BANDWIDTH SELECTION FOR SPATIAL HAC AND OTHER ROBUST COVARIANCE ESTIMATORS

by

Dayton M. Lambert, Raymond J.G.M. Florax and Seong-Hoon Cho

Working Paper # 08-10

October 2008

Dept. of Agricultural Economics

Purdue University

BANDWIDTH SELECTION FOR SPATIAL HAC AND OTHER ROBUST COVARIANCE ESTIMATORS

Dayton M. Lambert[1], Raymond J.G.M. Florax[2,3] and Seong-Hoon Cho[1]
dmlambert@tennessee.edu, rflorax@purdue.edu, scho9@utk.edu

[1] Department of Agricultural Economics
University of Tennessee
321 Morgan Hall
Knoxville, TN 37996-4518, USA

[2] Department of Agricultural Economics
Purdue University
403 West State Street
West Lafayette, IN 47907-2056, USA

[3] Department of Spatial Economics
VU University Amsterdam
De Boelelaan 1105
1085 HV Amsterdam, The Netherlands

Working Paper # 08-10
October 2008

Abstract

This research note documents estimation procedures and results for an empirical investigation of the performance of the recently developed spatial, heteroskedasticity and autocorrelation consistent (HAC) covariance estimator calibrated with different kernel bandwidths. The empirical example is concerned with a hedonic price model for residential property values. The first bandwidth approach varies an a priori determined plug-in bandwidth criterion. The second method is a data driven cross-validation approach to determine the optimal neighborhood. The third approach uses a robust semivariogram to determine the range over which residuals are spatially correlated. Inference becomes more conservative as the plug-in bandwidth is increased. The data-driven approaches prove valuable because they are capable of identifying the optimal spatial range, which can subsequently be used to inform the choice of an appropriate bandwidth value. In our empirical example, pertaining to a standard spatial model and ditto dataset, the results of the data driven procedures can only be reconciled with relatively high plug-in values ($n^{0.65}$ or $n^{0.75}$). The results for the semivariogram and the cross-validation approaches are very similar which, given its computational simplicity, gives the semivariogram approach an edge over the more flexible cross-validation approach.

Keywords: spatial HAC, semivariogram, bandwidth, hedonic model
JEL Codes: C13, C31, R21

1 Introduction

Recent attention in the spatial econometric literature has focused on the development of spatial heteroskedasticity and autocorrelation consistent (HAC) covariance estimators. Conley (1999) and Kelejian and Prucha (2007a), KP for short from here on, both suggest a nonparametric procedure to attend to non-zero covariances between cross-sectional units and unequal variances. While both approaches are mechanically similar, the underlying assumptions regarding the data generating process between spatial units is more general in the KP approach. Their approach covers the widely popular Cliff and Ord (1973, 1981) non-stationary spatial processes, as well as the stationary spatial processes represented by Conley's approach.

The extent to which cross-sectional units are allowed to be correlated in the spatial HAC estimator is determined by a kernel density function. Different specifications of the kernel density tend to produce similar results asymptotically, as is well-known from the nonparametric literature (Mittelhammer, Judge and Miller 2000), although it can be demonstrated that the Epanechnikov kernel has a slight advantage over other candidate functions in terms of yielding the lowest mean integrated squared error (Cameron and Trivedi 2005). More crucial in terms of the performance of the estimator is the kernel bandwidth selection. Not unlike many other nonparametric approaches, the spatial HAC estimator essentially estimates residual cross-dependence as a locally weighted average. There is an obvious trade-off between setting the bandwidth sufficiently small to reduce the inefficiency caused by spatial error dependence, and choosing a bandwidth large enough to ensure a smooth distance decay process. At the same time, the bandwidth should not be too large to avoid over-smoothing (Cameron and Trivedi 2005).

There are two generally accepted methods for selecting appropriate bandwidths. The first approach suggests a plug-in estimator, which is typically determined as a function of sample

size. Newey and West's (1987) approach is an example from the time series literature, where the plug-in estimator determining the optimal number of temporal lags is trunc($4[T^{-1}100]^{2/9}$), where T is the number of time periods. In non-spatial local regressions, Silverman's (1986) plug-in estimator is frequently applied. It is defined as $b = 1.3643\delta N^{-0.2}\min\{s, iqr/1.349\}$, where N is sample size, s the sample standard deviation, iqr the interquartile range, and δ a constant associated with a kernel estimator (Cameron and Trivedi 2005). The second approach is empirical and is based on minimization of some objective criterion such as the squared residuals using a cross-validation procedure (Härdle and Marron 1985), or the bias-corrected Aikaike information criterion (Hurvich, Simonoff and Tsai 1998).

Data driven optimal bandwidth selection procedures are well-developed in the non-spatial, nonparametric local regression literature, but good guesses (Conley 1999; Lambert et al. 2007; Anselin and Lozano-Gracia, 2008), and plug-in estimators (Lambert et al. 2007) have generally been applied in the empirical literature utilizing spatial HAC estimators. In this research note, we suggest two data driven approaches for determining an optimal bandwidth criterion for KP's spatial HAC estimator. The first approach is based on the estimation of a robust semivariogram, using geostatistical procedures to determine an optimal range for which residual spatial covariance is significantly different from zero (Cressie 1993). The second approach applies a cross-validation procedure used to calibrate local spatial regressions. The geostatistical approach has two immediate advantages. First, the extent to which disturbances across spatial units are correlated is oftentimes clear from visual inspection of the robust semivariogram. Second, fitting the empirical semivariogram with a distance decay function allows for hypothesis testing with respect to the statistical significance of the range over which spatial units are correlated.

The remainder of this research note proceeds as follows. In the next section, the basic assumptions behind KP's spatial HAC estimator are reviewed, along with the data driven procedure to estimate the kernel bandwidth. The third section outlines the premise behind nonparametric estimation of the semivariogram and discusses the nonlinear regression procedures commonly used to fit the semivariogram. In section 4, various common empirical models and estimators are discussed, followed by a description of the data in section 5. Results are discussed in the sixth section, and the final section concludes.

2 Spatial HAC and data driven bandwidth selection

Kelejian and Prucha (2007a) suggest a general cross-sectional disturbance process allowing for unknown forms of spatial autocorrelation and heteroskedasticity across spatial units. An n by 1 vector of disturbances is generated as $\mu = R\varepsilon$, where ε is a vector of independently and identically distributed disturbances with mean zero and covariance $\sigma^2\Phi$, R is an n by n non-stochastic matrix with unknown elements whose row and column sums are uniformly bounded in absolute value (i.e., the correlation between cross-sectional units is restricted), and Φ is a diagonal matrix with non-negative, uniformly bounded elements. The asymptotic distribution of the n by p non-stochastic set of exogenous instruments H is $\Psi = n^{-1}H'\Sigma H$, where Σ is the covariance matrix of μ.

The problem remains to find an asymptotically consistent estimator for Ψ. Given consistent estimates of any linear or nonlinear regression, the residual vector $\hat{\mu}$ is used in conjunction with a kernel density function, $K(d_{ij}/d_{\max})$, that generates weighted averages of the residual cross-products over a certain distance ($d_{\max}$) at a decaying rate. The (r, s) estimated elements of Ψ are:

$$(1) \qquad \hat{\psi}_{rs} = \sum_{i=1}^{n}\sum_{j=1}^{n} h_{ir} h_{js} \hat{\mu}_i \hat{\mu}_j K\left(d_{ij} / d_{\max}\right),$$

where d_{ij} is the distance between observations i and j, and for all $d_{ij} \geq d$, $K(d_{ij}/d_{\max}) = 0$. In general, the function must be a real, continuous, bounded, and symmetric function that integrates to unity. There are several functional forms for $K(.)$ that meet these criteria, for example, Bartlett, bi-square, tri-cube, Epanechnikov, or Parzen kernels (Cameron and Trivedi 2005). In this analysis we follow Kelejian and Prucha, and use the Parzen kernel. The kernel function is effectively "adaptive" in the sense that at every location the smoothing parameter takes on different values. That is, for every observation, the distances between a spatial unit and its neighbors are sorted from low to high. The first b neighbors are then identified in the sorted vector, and subsequently used to build the spatial HAC spectral density matrix assuming an admissible kernel function.

Kelejian and Prucha (2007a) suggest a plug-in estimator b^* to identify $d_{\max}$ for each observation, with $b^* = n^{2\tau}$, with $\tau \leq \frac{1}{3}$.[1] In their Monte Carlo study of the spatial HAC estimator, they used $\tau \leq \frac{1}{8}$. In this study, we assume a variety of values for τ, ranging from 0.125 to 0.375 with 0.05 increments.

We compare the performance of the spatial HAC estimator based on Kelejian and Prucha's plug-in estimator to a spatial HAC estimator whose kernel bandwidth is selected using a data-driven procedure. In the spirit of Andrews (1990), who estimates bandwidths for HAC estimators in the time series literature, we suggest a cross-validation procedure typically used to estimate kernel bandwidths in the Geographically Weighted Regression (GWR) literature

[1] A similar plug-in bandwidth value was deduced by Conley (1999).

(McMillen 1996; Fotheringham, Brunsdon and Charlton 2002), or generally in the nonparametric local spatial regression literature (Cleveland and Devlin 1988; McMillen 2004). The cross-validation function *CV*, used to select *b* for the spatial HAC kernel, is given by:

$$(2) \qquad CV = \min_{b} \sum_{i=1}^{n} \left(y_i - \hat{y}_{-i}(b)\right)^2 ,$$

where $\hat{y}_{-i}$ is the fitted value of y_i with location *i* omitted during the fitting process. The predicted values $\hat{y}_{-i}$ are estimated with the local regression estimator:

$$(3) \qquad \hat{\beta}_i = (X'A_iX)^{-1}X'A_iy ,$$

where $a_{ij} = K(d_{ij}/d_{\max};b)$. One should note that the matrix *A* changes at each location *i*, and $\hat{\beta}_i$ is a parameter vector estimated at target location *i*, given a neighborhood defined by *b*. The optimal number of neighbors *b* (the bandwidth) minimizes the cross-validation function. Thus, in the locally weighted regression model, only observations up to the nearest *q* neighbors are assigned non-zero weights with respect to location *i*. To re-iterate, for every observation *i*, the vector of distances between *i* and all other observations are sorted in ascending order, and the *b* nearest neighbors are selected from this vector to form the relational matrix. This value is used as a cutoff point in the sorted distance vector to select $d_{\max}$ for the target location, which corresponds to the last distance entry in the truncated vector corresponding to spatial unit *i*. The mechanism permits $K(d_{ij}/d_{\max})$ to expand or contract across cross-sectional units, conditional on

the number of neighbors surrounding a given observation, and thereby re-weighting residual cross-products according to a localized neighborhood structure.

The *CV* function may be evaluated g (= 1, …, n–k) times, starting at $b_1 = k$ and continuing to $b_{n-k} = n - k$, where k is the number of variables in the regression model and n the number of observations in the data set. The parameter β_i is estimated for every location along the $n - k$ sequence and subsequently stored. As a result, $n - k$ evaluations of the model, that is estimates of β_i and their corresponding predicted values $\hat{y}_{-i}$, are generated in the search for the value of b_g that minimizes equation (2). The *CV* score from each iteration, which is the sum of the squared residuals, is saved for every $n - k$ evaluations. Once the sequence is exhausted, the smallest *CV* score in the stored vector is identified along with the value of b producing that score. Note that as b approaches the $n - k$ limit, the parameter estimates of the local regressions approach the same results that would be obtained with a global regression. The value of b minimizing the *CV* objective identifies the "optimal" number of neighbors that minimizes the residual sum of squares of the local regressions. Given the bandwidth b minimizing the *CV* objective, the spatial HAC spectral density matrix Ψ is estimated for the kernel function selected during the cross-validation procedure.

3 Semivariogram and empirical fitting

As an alternative to the data driven cross-validation approach one could also utilize semivariogram analysis to determine an optimal bandwidth. Details of this geostatistical method can be found in Schabenberger and Pierce (2002). In the first step, the semivariogram γ of the

model residuals is estimated using Cressie and Hawkin's (1980) robust estimator.[2] In the second step, the range α, nugget c_0, and sill ξ of the semivariogram are estimated with nonlinear regression. The range is the distance after which residual correlation between locations is effectively zero (white noise). The sill is the upper covariance limit between locations. The nugget is the model variance in the absence of residual covariance, or the constant variance component identical for all observations. There are a variety of functional forms suitable for estimating these parameters. In this application, we used the exponential function to obtain estimates for the range parameter α:

$$(4) \qquad \gamma(d) = c_0 + (\xi - c_0)\left(1 - e^{-3\|d\|/\alpha}\right),$$

where $||d||$ is the Euclidean distance between two locations. Weighted nonlinear least squares is used to fit the exponential function to the sample data—in our case are the residuals of the regression model of interest. Rejection of the null hypothesis $\alpha = 0$ suggests significant residual correlation between locations.

The distance corresponding to the range parameter α is subsequently used as $d_{\max}$ in the spatial HAC kernel. One should note that the kernel function of this estimator is not adaptive because the cutoff point is defined as a distance common to all locations (i.e., there may be more or fewer observations determining the covariance within the radius), whereas the adaptive bandwidth determined using the cross-validation approach is based on a distance metric defined by a set of nearest b neighbors where the furthest neighbor may vary for each location.

[2] Cressie and Hawkin's (1980) robust semivariogram estimator is particularly useful in the case of large datasets potentially containing bothersome outliers (see Cressie 1993 for details).

4 Empirical models and estimators

The semivariogram technique has been incorporated in spatial regression models in, for instance, Dubin (1992) and more recently with a focus on directionality (anisotropy) in Bannerjee, Gelfand and Sirmans (2003). These so-called direct representation models are, however, not the focus of this research note. Instead, we investigate the impact of alternative bandwidth selectors on the magnitude of the estimated Student *t*-values in various spatial econometric specifications, using a dataset on residential property values described in more detail below.

Most hedonic pricing studies use a spatial process model going back to Whittle, in which an endogenous variable is specified to depend on spatial interactions between cross-sectional units plus a disturbance term. The spatial interactions are typically modeled as a weighted average of nearby cross-sectional units, and the endogenous variable comprising the interactions is usually referred to as a spatially lagged variable. The weights are grouped in a matrix identifying neighborhood connections, which forms the distinctive core of spatial process models. The model containing a spatially lagged dependent variable is termed a spatial autoregressive lag model in the terminology of Anselin and Florax (1995). Whittle's spatial autoregressive lag model (SAR) was popularized and extended by Cliff and Ord (1973, 1981), who also distinguished non-stationary models in which the disturbances follow a spatial autoregressive process, the so-called spatial autoregressive error model (SEM). The general model, which contains a spatially lagged endogenous variable as well as spatially autoregressive disturbances in addition to exogenous variables, is called a spatial autoregressive model with autoregressive disturbances (SARAR). This SARAR model reads as $y = \rho Wy + X\beta + \varepsilon$, $\varepsilon = \lambda M\varepsilon + \mu$, where $\mu \sim$ iid$(0, \Omega)$, and W and M are (possibly identical) matrices defining connectedness between spatial units (Anselin 1988).

Two different approaches to dealing with model uncertainty can be distinguished. The first approach rests on the premise that the researcher can make credible assumptions about the underlying spatial process and that the spatial process is not merely a "nuisance". As a result, the researcher specifies a SAR or SARAR model based on an a priori defined, non-stochastic weights matrix. It is usually much more difficult to make an informed choice about the nature of the heteroskedasticity of the process, so it is typical in recent applied research to "correct for" heteroskedasticity by using a Huber-Eicker-White transformation or to allow for a very general form of heteroskedasticity (Kelejian and Prucha 2007b, Lambert and Florax 2008).[3] It is also possible that the spatial process is actually a nuisance, and that the researcher specifies a spatial error model (SEM). As with the SAR model, one can allow for heteroskedasticity through a "robustification" approach á la Huber-Eicker-White (Lambert and Florax 2008) or use the heteroskedastic version of the methods of moments estimator suggested by Kelejian and Prucha (1999, 2007b).

The second approach is based on substantially less information, because the researcher may not have a well-founded idea about the nature of the heteroskedastic process, but neither about the specification of the spatial process. In that case a very general spatial correlation process in the error terms, $\mu = R\varepsilon$, can be allowed for that concurrently incorporates heteroskedasticity. This is the approach suggested with KP's spatial HAC estimator. Strictly speaking, there is no need for a spatial weights matrix as the extent of spatial correlation is inferred from an a priori specified kernel estimator.

[3] The crucial distinction between these two procedures is that the Huber-Eicker-White correction is based on an estimator that assumes homoskedasticity (e.g., ordinary least squares) and only uses an ex post adjustment of the error variance-covariance matrix to correct for heteroskedasticity. See Lambert and Florax (2008) for details on how this can be implemented with SAR and SEM models. The GMM/IV estimator for the SARAR model derived in Kelejian and Prucha (2007b) is explicitly designed to accommodate a general form of heteroskedasticity. See Arraiz, Kelejian and Prucha (2007) for a lucid explanation.

As reference cases, we can distinguish the situation in which both heteroskedasticity and spatial correlation are ignored and straightforward OLS is used, and the situation in which only heteroskedasticity is allowed for using a Huber-Eicker-White estimator for the variances. Table 1 provides an overview of the different models, the processes accounted for, and the respective estimators.

Table 1. Type of processes included in spatial process models and different estimators

Model includes	Processes		Estimator[a]
	Spatial autocorrelation	Heteroskedasticity	
X	—	—	OLS
	—	×	Robust OLS
	×	×	Spatial HAC
$W\varepsilon$	×	—	GM
	×	×	GM, robust or heteroskedastic
Wy	×	—	IV
	×	×	IV, robust or heteroskedastic
Wy and $W\varepsilon$	×	—	GS2SLS
	×	×	GS2SLS, heteroskedastic

[a] "Robust" refers to Huber-Eicker-White adjustments, and "heteroskedastic" signals that the estimator allows for a general form of heteroskedasticity (see Kelejian and Prucha 2007b). As an alternative to GM, IV and GS2SLS estimation, corresponding maximum likelihood estimators are available as well. Most of these are outlined in Anselin (2006), but they are not yet used very often in practice.

Below, we provide an empirical comparison in terms of *t*-values for the spatial HAC estimator with several different bandwidth selection procedures to a non-spatial model estimated with either OLS or robust OLS, and the error model estimated with GM or robust GM. A comparison for the type of models included in the bottom half of the table would be much more involved because of the spatial multiplier process implied in these specifications (see Anselin 2006).[4]

[4] The latter is deferred to a more extensive Monte Carlo comparison. One should note that interesting hybrid versions of the approaches outlined in Table 1 are possible as well. For instance, Anselin and Lozano-Gracia (2008) employ the spatial HAC estimator in combination with a spatial lag model.

In the empirical comparison, we supply results for two heteroskedasticity-robust approaches. The first approach (HC1) is the traditional bias-corrected Huber-White robust covariance estimator, and the second (HC2) the MacKinnon and Davidson (1993) "jackknife" covariance estimator weighted by leverage values (see Lambert and Florax 2008, for more details). The variants of the spatial HAC estimator are for the Parzen kernel with plug-in bandwidths ranging from $n^{0.25}$ to $n^{0.75}$ (τ = 0.125 to 0.375) with 0.10 increments for the exponent, and the cross-validation procedure and the semivariogram approach as described above.

5 Data

Data used in the hedonic housing price example are for single-family house sales during 2001 in Knox County, Tennessee. There were 2,889 observations, after eliminating observations with missing information. Four primary GIS data sets include individual parcel data, census-block group data, boundary data, and environmental feature data. Individual parcel data (sales price, lot size, and structural information) and boundary data (high school district and jurisdiction boundaries) are from county offices. The individual parcel data are from the Knoxville, Knox County, Knoxville Utilities Board Geographic Information System (KGIS 2007) and the Knox County Tax Assessor's Office. The boundary data are from the Knoxville-Knox County Metropolitan Planning Commission (MPC 2006). Environmental feature data, including water bodies and golf courses, are from the Environmental Systems Research Institute Data and Maps 2004 (ESRI 2004). Information from census-block groups were assigned to houses located within the boundaries of the block groups.

In the empirical illustration we use a row-standardized weights matrix based on Delauney

triangulation, which on average assigns 6 neighbors to every residential property. The minimum number of neighbors is 3 and the maximum 13.

6 Results

Table 2 provides an overview of the estimation results for a standard hedonic pricing model with the logarithm of transaction prices on the left hand side, and structural characteristics, distance to or size of amenities, high school district, dummy variables for the city of Knoxville and flooding, and a season indicator on the right hand side.[5] The table allows for a comparison of *t*-values of a non-spatial model and an error model, assuming either homoskedasticity or with heteroskedasticity robust standard errors, and the spatial HAC estimator for the recommended bandwidth $n^{0.35}$ (Kelejian and Prucha 2007a). As expected, the estimated coefficients of the non-spatial and the error model are nearly identical. Not accounting for heteroskedasticity and spatial autocorrelation generally leads to inflated *t*-values. The results for the spatial HAC estimator are by and large closer to the heteroskedasticity robust OLS results than to the more conservative heteroskedasticity robust results for the spatial error model. This inference is obviously dependent on the bandwidth selection criterion, and will be investigated in more detail below.

Misspecification tests based on the OLS residuals suggest that there is clear evidence for heteroskedasticity. The Breusch-Pagan test with random coefficients as the alternative hypothesis is 1,914 (df = 38, $p < 0.01$). Spatial Lagrange Multiplier tests on the OLS residuals suggest that the spatial lag or the SARAR model are attractive alternative models.[6] However, our aim in this paper is not to select the "best" specification, but rather to illustrate the impact of

[5] For a discussion of substantive issues see Cho, Poudyal and Lambert (2008).

[6] See Anselin et al. (1996) for the standard interpretation of spatial misspecification test results based on the Lagrange Multiplier principle. The χ^2-values for the different tests are: LMERR 113.22 ($p = 0.00$), LMLAG 218.82 ($p = 0.00$), LMERR-robust 0.37 ($p = 0.54$), LMLAG-robust 105.97 ($p = 0.00$), and LMARAR 219.19 ($p = 0.00$).

Table 2. Estimation results for a hedonic model of Knox County housing sales transactions in 2001

Variable/estimator	OLS Estimate	*t*-value	HC2[b]	GM Estimate	*t*-value	HC2[b]	OLS SHAC[b]
Constant	5.932	13.040	12.082	6.079	11.741	10.709	10.769
Structural variables							
Lot size[a]	0.061	7.136	5.500	0.063	7.198	5.309	5.605
Age	–0.004	–11.036	–8.576	–0.004	–10.449	–7.939	–8.257
Brick	0.056	4.174	3.615	0.049	3.642	3.078	3.413
Pool	0.048	2.048	1.418	0.045	1.967	1.362	1.451
Garage	0.096	7.535	7.937	0.090	7.162	7.488	7.859
Bedroom	0.016	1.548	1.451	0.016	1.562	1.432	1.441
Stories	0.099	6.662	6.330	0.094	6.307	5.949	5.938
Fireplaces	0.043	3.875	3.290	0.040	3.607	2.999	3.313
Construction quality	0.181	12.841	14.339	0.177	12.301	13.760	12.911
Condition of structure	0.095	6.406	6.931	0.098	6.634	7.148	6.730
Finished area[a]	0.523	23.972	17.600	0.505	23.257	16.536	18.268
Census block-group							
Vacancy rate	0.272	1.246	1.072	0.271	1.123	0.953	0.971
Unemployment rate	–0.015	–0.072	–0.071	0.005	0.024	0.023	–0.072
Travel time to work	0.006	2.946	2.542	0.006	2.365	2.016	2.484
Household income[a]	0.167	6.189	5.361	0.165	5.500	4.670	4.836
Housing density	–0.009	–0.881	–0.906	–0.013	–1.241	–1.244	–0.816
Distance to, or size							
CBD[a]	–0.046	–1.438	–1.225	–0.037	–1.005	–0.842	–1.145
Greenway[a]	–0.029	–3.736	–3.134	–0.027	–2.875	–2.349	–3.070
Railroad[a]	0.009	1.336	1.240	0.013	1.609	1.470	1.073
Sidewalk[a]	–0.011	–2.001	–1.692	–0.014	–2.160	–1.807	–1.596
Park[a]	–0.011	–1.428	–1.440	–0.013	–1.391	–1.350	–1.374
Size park[a]	0.018	2.426	1.893	0.015	1.759	1.337	1.840
Golf courses[a]	–0.031	–2.353	–2.104	–0.032	–2.051	–1.802	–1.884
Water body[a]	–0.018	–2.298	–1.615	–0.023	–2.465	–1.645	–1.480
Size of water body[a]	0.006	2.611	2.473	0.006	2.172	2.029	2.294
High school district							
Doyle	–0.142	–4.060	–3.463	–0.148	–3.618	–2.955	–3.119
Bearden	–0.105	–3.601	–3.054	–0.103	–2.992	–2.530	–2.532
Carter	–0.173	–3.846	–3.301	–0.196	–3.708	–3.106	–3.170
Central	–0.088	–3.089	–2.927	–0.087	–2.611	–2.458	–2.627
Fulton	–0.134	–3.672	–3.871	–0.137	–3.190	–3.340	–3.455
Gibbs	–0.115	–2.979	–2.948	–0.122	–2.715	–2.658	–2.646
Halls	–0.089	–2.367	–2.468	–0.091	–2.052	–2.130	–2.186
Karns	–0.076	–2.480	–2.318	–0.083	–2.297	–2.122	–2.042
Powell	–0.073	–2.126	–2.275	–0.082	–2.009	–2.111	–1.937
Farragut	–0.167	–4.529	–3.980	–0.165	–3.774	–3.295	–3.339
Austin	–0.218	–3.372	–3.244	–0.214	–2.880	–2.677	–3.162
Other spatial dummies							
Knoxville	–0.018	–0.856	–0.738	–0.021	–0.849	–0.722	–0.632
Flood	–0.010	–0.208	–0.266	–0.006	–0.121	–0.153	–0.244
Real estate market							
Season	0.016	1.549	1.497	0.016	1.578	1.526	1.536
λ				0.215	11.539		

[a] Logarithmic transformation.

[b] Except for the columns with coefficient estimates, the reported values are *t*- or *z*-values. The SHAC estimator is implemented with the plug-in bandwidth $n^{0.35}$, which is very close to the $n^{0.33}$ recommended in Kelejian and Prucha (2007a) and Conley (1999).

data driven bandwidth selection procedures in the context of an autoregressive heteroskedastic error model. We therefore proceed by estimating the spatial autoregressive error model.

Visual inspection of the residual semivariogram, which is not based on an a priori specified weights matrix, suggests significant spatial autocorrelation in the residuals across the Knox County housing market (Figure 1). The correlation between predicted and actual values shows that the exponential function explained about 47% of the variation in the semivariogram. The range, sill, and nugget parameters are strongly significant.

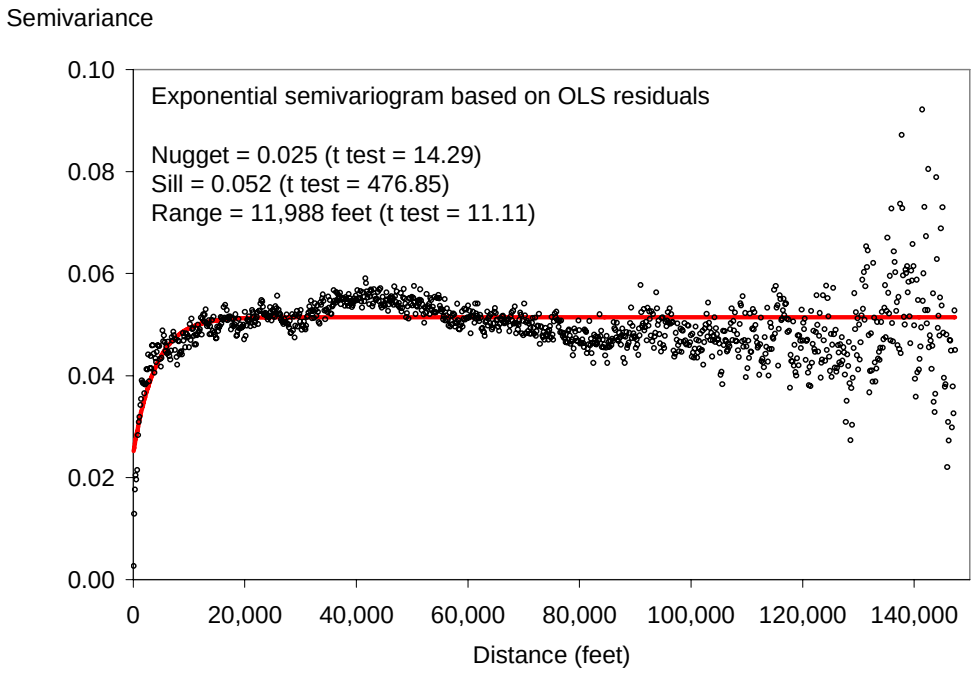

Figure 1. Semivariogram fit of OLS residuals of the hedonic price model for Knox County housing sales transactions in 2001.

The range estimated by the semivariogram was 11,988 feet (about 3.65 km), which implies an average number of neighbors of 179 ($\approx n^{0.65}$), about 6% of the observations. This distance was used as the neighborhood window in the spatial HAC estimator. The optimal bandwidth produced by the calibration procedure was 1,528 neighbors, or about 54% of the

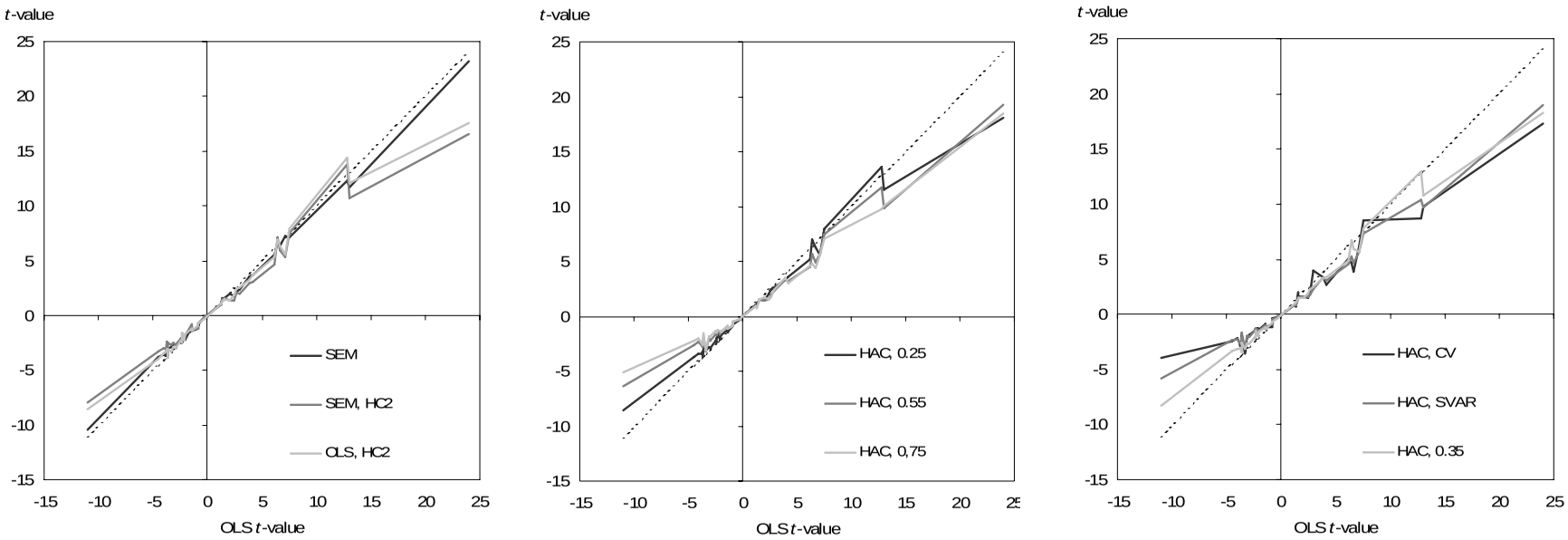

Figure 2. Spatial HAC and other (robust) estimators in comparison to OLS *t*-values.

number of housing transactions used in the regression.[7] For the series of plug-in estimators, the number of neighbors included in determination of the spatial HAC covariance estimator ranged between $n^{0.25}$ ($\tau = 0.125$, about 0.2% of the observations) and $n^{0.75}$ ($\tau = 0.375$, about 14% of the observations). Hence, we see that in this empirical example the plug-in values suggested in the literature implicitly use a substantially smaller spatial range than the data driven cross-validation and semivariogram approaches.

This obviously has implications for statistical inference. In Figure 2 we show the *t*-values for OLS on the diagonal (i.e., a one-to-one correspondence), and compare with the *t*-values for alternative estimators. The graph on the left hand side compares OLS and the error model estimated by GM, both with and without the jack-knife heteroskedasticity correction. It shows that on average the GM error result with heteroskedasticity robust standard errors leads to the most conservative inference. The graph in the middle provides a comparison with OLS for the spatial HAC estimator using different plug-in bandwidth values. By and large one can observe that a higher plug-in bandwidth value causes the absolute value of the *t*-statistics to be lower. The graph on the right hand side shows a comparison of the most conservative plug-in value for the HAC estimator with the two data driven approaches. It is obvious that in our empirical example the higher spatial ranges utilized in the data driven approaches cause the absolute *t*-values to be more conservative. It is not entirely clear how the cross-validation procedure compares to the semivariogram approach, but it seems like the cross-validation approach leads to slightly lower absolute *t*-values in the higher ranges, $|t| > 5$, whereas in the lower ranges the semivariogram approach has an edge.

[7] One should note that the value for b (= 1,528), which is the number of neighbors included around each observation, is constant across the entire dataset. Because the point pattern around each observation varies, the d_{max} differs by observation.

A more systematic comparison can be based on a simple ANOVA regression in which the absolute *t*-value is regressed on dummy variables for the various estimators, with OLS as the reference case. The results, which amount to a simple multivariate comparison of means, are summarized in Table 3, and provide additional support for the conclusions derived from the graphs in Figure 2.

Table 3. Estimation results for a comparison of absolute *t*-values for the different estimators

Diagnostics			
Multiple R	0.10	F-statistic	0.422
R^2	0.01	p-value	0.96
Standard error	3.696	Number of observations	560

ANOVA	Degrees of freedom	Sum of squares	Mean sum of squares
Regression	13	74.883	5.760
Residual	546	7456.579	13.657
Total	559	7531.462	

Regression results	Coefficient	*t*-value	*p*-value
Constant	4.249	7.272	0.00
OLS, HC1	–0.393	–0.476	0.63
OLS, HC2	–0.445	–0.538	0.59
GM	–0.300	–0.363	0.72
GM, HC1	–0.722	–0.874	0.38
GM, HC2	–0.773	–0.936	0.35
HAC, 0.25	–0.494	–0.598	0.55
HAC, 0.35	–0.661	–0.800	0.42
HAC, 0.45	–0.827	–1.001	0.32
HAC, 0.55	–1.010	–1.223	0.22
HAC, 0.65	–1.137	–1.376	0.17
HAC, 0.75	–1.246	–1.508	0.13
HAC, CV	–1.182	–1.431	0.15
HAC, SVAR	–1.147	–1.388	0.17

Table 3 shows that compared to OLS all estimators have on average lower absolute *t*-values, although the differences are so small that in conjunction with the relatively small sample size none of these differences is significantly different from zero. A small bandwidth plug-in value provides results very similar to the robust jackknifed OLS estimator, but is less conservative than the jackknifed GM estimator for the spatial error model. Obviously, the size of the absolute *t*-values for the spatial HAC estimator is negatively correlated with the magnitude of

the plug-in bandwidth criterion. In our empirical example, the spatial HAC estimator with plug-in value $n^{0.75}$ is the most conservative, and it provides results that are very similar to the results for the data driven cross-validation and semivariogram approaches. The difference between the latter two approaches is actually very small, even although the cross-validation approach is much more flexible in that it allows the bandwidth criterion to vary by location.

7 Conclusions

In this paper, we compared three methods for calibrating the recently developed spatial HAC estimator using results from a hedonic price regression for residential properties. The hedonic regression was estimated with OLS for a non-spatial model and with GM for an autoregressive error specification. We investigated the performance of two heteroskedasticity robust variance estimators, and three versions of a heteroskedasticity and autocorrelation robust covariance estimator developed by Kelejian and Prucha (2007a). The three versions of the spatial HAC estimator differed in the way the kernel bandwidth is determined. The first kernel calibration method varied a plug-in neighborhood bandwidth over a number of preset values suggested in the literature. The second approach used a data driven cross-validation procedure to determine an optimal window for the spatial HAC kernel function. The third method used the estimated distance range from an empirical semivariogram based on the OLS residuals.

Our empirical findings suggest that some of the plug-in bandwidths (specifically, the $n^{0.25}$) suggested in the literature are actually relatively small, even for moderately spatial correlated samples. In that case, the results for the spatial HAC estimator seem on average very close to the traditional jackknife estimator that corrects for heteroskedasticity. Slightly higher bandwidth values (e.g., $n^{0.35}$) produce results in terms of *t*-values that are closer to the results of a

GM estimated spatial autoregressive error model with heteroskedasticity correction. However, the data driven approaches we investigated in this paper, based on either cross-validation or a semivariogram, suggest that higher plug-in values are preferable (e.g., $n^{0.65}$ or $n^{0.75}$). This conclusion obviously depends on the spatial range implied in real-world spatial processes, but it does show that it is advisable to use an empirical, data driven approach to underpin the a priori selection of the bandwidth for the spatial HAC estimator.

The advantage of the data driven and semivariogram approaches is that they rely on the data to determine optimal neighborhood criteria for the spatial HAC. In addition, hypotheses can be tested with respect to error autocorrelation in the case of the semivariogram approach without reliance on assumptions regarding neighborhood structure, whereas conventional test statistics would require a prior designation of spatial neighborhoods through the weights matrix. Our results suggest that the cross-validation and the semivariogram approach are very similar in their outcome, with the cross-validation approach being slightly more conservative than the semivariogram approach. The advantage of the cross-validation approach is that it allows the bandwidth criterion to vary across space, but this obviously comes at the cost of the procedure being much more computationally intensive than the straightforward semivariogram approach. In general, it seems that there is only limited pay-off in terms of inference scrutiny to using the more flexible cross-validation approach.

Obviously, our results refer to a situation where $n = 1$ since we merely use an empirical example. However, the results with respect to data driven bandwidth selection are interesting and warrant further investigation in a Monte Carlo setting. Extending the experiments to a setting in which SAR, SARAR and hybrid models are also considered may be particularly rewarding.

References

Andrews, D.W.K. 1991. Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation, *Econometrica* 49, 817–58.

Anselin, L. 1988. *Spatial Econometrics: Methods and Models*. London, Kluwer Academic Publishers.

Anselin, L. 2006. Spatial Econometrics, in: T.C. Mills and K. Patterson (Eds.), *Palgrave Handbook of Econometrics: Volume 1, Econometric Theory*. Basingstoke, Palgrave Macmillan, 901–69.

Anselin, L. and R. Florax. 1995. *New Directions in Spatial Econometrics*. Berlin, Springer.

Anselin, L., and N. Lozano-Gracia, 2008. Error in Variables and Spatial Effects in Hedonic House Price Models of Ambient Air Quality, *Empirical Economics* 34(1), 5–34.

Anselin, L., A.K. Bera, R. Florax and M.J. Yoon. 1996. Simple Diagnostic Tests for Spatial Dependence, *Regional Science and Urban Economics* 26(1), 77–104.

Arraiz, I., D.M. Drukker, H.H. Kelejian and I.R. Prucha. 2008. A Spatial Cliff-Ord-type Model with Heteroskedastic Innovations: Small and Large Sample Results, *Regional Science and Urban Economics*, forthcoming.

Bannerjee, S., A.E. Gelfand and C.F. Sirmans. 2003. Directional Rates of Change under Spatial Process Models, *Journal of the American Statistical Association* 98(464), 946–54.

Cameron, A.C. and P. K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge, Cambridge University Press.

Cho, S.H., N. Poudyal and D.M. Lambert. 2008. Estimating Spatially Varying Effects of Urban Growth Boundaries on Land Development and Land Value, *Land Use Policy* 25, 320–29.

Cleveland, W.S. and S.J. Devlin. 1988. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting, *Journal of the American Statistical Association* 83, 596–610.

Cliff A.D. and J.K. Ord. 1973. *Spatial Autocorrelation*. London, Pion.

Cliff, A.D. and J.K. Ord. 1981 *Spatial Processes*. London, Pion.

Conley, T. 1999. GMM Estimation with Cross Section Dependence, *Journal of Econometrics* 92, 1–45.

Cressie, N. 1993. *Statistics for Spatial Data*. New York, Wiley Interscience.

Cressie, N. and D.M. Hawkins. 1980. Robust Estimation of the Variogram, *Mathematical Geology* 12(2), 115–125.

Dubin, R. 1992. Spatial Autocorrelation and Neighborhood Quality, *Regional Science and Urban Economics* 22, 433–52.

Fotheringham, A.S., C. Brunsdon and M. Charlton. 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. New Jersey, Wiley.

Härdle, W. and J.S. Marron. 1985. Optimal Bandwidth Selection in Nonparametric Regression Function Estimation, *Annals of Statistics* 13, 1465–81.

Hurvich, C.M., J.S. Simonoff and C.L. Tsai. 1998. Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion, *Journal of the Royal Statistical Society B* 60, 271–93.

Kelejian, H.H. and I.R. Prucha. 1999. A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model, *International Economic Review* 40(2), 509–33.

Kelejian, H.H. and I.R. Prucha. 2007a. HAC Estimation in a Spatial Framework, *Journal of Econometrics* 140(1), 131–54.

Kelejian, H.H. and I.R. Prucha. 2007b. Specification and Estimation of Spatial Autoregressive Models with Autoregressive and Heteroskedastic Disturbances, *Journal of Econometrics*, forthcoming, University of Maryland.

Lambert, D.M. and R.J.G.M. Florax. 2008. Heteroskedasticity-Robust Covariance Estimation with Spatial Autocorrelation: New Results and Monte Carlo Experiments, Selected Paper, Annual Meeting of the American Agricultural Economic Association, Orlando, Florida, July 28–30, 2008.

Lambert, D.M., C.D. Clark, M.D. Wilcox and W.M. Park. 2007. Do Migrating Retirees Affect Business Establishment and Job Growth? An Empirical Look at Southeastern Non-metropolitan Counties, 2000–2004, *Review of Regional Studies* 37(2), 251–78.

Davidson, R. and J.G. MacKinnon. 1993. *Estimation and Inference in Econometrics*. New York, Oxford University Press.

McMillen, D. P. 1996. One Hundred Fifty Years of Land Values in Chicago: A Nonparametric Approach, *Journal of Urban Economics* 40(1), 100–24.

McMillen, D.P. 2004. Employment Densities and Subcenters in Large Metropolitan Areas, *Journal of Regional Science* 44, 225–43.

Mittelhammer, R.C., G.G. Judge and D.J. Miller. 2000. *Econometric Foundations*. Cambridge, Cambridge University Press.

Newey, W.K. and K.D. West. 1987. A Simple Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix, *Econometrica* 55, 703–8.

Schabenberger, O. and F.J. Pierce. 2002. *Contemporary Statistical Models for the Plant and Soil Sciences*. New York, CRC Press.

Silverman, B.W. 1986. *Density Estimation for Statistics and Data Analysis*. London, Chapman and Hall.