms



Model 1: Basic (OLS)

Model 2: Submarkets (OLS)

Model 3: Submark. & interact. (OLS)

Model 4: SEM (GMM), kn5

Model 5: SAR (S2SLS), kn5

Model 6: SARAR (GMM), kn5

Note: Brown and orange: high-variance clusters significant at 5% and 10%, respectively; dark and light blue: low-variance clusters significant at 5% and 10%, respectively; white: rest of the observations.

APPENDICES:

Appendix 1

Proof Theorem 1.

Let $\mathrm{Z}\ \hat{\mathrm{I}}\,\mathrm{Q}$

$$\mathrm{Scan_s(Z)} = \max_{Z\in Q}[\mathrm{l}(H_A) - \mathrm{l}(H_0)] = \max_{Z\in Q}\left[2n\left(\ln(\frac{\hat{s}^2_{H_0}}{\hat{s}^2_{H_A}(\bar{Z})}) + \frac{n_Z}{n}\ln(\frac{\hat{s}^2_{H_A}(\bar{Z})}{\hat{s}^2_{H_A}(Z)})\right)\right]$$

Note that

$$\hat{s}^2_{H_0} = \frac{1}{n}\overset{n}{\underset{i=1}{\mathring{a}}}(x_i - \bar{x})^2 = \frac{1}{n}\left(n_Z\hat{s}^2_{H_A}(Z) + (n - n_Z)\hat{s}^2_{H_A}(\bar{Z})\right)$$

therefore

$$\ln\left(\frac{\hat{s}^2_{H_0}}{\hat{s}^2_{H_A}(\bar{Z})}\right) = \ln\left(\frac{n_Z}{n}\left(\frac{\hat{s}^2_{H_A}(Z)}{\hat{s}^2_{H_A}(\bar{Z})} - 1\right) + 1\right) @ \frac{n_Z}{n}\left(\frac{\hat{s}^2_{H_A}(Z)}{\hat{s}^2_{H_A}(\bar{Z})} - 1\right) \text{ if } \frac{n_Z}{n}\left(\frac{\hat{s}^2_{H_A}(Z)}{\hat{s}^2_{H_A}(\bar{Z})} - 1\right) \approx 0$$

and here

$$\mathrm{Scan}_s(Z) @ \max_{Z\in Q} 2n\left[\frac{n_Z}{n}\left(\frac{\hat{s}^2_{H_A}(Z)}{\hat{s}^2_{H_A}(\bar{Z})} - 1\right) + \frac{n_Z}{n}\ln(\frac{\hat{s}^2_{H_A}(\bar{Z})}{\hat{s}^2_{H_A}(Z)})\right] = \max_{Z\in Q} 2n_Z\left(\hat{d} - \ln(\hat{d}) - 1\right)$$

where we note:

$$\hat{d} = \hat{s}^2_{H_A}(Z)\Big/\hat{s}^2_{H_A}(\bar{Z})$$

Now, consider the case of the alternative hypothesis H$_A$: {x$_i$} i.i.d. normal processes with E[x$_i$]=μ, with a set Z∈Θ where $s^2_Z = d\,s^2_{\bar{Z}}$ with δ≠1 (note that if δ>1 the alternative hypothesis would indicate a high variance cluster (and if 0<δ< 1, a low variance cluster would be indicated). We prove the consistence of the *Scan*$_σ$ test.

Let 0<C<+∞ with C∈R. By definition, the *Scan*$_σ$ is consistent if:

$$\lim_{n\to\infty} P(\mathrm{Scan_s}(\hat{Z}) > C\,|_{H_A}) = 1\ ;\ \ "C > 0$$

Using the previous result:

$$\lim_{n\to\infty} P(\mathrm{Scan_s}(\hat{Z}) > C\,|_{H_A}) = \lim_{n\to\infty} P(\max_{Z\in Q} 2n_{\hat{Z}}\left(\hat{d} - \ln(\hat{d}) - 1\right) > C\,|_{H_A}) = \lim_{n\to\infty} P(\left(\hat{d} - \ln(\hat{d}) - 1\right) > \frac{C}{2n_Z}\,|_{H_A})$$

Note that $f(\hat{d}) = \left(\hat{d} - \ln(\hat{d}) - 1\right)$ is a positive function and reaches a minimum for $\hat{d}$=1 ($f(1) = 0$). Therefore, it is necessary to have $n_Z \to +\infty$ and 1/n an infinitesimal of high order than 1/n$_Z$ for the *Scan*$_\sigma$ test to be consistent. In this case, as $\hat{s}^2_{H_A}(\bar{Z})$ and $\hat{s}^2_{H_A}(Z)$ are consistent $\hat{d} \xrightarrow{n_Z\to+\infty,\ n\to+\infty} d$ and therefore

$$\lim_{\substack{n\to\infty\\ n_Z\to\infty}} P((\hat{d} - \ln(\hat{d}) - 1) > \frac{C}{2n_Z}|_{H_A}) = P((d - \ln(d) - 1) > 0) = 1$$

Appendix 2

Theorem 2. If the SAR process *y* is defined by $y = (I - rW)^{-1}(Xb + e);\ e \equiv N(0, s^2 W)$, where $W = (w_{ij}) = (I - rW)^{-1}(I - rW')^{-1}$, then the spatial process H defined as $H = j^{-1/2}y$, with $j = \mathrm{diag}(w_{11},...,w_{nn})$, is a homoscedastic SAR process (HSAR) and $H \equiv N(\tilde{X}\beta, \sigma^2(I-\rho\tilde{W})^{-1}\varphi^{-1}(I-\rho\tilde{W}')^{-1})$ with $\tilde{X} = \varphi^{-1/2}X$ and $\tilde{W} = \varphi^{-1/2}W\varphi^{1/2}$.

Proof.

The

$$E[H] = E[\varphi^{-1/2}y] = \varphi^{-1/2}X\beta = \tilde{X}\beta$$

and

$$\mathrm{Var}(H) = s^2 j^{-1/2}(I - rW)^{-1}(I - rW')^{-1} j^{-1/2} =$$

$$= s^2 j^{-1/2}\left(j^{1/2}(I - r j^{-1/2}W j^{1/2}) j^{-1/2}\right)^{-1}\left(j^{-1/2}(I - r j^{1/2}W' j^{-1/2}) j^{1/2}\right)^{-1} j^{-1/2} =$$

$$= s^2(I - r j^{-1/2}W j^{1/2})^{-1}(I - r j^{1/2}W' j^{-1/2})^{-1} =$$

$$= \sigma^2(I-\rho\tilde{W})^{-1}(I-\rho\tilde{W}')^{-1}$$

Highlights

We propose a test for the presence of spatial groupwise heteroskedasticity

The test output show spatial clusters of observations with higher/lower variances

The test output is vital to improve the spatial regression model specification

Spatial dependence and spatial groupwise heteroskedasticity tests can be used jointly

An application on houses prices in Madrid shows the usefulness of this test