Ekonomi och samhälle
Economics and Society

Skrifter utgivna vid Svenska handelshögskolan
Publications of the Hanken School of Economics

Nr 219

Linda Gerkman

Topics in Spatial Econometrics

With Applications to House Prices

Helsingfors 2010

Topics in Spatial Econometrics – With Applications to House Prices

Keywords: Spatial Econometrics, Spatial Autoregression, Weights Matrix, *k*-Nearest Neighbours, Small Scale Neighbourhood, Spatial *J*-Test, House Prices

Linda Gerkman
Svenska handelshögskolan
Institutionen för finansiell ekonomi och ekonomisk statistik
PB 479, 00101 Helsingfors, Finland

Distributör:

Biblioteket
Svenska handelshögskolan
PB 479
00101 Helsingfors, Finland

Telefon: +358-40-3521 376, +358-40-3521 265
Fax: +358-40-3521 425
E-post: publ@hanken.fi
http://www.hanken.fi

ISBN 978-952-232-104-6 (tryckt)
ISBN 978-952-232-105-3 (PDF)
ISSN 0424-7256

Edita Prima Ltd, Helsingfors 2010

ACKNOWLEDGEMENTS

I started studying for my Master's degree at Hanken in 1997, and already the first autumn I choose statistics as my major. The reason was that I felt that I was good at statistics and mathematics, and able to understand it rather easily, in comparison to so many of my fellow students who thought it difficult. But when I got further in my studies I realised that the more I learned, the broader the field of statistics got, and the more there was that I did not know. The same applies to my experience of writing this thesis in spatial econometrics. The deeper I go the more there is to learn.

I want to thank my thesis advisors Professor Gunnar Rosenqvist and Assistant Professor Niklas Ahlgren. Thank you for all the hours you have spent on reading the manuscript, and all constructive criticism you have delivered. Gunnar, I especially thank you for understanding that my family is important to me, too, and for your trust in me to let me take the time I have needed for this project. Without knowing your humane attitude towards other people I would never have begun the project in the first place. Niklas, I especially thank you for being the one to put the pressure on me when needed, without asking for the impossible. Without your encouragement I am not sure that I would ever have finished the project.

I am most grateful for all the comments and suggestions provided by my pre-examiners, Professor Roger Bivand from Norwegian School of Economics and Business Administration, and Markku Rahiala from University of Oulu. Thank you for all the hours spent on reading the manuscript. Your valuable notes have helped me improve the quality of the work during the final stage of the process.

Thank you Susanna for just being here and keeping me sane in this male dominant world, I could not have done this without you. Thank you, Johan, Andry, Paul, Olof and all other colleagues for all your encouragement during these years, whenever I have needed it.

Special thanks to my fantastic husband, Kristian, who has supported and encouraged me during the process, and to my children Jason and Jakob, who have learnt to manage a lot of things by themselves by now. I want to thank my mother, my father, all my sisters and the rest of the family for being there when I have needed you, and my mother-in-law, Kristina for always coming when we have needed baby-sitting. I also want to thank all my friends for listening to my problems, which not always must have been of great interest to you. But you did it anyway, thanks.

Financial support from Victoriastiftelsen, Bröderna Lars och Ernst Krogius forskningsfond, Waldemar von Frenkells stiftelse and the Foundation of Hanken School of Economics are gratefully acknowledged.

Helsinki, November 2010

Linda Gerkman

CONTENTS

THE ESSAYS

[1] Ahlgren, N. and Gerkman L. (2010), Unilateral Spatial Autoregression, Manuscript Hanken School of Economics.

[2] Gerkman, L. (2010), Small Scale Neighbourhood in Spatial Econometrics, Manuscript, Hanken School of Economics.

[3] Gerkman, L. (2010) A Practical Proposal to Specification Search of a *k*-Nearest Neighbours Weights Matrix, Manuscript, Hanken School of Economics.

[4] Ahlgren, N. and Gerkman L. (2010), Bootstrap Spatial *J*-Tests for *k*-Nearest Neighbours, Manuscript, Hanken School of Economics.

1 Introduction

Recent technical developments, geographical information systems (GIS) and global positioning systems (GPS) have brought about a renewed interest in spatial matters. For instance, it is today technically possible to incorporate the exact location of an observation into the data. Spatial analysis deals with these kinds of geocoded data. Haining (2003, p. 4) divides spatial analysis into cartographic modelling, mathematical modelling and spatial data analysis. Spatial data analysis is the part that includes statistical methods for analyzing the data. Standard references in the broad field of spatial data analysis are the books by Ripley (1981), Cressie (1993), and Schabenberger and Gotway (2005), among others.

In this thesis spatial econometrics is seen as a subset of spatial data analysis, i.e. the spatial aspects of the data are dealt with from an econometric perspective. We follow Anselin (2006, p. 902 and 2009, p. 3) who defines spatial econometrics as 'a subset of econometric methods that is concerned with spatial aspects present in cross-sectional and space-time observations. Variables related to location, distance and arrangement (topology) are treated in model specification, estimation diagnostic checking and prediction.' In his reflections on spatial econometrics over thirty years, Anselin (2009) sums the developments of the field by saying that 'the definition and scope of spatial econometrics has evolved substantially, moving from the "margins" of urban and regional modeling to the "mainstream" of econometric methodology'. The book by Anselin (1988) has been a central reference in the spatial econometric literature, whereas the one by LeSage and Pace (2009) presents more recent developments.

The thesis concentrates mainly on methodological issues, but the findings are illustrated by empirical studies on house price data. The thesis consists of an introductory chapter and four essays. The introductory chapter presents an overview of topics and problems in spatial econometrics that are essential for the essays. We begin by introducing spatial effects. Then spatial weights matrices and their specifications are discussed, especially, k-nearest neighbours weights matrices. We introduce various spatial econometric models, review estimation methods, and briefly touch upon the interpretation of the parameter estimates. We discuss the problem of omitted variables in spatial econometric models, and continue by some computational and empirical aspects, the bootstrap procedure and the spatial J-test (Kelejian 2008). Hedonic house price models are discussed to the extent that is relevant for

the applications. The introduction ends by overviews of the essays and a conclusion.

2 Spatial Effects

Spatial effects occur when geographical closeness of observations influences the relation between the observations. When two points on a map are close to each other, it is natural that also the observed values on a variable at those points are similar. Such points are defined as neighbours. The further away the two points are from each other, the less similar are the observations. Spatial effects do not necessarily have to be linked to geographical closeness. Instead of distance between units, we can consider some other measure of closeness. Neighbouring units can then be defined as units, which are similar or, which interact in a meaningful way. Examples of these measures of closeness or interactions are the level of education, the proportion of housing that is rental units, commuting times, commuting flows or trade flows. If there is no interaction between units they are considered spatially independent. This thesis concentrates on geographical closeness.

One can distinguish between two spatial effects, namely spatial dependence and spatial heterogeneity (Anselin 1988). Spatial dependence, or spatial autocorrelation, can be seen as dependence between observations on the dependent variable in a regression model. There are similarities between spatial dependence and autocorrelation in time series, which are explored in Essay 1. Observations on a time series can be seen as points on a line. If the series is causal, points further in time depends only on the previous ones. But observations in spatial data have a location in a two-dimensional plane. As a consequence, observations do not usually have a specific order, and the dependence between the observations may extend in all directions.

Spatial heterogeneity implies that the studied phenomenon, or relation lacks stability over the spatial plane. This means that the functions themselves, as well as their parameters, may vary depending on the location of the units in the spatial plane. In this thesis we focuses on spatial dependence.

3 Spatial Weights Matrices

In spatial econometrics the weights matrix has an important role. It contains the assumed spatial structure of the variables in the model. Usually the weights matrix is assumed to be exogenous to the model, and ideally, the structure of the weights matrix is based on relevant theory rather than on spatial patterns found in the data. Typically it is based on geographic arrangements of the observations.

In general, a weights matrix is given by

$$\mathbf{W} = \begin{pmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nn} \end{pmatrix}, \tag{1}$$

which is an $n \times n$ matrix, where n denotes the number of observations. The weights matrix specifies which of the other units in the system that affect the observed value at some particular location, i.e. which units are considered neighbours. It can also indicate how important each neighbour is to a particular unit. Suppose that unit i of the dependent variable only has neighbours j, k and m. Then in the ith row of $\mathbf{W}$ there will be only three non-zero elements, namely w_{ij}, w_{ik} and w_{im}. The simplest form of a weights matrix is a binary weight matrix. In this type of matrix w_{ij}, w_{ik} and w_{im} are all set to one if j, k and m are considered neighbours to i. All other elements in the ith row of $\mathbf{W}$ are zero. The diagonal of a weights matrix is always zero since a unit cannot depend on itself. The definition of neighbours can for instance be based on geographical distances or other differences between units, or on contiguity.

If the units in the data have a physical area, two units that share a border can be defined as neighbours. Then there are several definitions of contiguity. If the units are regularly spaced on a lattice, contiguity can alternatively be defined as follows (Anselin 1988, p. 18):

Linear contiguity: Define $w_{ij} = 1$ only if unit i has a common border immediately to the right or to the left with unit j.

Rook contiguity: Define $w_{ij} = 1$ only if unit i has a common border with unit j. The border can be up, down, to the left or to the right.

Bishop contiguity: Define $w_{ij} = 1$ only if unit i has a common vertex with unit j. The vertex can be to the north-east, south-east, south-west or north-west of unit j.

Queen contiguity: Define $w_{ij} = 1$ only if unit i has a common border or a common vertex with unit j.

When the neighbour relation is bilateral, a binary weights matrix that is based on contiguity will be symmetric. If unit i is the neighbour of j, then j is also the neighbour of i. If the dependence between units is restricted to exist in only one direction, the weights matrix is said to be unilateral.

Cliff and Ord (1973) suggested that the elements, w_{ij}, in $\mathbf{W}$ should be based on the Euclidian distance, d_{ij}, between the two spatial units and the proportion, β_{ij}, of the border of unit i that is common with unit j, formally

$$w_{ij} = [d_{ij}]^{-a} \times [\beta_{ij}]^{b}.$$

The parameters a and b are assumed to be greater than zero. This weights matrix can also be applied when the spatial units do not lay on a regular grid. But then, since the spatial units do not have exactly equal physical form, the proportion β will vary and therefore the weights matrix will not be symmetric.

If the spatial units are points on a map and therefore have no physical area, a weights matrix based on the above defined types of contiguity, or some form of the Cliff-Ord weights matrix is not applicable. An alternative is then to specify contiguity by a Voronoi diagram. Around each location of a spatial unit on the map a Voronoi polygon is drawn in such a way that inside each Voronoi polygon all points are closer to the particular spatial unit than to any other spatial unit on the map. The points of the spatial units in the contiguous Voronoi polygons are connected by segments that form the legs of Delauny triangles. The Delauny triangles thereby form a way to specify contiguity. If the spatial units are randomly distributed on a plane, a common contiguity weights matrix will have an average of approximately 6 neighbours for each spatial unit. For more on Voronoi polygons, also known as Dirichlet cells, and on Delauny triangularization see for instance Ripley (1981, section 4.3), or LeSage and Pace (2009, p. 118). In his Spatial Econometrics Toolbox for Matlab, LeSage provides code for forming contiguity weights matrices.

When the exact location of the units are known, the definition of neighbours can also be based on various functions of geographical distances between locations. The location in the plane can be defined by the Cartesian coordinates in longitudes and latitudes. The coordinates enable the calculation of the Euclidian distances between units. If the spatial dependence is assumed to be positive, it means that the dependence will diminish when

the distance grows. For instance, Pace and Gilley (1997) define their weights matrix as follows:

$$w_{ij} = \max[1 - (d_{ij}/d_{max}), 0].$$

The distance, d_{ij}, is the Euclidian distance between units i and j. The distance, d_{max}, is a predetermined maximum. By definition, $w_{ij} = 0$ if the distance between units i and j is greater than this maximum. Another common solution, for instance used by Wilhelmsson (2002), is

$$wij = 1/d_{ij}^2,$$

where the d_{ij}, is the Euclidian distance between units i and j.

In this thesis we focus on weights matrices based on nearest neighbours. In this type of weights matrices every unit i is assigned a given number, say k, of nearest neighbours according to the shortest distance in space. The resulting weights matrix is usually not symmetric, since even if the k nearest neighbours to unit i includes, say, unit j, the k nearest neighbours to unit j might not include unit i. Equal weights are given to all k neighbours, and the weights matrix is row-normalized. Therefore every non-zero element of the weights matrix will be equal to $1/k$. The spatial weights matrix can be seen as a spatial lag operator, and if $\mathbf{W}$ is a k-nearest neighbours weights matrix a spatial lag becomes a vector of averages of neighbouring observations. In this thesis we refer to a weights matrix that is based on a lower number of nearest neighbours than the weights matrix in the data generating process, as an underspecified weights matrix. An overspecified weights matrix has more neighbours than the weights matrix of the data generating process. A code for forming k-nearest neighbours weights matrices is provided by LeSage in his Spatial Econometrics Toolbox for Matlab.

An important practical issue is how to choose the weights matrix. According to Anselin (1988), the weights matrix is assumed to be exogenous and should be based on theoretical assumptions regarding the nature of the dependence structure. One should not create a purpose-specific, detailed description of spatial patterns in the actual data. However, in practice we may lack sufficient theory to support the choice a particular weights matrix. Relevant theory can also be subjectively interpreted and might therefore result in several weights matrices, with different structure, but equally realistic. Hence, there is not an unambiguous solution to this problem.

Models which are otherwise the same, but entertain different weights matrices, are non-nested (Kelejian 2008). LeSage and Pace (2009, page 162)

note, that it is not in general possible to use formal tests for significant differences between the log likelihood function values for models entertaining different weights matrices. The number of parameters are fixed and equal in the models. For instance, the usual specification search based on information criteria is then not available. An option is Bayesian model comparison (see LeSage and Pace, 2009, Section 6.3), where models entertaining different spatial weights matrices can be compared through log marginal likelihoods and associated model probabilities. An other alternative is the spatial J-test, suggested by Kelejian (2008). The spatial J-test can be applied in order to test a given spatial model against one or more non-nested alternative models. The test is discussed in Section 8, and it is applied and its properties are further studied in Essays 3 and 4.

In this thesis a lot of attention is given to k-nearest neighbours weights matrices, which appear through all essays. In Essay 1 the impact of under- and overspecified weights matrices on the spatial autoregressive parameter is examined. Essay 2 is an application, where the focus is on omitted variables, but the weights matrix that is applied is a k-nearest neighbours weights matrix with $k = 4$. The objective of Essay 3 is to suggest and examine two approaches for finding the number k for a general spatial econometric model. As the means of the specification search, we use the spatial J-test proposed by Kelejian (2008). Essay 4 continues on the subject by examining the properties of the asymptotic and a bootstrap spatial J-test when they are used for discriminating between different k-nearest neighbours weights matrices.

4 Spatial Econometric Models

In spatial econometric models the information about the dependence between the observations is incorporated into the weights matrix. When models are specified the weights matrices can be used to create spatial lags of the dependent variables, the explanatory variables or the error terms (Cliff and Ord, 1973, 1981, Ord, 1975). We can also specify models which are a combination of these lagged variables or error terms, resulting in a rich variety of models.

Assume that there exists spatial dependence between the observations on a random variable $\mathbf{y}$. When the sample size is n, some of the variation in the $n \times 1$ vector $\mathbf{y}$ is then explained by the neighboring observations to each

observation, for instance,

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \boldsymbol{\varepsilon}. \tag{2}$$

This is the first-order spatial autoregressive (FAR) model, and it has the solution

$$\mathbf{y} = (\mathbf{I}_n - \rho\mathbf{W})^{-1}\boldsymbol{\varepsilon}, \tag{3}$$

provided that $(\mathbf{I}_n - \rho\mathbf{W})$ is not singular. In the FAR model ρ is a spatial autoregressive parameter, and $\mathbf{y}$ is expressed as deviations from its mean, since no constant is included in the model. The error term $\boldsymbol{\varepsilon}$ is a vector of $IID(0, \sigma_\varepsilon^2)$. The FAR model, which lacks exogenous explanatory variables, is the model of interest in the study of Essay 1.

The model can be estimated by maximum likelihood (ML), given the following assumptions,

$$\begin{aligned} \boldsymbol{\varepsilon} &\sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n), \\ \text{all diagonal elements of } \mathbf{W} &= 0, \\ \text{all row sums of } \mathbf{W} &= 1, \\ \text{and the parameter space is } |\rho| &< 1. \end{aligned} \tag{4}$$

According to assumptions (4), the weights matrix is row-normalized when ML estimation procedures are used. This ensures that $(\mathbf{I} - \rho\mathbf{W})$ is non-singular so that $(\mathbf{I} - \rho\mathbf{W})^{-1}$exists for all $|\rho| < 1$. At $\rho = 1$, $(\mathbf{I} - \rho\mathbf{W})$ will be singular. Within the restricted parameter space, $|\rho| < 1$, the model is then solvable. According to Kelejian and Prucha (2010), problems will occur if the weights matrix is not row-normalized, since the parameter space will not be continuos. The matrix $(\mathbf{I} - \rho\mathbf{W})$ will be singular for certain values of $|\rho| < 1$. All the weights matrices considered in this thesis are row normalized.

The following general spatial econometric model,

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{W}_1\mathbf{y} + \mathbf{u}, \\ \mathbf{u} &= \lambda\mathbf{W}_2\mathbf{u} + \boldsymbol{\varepsilon}, \end{aligned} \tag{5}$$

is the model we focus on in Essay 3 and in Essay 4. In this model the dependent variable vector $\mathbf{y}$ is explained by the exogenous variables in the matrix $\mathbf{X}$ and a spatial lag $\mathbf{W}_1\mathbf{y}$. The error term $\mathbf{u}$ is spatially autocorrelated, and $\boldsymbol{\varepsilon}$ is a vector of $IID(0, \sigma_\varepsilon^2)$ error terms. Here $\boldsymbol{\beta}$ is a vector of regression coefficients and ρ and λ are the scalar spatial autoregressive parameters. The

structure of spatial dependence in the data is again formulated in the weights matrices $\mathbf{W}_1$ and $\mathbf{W}_2$. The solution of the model is

$$\mathbf{y} = (\mathbf{I} - \rho\mathbf{W}_1)^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \rho\mathbf{W}_1)^{-1}(\mathbf{I} - \lambda\mathbf{W}_2)^{-1}\boldsymbol{\varepsilon}, \tag{6}$$

provided that $(\mathbf{I} - \rho\mathbf{W}_1)$ and $(\mathbf{I} - \lambda\mathbf{W}_2)$ are non singular, which is the case when $|\rho| < 1$ and $|\lambda| < 1$ and the weights matrices, $\mathbf{W}_1$ and $\mathbf{W}_2$, are row-normalized. In practice it is often assumed that $\mathbf{W}_1 = \mathbf{W}_2 = \mathbf{W}$.

When $\rho = 0$ and $\lambda = 0$ model (5) is a traditional linear regression model, which is the starting point of the empirical analysis in Essay 2. In that essay we also estimate model (5) under the assumption that only $\rho = 0$, i.e. a spatial error model. If $\lambda = 0$ in (5), the model is referred to as a spatial lag model.

Another model considered in Essay 2 is the spatial Durbin model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{W}\mathbf{y} + \mathbf{W}\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}. \tag{7}$$

In this model $\mathbf{y}$ is explained by spatially lagged explanatory variables $\mathbf{WX}$, in addition to the variables in the matrix $\mathbf{X}$ and a spatial lag $\mathbf{Wy}$. Through the spatial lag $\mathbf{WX}$, an observation on the dependent variable is dependent on a weighted average of the observations on each explanatory variable of its neighbours.

5 Estimation Methods and Inference

When estimating spatial econometric models like (2), (5) and (7), the method of ordinary least squares is not an option. Although it gives unbiased and consistent estimates of the spatial error model, (i.e. model (5), when $\rho = 0$) the estimates of standard errors will be biased (Anselin 1988). When the model includes a spatially lagged dependent variable, as in the general spatial model, spatial lag model or spatial Durbin model, the least squares estimates of the parameters will be biased and inconsistent. According to Anselin (2009), there are two main approaches to estimate spatial econometric models, maximum likelihood (ML) and instrumental variables/general method of moments (GMM). For the ML estimation approach for the Cliff-Ord type of spatial econometric model, see Ord (1975), or Anselin (1988). The conditions that ensure consistency and asymptotic normality of the ML estimator for the spatial autoregressive model were given by Lee (2002, 2004). The

GMM estimator was suggested by Kelejian and Prucha (1998, 1999). They also demonstrated its consistency and asymptotic properties. Lee (2003) and Kelejian et al. (2004) further developed the analysis by including the use of ideal instruments.

The ML approach gives consistent estimates also when the models include spatially lagged variables, but it may be numerically challenging. In order to formulate the likelihood function an assumption about the distribution of the errors has to be made. The errors are usually assumed to be normal. The assumption of normality may not always be fulfilled. The GMM approach, on the other hand, gives consistent estimates without assuming that the errors have any particular distribution, except that they are independent and identically distributed. This may be a benefit in some cases. The GMM is also computationally easier. However, Pace et al. (2010) found that the performance of the instrumental variable techniques, is affected, when estimating a spatial lag model or a spatial Durbin model in the presence of spatially autocorrelated regressors. This may be a drawback in some applications.

Another alternative for estimating spatial econometric models are Bayesian methods. They are presented in LeSage (1997) and LeSage and Pace (2009). Anselin (2009) notes that the Bayesian approach has not been widely used in spatial econometric applications, with the exception of the work by LeSage.

In Essay 1 we estimate the model (2) by ML. Since the data are assumed to be generated by a spatial unilateral process with normal errors, it is straight forward to find the likelihood function. In Essay 2 one of the estimated models is a spatial Durbin model and there are spatially autocorrelated regressors present. We therefore choose the ML approach over GMM. In Essay 3 the objective is to find a strategy for specifying the number k for a general spatial econometric model by using the spatial J-test of Kelejian (2008). Since Kelejian developed the test in the IV/GMM framework, it is natural that we continue by using the same estimation method. This argument applies also to Essay 4.

Because of the complicated dependence structure between observations in a spatial econometric model the parameter estimates also contain information about the relationships between observations. A change in the value of a single observation on a certain explanatory variable may therefore affect the value of the outcome for the observation itself (direct impact), in addition to the outcomes for all other observations in the system (indirect impact). Traditionally in linear regression models the parameters are interpreted as partial derivatives of the dependent variable with respect to the explanatory

variable. When the model contains spatial lags of the dependent variable or the explanatory variables, as in model (5), $\rho \neq 0$, or model (7), the interpretation becomes richer and more complicated, since the partial derivatives of the dependent variable y_i, $i = 1...n$, with respect to an explanatory variable r and an other observation j, $\partial y_i / \partial x_{jr}$ is not zero as it is in the traditional case. We note that the parameter estimates of the spatial error model, i.e. model (5), $\rho = 0$, can be interpreted as in linear regression models, since this model does not contain a spatial lag on the dependent variable nor on the explanatory variable. Because of the different interpretations it is not meaningful to compare parameter estimates from a linear regression model or a spatial error model with estimates of a spatial lag model or a spatial Durbin model. The issue of interpretation of the parameter estimates of a spatial econometric model is briefly discussed in Essay 2. For a thorough presentation of the direct and indirect effects and how to calculate summary measures of the impacts, see LeSage and Pace (2009). It is worth noting that the applied literature contains a number of studies that misinterpret the parameter estimates as pointed out by LeSage and Pace (2009).

Code for estimating the models above is provided by LeSage in his Spatial Econometric Toolbox for Matlab. The routines include estimation by ML and by IV/GMM as well as by Bayesian methods. The routines are applied in the essays.

6 An Omitted Variables Argument

When modelling for instance house price data the model is very often lacking some location and neighbourhood variables. According to LeSage and Pace (2009), there is then a strong motivation for spatial econometric models, since these omitted location and neighbourhood variables are usually thought of as spatially autocorrelated. The following is based on arguments adapted from Pace and LeSage (2008) and LeSage and Pace (2009).

We consider a scenario where the assumptions are that there are important explanatory variables missing, the omitted variables are spatially autocorrelated and, in addition, they are correlated with the explanatory variables in the model. Below it is shown that these assumptions lead to a spatial Durbin model. Then we introduce some of the omitted variables. In this second scenario, we still assume that there are variables missing from the model despite the included uncovered new variables, and they are spatially

autocorrelated, but now we assume that the remaining omitted variables are uncorrelated with the explanatory variables in the model. These assumptions will lead to a spatial error model. In Essay 2 the results of the theoretical analysis below are used on real data to demonstrate the usefulness of the arguments. The objective of Essay 2 is to show that by uncovering some new spatially autocorrelated explanatory variables, we can obtain a simpler model, which is easier to interpret.

Begin by considering the following non-spatial model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_1 + \mathbf{Z}\boldsymbol{\beta}_2, \tag{8}$$

where the $n \times 1$ vector $\mathbf{y}$ is observations on the dependent variable, the $n \times k$ matrix $\mathbf{X}$ represents the available explanatory variables, the $n \times m$ matrix $\mathbf{Z}$ represents missing or unobserved explanatory variables, and $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are vectors containing regression coefficients. For ease of exposition there is no error term in the model, instead the variables in $\mathbf{X}$ and $\mathbf{Z}$ are assumed to completely explain the dependent variable. Further assume that the missing variables in the matrix $\mathbf{Z}$ are spatially autocorrelated , i.e. that we have

$$\mathbf{Z} = \rho\mathbf{W}\mathbf{Z} + \mathbf{R}, \quad E(\mathbf{R}) = \mathbf{0}. \tag{9}$$

The scalar parameter ρ is the spatial autocorrelation coefficient and $\mathbf{W}$ is a weights matrix. Solve $\mathbf{Z} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{R}$, and substitute into (8) to obtain

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_1 + (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{R}\boldsymbol{\beta}_2. \tag{10}$$

Define the vector of errors as $\mathbf{R}\boldsymbol{\beta}_2 = \mathbf{u}$. Further assume that the missing variables in $\mathbf{Z}$ correlate with the explanatory variables $\mathbf{X}$ in model (8). Assume a simple linear dependence,

$$\mathbf{u} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{v}, \quad \mathbf{v} \sim N(0, \sigma_v^2\mathbf{I}). \tag{11}$$

The parameters in the vector $\boldsymbol{\gamma}$ and the variance of the error term σ_v^2 describe the relation between $\mathbf{X}$ and $\mathbf{Z}$. Since $\mathbf{R}\boldsymbol{\beta}_2 = \mathbf{u}$, we can substitute (11) into (10) and obtain

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta}_1 + (\mathbf{I} - \rho\mathbf{W})^{-1}(\mathbf{X}\boldsymbol{\gamma} + \mathbf{v}) \\ &= \mathbf{X}\boldsymbol{\beta}_1 + (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\gamma} + (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{v}. \end{aligned} \tag{12}$$

Multiply from the left by $(\mathbf{I} - \rho\mathbf{W})$,

$$(\mathbf{I} - \rho\mathbf{W})\mathbf{y} = (\mathbf{I} - \rho\mathbf{W})\mathbf{X}\boldsymbol{\beta}_1 + \mathbf{X}\boldsymbol{\gamma} + \mathbf{v}, \tag{13}$$

and we get the spatial Durbin model (SDM),

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}(\boldsymbol{\beta}_1 + \boldsymbol{\gamma}) + \mathbf{W}\mathbf{X}(-\rho\boldsymbol{\beta}_1) + \mathbf{v}. \tag{14}$$

Now consider the scenario where we assume that some of the variables, say $\mathbf{Z}_1$, of the matrix $\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}^*]$ in (8) are revealed, and we get a new matrix $\mathbf{X}^* = [\mathbf{X}, \mathbf{Z}_1]$ of explanatory variables. Instead of model (8) we have

$$\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta}_1^* + \mathbf{Z}^*\boldsymbol{\beta}_2^*. \tag{15}$$

Assume that the remaining missing variables $\mathbf{Z}^*$ are spatially autocorrelated, as before, i.e. that we have

$$\mathbf{Z}^* = \rho^*\mathbf{W}\mathbf{Z}^* + \mathbf{R}^*, \quad E(\mathbf{R}^*) = \mathbf{0}, \tag{16}$$

but that the missing variables $\mathbf{Z}^*$ are now uncorrelated with the explanatory variables $\mathbf{X}^*$ in the model. Solve $\mathbf{Z}^* = (\mathbf{I} - \rho^*\mathbf{W})^{-1}\mathbf{R}^*$ and substitute into (15) to get

$$\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta}_1^* + (\mathbf{I} - \rho^*\mathbf{W})^{-1}\mathbf{R}^*\boldsymbol{\beta}_2^*. \tag{17}$$

Define $\mathbf{u}^* = \mathbf{R}^*\boldsymbol{\beta}_2^*$ and $\mathbf{q} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{u}^*$. Then (17) can be written as a spatial error model (SEM),

$$\begin{aligned} \mathbf{y} &= \mathbf{X}^*\boldsymbol{\beta}_1^* + \mathbf{q} \\ \mathbf{q} &= \rho^*\mathbf{W}\mathbf{q} + \mathbf{u}^*. \end{aligned}$$

7 A Few Computational Aspects

Applications in spatial econometrics sometimes involve large data sets, entailing that the spatial weights matrix is of a large dimension. Fortunately the weights matrix is typically sparse, i.e. the proportion of non-zero elements is small. For instance, the programming language Matlab offers an option to store sparse matrices in a way that decreases the use of memory substantially. In addition, it takes much less time to compute spatial lags when the weights matrix is sparse, as well as the log-determinant, inverse, and other quantities, which are needed for maximum likelihood estimation of the parameters and inference in the spatial econometric model. The unilateral weights matrices used in the simulations in Essay 1 are created using the sparse routine in Matlab.

Three of the essays in the thesis include simulation studies. In Essay 1 we examine the finite sample properties of the estimator of the spatial autoregressive parameter in a unilateral spatial autoregressive model when data are generated by an unilateral spatial autoregressive process. In Essay 3 we simulate rejection probabilities of the spatial J-test for different amounts of spatial autocorrelation in the data and for different k-nearest neighbours. We continue on the subject in Essay 4 by studying the properties of the asymptotic and a bootstrapped version of the test for different sample sizes. All simulations are executed in Matlab and routines from the Spatial Econometric Toolbox (versions 2005 and 2010) by LeSage are used in the estimations. In all three studies the function randn in Matlab, which uses Margasalia's ziggurat algorithm, is used to generate pseudo-random numbers. The spatial J-test, which is studied in Essay 3 and Essay 4, is programmed in Matlab following Kelejian (2008).

8 The Spatial J-test

The original J-test, which was introduced by Davidson and MacKinnon (1981), is a test for non-nested hypothesis. This means that neither of the null model and the alternative model is a special case of the other. The main idea of the J-test is to estimate a model which consists of the null model and the predictive value of the alternative model. Then the significance of the additional term is tested.

In Essay 3 we suggest and examine strategies for specification search for the number of neighbours in a k-nearest neighbours weights matrix of a general spatial econometric model. Since spatial models, which are otherwise equal but have different weights matrices, are non-nested we use a spatial J-test proposed by Kelejian (2008) as means of the specification search. This section briefly introduces the test. The test and its properties are described in detail in the appendix of Kelejian.

Following Kelejian, the spatial model of the null hypothesis is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{W}_1\mathbf{y} + \mathbf{u} = \mathbf{Z}\boldsymbol{\gamma} + \mathbf{u}, \tag{18}$$

$$\mathbf{u} = \lambda\mathbf{W}_2\mathbf{u} + \boldsymbol{\varepsilon}, \tag{19}$$

where $\mathbf{Z} = (\mathbf{X}, \mathbf{W}_1\mathbf{y})$, $\boldsymbol{\gamma} = (\boldsymbol{\beta}', \rho)'$. The vectors $\mathbf{y}, \mathbf{u}$ and $\boldsymbol{\varepsilon}$, the parameter vector $\boldsymbol{\beta}$, the matrices $\mathbf{X}, \mathbf{W}_1$ and $\mathbf{W}_2$ and the scalar parameters ρ and λ are as in Section 4. The sample size is n.

The alternative model ($G = 1$) or models ($G > 1$) are of the form:

$$\begin{aligned} \mathbf{y} &= \mathbf{X_i}\boldsymbol{\beta}_i + \rho_i \mathbf{W}_{1i}\mathbf{y} + \mathbf{u}_i = \mathbf{Z}_i\boldsymbol{\gamma}_i + \mathbf{u}_i, \\ \mathbf{u}_i &= \lambda_i \mathbf{W}_{2i}\mathbf{u}_i + \boldsymbol{\varepsilon}_i, \qquad\qquad i = 1, ..., G, \end{aligned} \tag{20}$$

where G is a finite constant equal to the number of possible alternatives to the null model.

Solve (19) and get $\mathbf{u} = (\mathbf{I} - \lambda\mathbf{W}_2)^{-1}\boldsymbol{\varepsilon}$. Insert the solution into (18) and pre-multiply by $(\mathbf{I} - \lambda\mathbf{W}_2)$ to get:

$$(\mathbf{I} - \lambda\mathbf{W}_2)\mathbf{y} = (\mathbf{I} - \lambda\mathbf{W}_2)\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}.$$

Define $\mathbf{y}(\lambda) = (\mathbf{I} - \lambda\mathbf{W}_2)\mathbf{y}$ and $\mathbf{Z}(\lambda) = (\mathbf{I} - \lambda\mathbf{W}_2)\mathbf{Z}$, and get

$$\mathbf{y}(\lambda) = \mathbf{Z}(\lambda)\boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \tag{21}$$

Likewise for the alternative model (20), define $\mathbf{Z}_i(\lambda_i) = (\mathbf{I} - \lambda_i\mathbf{W}_{2i})\mathbf{Z}_i$. Then for all $i = 1, ..., G$, estimate $\boldsymbol{\gamma}_i$ in the ith alternative model (20) by generalized spatial two stage least squares (Kelejian and Prucha 1998, 1999), and get $\widehat{\boldsymbol{\gamma}}_i$. Add the predictive power $\mathbf{Z}_i(\lambda_i)\widehat{\boldsymbol{\gamma}}$ of all $i = 1, ..., G$ alternative models to the null model (21) to get:

$$\mathbf{y}(\lambda) = \mathbf{Z}(\lambda)\boldsymbol{\gamma} + \sum_{i=1}^{G} \alpha_i[\mathbf{Z}_i(\lambda_i)\widehat{\boldsymbol{\gamma}}_i] + \boldsymbol{\varepsilon}. \tag{22}$$

The parameter α_i is a scalar. Given that the null model is true, $\alpha_i = 0$ for all $i = 1, ..., G$.

Insert $\mathbf{Z}_i(\lambda_i) = (\mathbf{I} - \lambda_i\mathbf{W}_{2i})\mathbf{Z}_i$ into (22) and define $\phi_i = -\alpha_i\lambda_i$ to get:

$$\mathbf{y}(\lambda) = \mathbf{Z}(\lambda)\boldsymbol{\gamma} + \sum_{i=1}^{G} \alpha_i[\mathbf{Z}_i\widehat{\boldsymbol{\gamma}}_i] + \sum_{i=1}^{G} \phi_i[\mathbf{W}_{2,i}Z_i\widehat{\boldsymbol{\gamma}}_i] + \boldsymbol{\varepsilon}.$$

Let $\boldsymbol{\delta} = (\alpha_1, ..., \alpha_G, \phi_1, ..., \phi_G)'$. Then the test of the null model (18) against the alternative models (20) is simply a Wald test of $\boldsymbol{\delta} = \mathbf{0}$. At the a level of significance the null hypothesis will be rejected if

$$J = \widehat{\boldsymbol{\delta}}'\widehat{\mathbf{V}}_{\widehat{\boldsymbol{\delta}}}^{-1}\widehat{\boldsymbol{\delta}} > \chi^2_{1-a}(2G),$$

where $\widehat{\mathbf{V}}_{\widehat{\boldsymbol{\delta}}}$ is the estimated small-sample variance-covariance matrix of $\widehat{\boldsymbol{\delta}}$.

9 The Bootstrap

In Essay 4 a bootstrap is applied to the spatial J-test, since the asymptotic J-test may suffer from size distortion in small samples (Piras and Lozano-Garcia 2008; Burridge and Fingleton 2009).

Under the assumption that the model is a general spatial autoregressive model, we use the following parametric bootstrap algorithm to perform a bootstrap spatial J-test on a sample of data. Bootstrap quantities are denoted by a '$*$'. For generating samples of bootstrap observations $\mathbf{y}^*$:

Algorithm 1 (Bootstrap J-test)

1. *Estimate the spatial autoregressive model by the generalized method of moments (GMM) procedure suggested in Kelejian and Prucha (1999) to obtain the estimates* $\widehat{\rho}$, $\widehat{\lambda}$ *and* $\widehat{\boldsymbol{\beta}}$.

2. *Check whether* $|\widehat{\rho}| < 1$ *and* $|\widehat{\lambda}| < 1$.

3. *If both conditions are satisfied, generate bootstrap observations* $\mathbf{y}^*$ *from*

$$\begin{aligned} \mathbf{y}^* &= \mathbf{X}\widehat{\boldsymbol{\beta}} + \widehat{\rho}\mathbf{W}\mathbf{y}^* + \mathbf{u}^*, \\ \mathbf{u}^* &= \widehat{\lambda}\mathbf{W}\mathbf{u}^* + \boldsymbol{\varepsilon}^*, \end{aligned}$$

using the reduced form

$$\mathbf{y}^* = (\mathbf{I} - \widehat{\rho}\mathbf{W})^{-1}\mathbf{X}\widehat{\boldsymbol{\beta}} + (\mathbf{I} - \widehat{\rho}\mathbf{W})^{-1}(\mathbf{I} - \widehat{\lambda}\mathbf{W})^{-1}\boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}^* \sim N_n(0, \mathbf{I})$.

We repeat the algorithm B times and get bootstrap samples $\mathbf{y}_j^*$, where $j = 1, ..., B$. We use the same weights matrix to generate the bootstrap samples, as when we estimate the model to obtain the estimates $\widehat{\rho}$, $\widehat{\lambda}$ and $\widehat{\boldsymbol{\beta}}$. Therefore the bootstrap sample size is the same as the number of observations in the original data. The bootstrap samples are then used to compute B bootstrap J-statistics J_j^*. The empirical distribution of the J_j^* is used to approximate the distribution of J under the null hypothesis. The bootstrap critical value at the significance level α is given by the $1 - \alpha$ quantile of the J_j^*.

Let $\widehat{J}$ denote the realized value of the J-statistic computed on the data. For a test at significance level α we reject the null hypothesis if $\widehat{J}$ is larger than the bootstrap critical value.

In many cases in practice it is more convenient to translate the test statistics into p-values. The bootstrap p-value is defined as

$$p^* = \frac{1}{B}\sum_{j=1}^{B} I(J_j^* > \widehat{J}), \tag{23}$$

i.e. the fraction of the bootstrap samples for which J_j^* is larger than $\widehat{J}$. For a test at significance level α we reject the null hypothesis if $p^* < \alpha$.

In Essay 4 the focus is on the size and power of the bootstrap J-test. The size and power can be estimated in simulation experiments (see Horowitz 1994; Davidson and MacKinnon 2006) in the following way. We begin by generating M samples $\mathbf{y}_m$, indexed by $m = 1, \ldots, M$. For each replication m, we compute the J-statistic $\widehat{J}_m$ and generate B bootstrap J-statistics J_{mj}^*, indexed by $j = 1, \ldots, B$. Then we compute $J_m^*(1-\alpha)$, which is the $1-\alpha$ quantile of the J_{mj}^*. The rejection probability of the bootstrap test is then estimated by

$$\frac{1}{M}\sum_{m=1}^{M} I(\widehat{J}_m > J_m^*(1-\alpha)), \tag{24}$$

i.e. the fraction of the M replications for which the value of the J statistic $\widehat{J}_m$ is larger than the bootstrap critical value $J_m^*(1-\alpha)$.

The procedure involves computing $M(B+1)$ J-statistics. The computational time for estimating the rejection probabilities becomes prohibitive when M and B are large. Commonly used values of M and B are $M = 10000$ and $B = 1000$, resulting in computing approximately 10 million J-statistics. For the values of M and B above the computational time for $n = 100$ is about 70 hours CPU time on a Core 2 Duo CPU, 2.4 GHz, 2.00 GB memory machine.

In Essay 4 we apply a procedure for estimating the size and power of a bootstrap test proposed by Davidson and MacKinnon (2006). This approach substantially reduces the computational burden. For each replication $m = 1, \ldots, M$, we compute the J-statistic $\widehat{J}_m$ as before, but we generate only one bootstrap J-statistic J_m^*. We then compute the bootstrap critical value $J^*(1-\alpha)$, which is the $1-\alpha$ quantile of the J_m^*. The approximate rejection probability of the bootstrap test is then estimated by

$$\frac{1}{M}\sum_{m=1}^{M} I(\widehat{J}_m > J^*(1-\alpha)). \tag{25}$$

See Davidson and MacKinnon (2006) for more details.

Both procedures estimate the nominal power of the bootstrap test, i.e. the power of the bootstrap test at the nominal level α. The difference between (24) and (25) is in the estimation of the $1-\alpha$ quantile of the bootstrap replications. In (24) the $\widehat{J}_m$ are compared to different estimated critical values, whereas in (25) the $\widehat{J}_m$ are compared to the same estimated critical value.

The amount of computation required to estimate (25) is $2M$ J-statistics, which is much smaller than $M(B+1)$. Davidson and MacKinnon (2006) establish the validity of (25) under the assumption of independence between the bootstrap data generating process (DGP) and the test statistic. For the J-test, the bootstrap DGP and the J-statistic are asymptotically independent.

10 Empirical Applications to House Prices

According to real estate agents there are three things that affect the price of a house the most: location, location and location. This remark suggests that location and small scale neighbourhood have a substantial impact on the price of a house or an apartment.

A hedonic price model is defined by Rosen (1974) as a regression model, where the price of a commodity is explained by the attributes of the commodity in question. Earlier studies have confirmed that the geographical location of the observed data are important when modelling house prices and it has been pointed out that the small scale neighbourhood have an substantial effect on the price (Laakso 1997, Karakozova 2005, Turnbull, Dombrow and Sirmans 2006, Kiel and Zabel 2007). However, a problem is how to capture and formulate location and small scale neighbourhood in an econometric model. Spatial econometrics offers tools for this. Spatial econometrics applications to house prices are included in Dubin(1988), Pace and Gilley (1997), Pace, Barry and Sirmans (1998), Berg (2002), Wilhelmsson (2002) and Case, Clapp, Dubin and Rodriguez (2004), among others.

Relevant variables for a hedonic house price model can be grouped into three broad categories: location variables, structural variables and neighbourhood variables (Dubin 1988). Location variables are attributes describing the geographical location of the object. They are usually easily measured and can, for instance, be the presence or absence of sea view, the distance to the

central business district or the nearest train station. Structural variables are attributes of the house or apartment itself. Typical examples are the size of the house or apartment, its age, number of rooms, the condition of the interior, lot size and the presence or absence of garage, balcony, fireplaces, and so on. Whereas the first two categories are usually easily measured, the third group of variables, neighbourhood variables, can be thought of as latent characteristics of the neighbourhood. By this type of variables we try to capture the spirit that makes some neighbourhoods more appreciated than others, even when the location and structural attributes are similar for the neighbourhoods. They are the most difficult ones to obtain. Commonly used variables for this purpose are the local crime rate, socioeconomic characteristics of residents, pollution, or noise levels. These variables may be thought of as proxies of the unobservable quality of the small scale neighbourhood.

The focus of Essay 2 is on modelling small scale neighbourhood. In this paper the coordinates of the observations and the information in the original data are used to create new variables by following ideas in Turnbull, Dombrow and Sirmans (2006). The purpose is to examine whether these coordinate-based variables, which measure small scale neighbourhood conditions, can replace some of the spatial structure in a spatial hedonic house price model.

A k-nearest neighbours specification of the weights matrix, which is applied in all essays of the thesis, is particularly suitable in house price models. Assume that the buyer has chosen a certain neighbourhood where he or she intends to purchase a house. The buyer will then compare the available houses in that area. The price that he or she is willing to pay for a particular house is affected by the price level for houses in the neighbourhood. Also the seller's expectation of the price is affected by the selling prices of the other houses in the neighbourhood. The seller will be satisfied with a price that is at least as good as the average in the neighbourhood at the moment of the transaction. The average price in the neighbourhood is on the other hand affected by the attributes of the neighbourhood itself, like available service, safety, or simply good reputation. Also the supply and demand in a particular neighbourhood affect the price. A practical issue is how many of the nearby observations should be considered as nearest neighbours when the weights matrix is specified. This problem is addressed in Essay 3, where we suggest a specification search for the number of nearest neighbours k and illustrate it on real house price data.

Three of the four essays include empirical applications to house price data. Essay 1 and Essay 3 use data from the county of Stockholm in Swe-

den to demonstrate the findings of the essays. The data are based on 1377 transactions of single-family houses between January 2000 and May 2001. The data have been analyzed by Wilhelmsson (2002), and include the selling price, spatial coordinates for observation and, in addition, information about the size of the house in square metres as well as some other characteristics. The empirical analysis in Essay 2 is based on new and unique cross-sectional data on the free from debt selling price of apartments in residential buildings in Helsinki, the capital of Finland, in January, February and March 2002. The data include 649 transactions and are based on information provided by real estate agents in the area. The data were originally collected to be used only by real estate agents. The idea was that only agents that gave particulars of their own transactions could get the particulars of the transactions of their competitors. The original data include 15 attributes describing the apartment, its location or the transaction. Statistics Finland have also data on the selling prices of apartments, but the data lack the exact location of the observations, i.e. the coordinates, which are essential in spatial econometric applications. Statistics Finland can offer postal codes of the apartments or aggregated information in squares as small as $250m \times 250m$. The exact location for individual apartments are not provided, since the information is considered sensitive and therefore protected by legislation.

11 Overview of the Essays

This section contains summaries of the essays focusing on the motivation, the findings and contribution. A common theme of the essays is the spatial econometric perspective. In particular, the focus is on k-nearest neighbours weights matrices, and all applications are on house price data.

11.1 Essay 1: Unilateral Spatial Autoregression

This essay is motivated by the similarities between spatial dependence in spatial econometrics and autocorrelation in time series econometrics. Spatial dependence, or spatial autocorrelation, can be seen as dependence between observations on the dependent variable in a regression model. Observations in spatial data have a location in a two-dimensional plane. As a consequence, observations do not usually have a natural order, and the dependence between the observations may be bilateral. By bilateral, is meant that if one

observation is dependent on another observation, the latter may also depend on the first one. On the other hand, observations on a time series can be seen as points on a line. In time series econometrics the dependence is unilateral, because in causal models an observation only depends on earlier observations, not on future ones.

The distinction between bilateral dependence in spatial econometric models and unilateral dependence in time series processes was pointed out by Cliff and Ord (1969). Despite this there are some similarities, which are recognized by Fingleton (2009), who brings the two fields of econometrics together. He utilize a binary contiguity weights matrix to make the connection between temporal and spatial series. In this essay we further develop the ideas in Fingleton. By adapting results from the probability literature (Whittle 1954, Tjøstheim 1978, 1981, 1983, and Yao and Brockwell 2006) we show that by considering a unilateral spatial autoregressive process, we may define a unilateral spatial autoregression, which have similar properties as an autoregression with time series.

The following process is an example of a first-order unilateral spatial autoregressive process defined in a two-dimensional plane (Whittle 1954):

$$y_{ij} = \alpha_1 y_{i-1,j} + \alpha_2 y_{i,j-1} + \varepsilon_{ij}, \tag{26}$$

whereas a first-order spatial autoregression (cf. model (2) in Section 4) is

$$\mathbf{y} = \rho \mathbf{W} \mathbf{y} + \boldsymbol{\varepsilon}. \tag{27}$$

Here $\mathbf{y} = [y_{ij}]$ is an $N \times 1$ vector of observations on a spatial series y_{ij} on a regularly spaced lattice, $\mathbf{W} = [w_{ij}]$ is an $N \times N$ spatial weights matrix, ρ is the spatial autoregressive parameter and $\boldsymbol{\varepsilon} = [\varepsilon_{ij}]$ is an $N \times 1$ vector of errors assumed to be independent and identically distributed with $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_N$.

In the essay we begin by deriving the representation of a unilateral spatial autoregressive process, for instance the one given by (26), as a unilateral spatial autoregression (27). We assume that the weights matrix $\mathbf{W}$ is unilateral and that it has two non-zero entries on each row, in order to correspond to the process (26).

To define the parameter space for ρ we begin by giving a causality condition for unilateral spatial autoregressive models based on the theory of unilateral spatial autoregressive processes. For the unilateral first-order spatial autoregressive process (26) the causality condition is given by Whittle

(1954), and it is simply $|\alpha_1| + |\alpha_2| < 1$. We find that for model (27) when ρ is positive, and it is entertaining a unilateral row-normalized weights matrix, the condition boils down to $\rho < 1$.

We show consistency and asymptotic normality in unilateral spatial autoregression by using results obtained by Yao and Brockwell (2006). Similar results on consistency and asymptotic normality in spatial econometric models have been obtained by Kelejian and Prucha (1998, 1999), Lee (2002, 2003, 2004), and Mynbaev (2010) using different sets of assumptions on the spatial process, not restricting their studies to the unilateral case. However, the results obtained by Yao and Brockwell (2006) are in some sense more elementary, because they do not require high level assumptions about moment conditions, ergodicity and mixing.

Further, the relation between the spatial autoregressive parameter ρ in (27) and the coefficients α_1 and α_2 in (26) is studied. Through simulations we examine the finite sample properties of the maximum likelihood estimator of ρ when the data are generated by the unilateral spatial autoregressive process (26). We find that $\hat{\rho} \approx 0.5(\alpha_1 + \alpha_2)$, except at the boundary of the parameter space. The connection is helpful for understanding the similarities between unilateral processes and spatial econometric models. The simulation results show that for large samples the estimated autoregressive parameter is nearly unbiased. For small samples there is a downward bias in the estimator. In addition, we find that underspecifying the weights matrix results in an underestimated spatial autoregressive parameter, and overspecifying the weights matrix results in an overestimated spatial autoregressive parameter.

In a time series context, the unilateral dependence arises naturally in stationary time series models. For spatial series the unilateral ordering may be viewed as an artefact, which limits its use in applications with real data. Nevertheless, there are some applications in which the unilateral ordering arises naturally. In the field of water management and downstream pollution in drainage basins, the spatial process is unilateral, to mention only one potential application. More generally, according to results of Tjøstheim (1981), a unilateral spatial autoregressive process can be seen as an approximation to a bilateral spatial process. A related argument is made by Pace et al. (2009) in the context of spatial modelling of house prices. Their argument is that in large samples the exact specification of the weights matrix is not so important, since different weights matrices do not yield materially different coefficient estimates. It may therefore be a moot point whether a unilateral or bilateral weights matrix should be preferred, and then we may use

a unilateral one, because it simplifies the asymptotic analysis. In the essay these arguments are demonstrated by an empirical application involving house prices in the county of Stockholm in Sweden.

The essay attempts to contribute to the spatial econometric literature by taking a time series point of view to spatial data. We use a unilateral approach, and show how the parameters of the studied models are connected to each other. Bringing together the two fields of econometrics may deepen the understanding between them, and therefore be beneficial to both.

11.2 Essay 2: Small Scale Neighbourhood in Spatial Econometrics

It is a common and reasonable assumption for house price data that objects situated near each other have similar values on variables describing the location and immediate surroundings, i.e. that these variables are spatially autocorrelated. In practice it is seldom possible to obtain all the desired variables in a model. When modelling house prices it is especially hard to find variables capturing the small scale neighbourhood conditions. According to LeSage and Pace (2009) there is a strong motivation for spatial econometric modelling when there are omitted variables which are assumed to be spatially autocorrelated.

The purpose of the essay is to examine whether coordinate-based variables, which are intended to measure small scale neighbourhood conditions, can replace the spatial structure in a spatial hedonic house price model. The motivation is that a model with a simpler spatial structure is easier to interpret.

In order to apply a spatial econometric model for house prices the exact location of every unit has to be known. The location is for instance determined by coordinates in a two-dimensional plane. Usually the coordinates are then utilized to create a spatial weights matrix. Here we take a different view. The coordinates, combined with information in the original data, are used to create new variables by following suggestions by Turnbull, Dombrow and Sirmans (2006).

The starting point of this study is a non-spatial hedonic house price model. We assume that there are important explanatory variables missing from the model, and that the omitted variables are spatially autocorrelated, as well as correlated with the explanatory variables included in the model.

These assumptions are shown to lead to a spatial Durbin model. We estimate the model and find that the spatial lags are significant. We then uncover candidates for some of the omitted variables, following the suggestions by Turnbull, Dombrow and Sirmans. The new variables for the small scale neighbourhood conditions are shown to be spatially autocorrelated and they are also shown to correlate with the original variables included in the model. They are included among the explanatory variables and the spatial Durbin model is re-estimated. We now find that the spatial lags of the Durbin model become insignificant. The model is reduced to a conventional non-spatial regression model. It seems reasonable to assume that there are other explanatory variables missing despite the included uncovered new variables. We assume that the remaining omitted variables are still spatially autocorrelated, but that they are uncorrelated with the explanatory variables included in the model. These assumptions are shown to lead to a spatial error model. When this model is estimated we find that the spatial lag of the error term is highly significant. In this final model the regression coefficients have their usual interpretation as partial derivatives. This is not the case if the model involves lags on the dependent variable or on the explanatory variables. Hence, an easy interpretation of the estimates in the final model is a benefit, compared to the interpretation of the estimates in the structurally richer spatial Durbin model.

Our empirical results show that for this particular data set it is possible to find explanatory variables that capture some of the small scale neighbourhood conditions, and thereby obtain a simplified model. However, the set of small scale neighbourhood variables cannot entirely replace a spatial econometric structure. The attempted contribution of this essay is methodological in the sense that it demonstrates the usefulness of the suggestions of Turnbull, Dombrow and Sirman in a spatial econometric setting. In addition, the essay has an empirical contribution, since the Finnish data is unique and has not previously been used.

11.3 Essay 3: A Practical Proposal to Specification Search of a k-Nearest Neighbours Weights Matrix

The focus of model specification search in the spatial econometrics literature has been on procedures for comparative testing of alternative model specifications (LeSage and Pace 2009), i.e. the choice between different structural

forms for the spatial econometric model. Examples of this are Florax, Folmer and Rey (2003, 2006) and Hendry (2006).

This essay has a different perspective, as it is motivated by a question often asked by practitioners — How should one determine the number of neighbours in a k-nearest neighbours weights matrix, or justify a particular choice? Hence, the purpose of the essay is to find a strategy for specification search for the number of neighbours in a k-nearest neighbours weights matrix. As the means of the specification search we use the spatial J-test proposed by Kelejian (2008).

Spatial models with different weights matrices are non-nested. A consequence, as LeSage and Pace (2009, page 162) notes, is that it is not in general possible to use formal tests for significant differences between the log likelihood function values for models based on different weights matrices. The number of parameters is fixed and equal in the models. The usual specification search based on information criteria is then not available either.

Recently Piras and Lozano-Garcia (2008), and Burridge and Fingleton (2010) have shown that the spatial J-test can be used in order to discriminate between different types of weights matrices, for instance, when contiguity, inverse distance or k-nearest neighbours weights matrices are pair-wise compared to each other. Here we limit the study to one type of spatial weights matrices, k-nearest neighbours weights matrices. If there are k-nearest neighbours weights matrices in both of the models that are compared, the models are relatively closer to each other than they are in the case when the objective is to discriminate between different types of weights matrices. This is studied as a possible problem, since usually, in a set-up where the null and the alternative model are close, a test may loose power.

In the essay we suggest and examine two approaches for finding k, the increasing and the decreasing number of neighbours approaches. In the increasing neighbours approach we begin the testing sequence with a spatial J-test of a null model where k is small. In the alternative model the number of neighbours in the weights matrix is $k+1$. Then k is gradually increased, and we continue as long as the null model is rejected. In the resulting sequence of spatial J-tests k is always smaller in the null model than it is in the alternative model. In the decreasing number of neighbours approach we begin with a large k, which is gradually decreased. Hence, in the resulting sequence of tests k is always larger in the null model than in the alternative model.

The results of a simulation study show that the spatial J-test can be used

for distinguishing between general spatial models with different k-nearest neighbours weights matrices. The size of the test is acceptable and the power of the test is high when the weights matrix of the null model is underspecified, especially if the amount of spatial autocorrelation in the dependent variable is at least moderate. If the weights matrix of the null model is overspecified, the power of the test is not equally high. The power of the test is low when the amount of spatial autocorrelation in the dependent variable is small and the weights matrix of the null model is overspecified. If the amount of spatial autocorrelation in the dependent variable is increased, the power of the test increases, although it never gets as high as when the weights matrix of the null model is underspecified. The results are therefore clearly in favour of the increasing neighbours approach.

When conducting tests in a sequence, ideally all rejection probabilities for false null hypothesis should be one, otherwise there is no guarantee that the sequence will come to the test where the null model corresponds to the data generating process. When we examine the rejection probabilities of the tests in the sequences of the two considered approaches, we find a clear difference between them. When the increasing neighbours approach is applied, and as long as the spatial autocorrelation in the dependent variable is at least intermediate, all the false models tend to be rejected and the significance level of the final test is unaffected by the sequential testing. However, if we conduct the specification search by the decreasing neighbours approach, all false models might not be rejected, since the sequence contains tests whose rejection probabilities are less than one. The implication is that by this method we might stop the sequence of tests too early and the outcome is too large a k. Based on the results, we suggest that the increasing neighbours approach should be used as a strategy for specification search for the number of neighbours in a k-nearest neighbours weights matrix.

The paper includes a specification search for the number of nearest neighbours on real data, which illustrate the findings in the simulations, and how the suggested strategy is applied in practice. The empirical illustration is concluded by a comparison between the proposed strategy and Bayesian approach to model comparison. In a Bayesian setting different spatial weights matrices can be compared through log marginal likelihoods and associated model probabilities (see LeSage and Pace 2009, Section 6.3). We find that the outcomes of the suggested strategy and Bayesian model comparison are almost the same.

The essay attempts to contribute to the spatial econometric literature

by suggesting a formal strategy to specification search which motivates the choice of a particular k in a k-nearest neighbours weights matrix. As shown in the illustration it can easily be applied in practice. A contribution to the literature about the spatial J-test is that the test can be applied in order to distinguish between spatial models entertaining different k-nearest neighbours weights matrices. There seems to be no loss of power, as long as the weights matrix of the null model is the one based on a smaller number of neighbours than the alternative. This is in line with the findings of Essay 4.

11.4 Essay 4: Bootstrap Spatial J-Tests for k-Nearest Neighbours

In this essay we study the properties of the asymptotic spatial J-test, proposed by Kelejian (2008). We suggest and examine a bootstrap spatial J-test, to see whether the properties of the asymptotic test can be improved.

Piras and Lozano-Garcia (2008), and Burridge and Fingleton (2010) have studied the spatial J-test when it used to discriminate between different types of weights matrices, for instance, when contiguity, inverse distance or k-nearest neighbours weights matrices are pair-wise compared to each other. They find that the asymptotic J-test may suffer from size distortion in small samples. Burridge and Fingleton (2010) also study the performance of bootstrap spatial J-tests in the same context. They find that the bootstrap test is superior in most cases.

Here the focus is different to the previous studies, as we study the properties of the spatial J-test when it is used to distinguish between spatial models with different k-nearest neighbours weights matrices. The motivation for the study is that when both the null and the alternative model entertain k-nearest neighbours weights matrices, the models are relatively closer to each other than they are in the case when the objective is to discriminate between different types of weights matrices. In a set-up where the null and the alternative model are close, the properties of the test may be affected. The finding that the asymptotic test is oversized in small samples gives a motivation for the bootstrap approach.

We begin with generating data by a spatial econometric model for different sample sizes, and different values on the autoregressive parameters. The parameters are chosen in order to represent small, medium and large

positive spatial autocorrelation. When the spatial J-tests are conducted, we compare models entertaining both under- and overspecified k-nearest neighbours weights matrices to the model of the data generating process. By underspecified we mean that the weighs matrix is based on a smaller number of nearest neighbours than in the data generating process, and by overspecified we mean that the weighs matrix is based on a larger number of nearest neighbours than in the data generating process.

The simulation study shows that the asymptotic spatial J-test can be used for distinguishing between general spatial econometric models entertaining different k-nearest neighbours weights matrices, but that the asymptotic test is oversized in small samples. This finding motivates the examination of the bootstrap spatial J-test. We find that the bootstrap is useful for correcting the size of the asymptotic test in small samples.

We find that the power of the bootstrap spatial J-test is very close to the power of asymptotic test. When the sample size is small, the amount of spatial autocorrelation in the dependent variable is small, and the weights matrix of the null model is overspecified, the power of the test is low. In contrast, the power of the test is high when the spatial autocorrelation in the dependent variable is at least intermediate, and the weights matrix of the null model is underspecified. When the sample size is increased the power increases. For large samples we find that the power of the test is very high when the null model entertains an underspecified weights matrix.

The essay attempts to contribute to the literature concerning the spatial J-test by studying the properties of the test when it is applied in order to distinguish between spatial models entertaining different k-nearest neighbours weights matrices. The finding that the asymptotic test is oversized in small samples, which motivates the bootstrap, is in line with previous studies. The finding that the power is increasing with the sample size, but is higher when the null model entertains an underspecified weights matrix, than in the case of an over specified weights matrix, supports the conclusions of Essay 3.

12 Conclusion

In the first essay we clarify the relation between a first-order spatial econometric model entertaining a unilateral weights matrix, and a unilateral spatial autoregressive process. In the second we show by an empirical study on house price data that it is possible to form coordinate-based, spatially autoregres-

sive variables, which at least to some extent can replace the spatial structure in a spatial econometric model. In the third essay we suggest, study and demonstrate a strategy for specifying a k-nearest neighbours weights matrix by applying the spatial J-test, suggested by Kelejian (2008). In the final essay we examine the properties of the asymptotic spatial J-test and suggest and study the properties of a bootstrap spatial J-test when the tests are used in order to distinguish between k-nearest neighbours weights matrices.

The attempted contribution of the thesis to the spatial econometric literature is mainly methodological, but the findings are illustrated by empirical studies on house price data. It is well known that spatial econometric models are suitable for, and therefore often applied to house price data. This motivates the choice of data for the empirical illustrations. Another theme, except the application to house price data, that binds the four essays together is the focus on k-nearest neighbours spatial weights matrices. In the context studied here, k-nearest neighbours weights matrices have not got much attention in the spatial econometric literature, which motivates the choice of examining this type of matrices.

References

[1] Anselin, L. (1980), *Estimation Methods for Spatial autoregressive Structures*, Ithaca, NY: Cornell University Regional Science Dissertation and Monographs Series # 8.

[2] Anselin, L. (1988), *Spatial Econometrics: Methods and Models*, Dordrecht: Kluver Academic Publishers.

[3] Anselin, L. (2003), Spatial Externalities, Spatial Multipliers and Spatial Econometrics, *International Regional Science Review*, 26 (2), p. 153–166.

[4] Anselin, L. (2006). Spatial Econometrics. In Mills, T. and Patterson, K., editors, Palgrave Handbook of Econometrics: Volume 1, *Econometric Theory*, p. 901–969. Palgrave Macmillan, Basingstoke.

[5] Anselin, L. (2009), Thirty Years of Spatial Econometrics, GeoDa Center for Geospatial Analysis and Computation, School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ 84287-5302, Working Paper 2009–2.

[6] Berg, L. (2002) Prices and Constant Quality Price Indexes for Multi–Dwelling and Commercial Buildings in Sweden, Uppsala, Uppsala universitet, Department of Economics, Working Paper 2002:2.

[7] Burridge, P. and Fingleton, B. (2010) Bootstrap Inference in Spatial Econometrics: The J Tes,. *Spatial Econometric Analysis,* 5, 1, p.93–119.

[8] Case, B., Clapp, J., Dubin, R. and M. Rodriguez (2004) Modeling Spatial and Temporal House Price Patterns: A Comparison of Four Models, *Journal of Real Estate Finance and Economics,* 29:2, 167–191.

[9] Cressie, N. A. (1993) *Statistics for Spatial Data*. John Wiley & Sons, Inc.

[10] Cliff, A. D. and Ord, J. (1969) The Problem of Spatial Autocorrelation. In *London Papers in Regional Science*, edited by A. Scott, p. 25–55. London: Pion.

[11] Cliff, A. D. and Ord, J.(1973). *Spatial Autocorrelation*. London: Pion, 1973.

[12] Cliff, A. D. and Ord, J.(1981). *Spatial Processes, Models and Applications*. London: Pion, 1981.

[13] Davidson, R. and MacKinnon, J. G. (1981) Several Tests for Model Specification in the Presence of Alternative Hypothesis, *Econometrica,* 49, p. 781–794.

[14] Davidson, R. and MacKinnon, J. G. (2006) The Power of Bootstrap and Asymptotic Tests, *Journal of Econometrics,* 133, p. 421–441.

[15] Dubin, R. (1988) Estimation of Regression Coefficients in the Presence of Spatially Autocorrelated Error Terms, *The Review of Economics and Statistics*, Vol. 70, Issue 3, p. 466–474.

[16] Fingleton, B. (2009) Spatial Autoregression, *Geographical Analysis,* 41, p. 385–391.

[17] Florax, R.J.G.M., Folmer, H. and Rey, S.J. (2003) Specification Searches in Spatial Econometrics: The Relevance of Hendry's Methodology, *Regional Science and Urban Economics*, 33, p. 557–579..

[18] Florax, R.J.G.M., Folmer, H. and Rey, S.J. (2006) A Comment on Specification Searches in Spatial Econometrics: The Relevance of Hendry's Methodology: A Reply, *Regional Science and Urban Economics*, Volume 36, Issue 2, March 2006, p. 300–308.

[19] Haining, R. P.(2003) *Spatial Data Analysis: Theory and Practice*, Cambridge University Press.

[20] Hendry, D. (2006) A Comment on "Specification Searches in Spatial Econometrics: The Relevance of Hendry's Methodology", *Regional Science and Urban Economics,* Volume 36, Issue 2, March 2006, p. 309–312

[21] Horowitz (1994) Bootstrap-Based Critical Values for the Information Matrix Test, *Journal of Econometrics,* 61, p. 395–411.

[22] Karakozova, O. (2005) Modelling and Forecasting Property Rents and Returns. Ekonomi och Samhälle, Publications of the Swedish School of Economics and Business Administration, Nr 149, Helsinki.

[23] Kelejian, H.H. (2008) A Spatial J-Test for Model Specification Against a Single or a Set of Non-Nested Alternatives, *Letters in Spatial and Resource Sciences,* 1, 1, p. 3–11.

[24] Kelejian, H. H. and Prucha, I. R. (1998), A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances, *Journal of Real Estate Finance and Economics*, 17, p. 99–121.

[25] Kelejian, H. H. and Prucha, I. R. (1999), A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model, *International Economic Review,* 40, p. 509–533.

[26] Kelejian, H. H. and Prucha, I. R. (2010), Specification and Estimation of Spatial Autoregressive Models with Autoregressive and Heteroscedastic Disturbances, *Journal of Econometrics,* 157, 1, p. 53–67.

[27] Kelejian, H. H., Prucha, I. R. and Yuzefovich, E. (2004), Instrumental Variable Estimation of a Spatial Autoregressive Model with Autoregressive Disturbances: Large and Small Sample Results. In: LeSage, J. and Pace, K. (Eds.), *Advances in Econometrics: Spatial and Spatiotemporal Econometrics*. Elsevier, New York. p. 163–198.

[28] Kiel, K. A. and Zabel, J. E. (2007) Location, Location, Location: The 3L Approach to House Price Determination, *Journal of Housing Economics,* 17, 2, p. 175–190.

[29] Laakso, S. (1997) Urban Housing Prices and the Demand for Housing Characteristics. A Study on Housing Prices and Willingness to Pay for Housing Characteristics and Local Public Goods in the Helsinki Metropolitan Area. Helsinki, Elinkeinoelämän tutkimuslaitos, 27.

[30] Lee, L.–F. (2002) Consistency and Efficiency of Least Squares Estimation for Mixed Regressive, Spatial Autoregressive Models, *Econometric Theory,* 18, p. 252–277.

[31] Lee, L.–F. (2003) Best Spatial Two-Stage Least Squares Estimators for a Spatial Autoregressive Model with Autoregressive Disturbances, *Econometric Reviews,* 22, p. 307–335.

[32] Lee, L.–F. (2004) Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Autoregressive Models, *Econometrica,* 72, 1899–1925.

[33] LeSage, J. P. (1997) Bayesian Estimatimation of Spatial Autoregressive Models, *International Regional Science Review,* 20, 113-129.

[34] LeSage, J.P. (2005, 2010) Spatial Econometric Toolbox for Matlab, http://www.spatial-econometrics.com.

[35] LeSage, J.P. and Pace, R. K. (2009) *Introduction to Spatial Econometrics.* CRC Press, Taylor and Francis Group.

[36] Mynbaev, K. T. (2010) Asymptotic Distribution of the OLS Estimator for a Mixed Spatial Model, *Journal of Multivariate Analysis,* 101, 3, March 2010, 733-748.

[37] Ord, J.K. (1975) Estimation Methods for Models of Spatial Interaction, *Journal of the American Statistical Association,* 70, 120-126.

[38] Pace, R. K., and O. W. Gilley (1997) Using the Spatial Configuration of the Data to Improve Estimation, *Journal of Real Estate Finance and Economics,* Vol. 14, Number 3, 333-340.

[39] Pace, R.K. and LeSage, J. P.(2008) Biases of OLS and Spatial Lag Models in the Presence of an Omitted Variable and Spatially Dependent Variables (February 19, 2008). Available at SSRN: http://ssrn.com/abstract=1133438

[40] Pace, R. K., LeSage, J. P. and Zhu, S (2009) Impact of Cliff and Ord on the Housing and Real Estate Literature, *Geographical Analysis,* 41, 418-424.

[41] Pace, R. K., LeSage, J. P. and Zhu, S. (2010, June 2nd) Spatial Dependence in Regressors and its Effect on Estimator Performance, Conference paper, presented at the IVth Conference of the Spatial Econometric Association, SEA 2010, in Chicago, USA.

[42] Pace, R. K., Barry, R. and C. F. Sirmans (1998) Spatial Statistics and Real Estate, *Journal of Real Estate Finance and Economics,* Vol. 17, Number 1, 5-13.

[43] Piras, G. and Lozano-Garcia N. (2008) Spatial J-test: Some Monte Carlo Evidence. Conference paper, presented at the Annual Meeting of the RSAI in New York, November 2008.

[44] Ripley, B. (1981) *Spatial Statistics.* John Wiley & Sons, Inc.

[45] Rosen, S. (1974) Hedonic Prices and Implicit Markets: Product Differentation in Pure Competition, *Journal of Political Economy,* 82, 34-55.

[46] Schabenberger, O. and Gotway, C. A. (2005) *Statistical Methods for Spatial Data Analysis.* Chapman &Hall / CRC Press.

[47] Tjøstheim, D. (1978), Statistical Spatial Series Modelling, *Advances in Applied Probability,* 10, 130–154.

[48] Tjøstheim, D. (1981) Autoregressive Modeling and Spectral Analysis of Array Data in the Plane, *IEEE Transactions on Geoscience and Remote Sensing,* 19, 15-24.

[49] Tjøstheim, D. (1983) Statistical Spatial Series Modelling II: Some further Results in Unilateral Processes, *Advances in Applied Probability,* 15, 562–684.

[50] Turnbull, G. K., Dombrow, J. and Sirmans, C. F. (2006) Big House, Little House: Relative Size and Value, *Real Estate Economics,* 34, 3: 439-456.

[51] Whittle, P. (1954) On Stationary Processes in the Plane, *Biometrika*, 41, 434-449.

[52] Wilhelmsson, M. (2002) Spatial Models in Real Estate Economics, *Housing, Theory and Society,* 19, 92-101.

[53] Yao, Q. and Brockwell, P. J. (2006) Gaussian Maximum Likelihood Estimation for ARMA Models II: Spatial Processes, *Bernoulli*, 12, 403–429.

Unilateral Spatial Autoregression*

Niklas Ahlgren Linda Gerkman
Hanken School of Economics,
Department of Finance and Statistics

Abstract

In spatial econometric models the dependence between the observations may extend in all directions. This is in marked contrast to time series processes, where the dependence is unilateral. The two fields may be brought together by defining spatial autoregression through a binary weights matrix with a special structure. We show that by restricting the attention to unilateral spatial autoregressive processes, we may define a unilateral spatial autoregression which enjoys similar properties as an autoregression with time series. For spatial series the unilateral ordering may be viewed as an artefact, although there are some applications in which the unilateral ordering arises naturally. More generally, a unilateral spatial autoregressive process may be seen as an approximation to a bilateral spatial process. The causality condition for unilateral spatial autoregression is given. Inference in unilateral spatial autoregression is considered. An empirical application illustrates the use of unilateral spatial autoregressive models with real data.

Key words: Maximum likelihood estimator, Spatial autoregression, Unilateral spatial autoregressive process, Weights matrix.

1 Introduction

In spatial econometric models the dependence between the observations may extend in all directions. This is in marked contrast to time series processes,

*We thank Markku Rahiala for useful comments, and Mats Wilhelmsson for providing the data used in the empirical application in Section 4. Niklas Ahlgren acknowledges financial support from The Finnish Society of Sciences and Letters.

where the dependence is unilateral (Cliff and Ord 1969). Fingleton (2009) brings the two fields together by defining spatial autoregression through a binary weights matrix with a special structure. In this paper we show that by restricting the attention to unilateral spatial autoregressive processes, we may define a unilateral spatial autoregression which enjoys similar properties as an autoregression with time series. We extend the ideas in Fingleton and put them on a more formal footing by adapting results from the theory of unilateral spatial autoregressive processes in the probability literature (Whittle 1954, Tjøstheim 1978, 1981, 1983, and Yao and Brockwell 2006).

We derive the representation of a unilateral spatial autoregressive process as a unilateral spatial autoregression. The causality condition for unilateral spatial autoregression is given. Inference in unilateral spatial autoregression is considered. We examine the maximum likelihood (ML) estimator of the spatial autoregressive parameter in the context of spatial econometric models when the data are generated by a unilateral spatial autoregressive process of finite order. By exploring the fact that the spatial autoregressive (SAR) model formally resembles a spatial autoregressive process with a special structure, we show that the results in Yao and Brockwell (2006) on the consistency and asymptotic normality of the ML estimator also hold for unilateral spatial autoregression. The process is only required to have finite second moments, and no assumptions on ergodicity and mixing are required.

We remark that similar results on consistency and asymptotic normality in spatial econometric models have been obtained by Kelejian and Prucha (1998, 1999), Lee (2002, 2003, 2004), and Mynbaev (2010) using different sets of assumptions on a broader class of spatial autoregressive processes than the unilateral case. The results obtained by Yao and Brockwell (2006) are in some sense more elementary, because they do not require high level assumptions about moment conditions, ergodicity and mixing.

Some simulation results to examine the finite sample properties of the estimator show that it is nearly unbiased, except at the boundary of the parameter space. The specification of spatial econometric models is based on the assumption that the weights matrix is known. In many cases in practice the weights matrix may be unknown. We show that underspecifying the weights matrix results in an underestimated spatial autoregressive parameter, whereas overspecifying the weights matrix results in an overestimated spatial autoregressive parameter.

In a time series context, the unilateral dependence arises naturally in stationary time series models. For spatial series the unilateral ordering may be

viewed as an artefact, which limits its use in applications with real data. Nevertheless, there are some applications in which the unilateral ordering arises naturally. In the field of water management and downstream pollution in drainage basins, the spatial process in unilateral, to mention only one potential application. More generally, a unilateral spatial autoregressive process may be seen as an approximation to a bilateral spatial process. Tjøstheim (1981) establishes the existence of a unilateral approximation to a spatial series under the relatively mild condition that the spatial autoregressive series has a continuous and positive spectral density. Under this condition a spatial autoregressive series can be approximated arbitrarily close in quadratic mean by a unilateral spatial autoregressive process. A related argument is made by Pace et al. (2009) in the context of spatial modelling of house prices. Their argument is that in large samples the exact specification of the weights matrix is not so important, since different weights matrices do not yield materially different coefficient estimates. It may therefore be a moot point whether a unilateral or bilateral weights matrix should be preferred, and then we may use a unilateral one, because it simplifies the analysis to a great extent. To be specific, we will demonstrate this argument in the context of an empirical application involving house prices in the county of Stockholm in Sweden.

The paper is structured as follows. Section 2 introduces the unilateral spatial autoregressive model and gives the limiting distribution of the spatial autoregressive parameter. Section 3 investigates the finite sample properties of the ML estimator of the spatial autoregressive parameter. Section 4 contains an empirical application. Section 5 concludes.

We use the following notation in the paper. The matrix $\mathbf{Y} = [y_{ij}]$ is an $N_1 \times N_2$ matrix of observations on a spatial series on a regularly spaced lattice. We let $\mathbf{y} = \text{vec}(\mathbf{Y}')$ denote the $N \times 1$ row vectorization of the matrix $\mathbf{Y}$, $N = N_1 N_2$. The $N \times N$ matrix $\mathbf{W}_k$ denotes a general weights matrix and the subindex k refers to the number of nonzero elements in each row of the matrix.

2 Unilateral Spatial Autoregressive Models

The model we work with is a spatial autoregressive (SAR) model (Anselin 2003)

$$\mathbf{y} = \rho \mathbf{W} \mathbf{y} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y} = [y_{ij}]$ is an $N \times 1$ vector of observations on a spatial series y_{ij}, $\mathbf{W}$ is an $N \times N$ spatial weights matrix, ρ is a spatial autoregressive parameter and $\boldsymbol{\varepsilon} = [\varepsilon_{ij}]$ is an $N \times 1$ vector of errors assumed to be independent and identically distributed with $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_N$.

The spatial weights matrix is given by

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_{11} & \cdots & \mathbf{W}_{1N_1} \\ \vdots & \ddots & \vdots \\ \mathbf{W}_{N_1 1} & \cdots & \mathbf{W}_{N_1 N_1} \end{pmatrix}, \tag{2}$$

which is an $N \times N$ block matrix. The blocks $\mathbf{W}_{ij}$, $i = 1, \ldots, N_1$, $j = 1, \ldots, N_1$, are $N_2 \times N_2$ matrices. See LeSage and Pace (2009) for the motivation and interpretation of spatial econometric models.

2.1 Unilateral Spatial Autoregressive Processes

We assume that the data are generated from a unilateral spatial autoregressive process defined on a regular lattice (Basu and Reinsel 1993)

$$y_{ij} = \sum_{k=0}^{p_1} \sum_{l=0}^{p_2} \alpha_{kl} y_{i-k,j-l} + \varepsilon_{ij}, \quad \alpha_{00} = 0. \tag{3}$$

In this model the value at site (i, j) is a finite autoregression on the values at the sites in the lower quadrant of (i, j).

The SAR process in (3) is defined in terms of the quarter-plane order, and is a special case of the more general specification of Yao and Brockwell (2006), which includes both half-plane and quarter-plane spatial autoregressive moving average (ARMA) processes.

A general spatial ARMA process is said to be causal if it admits a pure MA representation in terms of $\{\varepsilon_{ij}\}$ in the quarter plane with absolutely summable coefficients (see e.g. Yao and Brockwell 2006, p. 406).

The causality condition for the unilateral SAR process (3) is given by Tjøstheim (1978). Define the characteristic polynomial

$$\Phi(z_1, z_2) = 1 - \sum_{k=0}^{p_1} \sum_{l=0}^{p_2} \alpha_{kl} z_1^k z_2^l. \tag{4}$$

The causality condition requires that

$$\Phi(z_1, z_2) \neq 0 \quad \text{for all} \quad |z_1| \leq 1 \quad \text{and} \quad |z_2| \leq 1. \tag{5}$$

2.2 The Structure of Unilateral Spatial Autoregressive Models

We consider again (1) and assume that the data are generated from (3). We make the following assumption throughout:

Assumption 1 *The spatial weights matrix* $\mathbf{W}$ *is an* $N \times N$ *(strictly) lower triangular matrix, and such that the dependence in* $\mathbf{y}$ *is unilateral.*

It is worth noting that the condition on $\mathbf{W}$ depends on how $\mathbf{Y}$ is vectorized. Here we have defined $\mathbf{y} = \text{vec}(\mathbf{Y}')$, i.e. $\mathbf{y}$ is the $N \times 1$ row vectorization of the matrix $\mathbf{Y} = [y_{ij}]$. We will refer to model (1) with the weights matrix satisfying Assumption 1 as a unilateral spatial autoregressive model.

In addition, we assume that the weights matrix (2) is binary. The unilateral SAR model (1) with a binary weights matrix formally resembles a unilateral spatial autoregressive process with a special structure. We may write the model as

$$y_{ij} = \rho \sum_{k,l} \sum_{k=l\neq 0} y_{i-k,j-l} + \varepsilon_{ij}.$$

If the weights matrix $\mathbf{W}$ is row-normalised so that the elements in each row sum to one, then

$$y_{ij} = \frac{\rho}{k} \sum_{k,l} \sum_{k=l\neq 0} y_{i-k,j-l} + \varepsilon_{ij},$$

where $k = (p_1 + 1)(p_2 + 1) - 1$. In the model all the autoregressive parameters are equal. If the autoregressive parameters are not equal, it can be incorporated into the weights matrix by different weights. The SAR model is a convenient statistical model, in which $\sum_{k,l} \sum_{k=l\neq 0} y_{i-k,j-l}$ (possibly multiplied by $1/k$) is a spatial lag, and ρ is a spatial autoregressive parameter.

The following proposition is an immediate consequence of the causality condition for the unilateral SAR process. The causality condition in the SAR model depends only on the autoregressive parameter ρ.

Proposition 1 *Assume that the data are generated from (3). The unilateral SAR model (1) with* $\rho > 0$ *is causal if and only if* $\rho < 1/k$*, where* k *is the number of non-zero elements in each row of* $\mathbf{W}$*.*

Proof. When $\rho < 1/k$, by the triangle inequality,

$$\left| \sum_k \sum_l \alpha_{kl} z_1^k z_2^l \right| \leq \sum_k \sum_l |\alpha_{kl}| \left| z_1^k \right| \left| z_2^l \right| \leq \sum_k \sum_l |\alpha_{kl}| < 1$$

for all $|z_1| \leq 1$, $|z_2| \leq 1$, and all the roots of $\Phi(z_1, z_2) = 0$ lie outside the unit discs. On the other hand, if all the roots lie outside the unit discs, $\Phi(z_1, z_2) = 0$ implies that $|k\rho| < 1$. Hence, $\rho < 1/k$. ■

We remark that when the SAR model is estimated with the weights matrix row-normalised, the causality condition reduces to $\rho < 1$, which is the condition for the invertibility of the SAR model.

2.3 Inference in Unilateral Spatial Autoregressive Models

The unilateral SAR model (1) can be solved in terms of $\boldsymbol{\varepsilon}$:

$$\mathbf{y} = (\mathbf{I}_N - \rho\mathbf{W})^{-1}\boldsymbol{\varepsilon}, \tag{6}$$

provided that $(\mathbf{I}_N - \rho\mathbf{W})$ is non-singular, from which it is seen that the errors of the reduced form model are not IID.

The model can be estimated by maximum likelihood by maximizing the Gaussian log likelihood function

$$l(\rho, \sigma^2) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\sigma^2 + \ln|\mathbf{I}_N - \rho\mathbf{W}| - \frac{1}{2\sigma^2}\mathbf{y}'(\mathbf{I}_N - \rho\mathbf{W})'(\mathbf{I}_N - \rho\mathbf{W})\mathbf{y}. \tag{7}$$

In estimating the model the weights matrix is row-normalised (the elements in each row sum to one).

Let ρ_0 denote the true value of ρ. Following Yao and Brockwell (2006), we assume that the following condition holds:

(C1) The parameter space Θ is a compact set containing the true value ρ_0 as an interior point. Further, for any $\rho \in \Theta$, the causality condition holds.

Proposition 2 *Let* $\{\varepsilon_{ij}\} \sim$ *IID*$(0, \sigma^2)$ *and condition (C1) holds. Then as both* N_1 *and* $N_2 \to \infty$, $\widehat{\rho} \xrightarrow{P} \rho_0$.

Following Yao and Brockwell (2006), we assume the additional condition (C2).

(C2) One of the following three conditions holds:
(i) $N_1 \to \infty$ and N_1/N_2 has a limit $d \in (0, \infty)$,
(ii) $N_2 \to \infty$ and $N_1/N_2 \to \infty$,
(iii) $N_1 \to \infty$ and $N_1/N_2 \to 0$.

Proposition 3 *Let* $\{\varepsilon_{ij}\} \sim IID(0, \sigma^2)$ *and conditions (C1) and (C2) hold. Then* $N^{1/2}(\widehat{\rho} - \rho_0) \xrightarrow{D} N(0, v(\rho_0))$*, where the asymptotic variance is given by*

$$v(\theta) = \left\{ E\left[\sum_{k,l} \sum_{k=l\neq 0} y_{i-k,j-l} \right]^2 \right\}^{-1} . \tag{8}$$

Propositions 2 and 3 are immediate consequences of Yao and Brockwell's (2006) Theorems 1 and 2. Their proofs are sketched in the Appendix.

3 Finite Sample Properties

We investigate by simulation the finite sample properties of the ML estimator of the spatial autoregressive parameter. The unilateral first-order SAR process

$$y_{ij} = \alpha_1 y_{i-1,j} + \alpha_2 y_{i,j-1} + \varepsilon_{ij}, \quad i = 1, \ldots, N_1, \quad j = 1, \ldots, N_2 \tag{9}$$

is used as the data-generation process (DGP) in the simulation experiments. For this model the causality condition is simply (Whittle 1954): $|\alpha_1| + |\alpha_2| < 1$. We choose for α_1 and α_2 the values 0, 0.05, 0.10, 0.20 and 0.40. The errors ε_{ij} are simulated as NID$(0, 1)$. The number of observations is $N_1 = N_2 = 10$, 20, 30 and 50, so that $N = N_1 N_2 = 100$, 400, 900 and 2500. The number of replications is 1000.

The SAR model (1) is estimated with the unilateral weights matrix $\mathbf{W}_2$, which is given by (2), where the N_1 blocks of matrices $\mathbf{W}_{ii}$ on the main diagonal are $N_2 \times N_2$ matrices with ones on the subdiagonal, and all other elements are zeroes, and the N_1 blocks of matrices $\mathbf{W}_{ij}$ on the subdiagonal are $N_2 \times N_2$ identity matrices. All other blocks are zero matrices. The weights matrix $\mathbf{W}_2$ has then two non-zero elements in each row, which capture the unilateral autoregressive structure in the DGP.

Table 1 reports the mean of $\widehat{\rho}$ as an estimator of ρ, and Table 2 the mean of the bias of $\widehat{\rho}$. Note that the parameter space is symmetric in α_1 and α_2, so the entries above and below the diagonal are from a single experiment. In the cases where $\alpha_1 \neq \alpha_2$, because the weights on $y_{i-1,j}$ and $y_{i,j-1}$ are equal, $\widehat{\rho} \xrightarrow{P} 0.5(\alpha_1 + \alpha_2)$, which can be interpreted as the pseudo-true value of ρ. The bias is defined as $\text{bias}(\widehat{\rho}) = \widehat{\rho} - 0.5(\alpha_1 + \alpha_2)$. The bias is negative for all values of α_1 and α_2, except at the boundary of the parameter space

(represented by $\alpha_1 = \alpha_2 = 0.40$). For the sample size $n = 100$ the bias is quite large. For example, when $\alpha_1 = \alpha_2 = 0.2$ the bias is -0.040. The bias decreases fast with the sample size. For the sample size $n = 2500$ the bias is small, and the estimator is nearly unbiased, except at the boundary of the parameter space. For example, when $\alpha_1 = \alpha_2 = 0.2$ the bias is less than -0.001.

We now consider the cases where the weights matrix is under- or over-specified. In the case of an underspecified weights matrix, the SAR model is estimated with a unilateral weights matrix $\mathbf{W}_1$. The weights matrix $\mathbf{W}_1$ is again given by (2), where now the blocks of matrices $\mathbf{W}_{ii}$ on the main diagonal are $N_1 \times N_1$ matrices with ones on the subdiagonal, and all other elements are zeroes. All other blocks are zero matrices. The weights matrix $\mathbf{W}_1$ has only one non-zero element in each row. The bias of $\widehat{\rho}$ as an estimator of ρ is negative for all values of α_1 and α_2. Underspecifying the weights matrix results in an underestimated autoregressive parameter. It is interesting to note that the estimate of the autoregressive parameter is about half its value in the population.

In the case of an overspecified weights matrix, the SAR model is estimated with the weights matrix $\mathbf{W}_4$. The weights matrix $\mathbf{W}_4$ is again given by (2), where now the blocks of matrices $\mathbf{W}_{ii}$ on the main diagonal are $N_1 \times N_1$ matrices with ones both on the subdiagonal and on the superdiagonal, and all other elements are zeroes. The blocks of matrices $\mathbf{W}_{ij}$ on the subdiagonal and superdiagonal are $N_1 \times N_1$ identity matrices. All other blocks are zero matrices. The weights matrix $\mathbf{W}_4$ has four non-zero elements in each row. It represents a bilateral autoregressive process, but the DGP is unilateral. The bilateral process is used here solely to define an overspecified weights matrix. No conclusions may in general be drawn from this simulation experiment about the properties of the estimator of the autoregressive parameter when the data are generated by a unilateral spatial model but the model fitted to the data is bilateral. The bias $\widehat{\rho}$ as an estimator of ρ is positive for all values of α_1 and α_2. Overspecifying the weights matrix results in an overestimated autoregressive parameter. It is interesting to note that the estimate of the autoregressive parameter is about twice its value in the population.

Table 1: The mean of $\hat{\rho}$ in the SAR model (1). The DGP is given by (9). The number of replications is 1000.

α_1/α_2	0	0.05	0.10	0.20	0.40
			$N = 100$		
0	−0.006	0.014	0.035	0.076	0.160
0.05	0.014	0.034	0.054	0.096	0.183
0.10	0.035	0.054	0.075	0.117	0.207
0.20	0.076	0.096	0.117	0.160	0.257
0.40	0.160	0.183	0.207	0.257	0.377
			$N = 400$		
0	−0.003	0.020	0.042	0.087	0.182
0.05	0.020	0.042	0.065	0.110	0.208
0.10	0.042	0.065	0.087	0.134	0.236
0.20	0.087	0.110	0.134	0.184	0.298
0.40	0.182	0.208	0.236	0.298	0.456
			$N = 900$		
0	−0.001	0.023	0.046	0.093	0.190
0.05	0.023	0.046	0.069	0.117	0.218
0.10	0.046	0.069	0.093	0.142	0.248
0.20	0.093	0.117	0.142	0.194	0.313
0.40	0.190	0.218	0.248	0.313	0.483
			$N = 2500$		
0	−0.001	0.023	0.047	0.095	0.196
0.05	0.023	0.047	0.071	0.120	0.224
0.10	0.047	0.071	0.096	0.146	0.257
0.20	0.095	0.120	0.146	0.200	0.323
0.40	0.196	0.224	0.257	0.323	0.503

Table 2: The mean of the bias of $\hat{\rho}$ in the SAR model (1). The DGP is given by (9). The number of replications is 1000.

α_1/α_2	0	0.05	0.10	0.20	0.40
			$N = 100$		
0	−0.006	−0.011	−0.016	−0.025	−0.040
0.05	−0.011	−0.016	−0.021	−0.029	−0.042
0.10	−0.016	−0.021	−0.026	−0.033	−0.043
0.20	−0.025	−0.029	−0.033	−0.040	−0.043
0.40	−0.040	−0.042	−0.043	−0.043	−0.023
			$N = 400$		
0	−0.003	−0.005	−0.008	−0.013	−0.019
0.05	−0.005	−0.008	−0.011	−0.015	−0.017
0.10	−0.008	−0.011	−0.013	−0.016	−0.014
0.20	−0.013	−0.015	−0.016	−0.016	−0.002
0.40	−0.019	−0.017	−0.014	−0.002	0.056
			$N = 900$		
0	−0.001	−0.003	−0.004	−0.007	−0.010
0.05	−0.003	−0.004	−0.006	−0.008	−0.007
0.10	−0.004	−0.006	−0.007	−0.008	−0.002
0.20	−0.007	−0.008	−0.008	−0.006	0.013
0.40	−0.010	−0.007	−0.002	0.013	0.083
			$N = 2500$		
0	−0.005	−0.002	−0.003	−0.005	−0.004
0.05	−0.002	−0.003	−0.004	−0.005	−0.001
0.10	−0.003	−0.004	−0.004	−0.004	0.007
0.20	−0.005	−0.005	−0.004	−0.000	0.023
0.40	−0.004	−0.001	0.007	0.023	0.103

Table 3: The mean of $\hat{\rho}$ (upper panels) and the mean of the bias of $\hat{\rho}$ (lower panels) in the SAR model (1) when the weights matrix is underspecified, or overspecified. The DGP is given by (9). The number of observations is $N = 2500$ and the number of replications is 1000.

α_1/α_2	0	0.05	0.10	0.20	0.40
			Underspecified weights matrix		
0	0	−0.000	−0.00	−0.000	−0.001
0.05	0.025	0.025	0.025	0.026	0.029
0.10	0.050	0.050	0.050	0.052	0.058
0.20	0.100	0.100	0.100	0.103	0.119
0.40	0.212	0.213	0.215	0.223	0.272
0	0.000	−0.025	−0.050	−0.100	−0.201
0.05	0.000	−0.025	−0.050	−0.099	−0.196
0.10	−0.000	−0.025	−0.050	−0.099	−0.192
0.20	−0.001	−0.025	−0.050	−0.097	−0.181
0.40	0.012	−0.012	−0.035	−0.077	−0.128
			Overspecified weights matrix		
0	−0.001	0.046	0.094	0.190	0.376
0.05	0.046	0.094	0.143	0.237	0.424
0.10	0.094	0.143	0.191	0.285	0.469
0.20	0.191	0.238	0.285	0.377	0.552
0.40	0.376	0.424	0.470	0.552	0.723
			0		
0	−0.001	0.021	0.044	0.090	0.176
0.05	0.021	0.044	0.068	0.112	0.199
0.10	0.044	0.068	0.091	0.135	0.219
0.20	0.091	0.113	0.135	0.177	0.252
0.40	0.176	0.199	0.220	0.252	0.323

4 Application to House Prices

In this section we provide an example of the use of unilateral spatial econometric models with real data. Houses located near one another may have similar unobservable attributes. This type of spatial dependence may be modelled by a spatial error model (SEM) (Dubin 1988, Sedgley et al. 2008, Wilhelmsson 2002). In the SEM model the spatial dependence in the data is modelled through the error term:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \\ \mathbf{u} &= \lambda \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}, \end{aligned} \tag{10}$$

where $\mathbf{y} = [y_{ij}]$ an $N \times 1$ vector of observations on the dependent variable, $\mathbf{X} = [x_{ij1}, \ldots, x_{ijm}]$ is an $N \times m$ matrix of explanatory variables, $\mathbf{u} = [u_{ij}]$ is an $N \times 1$ vector of errors, $\boldsymbol{\beta}$ is an $m \times 1$ parameter vector, λ is a spatial autoregressive parameter, and $\mathbf{W}$ and $\boldsymbol{\varepsilon}$ are as before.

The LS estimator of $\boldsymbol{\beta}$ is unbiased and consistent in the SEM model, but is not efficient, and the standard errors will be biased. The model is therefore estimated by ML.

We use the data from Wilhelmsson (2002). The data consist of 1377 transactions of single-family houses between January 2000 and May 2001 in the county of Stockholm in Sweden. The data include the selling price, spatial coordinates, the size of the house in square metres, as well as other characteristics. See Wilhelmsson for a detailed description of the data.

The data do not lie equally spaced on a grid, as we have assumed in the previous sections, so modelling the spatial dependence is not as straightforward as before. The data include coordinates (longitudes and latitudes) for every observation. Figure 1 plots the coordinates of the observations. By computing distances between the observations, we can find the (say) k nearest neighbours of an observation. Following the analysis in the previous sections, we choose the weights matrix to be unilateral. Consequently, the dependence is allowed to go in only one direction, i.e. if the observation kl depends on ij, then ij is not allowed to depend on kl, and vice versa. The somewhat subjective choices about the direction of dependence and the number of neighbours can be made in many ways, and should depend on the application at hand. For the choice of the weights matrix in a k-nearest neighbours setting, see Gerkman (2010). Here we choose the observations to

Figure 1: Plot of the coordinates for the house prices data in the county of Stockholm in Sweden. Each point correspons to one observation. The number of observations is $n = 1377$.

have neighbours with smaller coordinates in latitude and larger coordinates in longitude. The direction of dependence is thus chosen to be from north-west to south-east in Figure 1. The objective is to ensure that the neighbours of an observation are as close as possible to the observation in question, given the restriction of unilateral dependence. However, choosing other directions of dependence gives very similar results, and are therefore not reported.

We estimate the SEM model entertaining the unilateral weights matrices $\mathbf{W}_k$, where $k = 0, 2, 4, 8, 12, 14, 16, 20$. All weights matrices are row-normalised. The weights matrix $\mathbf{W}_0$ refers to the model without spatial errors, and then the maximum likelihood estimator of $\boldsymbol{\beta}$ is equivalent to the LS estimator. For comparison, the corresponding k-nearest bilateral models are also reported.

Table 4 reports the estimated autoregressive parameter and model summary statistics. We find that the magnitude of the estimated spatial autoregressive parameter λ crucially depends on the choice of the weights matrix. This result is not a surprise, given the simulation results in Section 3. If we compare the models with two and four neighbours, we find that in the

model with two neighbors $\widehat{\lambda} = 0.288$ and in the model with four neighbours $\widehat{\lambda} = 0.475$, i.e. almost than twice as large. When more neighbours are added, $\widehat{\lambda}$ increases further, but with a decreasing rate. All estimates $\widehat{\lambda}$ in the unilateral models are such that the causality condition ($\lambda < 1$) is satisfied.

Comparing the unilateral models with the bilateral models, we find that for all values of k the estimated autoregressive parameter λ and the model summary statistics are very similar. The bilateral models provide a marginally better fit, though. Also the estimates of the parameter $\boldsymbol{\beta}$ (not reported) are very similar. We note that, if the strategy for selecting the weights matrix proposed in Gerkman (2010) is used, the value $k = 14$ is found. Then we find that $\widehat{\lambda} = 0.722$ in the unilateral model and $\widehat{\lambda} = 0.746$ in the bilateral model.

The main conclusion is that the Swedish house prices data can successfully be modelled by a unilateral spatial error model, and that a model with a unilateral weights matrix gives almost identical empirical results as a model with a bilateral weights matrix. In fact, Pace et al. (2009) argue that in large samples the exact specification of the weights matrix is not so important, since different weights matrices do not yield materially different coefficient estimates. It may therefore be a moot point whether a unilateral or bilateral weights matrix should be preferred, and then we may use a unilateral one, because it simplifies the analysis to a great extent.

5 Conclusions

In this paper we show that by restricting the attention to unilateral spatial autoregressive processes, we may define a unilateral spatial autoregression which enjoys similar properties as an autoregression with time series. First, we give a causality condition for unilateral spatial autoregression based on the theory of unilateral spatial autoregressive processes. Second, we investigate the properties of the maximum likelihood estimator of the spatial autoregressive parameter. The estimator is shown to be consistent and asymptotically normal, without requiring high level assumptions about moment conditions, ergodicity and mixing. Some simulation results to examine the finite sample properties of the estimator show that it is nearly unbiased, except at the boundary of the parameter space. Third, we investigate the impact of the choice of the weights matrix on the estimated spatial autoregressive parame-

Table 4: The estimates of the spatial autoregressive parameter and model summary statistics for unilateral and bilateral spatial error models entertaining different k-nearest neighbours weights matrices. The table shows the results for the unilateral models over the results for the bilateral models.

	$\mathbf{W}_0$	$\mathbf{W}_2$	$\mathbf{W}_4$	$\mathbf{W}_8$	$\mathbf{W}_{12}$	$\mathbf{W}_{14}$	$\mathbf{W}_{16}$	$\mathbf{W}_{20}$
$\widehat{\lambda}$		0.288	0.475	0.640	0.699	0.722	0.739	0.767
		0.307	0.482	0.650	0.719	0.746	0.779	0.854
t-stat		21.3	45.8	89.9	118.2	133.6	146.6	173.4
		32.5	46.3	90.0	125.9	138.2	146.7	126.0
$\widehat{\sigma}_\varepsilon$	0.074	0.065	0.061	0.059	0.059	0.060	0.060	0.060
		0.064	0.060	0.056	0.055	0.055	0.055	0.055
–log lik		373.6	408.0	439.4	439.6	439.5	438.7	439.0
		377.9	414.2	461.8	475.1	480.3	483.5	482.6
R^2	0.581	0.627	0.651	0.665	0.661	0.660	0.658	0.657
		0.633	0.658	0.682	0.686	0.686	0.687	0.685

ter. We find that the estimated spatial autoregressive parameter crucially depends on the weights matrix. Underspecifying the weights matrix results in an underestimated spatial autoregressive parameter, and overspecifying the weights matrix results in an overestimated spatial autoregressive parameter. Finally, the use of unilateral spatial autoregression with real data is illustrated by an empirical application to house prices.

References

[1] Anselin, L. (1988) *Spatial Econometrics: Methods and Models*, Dordrecht: Kluver Academic Publishers.

[2] Anselin, L. (2003), Ch. 14, Spatial Econometrics, in Baltagi, B. H. (ed.), *A Companion to Theoretical Econometrics*, Blackwell Publishing, Oxford, 310–330.

[3] Basu, S. and Reinsel, G. C. (1993), Properties of the Spatial Unilateral First-Order ARMA Model, *Advances in Applied Probability*, 25, 631–648.

[4] Brockwell, P.J. and Davis, R.A. (1991), *Time Series: Theory and Methods*, NY, Springer 2nd ed.

[5] Cliff, A. D. and Ord, J. (1969) The Problem of Spatial Autocorrelation. In *London Papers in Regional Science*, edited by A. Scott, 25–55. London: Pion.

[6] Dubin, R. (1988) Estimation of Regression Coefficients in the Presence of Spatially Autocorrelated Error Terms, *The Review of Economics and Statistics,* Vol. 70, Issue 3, 466–474.

[7] Fingleton, B. (2009) Spatial Autoregression, *Geographical Analysis,* 41, 385–391.

[8] Gerkman, L. (2010), A Practical Proposal to Specification Search of a k-Nearest Neighbours Weights Matrix, Manuscript, Hanken School of Economics, Helsinki.

[9] Hannan, E.J. (1973), The Asymptotic Theory of Linear Time-Series Models, *Journal of Applied Probability*, 10, 130–145.

[10] Kelejian, H. H. and Prucha, I. R. (1998), A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances, *Journal of Real Estate Finance and Economics*, 17, 99–121.

[11] Kelejian, H. H. and Prucha, I. R. (1999), A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model, *International Economic Review*, 40, 2, 509–533.

[12] Lee, L.–F. (2002), Consistency and Efficiency of Least Squares Estimation for Mixed Regressive, Spatial Autoregressive Models, *Econometric Theory*, 18, 252–277.

[13] Lee, L.–F. (2003), Best Spatial Two-Stage Least Squares Estimators for a Spatial Autoregressive Model with Autoregressive Disturbances, *Econometric Reviews*, 22, 307–335.

[14] Lee, L.–F. (2004), Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Autoregressive Models, *Econometrica*, 72, 1899–1925.

[15] LeSage, J.P. and Pace, R. K. (2009) *Introduction to Spatial Econometrics.*, CRC Press, Taylor and Francis Group.

[16] Mynbaev, K. T. (2010), Asymptotic distribution of the OLS estimator for a mixed spatial model, *Journal of Multivariate Analysis*, 101, 3, March 2010, 733–748.

[17] Pace, R. K., LeSage, J. P. and Zhu, S (2009) Impact of Cliff and Ord on the Housing and Real Estate Literature, *Geographical Analysis*, 41, 418–424.

[18] Sedgley, N. H., Williams, N. A. and Derrick, F. W. (2008) The Effect of Educational Test Scores on House Prices in a Model with Spatial dependence, *Journal of Housing Economics*, 17, 191–200.

[19] Tjøstheim, D. (1978), Statistical Spatial Series Modelling, *Advances in Applied Probability*, 10, 130–154.

[20] Tjøstheim, D. (1981) Autoregressive Modeling and Spectral Analysis of Array Data in the Plane, *IEEE Transactions on Geoscience and Remote Sensing,* 19, 15–24.

[21] Tjøstheim, D. (1983) Statistical Spatial Series Modelling II: Some further Results in Unilateral Processes, *Advances in Applied Probability,* 15, 562–684.

[22] Whittle, P. (1954), On Stationary Processes in the Plane, *Biometrika*, 41, 434–449.

[23] Wilhelmsson, M. (2002), Spatial Models in Real Estate Economics, *Housing, Theory and Society*, 19, 92–101.

[24] Yao, Q. and Brockwell, P. J. (2006), Gaussian Maximum Likelihood Estimation for ARMA Models II: Spatial Processes, *Bernoulli*, 12, 403–429.

6 Appendix: Proofs

In Section 2 we have seen that the unilateral SAR model (1) satisfying Assumption 1 is a special case of a unilateral SAR process in the plane (3). The consictency and asymptotic normality of the estimator $\widehat{\rho}$ obtained from maximizing (7) may be obtained from Theorems 1 and 2 of Yao and Brockwell (2006) as a special case.

We begin by reviewing the results obtained by Yao and Brockwell for a general causal and invertible spatial autoregressive and moving average process in the two-dimensional plane. They establish consistency and asymptotic normality of the Gaussian maximum likelihood estimator. The limiting distribution is determined by two AR models defined by the AR and MA forms in the original model. Their result is an analogue of Hannan's (1973) classic result for time series in the context of spatial processes (see also Brockwell and Davis 1991, Section 10.8).

For a causal spatial autoregressive processes their result is given below as a special case.

Let $\mathbf{y}_{-1} = (y_{i-1,j}, y_{i,j-1}, \ldots, y_{i-p_1,j-p_2})'$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{p_1+p_2})'$.

Theorem 1 *Let* $\{\varepsilon_{ij}\} \sim$ *IID*$(0, \sigma^2)$ *and conditions (C1) and (C2) hold. Then*

$$N^{1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{D} N(\mathbf{0}, \mathbf{W}(\boldsymbol{\theta}_0)),$$

where the asymptotic covariance matrix is given by

$$\mathbf{W}(\boldsymbol{\theta}_0) = [E(\mathbf{y}_{-1}\mathbf{y}'_{-1})]^{-1}. \tag{11}$$

Proof of Proposition 3. Write the SAR model (1) as a unilateral SAR process

$$y_{ij} = \rho \sum_{k,l} \sum_{k=l\neq 0} y_{i-k,j-l}. \tag{12}$$

Applying Theorem 1 with y_{ij} as in (12), we obtain the conclusion

$$N^{1/2}(\widehat{\rho} - \rho_0) \xrightarrow{D} \mathrm{N}(\rho_0, v(\theta_0)),$$

where the asymptotic variance $v(\theta)$ is given by

$$v(\theta) = \left\{ \mathrm{E} \left[\sum_{k,l} \sum_{k=l\neq 0} y_{i-k,j-l} \right]^2 \right\}^{-1}. \tag{13}$$

■

Small Scale Neighbourhood in Spatial Econometrics

Linda Gerkman
Hanken School of Economics,
Department of Finance and Statistics

Abstract

This paper studies to what extent a set of coordinate-based variables describing the small scale neighbourhood conditions can replace the spatial econometric structure in a house price model. We make assumptions that motivate a spatial Durbin model which has a rich spatial structure. When it is estimated without the small scale neighbourhood variables, its spatial lags are found to be significant. However, when the additional set of variables is included, the spatial lags become insignificant. Hence, the model would be reduced to a conventional non-spatial regression model. The assumptions still motivate a Spatial Error model. The spatial lag of the error term of this model is found to be highly significant. The paper gives a formal motivation for the empirical analysis and the arguments are implemented on real data for apartments sold in Helsinki in Finland, during the first quarter of 2002. In the empirical analysis we find that the new explanatory variables capture some of the small scale neighbourhood conditions, and thereby we obtain a simplified model. However, these small scale neighbourhood variables cannot entirely replace the spatial econometric structure.

Key Words: House prices, Omitted variables, Small scale neighbourhood, Spatial econometrics.

1 Introduction

According to real estate agents there are three things that affect the price of a house the most: location, location and location. This remark suggests

that location and small scale neighbourhood have a substantial impact on the price of a house or an apartment. Earlier studies of the housing market have found that geographical location is important. Especially the small scale neighbourhood has a substantial effect on the price a house or an apartment. But how should one measure small scale neighbourhood attributes such as well maintained surroundings, the availability of parks for recreation, the level of service, communication networks, good schools or simply nice neighbours? These variables are often both hard to define and measure. Further, small scale neighbourhood variables are spatially dependent, which means that observations are very similar for objects located close to each other. The problem is how to capture and formulate this type of spatial dependence in an econometric model.

Recent technical developments of geographical information systems (GIS) have brought about a renewed interest in spatial matters. For instance, it is common to attach coordinates to data. The coordinates can be used to create a spatial weights matrix in a spatial econometric model. In this paper a different view is taken. The coordinates of the observations and the information in the original data are used to create new variables by following the ideas in Turnbull, Dombrow and Sirmans (2006). The purpose is to examine whether these coordinate-based variables, which measure small scale neighbourhood market conditions, can replace the spatial structure in a hedonic house price model. We begin by considering a scenario where the assumptions are that there are important explanatory variables missing from the model, the omitted variables are spatially autocorrelated and, in addition, they are correlated with the explanatory variables in the model. These assumptions lead to a spatial Durbin model. Then we uncover some of the omitted variables, namely the coordinate-based variables capturing the small scale neighbourhood market conditions. Despite the uncovered new variables, we expect that there still are spatially autocorrelated explanatory variables missing from the model, but we find reason to assume that the remaining omitted variables are uncorrelated with the explanatory variables in the model. These assumptions lead to a spatial error model. The arguments of the theoretical analysis are implemented on real house price data from Helsinki, Finland, to demonstrate their usefulness in practice. The empirical results show that it is possible to find new explanatory variables and obtain a simplified model. The benefit of the final model is that it is easier to interpret than a spatial econometric model with a richer structure.

The paper is organized as follows. The next section describes hedonic

price models for housing. The section also briefly introduces spatial econometric models. The third section discusses the assumptions and the implications of them, and explains the arguments and the theoretical motivations for the models which are used in the empirical analysis. In the fourth section the data and the variables, especially the small scale neighbourhood variables, are discussed. The fifth section contains the empirical analysis and the last section concludes.

2 Hedonic House Price Models and Spatial Econometrics

A hedonic price model is a regression model where the price of a commodity is explained by the attributes of the commodity in question (Rosen 1974). The relevant attributes for a hedonic house price model can be grouped into three broad categories: location variables, structural variables and neighbourhood variables (Dubin 1988). Location variables are attributes describing the geographical location of the object. They are usually easily measured and can, for instance, be the presence or absence of sea view, the distance to the central business district or the nearest train station. Structural variables are attributes of the house or apartment itself. Typical examples are the size of the house or apartment, its age, number of rooms, the condition of the interior, lot size and the presence or absence of garage, balcony, fireplaces, and so on. Whereas the first two categories are usually easily measured, the third group of variables, neighbourhood variables, can be thought of as latent characteristics of a neighbourhood. By these type of variables we try to capture the spirit that makes some neighbourhoods more appreciated than others, even when the location and structural attributes are similar for the neighbourhoods. They are the most difficult ones to obtain. Commonly used variables for this purpose are the local crime rate, socioeconomic characteristics of residents, pollution or noise levels. These variables may be thought of as proxies of the unobservable quality of the small scale neighbourhood.

Price models for housing are often of the loglinear form (Wilhelmsson 2002) or semilog form (Pace and Gilly 1997). A hedonic regression model in semi-log form is given by

$$\ln P_i = \alpha + \sum_{m=1}^{k} \beta_m x_{mi} + \sum_{d=1}^{l} \gamma_d D_{di} + \varepsilon_i, \qquad i = 1, ..., N, \tag{1}$$

where the logarithm of the price P_i is explained by the explanatory variables $x_1, x_2, ...x_k$ and dummy variables.

In practice it is seldom possible to obtain all the variables needed according to the theory. If we lack some important variables, the least squares estimates of the parameters in (1) will be biased and inconsistent, since the estimated coefficients of the variables in the model will pick up the part of the influence of the omitted variables that is correlated with the variables in the model. A reasonable assumption is that objects situated near each other have similar values on location and neighbourhood variables, i.e. these variables are spatially autocorrelated. If the omitted variables are spatially autocorrelated, Pace and LeSage (2008) show that the bias is going to be even larger than in the conventional least squares case. When the omitted variables are spatially correlated, spatial econometric models can be used in order to reduce the bias of the estimates.

Spatial effects occur when the geographical closeness of observations affect the degree of dependence between the observations. When two points on a map are close to each other, it is natural that also the observations made at those points are similar. The further away the two points are from each other, the less similar the observations made at those points tend to be. In spatial econometric models the weights matrix plays an important role. A weights matrix is usually assumed to be exogenous to the model and is typically based on the geographic arrangements of the observations. It specifies which of the other locations in the system that are assumed to affect the observed value at some location. The weights matrix can be based on distances between the observations or on contiguity.

In the empirical analysis in Section 5, we use a row normalized weights matrix that is based on nearest neighbours. The coordinates of every observation give the exact position of the apartment and hence, the exact position of all neighbouring apartments. When creating the weights matrix every observation is assumed to have four nearest neighbours. This is one of the standard ways to specify a weights matrix in the spatial econometric literature, see for instance Anselin (1988). An observation is also assumed to be equally spatially dependent of all four neighbours. The dependence structure does not consider the difference in distance between the observation and its neighbours. Since every observation is assumed to have exactly four neighbours, it is possible that false neighbours are imposed on some observations. On the other hand, some observations may lack important neighbours. It would of course be possible to analyze other weights matrices, or the impact

of different numbers of nearest neighbours on the estimates, but this is not the focus of this paper and therefore left to future studies.

A standard spatial econometric model, see for instance Anselin (1988), is the general spatial autoregressive model which given in matrix notation is

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{W}_1\mathbf{y} + \mathbf{u}, \\ \mathbf{u} &= \lambda\mathbf{W}_2\mathbf{u} + \boldsymbol{\varepsilon}, \\ \boldsymbol{\varepsilon} &\sim N(\mathbf{0}, \sigma^2\mathbf{I}), \end{aligned} \tag{2}$$

where the dependent variable $\mathbf{y}$ is explained by the exogenous variables in the matrix $\mathbf{X}$ and a spatial lag $\mathbf{W}_1\mathbf{y}$. The Matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ are weights matrices. The error term $\mathbf{u}$ is spatially autocorrelated.

Applied to the hedonic price model (1) and with $\mathbf{y} = [\ln P_i]$, the model (2) becomes

$$\begin{aligned} \ln P_i &= \alpha + \sum_{m=1}^{k} \beta_m x_{mi} + \sum_{d=1}^{l} \gamma_d D_{di} + \rho \sum_{j=1}^{n} w_{1ij} \ln P_j + u_i, \\ u_i &= \lambda \sum_{j=1}^{n} w_{2ij} u_j + \varepsilon_i. \end{aligned} \tag{3}$$

In this model w_{1ij} and w_{2ij} are elements in the ith row and jth column of the weights matrices $\mathbf{W}_1$ and $\mathbf{W}_2$, respectively. When the weight matrix is row normalized, the spatial lag of the dependent variable, $w_{1ij} \ln P_j$, can be interpreted as a weighted average of the neighbouring observations on the dependent variable.

The solution of the model (2) is

$$\mathbf{y} = (\mathbf{I} - \rho\mathbf{W}_1)^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \rho\mathbf{W}_1)^{-1}(\mathbf{I} - \lambda\mathbf{W}_2)^{-1}\boldsymbol{\varepsilon}, \tag{4}$$

provided that $(\mathbf{I} - \rho\mathbf{W}_1)$ and $(\mathbf{I} - \lambda\mathbf{W}_2)$ are non-singular, which is the case when $|\rho| < 1$, $|\lambda| < 1$ and the weights matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ are row normalized. In practice it is often assumed that $\mathbf{W}_1 = \mathbf{W}_2$. Kelejian (2008) points out that the associated identification problem occurs only when $\boldsymbol{\beta} = \mathbf{0}$. Then the parameters ρ and λ are not separately identified. In practice this is almost never the case, since there are always exogenous regressors in the model. If $\rho = 0$, the model (2) is called a spatial error model and if $\lambda = 0$, the model is referred to as a spatial lag model. If both $\rho = 0$ and $\lambda = 0$, the model is a linear regression model.

3 Implications of Spatially Autocorrelated Omitted Variables

When modelling house price data the model is very often lacking some location and neighbourhood variables. According to LeSage and Pace (2009), there is then a strong motivation for spatial econometric models, since these omitted location and neighbourhood variables are usually thought of as spatially autocorrelated. The following analysis is based on arguments adapted from LeSage and Pace (2009).

First we will consider a scenario where the assumptions are that there are important explanatory variables missing from the model, the omitted variables are spatially autocorrelated and, in addition, that they are correlated with the explanatory variables in the model. These assumptions will lead to a spatial Durbin model. Then we will assume that we uncover some of the omitted variables. In this second scenario, we still assume that there are variables missing from the model despite the included uncovered new variables, and that they are spatially autocorrelated, but now we assume that the remaining omitted variables are uncorrelated with the explanatory variables in the model. These assumptions will lead to a spatial error model. In Section 5 the results of the theoretical analysis below are used on real data to demonstrate the usefulness of the arguments. The objective is to show that by uncovering new spatially autocorrelated explanatory variables with certain properties, we can obtain a simpler model, which is easier to interpret.

Begin by considering the following non-spatial model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_1 + \mathbf{Z}\boldsymbol{\beta}_2, \tag{5}$$

where the $n \times 1$ vector $\mathbf{y}$ is the dependent variable, the $n \times k$ matrix $\mathbf{X}$ represents the available explanatory variables, the $n \times m$ matrix $\mathbf{Z}$ represents missing or unobserved explanatory variables, and $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are vectors containing regression coefficients. For ease of exposition there is no error term in the model, instead the variables in $\mathbf{X}$ and in $\mathbf{Z}$ are assumed to completely explain the dependent variable. Then assume that the missing variables in the matrix $\mathbf{Z}$ are spatially autocorrelated , i.e. that we have

$$\mathbf{Z} = \rho\mathbf{W}\mathbf{Z} + \mathbf{R}, \quad E(\mathbf{R}) = \mathbf{0}. \tag{6}$$

The scalar parameter ρ in (6) is the spatial autocorrelation coefficient and $\mathbf{W}$ is a weights matrix. From (6) solve $\mathbf{Z} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{R}$, and substitute it

into (5) to obtain

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_1 + (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{R}\boldsymbol{\beta}_2. \tag{7}$$

Define the vector $\mathbf{u} = \mathbf{R}\boldsymbol{\beta}_2$. Further assume that the missing variables in $\mathbf{Z}$ correlate with the explanatory variables $\mathbf{X}$ in model (5). Assume a simple linear dependence,

$$\mathbf{u} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{v}, \quad \mathbf{v} \sim N(0, \sigma_v^2\mathbf{I}). \tag{8}$$

The parameters in the vector $\boldsymbol{\gamma}$ and the variance of the error term σ_v^2 in (8) describes the relation between $\mathbf{X}$ and $\mathbf{Z}$. Since $\mathbf{R}\boldsymbol{\beta}_2 = \mathbf{u}$, we can substitute (8) into (7) and obtain

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta}_1 + (\mathbf{I} - \rho\mathbf{W})^{-1}(\mathbf{X}\boldsymbol{\gamma} + \mathbf{v}) \\ &= \mathbf{X}\boldsymbol{\beta}_1 + (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\gamma} + (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{v}. \end{aligned} \tag{9}$$

Multiply from the left by $(\mathbf{I} - \rho\mathbf{W})$,

$$(\mathbf{I} - \rho\mathbf{W})\mathbf{y} = (\mathbf{I} - \rho\mathbf{W})\mathbf{X}\boldsymbol{\beta}_1 + \mathbf{X}\boldsymbol{\gamma} + \mathbf{v}, \tag{10}$$

and hence,

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}(\boldsymbol{\beta}_1 + \boldsymbol{\gamma}) + \mathbf{W}\mathbf{X}(-\rho\boldsymbol{\beta}_1) + \mathbf{v}. \tag{11}$$

Model (11) is called a spatial Durbin model (SDM). LeSage and Pace (2009) point out that the interpretation of the parameters in models which includes spatial lags of the explanatory or dependent variable becomes richer and more complicated. For instance, the estimated regression coefficients can not be interpreted in the conventional regression sense as partial derivatives.

Now consider the scenario where we assume that some of the variables, say $\mathbf{Z}_1$, of the matrix $\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}^*]$ in (5) are revealed and we get a new matrix $\mathbf{X}^* = [\mathbf{X}, \mathbf{Z}_1]$ of explanatory variables. Instead of model (5) we have

$$\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta}_1^* + \mathbf{Z}^*\boldsymbol{\beta}_2^*. \tag{12}$$

Assume that the remaining missing variables $\mathbf{Z}^*$ are spatially autocorrelated, as before, i.e. that we have

$$\mathbf{Z}^* = \rho^*\mathbf{W}\mathbf{Z}^* + \mathbf{R}^*, \quad E(\mathbf{R}^*) = \mathbf{0}, \tag{13}$$

but that the missing variables $\mathbf{Z}^*$ are now uncorrelated with the explanatory variables $\mathbf{X}^*$ in the model. Solve $\mathbf{Z}^*$ from (13) and substitute into (12) to get

$$\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta}_1^* + (\mathbf{I} - \rho^*\mathbf{W})^{-1}\mathbf{R}^*\boldsymbol{\beta}_2^*. \tag{14}$$

Define $\mathbf{u}^* = \mathbf{R}^*\boldsymbol{\beta}_2^*$ and $\mathbf{q} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{u}^*$. Then (14) can be written as a spatial error model (SEM),

$$\begin{aligned} \mathbf{y} &= \mathbf{X}^*\boldsymbol{\beta}_1^* + \mathbf{q} \\ \mathbf{q} &= \rho^*\mathbf{W}\mathbf{q} + \mathbf{u}^*. \end{aligned}$$

In the SEM the regression coefficients have their usual interpretation as partial derivatives, since the model does not involve any lag on the dependent variable or the explanatory variables. If the coefficients are estimated by the method of ordinary least squares they will be unbiased, but not efficient, since $\boldsymbol{\beta}_2^*$ will increase the variance of $\mathbf{R}^*$. Maximum likelihood estimates will be unbiased and efficient.

In this section we have seen that if some new explanatory variables with certain properties are uncovered the model changes from a SDM to a simpler SEM. In the empirical analysis of this paper we will demonstrate the scenarios mentioned above. We will see how house price models are changed when new explanatory variables $\mathbf{Z_1}$ are revealed. In the next section the data are presented and special attention is given to the formal definition of our candidates for $\mathbf{Z_1}$, the new small scale neighbourhood variables mentioned in Section 1.

4 Data

The empirical analysis is based on cross-sectional data on the free from debt selling price of apartments in residential buildings in Helsinki, the capital of Finland, in January, February and March 2002. The data include 649 transactions and are based on information provided by real estate agents in the area. The data were originally collected to be used only by real estate agents. The idea was that only agents that gave particulars of their own transactions could get the particulars of the transactions of their competitors. Since not all real estate agents participated in the service, the data do not include all transactions during the period.

The original data include 15 attributes describing the apartment, its location or the transaction. All variables are presented in *Table 1.* The dependent variable is the selling price ($Price$). During the first quarter of 2002 the average apartment in Helsinki was sold for €130000. The attributes are the size of the apartment in square metre ($M2$), the number of rooms excluding kitchen ($Rooms$) and the age of the apartment (Age). Apartments that are

less than one year old are not available in this data set and are therefore not included. This is also typical for house price data from Statistics Finland and the motivation is that the selling prices for new apartments are strongly affected by building costs. A variable age squared ($Age2$) is created since the effect of age is not assumed to be linear. The average apartment is 55 square metres, has 2 rooms and is 49 years old. Note the large standard deviations of these variables. The data also include information about the housing loan ($Loan$), the maintenance charge per month ($Maintenance$), the floor number of the apartment ($Floor$) and the number of floors in the building ($FloorB$). The number of rooms, the floor number and the number of floors in the building are all transformed into dummy variables before entering the model. The average number of floors is 4.6, which is reasonable since there are no skyscrapers in Finland. The highest building has only 14 floors. The time an apartment is for sale (WOM) fluctuated between zero and 74 weeks, and the average time on the market is about 8 weeks. The data also include dummy variables for sauna, balcony, alcove and elevator. It can be seen that 5% of the apartments had a sauna, 18% had a balcony and about 50% of the apartments were situated in buildings with elevator. Before entered in the model the dummy variable for elevator is combined with the information whether the apartment is situated above the second floor or not, since it is expected that the elevator is relevant only when an apartment is situated at a higher floor. There is also information on the condition of the apartment. The condition of the apartments are graded *good*, $satisfactory$ or *bad*. This information may not be as reliable as for the other attributes, since it is based on real estate agents' subjective grading of the apartments. Note also that in 21% of the cases the information on the condition is missing. Finally there is information about wether the apartment was sold in January, February or March 2002. The number of apartments sold are about the same during all three months. The data set contains the addresses of the apartments. The address for each location is changed into coordinates, given in longitudes and latitudes.

The last four variables in *Table 1* are of special interest in this study, since they are our candidates for the missing variables in $\mathbf{Z_1}$. The variables $Smaller$ and $Larger$ capture the relative size effects. Following an idea suggested by Turnbull, Dombrow and Sirmans (2006), the variables measure the negative and positive deviations between the size of an apartment i and the mean of the apartment size in the small scale neighbourhood, relative to the mean of the apartment size in the small scale neighbourhood. Turnbull,

Table 1: Descriptive statistics for transactions of apartments in residental buildings sold in Helsinki January 2002 to March 2002.

Variable	Minimum	Maximum	Mean	Std. Dev.
Price (€)	31500.00	1086750.00	129959.00	85003.00
M2 (m^2)	14.00	189.00	54.65	24.48
Age (Years)	1.00	116.00	49.06	22.82
Loan (€)	0.00	64802.00	1469.15	5351.14
Maintenance (€)	0.00	734.35	119.56	89.89
Rooms	1.00	6.00	2.08	0.95
Floor (number)	0	11.00	2.63	1.75
FloorsB	0	14.00	4.60	1.93
WOM (weeks)	0	74.00	7.96	9.92
Sauna	binary		0.05	
Balcony	binary		0.18	
Alcove	binary		0.04	
Elevator	binary		0.51	
Condition/good	binary		0.43	
Condition/satisf.	binary		0.31	
Condition/bad	binary		0.05	
Condition/missing	binary		0.21	
January	binary		0.32	
February	binary		0.35	
March	binary		0.33	
Valid N=649				

Small Scale Neighbourhood Variables				
Smaller	0	0.78	0.28	0.22
Larger	0	1.78	0.07	0.22
Comp	0	672.42	51.90	79.39
Avgcomp	0	47.48	7.54	7.30

Dombrow and Sirmans define the immediate neighbourhood to be all houses within one-half-mile radius of house i, and calculate the mean of the size of the apartments. In order to utilize more information, this study uses an other data set from Statistics Finland in addition to the data set described above. It contains all sales for the year 1998 for Helsinki, and has 3305 observations. The data set for the year 2002 include 77 different postal districts. The mean of the apartment size in each postal district is calculated from the pooled data set and the mean is then used as a proxy for the small scale neighbourhood mean.

For an apartment i that belongs to postal district J, the standardized measure of relative size is

$$Localsize_i = \frac{M2_i - Mean(M2 \textit{ in postal district } J)}{Mean(M2 \textit{ in postal district } J)}.$$

The relative size variables $Smaller_i$ and $Larger_i$ are defined as follows:

$$\begin{aligned} Smaller_i &= |Localsize_i| \text{ for } Localsize_i < 0, \\ &= 0 \text{ otherwise,} \end{aligned}$$

and

$$\begin{aligned} Larger_i &= |Localsize_i| \text{ for } Localsize_i > 0, \\ &= 0 \text{ otherwise.} \end{aligned}$$

The small scale neighbourhood is also assumed to be reflected by the amount of competing sales in the immediate surroundings. Competition is measured by the number of apartments that are for sale at the same time. The variables which are used for the purpose are the cumulative amount of competition ($Comp$) and its average intensity ($Avgcomp$). They are modifications of the variables suggested by Turnbull, Dombrow and Sirmans. The definition of a competing apartment j is one that has a living area ($M2$) that is no more than 20% smaller or no more than 20% larger than the living area of the apartment for sale i. The competition is restricted to apartments within a distance of at most two kilometers. Since the data contain the coordinates, the competing apartments are easily found. The data do not include the exact listing and selling dates and therefore a proxy variable is used for the overlapping marketing time. The data contain the marketing time in weeks (WOM) for each apartment and information on the month it is sold. These

two variables are combined to get a proxy for the overlapping marketing time. For all apartments that are sold in January the selling date is set to 31 January and for February and March, 28 February and 31 March, respective. Keeping selling dates fixed, it is possible to get a proxy for overlapping marketing time for each pair of competing apartments by looking at the number of weeks on market for both i and j.

The variable $Comp_i$ measures the cumulative competition (in weeks) from other apartments $j \in I$ in the small scale neighbourhood I of apartment i :

$$Comp_i = \sum_{j \in I} (2 - D(i,j))^2 O(i,j).$$

Here $D(i,j)$ is the absolute distance in kilometers between observations i and j, if i and j are competing apartments within at most two kilometers from each other, otherwise $D(i,j)$ is set to zero. The variable $O(i,j)$ is the overlapping marketing time in weeks for each pair of competing apartments. If the overlapping marketing time is zero, or if i and j are more than two kilometers apart, $O(i,j)$ is set to zero. The factor $(2 - D(i,j))^2$ ensures that apartments further away from apartment i get less weight than apartments closer to i.

The other measure of competition in the small scale neighbourhood is *Avgcomp*. The variable measures the average intensity of the competition as the cumulative competition in weeks divided by the weeks of marketing time. For apartment i

$$Avgcomp_i = \sum_{j \in I} \frac{(2 - D(i,j))^2 O(i,j)}{WOM_i}.$$

According to Turnbull, Dombrow and Sirmans these two variables control for the window of opportunity for buyers who might be interested in any of the competing houses. Thereby they should also be able to capture the spatial dependence in the data arising from the fact that nearby observations have similar or the same market conditions. An interesting matter is that the hypothesis concerning the sign of the impact of competition on the selling price is not straightforward. When there are lots of similar apartments for sale in a neighbourhood, the buyer tends to buy the most reasonably priced apartment. In order to stay in the competition other sellers will have to lower their prices. So the conclusion is that competition in the amount of apartments will lower the selling price. On the other hand, buyers who are

attending a showing of an apartment also tend to attend showings of nearby apartments, because they are already in the neighbourhood. This means that if there is a lot of apartments for sale in a neighbourhood, it will attract more potential buyers. And if there are lots of buyers competing for the same apartments the prices will go up.

The small scale neighbourhood variables, *Smaller*, *Larger*, *Comp* and *Avgcomp* are our candidates for the missing explanatory variables $\mathbf{Z_1}$, mentioned in the previous section. Based on the definitions given above, we have reason to assume that the variables will have the desired properties, which are, that they are spatially autocorrelated and that they are correlated with the explanatory variables in the model. In the next section these assumptions are tested and found to hold.

5 The Empirical Analysis

In the empirical analysis we will concentrate on the scenarios mentioned in Section 3. The data are described in the previous section. We assume that there are important spatially dependent explanatory variables missing in the house price models, and we will analyze how the models are affected when we uncover some of the omitted variables. All estimation is done in Matlab, and routines in the Spatial Econometrics Toolbox by LeSage are used. (For further details and to download the toolbox, see www.spatial-econometrics.com.)

To begin with, we have only the original variables in matrix $\mathbf{X}$, which includes the variables *M2, Age, Age2, Loan, Maintenance, WOM,* and dummy variables for *Rooms, Floor, FloorB, Sauna, Balcony, Alcove, Elevator&Floor* > 2, *Condition*, *Feb* and *Mar*. The variables capturing the relative size effects and the competition in the nearby surroundings are assumed not available. The data are therefore assumed to lack important location and neighbourhood variables, and these omitted variables are assumed to be spatially autocorrelated. We also assume that these missing variables are correlated with the available explanatory variables. We estimate the following non-spatial hedonic price model of semi-log form:

$$\ln Price = \mathbf{X}\boldsymbol{\beta}_I + \boldsymbol{\varepsilon}_I. \qquad \text{(I)}$$

We name it model (I). The variation in the natural logarithm of the selling price of an apartment ($\ln Price$) is explained by the variables in matrix $\mathbf{X}$.

Table 2: Estimation Results for Model I

Variable	Coeff.	t-stat	t-prob.	Variable	Coeff.	t-stat	t-prob.
Constant	10.423	164.62	0.00	FloorsB=3	-0.051	-1.48	0.14
M2	0.012	18.60	0.00	FloorsB=4	-0.017	-0.49	0.62
Age	0.007	4.92	0.00	FloorsB=5	0.014	0.38	0.71
Age2	0.000	0.09	0.92	FloorsB=6	-0.006	-0.16	0.87
Loan	0.000	3.20	0.00	FloorsB=7	-0.017	-0.38	0.71
Maintenance	-0.000	-2.82	0.00	FloorsB>7	-0.069	-1.44	0.15
WOM	0.001	0.67	0.50	Sauna	0.131	3.11	0.00
Rooms=2	0.154	5.84	0.00	Balcony	0.020	0.84	0.40
Rooms=3	0.161	4.06	0.00	Alcove	0.037	0.86	0.39
Rooms>3	0.207	3.26	0.00	Elev.&Floor>2	0.061	2.05	0.04
Floor=3	-0.018	-0.70	0.48	Cgood	0.093	3.99	0.00
Floor=4	-0.017	-0.52	0.60	Csatisf	-0.026	-1.06	0.29
Floor=5	0.030	0.71	0.47	Cbad	-0.100	-2.32	0.02
Floor>5	0.070	1.53	0.13	Feb	0.044	2.11	0.04
				Mar	0.086	4.04	0.00

Moran I-test for Spatial Correlation in Residuals	
Moran I	0.399
Moran I-statistic	16.03
Marginal Probability	0.000

The estimation results are presented in *Table 2*. By assumptions the omitted explanatory variables are spatially dependent, and there will be an amplified bias in these traditional least squares estimates compared to the situation if the missing variables were not to be spatially dependent (LeSage and Pace, 2009, page 64). The residuals of model (I), $\hat{\boldsymbol{\varepsilon}}_I$, are tested for spatial autocorrelation by a Moran I test (*Table 2)* and the null hypothesis of no spatial autocorrelation is rejected.

In Section 3 we have seen that the assumptions that there are important spatially dependent explanatory variables missing from the model, and that they are correlated with the included explanatory variables, lead to a spatial Durbin model (SDM). Therefore we estimate the following SDM:

$$\ln Price = \rho \mathbf{W} \ln Price + \mathbf{X}\boldsymbol{\beta}_{II_1} + \mathbf{W}\mathbf{X}\boldsymbol{\beta}_{II_2} + \boldsymbol{\varepsilon}_{II}. \qquad \text{(II)}$$

We name it model (II). According to what was shown in section three $\boldsymbol{\beta}_{II_1} =$

$(\boldsymbol{\beta}_I + \boldsymbol{\gamma})$ and $\boldsymbol{\beta}_{II_2} = (-\rho\boldsymbol{\beta}_I)$, where the parameter vector $\boldsymbol{\gamma}$ describes the relation between the variables in $\mathbf{X}$ and the missing variables. The scalar parameter ρ in model (II) is the spatial autocorrelation coefficient and the matrix $\mathbf{W}$ is a row-normalized weights matrix that is based on the four nearest neighbours to each observation. The model is estimated by the method of maximum likelihood and the results are reported in *Table 3*. Especially note that ρ is significantly different from zero.

Let us then assume that we uncover some of the missing variables, say $\mathbf{Z}_1$, namely the variables capturing the relative size effects ($Larger$ and $Smaller$) and the competition in the small scale neighbourhood ($Comp$ and $Avgcomp$). In the previous section we showed how these variables were created. Since they are based on information in the existing data and the coordinates, it is natural to assume that these variables are both spatially autocorrelated, and correlated with the explanatory variables in $\mathbf{X}$. To verify these assumptions, we calculate the canonical correlation between $\mathbf{Z}_1$ and $\mathbf{X}$, and estimate a first order spatial econometric model (FAR), $\mathbf{z}_k = \rho_k \mathbf{W}\mathbf{z}_k + \boldsymbol{\varepsilon}_k$, for each of the four new small scale neighbourhood variables, $z_k = \{Smaller, Larger, Comp, Avgcomp\}$. The results are in *Table 4*. We can see that the first canonical correlation is 0.918 and is highly significant according to Wilks' test. Also the other canonical correlations are significantly different from zero. Further, all four estimates of the spatial autocorrelation parameters, ρ_k, are found to be significantly different from zero.

We include the new variables in the matrix of explanatory variables, $\mathbf{X}^* = [\mathbf{X}, \mathbf{Z}_1]$, and estimate the following SDM:

$$\ln Price = \rho^* \mathbf{W} \ln Price + \mathbf{X}^* \boldsymbol{\beta}_{III_1} + \mathbf{W}\mathbf{X}^* \boldsymbol{\beta}_{III_2} + \boldsymbol{\varepsilon}_{III}. \qquad \text{(III)}$$

We find that when the four new variables are included, the spatial lag of the dependent variable becomes insignificant ($\hat{\rho}^* = 0.529$, $Z - prob. = 0.453$). Most of the spatially lagged explanatory variables are also insignificant (not reported). According to the assumptions of this model there are important spatially dependent explanatory variables missing and that these variables correlate with the included explanatory variables. If the assumptions are true, the spatial lags of the spatial Durbin model should be significant, but this seems to happen only when the set of small scale variables are omitted, as in model (II). When the variables $\mathbf{Z}_1$are included in the matrix of explanatory variables, as in model (III), the lags becomes insignificant, and we end up with a conventional non-spatial regression model.

Table 3: Estimation Results for Model II

Variable	Coeff.	As.t	Z-prob.	Variable	Coeff.	As.t	Z-prob.
Constant	4.577	13.52	0.00	WM2	-0.005	-5.03	0.00
M2	0.013	24.85	0.00	WAge	0.005	2.66	0.00
Age	0.000	0.23	0.82	WAge2	-0.000	-1.10	0.27
Age2	0.000	0.66	0.51	WLoan	-0.000	-3.34	0.00
Loan	0.000	5.00	0.00	WMaintenance	-0.000	-1.56	0.12
Maintenance	-0.000	-2.27	0.02	WWOM	0.002	1.31	0.19
WOM	0.001	1.38	0.17	WRooms=2	-0.129	-3.43	0.00
Rooms=2	0.60	8.43	0.00	WRooms=3	-0.201	-3.55	0.00
Rooms=3	0.169	5.96	0.00	WRooms>3	-0.151	-1.61	0.11
Rooms>3	0.182	3.00	0.00	WFloor=3	-0.065	-1.67	0.10
Floor=3	0.002	0.08	0.93	WFloor=4	-0.048	-1.02	0.31
Floor=4	0.012	0.51	0.61	WFloor=5	-0.011	-0.18	0.86
Floor=5	0.031	1.05	0.29	WFloor>5	-0.029	-0.45	0.66
Floor>5	0.058	1.76	0.08	WFloorsB=3	0.013	0.28	0.78
FloorsB=3	-0.026	-1.02	0.31	WFloorsB=4	-0.005	-0.09	0.93
FloorsB=4	-0.017	-0.67	0.50	WFloorsB=5	-0.21	-0.40	0.69
FloorsB=5	-0.014	-0.52	0.61	WFloorsB=6	-0.028	-0.52	0.60
FloorsB=6	-0.044	-1.44	0.15	WFloorsB=7	-0.103	-1.66	0.10
FloorsB=7	-0.018	-0.52	0.61	WFloorsB>7	-0.056	-0.78	0.44
FloorsB>7	-0.071	-1.92	0.05	WSauna	0.023	0.36	0.72
Sauna	0.084	2.76	0.01	WBalcony	0.010	0.31	0.75
Balcony	0.020	1.17	0.24	WAlcove	-0.023	-0.36	0.72
Alcove	0.029	0.94	0.34	WElev&Floor>2	0.063	1.47	0.14
Elev&Floor>2	0.012	0.56	0.57	WCgood	0.023	0.64	0.52
Cgood	0.073	4.26	0.00	WCsatisf	0.029	0.83	0.41
Csatisf	-0.036	-1.99	0.05	WCbad	0.004	0.06	0.95
Cbad	-0.110	-3.52	0.00	WFeb	0.049	1.43	0.15
Feb	0.048	3.19	0.00	WMar	0.045	1.37	0.17
Mar	0.072	4.65	0.00	ρ	0.547	17.90	0.00

Table 4: Canonical correlations and FAR model estimates

Canonical Correlations		Wilks'	Chi-Sq	df	Prob.
1	0.918	0.036	2093.979	112	0.000
2	0.702	0.231	926.409	81	0.000
3	0.600	0.455	497.414	52	0.000
4	0.538	0.711	215.44	25	0.000
First Order Spatial Autoregressive Model Estimates					
$\mathbf{z}_k$	$\hat{\rho}_k$	As.t-stat	Z-prob.		
Smaller	0.208	6.787	0.000		
Larger	0.345	7.932	0.000		
Comp	0.389	5.582	0.000		
Avgcomp	0.528	3.810	0.000		

Let us then consider the other scenario in section three, where the spatially autocorrelated omitted variables are uncorrelated with the explanatory variables in the model. We begin by testing the least-squares residuals of the non-spatial model (I), with $\mathbf{X}^*$ in place of $\mathbf{X}$, for spatial autocorrelation by a Moran I test. The Moran I is 0.383, the Moran I-statistic is 15.58 and the marginal probability of the test is 0.000, which means that the null hypothesis of no spatial autocorrelation is rejected. The natural implication is that despite the set of small scale neighbourhood variables that is included in the explanatory variables $\mathbf{X}^*$, there are still spatially autocorrelated explanatory variables missing from the model. Since the set of new small scale variables is found to be correlated with the other explanatory variables in $\mathbf{X}^*$, and the spatial lags in the SDM model (III) is found insignificant, we assume that the remaining missing variables are uncorrelated with the included explanatory variables in $\mathbf{X}^*$. In Section 3 we showed that these assumptions will result in the following spatial error model (SEM),

$$\begin{aligned} \ln Price &= \mathbf{X}^*\boldsymbol{\beta}^*_{IV} + \mathbf{u} \\ \mathbf{u} &= \lambda \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}_{IV}. \end{aligned} \tag{IV}$$

The results of the estimation of model (IV) are in *Table 5*.

We find that the estimated spatial autocorrelation coefficient is significantly different from zero, so the SEM cannot be reduced to a non-spatial hedonic regression model. The adjusted coefficient of determination $\bar{R}^2$ for the SEM (IV) is 0.89. For the SDM (II), which does not include the new small

Table 5: Estimation Results for Model IV

Variable	Coeff.	As.t	Z-prob.	Variable	Coeff.	As.t	Z-prob.
Constant	11.189	194.47	0.00	FloorsB=6	-0.037	-1.27	0.20
M2	0.007	7.54	0.00	FloorsB=7	-0.026	-0.77	0.44
Age	0.001	0.98	0.33	FloorsB>7	-0.068	-1.89	0.06
Age2	0.000	2.58	0.01	Sauna	0.086	2.94	0.00
Loan	0.000	13.11	0.00	Balcony	0.024	1.43	0.15
Maintenance	-0.000	-2.84	0.00	Alcove	0.032	1.10	0.27
WOM	-0.000	-0.08	0.93	Elev&Floor>2	0.019	0.95	0.34
Rooms=2	0.108	5.15	0.00	Cgood	0.055	3.50	0.00
Rooms=3	0.116	3.68	0.00	Csatisf	-0.035	-2.06	0.04
Rooms>3	0.178	3.96	0.00	Cbad	-0.094	-3.25	0.00
Floor=3	0.009	0.54	0.59	Feb	0.029	2.08	0.04
Floor=4	0.007	0.32	0.75	Mar	0.058	4.09	0.00
Floor=5	0.037	1.30	0.19	Smaller	-0.497	-6.67	0.00
Floor>5	0.054	1.71	0.09	Larger	0.278	4.87	0.00
FloorsB=3	-0.039	-1.59	0.11	Comp	0.000	1.64	0.10
FloorsB=4	-0.021	-0.87	0.39	Avgcomp	0.003	0.00	0.02
FloorsB=5	-0.018	-0.68	0.50	λ	0.718	192.41	0.00

scale neighbourhood variables, the $\bar{R}^2$ is 0.84. Thereby the variables in SEM explain more of the variation in the dependent variable than the variables in the SDM. A great benefit of the spatial error model compared to a spatial Durbin model is that the regression coefficients have their usual interpretation as partial derivatives. As model (IV) is of semilog form, an estimate multiplied by 100 is interpreted as the percentage change in the dependent variable, the selling price ($Price$), when the corresponding explanatory variable is changed one unit and all other variables are kept unchanged. As expected, the selling price is found to be higher if the living area ($M2$) is larger. The coefficient on Age is expected to be negative and the coefficient on $Age2$ is expected to be positive. The combination is interpreted that as a building gets older, the price of the apartment falls with a decreasing rate. The results show that the variable age squared ($Age2$) is positive and significant. The estimate on the coefficient for the variable $Loan$ is positive and significantly different from zero. If there is a housing loan for the apartment, the corresponding amount can be deducted from the selling price before tax. This is beneficiary to the buyer. The variable $Maintenance$ is significant and the estimate is negative, as expected. It is natural that a higher maintenance charge translates into a lower price for the apartment, since it usually reflects the financial position of the residential building. The reference group for the dummy variable for the number of rooms ($Rooms$) is single-room apartments. We find that apartments with two rooms have on the average an 11% higher selling price than single-room apartments. The selling price of apartments with three rooms are on the average 12% higher, and the selling price of apartments with more than three rooms are on the average 18% higher than the selling price of single-room apartments. Also the floor number of the apartment ($Floor$) is entered as a dummy variable. Several different specifications of the variable were considered, but it always came out insignificant. The expectation is that apartments at higher floors are sold at higher prices. The same type of problem occurred when the number of floors in the building ($FloorsB$) was transformed into a dummy variable. Several possibilities were examined, but there was no solution where the variable came out significant.

Further, the dummy variable for sauna ($Sauna$) is significant. If two otherwise identical apartments are sold, the one including a sauna is sold for about 9% more than the one without sauna. This is not a surprising finding, since Finns do love their saunas. On the other hand, it is a bit surprising that the dummy variable for elevator ($Elevator\&Floor > 2$) is insignificant. Recently, the municipality of Helsinki has started to support the building of

elevators in old residential buildings. One of the main argument for residents to pay their share of the costs is that it is claimed that the values of the apartments will go up. This argument is not significantly supported by the present study. The variable was set to one only if the apartment building have an elevator and the apartment is situated above second floor. The assumption is that residents in the first and second floor do not care whether there is an elevator or not. The condition of the interior of the apartment (*Cond*: *good*, $satisfactory$ or *bad*) is significant for all three categories. In line with expectations, an apartment that is graded *bad* sells for less than one that is graded $satisfactory$ and an apartment that is graded *good* sells for more than one that is graded $satisfactory$. Notice that the dummy variable is compared to the case of missing observations. The assumption about a missing grade is that the agent thought that there was nothing special to mention about the apartment and therefore did not grade its condition. Therefore one should be a bit careful when interpreting these estimates. The dummy variables for the month of sale are significant for both February and for March. The estimate indicates that the selling price for apartments sold in February is on average 3% higher than for identical apartments sold in January. Apartments sold in March are sold for 6% more than identical apartments sold in January. The data are not corrected for inflation, since the time period is so short.

The small scale neighbourhood variables were of special interest in this study, since they enabled the choice of a simpler spatial error model over a richer spatial Durbin model. The variables measuring the relative size of an apartment in its small scale neighbourhood (*Smaller* and *Larger*) are highly significant. The coefficient on *Smaller* is negative. This indicates that the smaller an apartment is relative to the average apartment in the small scale neighbourhood, i.e. the deviation from the mean of the apartment size is growing, the lower is the price. On the other hand, the coefficient on the variable *Larger* is positive. If an apartment is larger than the average apartment in the small scale neighbourhood, the apartment is sold for more than an otherwise identical apartment in a homogenous neighbourhood. The same apartment is lower valued if it is the smallest one in a heterogenous neighbourhood than if it is situated in a homogenous neighbourhood. An apartment has a higher selling price if it is the biggest one in a heterogenous neighbourhood than if it is situated in a neighbourhood with equally big apartments.

The variable *Avgcomp* measures the average intensity of the competition

in the small scale neighbourhood and is significant. The sign of the estimate is positive, which indicates that the more apartments are for sale at the same time in a neighbourhood, the higher the price. This supports the theory that when there are lots of objects for sale at the same time, the showing of one apartment will also attract potential buyers to nearby showings, and when there are a lot of potential buyers the price will go up. The estimate of the coefficient on the variable measuring the cumulative competition in a small scale neighbourhood ($Comp$) has also a positive sign, but it is not significant in model (IV). We might expect that this variable competes with the variable reporting the number of weeks the apartment was for sale on the market (WOM), since the information about the selling time is used in the calculations of $Comp$. The model is therefore reestimated without WOM and then $Comp$ becomes significant. Its parameter estimate does not change, it is still positive, which supports the theory of prices going up with more competition on the small scale neighbourhood market. Also the other estimates are robust when WOM is omitted.

When we compare the estimation results of the initial non-spatial hedonic regression model (I), which does not include the set of small scale variables, and the final SEM (IV), we find that several estimates are different, for instance, the estimates for the parameters of $M2$, $Room$, $Sauna$ and $Condition$. A bias is expected in the least-squares estimates of model (I), since we are working under the assumptions that the model is lacking important explanatory variables and they are correlated with the explanatory variables. As noted before, Pace and LeSage (2008) shows that this bias is amplified if the omitted variables are spatially autocorrelated. In the final SEM (IV), which includes the set of small scale neighbourhood variables, we have captured some of the effects of the omitted variables and the estimates should then be less biased.

In this empirical analysis we have shown that for this particular data set, we are able to construct some of the missing spatially autocorrelated variables. These variables are shown to have the desired properties and they control for some of the small scale neighbourhood conditions. By including them in the model the result is a simpler spatial econometric model, which is easier to interpret than a spatial model not including these variables.

6 Conclusions

A common problem the researcher faces when modelling house prices is that in practice it is seldom possible to obtain all the desired variables. Especially variables capturing the small scale neighbourhood conditions are hard to find. For house price data it is a common and reasonable assumption, that objects situated near each other have similar values on variables describing the location and immediate surroundings, i.e. these variables are spatially autocorrelated. According to LeSage and Pace (2009) there is a strong motivation for spatial econometric modelling when the omitted variables are spatially autocorrelated.

The starting point of this study is a non spatial hedonic house price model. We assume that there are important explanatory variables missing from the model, the omitted variables are spatially autocorrelated, and that they are correlated with the explanatory variables included in the model. These assumptions are shown to lead to a spatial Durbin model. We estimate the model and find that its spatial structure fits, since the spatial lags are significant. We then assume that some of the omitted variables are uncovered following suggestions by Turnbull, Dombrow and Sirmans (2006). The set of new variables for the small scale neighbourhood market conditions are shown to be spatially autocorrelated and they are also shown to correlate with the original variables included in the model. The set of the uncovered variables are included in the explanatory variables and the spatial Durbin model is re-estimated. We now find that its spatial structure no longer fits, since the spatial lags become insignificant. The model is reduced to a conventional non-spatial regression model. It seems reasonable to assume that there still are other explanatory variables missing despite the included uncovered set of new variables. Further we assume that remaining omitted variables are still spatially autocorrelated, but now we assume that they are uncorrelated with the explanatory variables included in the model. These assumptions are shown to lead to a spatial error model. When this model is estimated we find that its spatial structure fits, since the spatial lag of the error term is highly significant. In this final model the regression coefficients have their usual interpretation as partial derivatives, since the model does not involve any lag on the dependent variable or on the explanatory variables. An easy interpretation of the estimates is a benefit compared to the structurally richer spatial Durbin model that fitted the original data. These empirical results show that for this particular data set it is possible to find

explanatory variables that capture some of the small scale neighbourhood conditions, and thereby obtain a simplified model. However, the set of small scale neighbourhood variables cannot entirely replace a spatial econometric structure.

References

[1] Anselin, L. (1988), *Spatial Econometrics: Methods and Models,* Kluwer Academic Publishers, Dordrecht.

[2] Dubin, R. (1988) Estimation of Regression Coefficients in the Presence of Spatially Autocorrelated Error Terms, *The Review of Economics and Statistics,* Vol. 70, Issue 3, 466-474.

[3] Kelejian, H.H. (2008), lecture handouts from Spatial Econometrics Association Institute 2008, based on forthcoming articles by the author.

[4] LeSage, J.P. and Pace, R. K. (2009) Introduction to Spatial Econometrics., CRC Press, Taylor and Francis Group.

[5] Lee, L.F.(2004) Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Autoregressive Models, *Econometrica,* 72, 2004, 1899-1925.

[6] Pace, R. K., and O. W. Gilley (1997) Using the Spatial Configuration of the Data to Improve Estimation, *Journal of Real Estate Finance and Economics,* Vol. 14, Number 3, 333-340.

[7] Pace, R.K. and LeSage, J. P., Biases of OLS and Spatial Lag Models in the Presence of an Omitted Variable and Spatially Dependent Variables (February 19, 2008). Available at SSRN: http://ssrn.com/abstract=1133438

[8] Rosen, S. (1974) Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition, *Journal of Political Economy,* 82, 34-55.

[9] Turnbull, G. K., Dombrow, J. and Sirmans, C. F. (2006) Big House, Little House: Relative Size and Value, *Real Estate Economics,* V 34, 3: 439-456.

[10] Wilhelmsson, M. (2002), Spatial Models in Real Estate Economics, *Housing, Theory and Society*, 19, 92-101.

A Practical Proposal to Specification Search of a k-Nearest Neighbours Weights Matrix

Linda Gerkman
Hanken School of Economics,
Department of Finance and Statistics

Abstract

A practical issue is how to choose the weights matrices of a spatial econometric model. Focusing on k-nearest neighbours weights matrices, this paper proposes a strategy for specification search. The proposed strategy gives formal justification for the choice of the number of nearest neighbours. Applying the spatial J-test as the means of specification search, two approaches, an increasing and a decreasing number of neighbours approach, are suggested and examined. The results are clearly in favour of the increasing number of neighbours approach. We find that as long as the spatial dependence in the dependent variable is at least moderate, all false null models are rejected, and that the size of the final test equals the nominal size. House price data from Stockholm, Sweden, are used to illustrate the strategy in practice.

Keywords: k-nearest neighbours; Model Specification; Spatial J-test; Weights matrix

1 Introduction

In spatial econometrics the weights matrix has an important role. It contains the assumed spatial structure of the variables in the model. An important practical issue is how to choose the weights matrix. A problem is that theory may be subjectively interpreted and might therefore result in several weights matrices, with different structures but equally realistic. Focusing on k-nearest neighbours weights matrices, this paper proposes a practical testing strategy that gives formal justification for specification.

The assumed model is a general spatial autoregressive model

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{W}_1\mathbf{y} + \mathbf{u}, \\ \mathbf{u} &= \lambda\mathbf{W}_2\mathbf{u} + \boldsymbol{\varepsilon}. \end{aligned} \tag{1}$$

Here the variation in the $n \times 1$ dependent variable vector $\mathbf{y}$ is explained by the exogenous variables in the matrix $\mathbf{X}$ and a spatial lag $\mathbf{W}_1\mathbf{y}$. The error term $\mathbf{u}$ is spatially autocorrelated and $\boldsymbol{\varepsilon}$ is a vector of $IID(0, \sigma_\varepsilon^2)$ error terms. Here $\boldsymbol{\beta}$ is a vector of regression coefficients, and ρ and λ are scalar spatial autoregressive parameters. If $\rho = 0$, model (1) is called a spatial error model and if $\lambda = 0$, the model is referred to as a spatial lag model. If both $\rho = 0$ and $\lambda = 0$, the model is a linear regression model. The structure of spatial dependence in the data is formulated in the weights matrices $\mathbf{W}_1$ and $\mathbf{W}_2$. In practice it often is assumed that $\mathbf{W}_1 = \mathbf{W}_2 = \mathbf{W}$. In general, the weights matrix is then given by

$$\mathbf{W} = \begin{pmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nn} \end{pmatrix}, \tag{2}$$

which is an $n \times n$ matrix, where n denotes the number of observations on the dependent variable $\mathbf{y}$. The weights matrix specifies which of the other units in the system that affect the observed value at some particular location. Suppose that unit y_i of the dependent variable only has neighbours y_j, y_k and y_m. Then in the ith row of $\mathbf{W}$ there will be only three non-zero elements, namely w_{ij}, w_{ik} and w_{im}. The simplest form of a weights matrix is a binary weight matrix. In this type of matrix w_{ij}, w_{ik} and w_{im} are all set to one if y_j, y_k and y_m are neighbours to y_i. All other elements in the ith row of $\mathbf{W}$ are zero. The diagonal of a weights matrix is always set to zero since,

by assumption, a unit does not directly depend on itself. The definition of neighbours may also be based on geographical distances or other differences between units, or on contiguity. See for instance Anselin (1988) for different alternatives.

According to Anselin (1988), the weights matrix is assumed to be exogenous to the model. Ideally, the structure of the weights matrix is based on relevant theory rather than on spatial patterns found in the data. Typically it is based on geographic arrangements of the observations. When different solutions are compared, a problem is that there are no formal tests for significant differences between log likelihood values of the models. The reason is that spatial models that are otherwise equal, but have different weights matrices, are non-nested (LeSage and Pace 2009, p. 162). In a Bayesian setting different spatial weights matrices can be compared through log marginal likelihoods and associated model probabilities (LeSage and Pace 2009, Section 6.3).

Kelejian (2008) suggests a spatial J-test to test a given spatial model against one or more non-nested alternative models. The suggested test uses estimation by two stages least squares and general method of moments for spatial econometric models (Kelejian and Prucha 1998, 1999). In previous simulation studies by Piras and Lozano-Garcia (2008) and Burridge and Fingleton (2010) it is shown that the spatial J-test can be used in order to discriminate between different types of weights matrices, for instance, when contiguity, inverse distance or k-nearest neighbours weights matrices are pairwise compared to each other.

In this paper we focus on weights matrices based on a set of nearest neighbours, say k. Every unit i is assigned k nearest neighbours according to the shortest distance in space. The resulting weights matrix is usually not symmetric, since even if the k nearest neighbours of unit i includes, say, unit j, the k nearest neighbours of unit j might not include unit i. The row sums of the resulting weights matrix will be equal, though. Equal weights are given to all k neighbours and the weights matrix is row normalized. Therefore every non-zero element of the weights matrix will be equal to $1/k$. The spatial lags $\mathbf{W_1 y}$ and $\mathbf{W_2 u}$ of (1) then become $n \times 1$ vectors of weighted averages of neighbouring observations.

In the case that the choice is a k-nearest neighbours-type of weights matrix, the number of neighbours, k, has to be decided. This paper is motivated by the lack of formal justification for the choice of k in a k-nearest neighbours weights matrix in the classical spatial econometric literature. If there

are only k-nearest neighbours weights matrices in the models that are compared, the models are relatively closer to each other than they are in the case when the objective is to discriminate between different types of weights matrices. This is a possible problem, since usually, in a set-up where the null and the alternative model are close, a test may loose its power.

The purpose of the paper is to find a strategy for specification search for the number of neighbours in a k-nearest neighbours weights matrix. As the means of the specification search we use the spatial J-test proposed by Kelejian (2008). First the spatial J-test is introduced following Kelejian. We then suggest and examine two approaches for finding the number k, an increasing and a decreasing number of neighbours approach. The properties of the spatial J-test are studied in the context of the two suggested approaches and then a strategy for conducting a sequence of tests in practice is proposed. Simulations show that when the proposed strategy for specification search is followed loss of power can be avoided. House price data from Stockholm, Sweden, are used to illustrate the results and the conclusions. The empirical illustration also contains a comparison between the proposed strategy and the Bayesian approach to model comparison.

The paper is organized as follows. The next section presents the spatial J-test. The third section focuses on specification search and suggests approaches for finding the number of nearest neighbours. The fourth section contains the simulation study, whereas the fifth section demonstrates the usefulness of the suggested strategy by an empirical example and discusses the results. The last section concludes.

2 The Spatial J-Test

The original J-test, which was introduced by Davidson and MacKinnon (1981), is a test for non-nested hypotheses. The main idea of a J-test is to estimate a model which consists of the null model and the predicted value of the alternative model. Then the significance of the additional term is tested.

Otherwise equal spatial models are non-nested if they entertain different weights matrices, even in the case that the weights matrices are of the k-nearest neighbours-type. In a k-nearest neighbour weights matrix every row has k non-zero elements. When the matrix is row standardized each of these elements get the value $1/k$. In a spatial model the spatial lag $\mathbf{Wy}$ (or $\mathbf{Wu}$)

therefore becomes a $n \times 1$ vector of averages of the k-nearest observations on the dependent variable (or the disturbances). Even if the average of, say, the four nearest neighbours contains exactly the same neighbours (and two additional) as the average of the two nearest neighbours, a neighbour is never tested separately and the neighbours are weighted differently by a weights matrix with $k = 4$ than by a weights matrix with $k = 2$.

Kelejian (2008) suggests a spatial J-test to test a given spatial model against one or more non-nested alternative models. The spatial model of the null hypothesis is

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{W}_1\mathbf{y} + \mathbf{u} = \mathbf{Z}\boldsymbol{\gamma} + \mathbf{u}, & (3) \\ \mathbf{u} &= \lambda\mathbf{W}_2\mathbf{u} + \boldsymbol{\varepsilon}, & (4) \end{aligned}$$

where $\mathbf{Z} = (\mathbf{X}, \mathbf{W}_1\mathbf{y})$, $\boldsymbol{\gamma} = (\boldsymbol{\beta}', \rho)'$. The vectors $\mathbf{y}, \mathbf{u}$ and $\boldsymbol{\varepsilon}$, the matrices $\mathbf{X}, \mathbf{W}_1$ and $\mathbf{W}_2$, the parameter vector $\boldsymbol{\beta}$, and the scalar parameters ρ and λ are as before. The sample size is n.

The alternative models are of the form

$$\begin{aligned} \mathbf{y} &= \mathbf{X_i}\boldsymbol{\beta}_i + \rho_i\mathbf{W}_{1i}\mathbf{y} + \mathbf{u}_i = \mathbf{Z}_i\boldsymbol{\gamma}_i + \mathbf{u}_i, & (5) \\ \mathbf{u}_i &= \lambda_i\mathbf{W}_{2i}\mathbf{u}_i + \boldsymbol{\varepsilon}_i, \qquad i = 1, ..., G, \end{aligned}$$

where G is a finite constant defining the number of possible alternatives to the null model.

The J-test and its properties are described in detail in the appendix of Kelejian (2008). Our discussion is based on Kelejian. To derive the J-statistic, first solve (4), and obtain $\mathbf{u} = (\mathbf{I} - \lambda\mathbf{W}_2)^{-1}\boldsymbol{\varepsilon}$. Insert the solution into (3) and pre-multiply the expression by $(\mathbf{I} - \lambda\mathbf{W}_2)$ to get

$$(\mathbf{I} - \lambda\mathbf{W}_2)\mathbf{y} = (\mathbf{I} - \lambda\mathbf{W}_2)\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}.$$

Define $\mathbf{y}(\lambda) = (\mathbf{I} - \lambda\mathbf{W}_2)\mathbf{y}$ and $\mathbf{Z}(\lambda) = (\mathbf{I} - \lambda\mathbf{W}_2)\mathbf{Z}$, and write

$$\mathbf{y}(\lambda) = \mathbf{Z}(\lambda)\boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \quad (6)$$

Likewise for the alternative models (5), define $\mathbf{Z}_i(\lambda_i) = (\mathbf{I} - \lambda_i\mathbf{W}_{2i})\mathbf{Z}_i$. Then for all $i = 1, ..., G$, estimate $\boldsymbol{\gamma}_i$ in the ith alternative model of (5) by generalized spatial two stage least squares (Kelejian and Prucha 1998, 1999), and get $\widehat{\boldsymbol{\gamma}}_i$. Add the predictive power $\mathbf{Z}_i(\lambda_i)\widehat{\boldsymbol{\gamma}}$ of all $i = 1, ..., G$ alternative models to the null model (6) to get:

$$\mathbf{y}(\lambda) = \mathbf{Z}(\lambda)\boldsymbol{\gamma} + \sum_{i=1}^{G} \alpha_i[\mathbf{Z}_i(\lambda_i)\widehat{\boldsymbol{\gamma}}_i] + \boldsymbol{\varepsilon}, \quad (7)$$

where $\widehat{\boldsymbol{\gamma}}_i$ is the 2SLS estimate of $\boldsymbol{\gamma}_i$. The parameter α_i is a scalar. Given that the null model is true, $\alpha_i = 0$ for all $i = 1, ..., G$. Insert $\mathbf{Z}_i(\lambda_i) = (\mathbf{I} - \lambda_i \mathbf{W}_{2i})\mathbf{Z}_i$ into (7) and define $\phi_i = -\alpha_i \lambda_i$ to get

$$\mathbf{y}(\lambda) = \mathbf{Z}(\lambda)\boldsymbol{\gamma} + \sum_{i=1}^{G} \alpha_i [\mathbf{Z}_i \widehat{\boldsymbol{\gamma}}_i] + \sum_{i=1}^{G} \phi_i [\mathbf{W}_{2,i} Z_i \widehat{\boldsymbol{\gamma}}_i] + \boldsymbol{\varepsilon}.$$

Then let $\boldsymbol{\delta} = (\alpha_1, ..., \alpha_G, \phi_1, ..., \phi_G)'$. The test of the null model (3) against the alternative models (5) is simply a Wald test of $\boldsymbol{\delta} = \mathbf{0}$. At the significance level α, reject the null hypothesis if

$$J = \widehat{\boldsymbol{\delta}}' \widehat{\mathbf{V}}_{\widehat{\boldsymbol{\delta}}}^{-1} \widehat{\boldsymbol{\delta}} > \chi^2_{1-a}(2G),$$

where $\widehat{\mathbf{V}}_{\widehat{\boldsymbol{\delta}}}$ is the estimated small-sample variance-covariance matrix of $\widehat{\boldsymbol{\delta}}$. In the next section the spatial J-test is used for specification search.

3 Two Approaches to Specification Search

Model specification search in spatial econometrics is studied by Florax, Folmer and Rey (2003). They concentrate on discriminating between a spatial lag model and a spatial error model and the conclusion of their simulation study is that the classical step-wise (specific-to-general) approach performs better than Hendry's (general-to-specific) strategy. Hendry (2006) comments on the study and a reply to the comments is presented in Florax et al (2006). As LeSage and Pace (2009, page 74) point out, an excessive focus in the spatial econometrics literature has been on the procedures for comparative testing of alternative model specifications, i.e. the choice between different structural forms for the spatial econometric model.

In this study a different view is taken. We assume the structural form of a general spatial econometric model and concentrate the specification search only to the weights matrices. As mentioned before, otherwise equal spatial models, but with different weights matrices, are non-nested models. The consequence is, as LeSage and Pace (2009, page 162) note, that it is not in general possible to use formal tests for significant differences between the log likelihood function values for models entertaining different weights matrices. The number of parameters are fixed and equal in the models. For instance, the usual specification search based on information criteria is then not available. See Maddala (2001, page 485), for the information criterion and a list

of other similar criteria. An option is, however, the approach to specification search by sequential testing. It is explored here. We use the spatial J-test as the means of the specification search. The study is focused on a special kind of spatial weights matrix, namely the k-nearest neighbour weights matrix, and the objective is to determine the number of nearest neighbours, k. For this purpose we consider two alternative approaches, which we call the increasing neighbours approach and the decreasing neighbours approach.

Assume that we have a model M_1 and another model M_2, which are non-nested and that we conduct two J-tests, M_1 against M_2 and M_2 against M_1. Adapted from a discussion in Maddala (2001) concerning tests for non-nested hypothesis, the four possible outcomes of the two tests are:

	$H_0 : M_2$, $H_1 : M_1$	
$H_0 : M_1, H_1 : M_2$	not rejected	rejected
not rejected	both models acceptable	M_1 acceptable, M_2 not
rejected	M_2 acceptable, M_1 not	neither model acceptable

Generally a rejection of a null hypothesis can mean two things. The alternative model may have some significant additional predictive power, or neither of the models are acceptable. If the null hypothesis is not rejected the reason is either that the null model is preferred over the alternative model, or that both models are acceptable.

Assume that we have a general spatial model as model (1), which we call M_k. For simplicity we assume that the weights matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ of (1) are equal, k-nearest neighbour weights matrices, $\mathbf{W}_k$. We want to specify the number of nearest neighbours k in the weights matrix $\mathbf{W}_k$ for some particular data with n observations. The possible range is $k =]0, n[$. An alternative approach to specification search is to start with a model where the number of neighbours for each observation is small, for instance $k = 2$, and then test this model against models with more neighbours in their weights matrices to see if there is any additional predictive power when the number of neighbours in the weights matrix is gradually increased. This is the increasing neighbours approach. Another alternative is a decreasing neighbours approach, in which we begin with a large number of neighbours in the weights matrix of the null model. The model is then tested against models which have gradually smaller numbers of neighbours.

First consider the increasing neighbours approach. Start with model M_p as the null model. This model entertains a weights matrix which has the

smallest possible number of nearest neighbours that we will entertain, say, p. Test this model against model M_{p+1}, where $k = p + 1$, by a spatial J-test. Assume that the null model is rejected. The conclusion is that the alternative model has some significant additional predictive power, or neither one is acceptable. We therefore add another neighbour in the weights matrix of the alternative model, and continue by testing the previous alternative model M_{p+1} against the new model M_{p+2}. As long as the null model is rejected, the previous alternative model becomes the next null model and we gradually increase the number of neighbours in the weights matrix of the new alternative model. Assume that finally, when the null model is M_{p+m}, which has a weights matrix that has $k = p + m$ number of nearest neighbours, we cannot reject it. The outcome of the last J-test means that there is no additional predictive power in a model that includes a weights matrix with one more neighbour, i.e. M_{p+m} is preferred over M_{p+m+1}, or both M_{p+m} and M_{p+m+1} are acceptable. The model M_{p+m} is also preferred over M_{p+m-1}, since the test previous to the last one rejected the null hypothesis of no additional predictive power of model M_{p+m}. The number of neighbours in the weights matrix of model M_k is then $k = p + m$.

Another considered approach to the specification search is the decreasing neighbours approach. Here begin with model M_q. This model includes a weights matrix that has the largest number of nearest neighbours that we will entertain, $k = q$. Test this model M_q against model M_{q-1} by a spatial J-test. Assume that the null model M_q is rejected. Decrease the number of neighbours by one in the previous alternative model and get model M_{q-2}. Then test M_{q-1} against M_{q-2}. As long as we reject the null model, we continue by always setting the previous alternative model as the new null model. The new alternative model is set by decreasing the number of neighbours in the weights matrix of the previous alternative model. Assume that the null model is finally accepted when model M_{q-h}, which has a weights matrix that has $k = q - h$ number of nearest neighbours, is tested against M_{q-h-1}. So M_{q-h} is preferred over M_{q-h-1}, or at least they are both acceptable models. M_{q-h} is also preferred over M_{q-h+1}, according to the test previous to the last one. The conclusion of specification search by this approach is that the number of neighbours in the weights matrix of model M_k is $k = q - h$.

Ideally $p + m = q - h$, which means that the increasing neighbours approach gives the same result as the decreasing neighbours approach. In practice one may suspect that in finite samples $p + m \neq q + h$ and then the question arises, which approach should be preferred? There is also a concern

about the significance levels, since we conduct a sequence of tests. In the next section we address these questions by a simulation study of the two approaches to specification search described above.

4 Simulation Study

This section presents a simulation study of the properties of the spatial J-test when it is used in order to distinguish between k-nearest neighbours weights matrices. We compare the performances of the increasing and the decreasing neighbours approach to specification search. The focus is on the rejection probabilities of the spatial J-test in the sequences of tests conducted in the two approaches.

The simulations are executed in Matlab and routines of the Spatial Econometric Toolbox (version 2005) by LeSage are used in the estimations. The spatial J-test is programmed in Matlab following Kelejian (2008), and he uses estimation by two stages least squares and general method of moments for spatial econometric models (Kelejian and Prucha 1998, 1999). These estimation methods are therefore applied in the simulations in this section, as well as in the empirical illustration in the next section.

As in the previous section, the model of interest is

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{W}_k\mathbf{y} + \mathbf{u}, \\ \mathbf{u} &= \lambda\mathbf{W}_k\mathbf{u} + \boldsymbol{\varepsilon}. \end{aligned} \tag{8}$$

This is a general spatial autoregressive model with the same weights matrix $\mathbf{W}_k$ in the spatial lag $\mathbf{W}_k\mathbf{y}$ as in the spatially correlated error term $\mathbf{u} = \lambda\mathbf{W}_k\mathbf{u} + \boldsymbol{\varepsilon}$. We restrict the study to five non-nested models, which are the same except that they have different k-nearest neighbour weights matrices $\mathbf{W}_k$, where $k = \{4, 6, 8, 10, 12\}$. The weights matrices are based on coordinates on the two-dimensional plane. The coordinates (xc, yc) are randomly drawn integers from a uniform $[1, 10000]^2$ distribution. The studied sample size is $n = 900$.

By solving (8) and setting $\mathbf{W}_k = \mathbf{W}_8$ we get the data generating process,

$$\mathbf{y} = (\mathbf{I} - \rho\mathbf{W}_8)^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \rho\mathbf{W}_8)^{-1}(\mathbf{I} - \lambda\mathbf{W}_8)^{-1}\boldsymbol{\varepsilon}. \tag{9}$$

The matrix $\mathbf{W}_8$ in (9) is the weights matrix where every observation has $k = 8$ nearest neighbours. The matrix $\mathbf{X}$ contains a column of ones and three

Table 1: The size and power of the spatial J-test when $n = 900$ and $MC = 1000$.

		Size				Power			
	$H_0:$	M_8	M_8	M_8	M_8	M_4	M_6	M_{10}	M_{12}
	$H_1:$	M_4	M_6	M_{10}	M_{12}	M_8	M_8	M_8	M_8
ρ	λ								
0.2	0.2	0.041	0.061	0.057	0.068	1.000	0.972	0.317	0.606
0.2	0.5	0.046	0.050	0.047	0.048	0.999	0.933	0.218	0.560
0.2	0.8	0.043	0.048	0.048	0.048	0.985	0.733	0.208	0.376
0.5	0.2	0.043	0.035	0.055	0.053	1.000	1.000	0.762	0.953
0.5	0.5	0.059	0.047	0.048	0.046	1.000	1.000	0.811	0.965
0.5	0.8	0.053	0.058	0.049	0.049	1.000	1.000	0.728	0.941
0.8	0.2	0.051	0.053	0.061	0.059	1.000	1.000	0.717	0.834
0.8	0.5	0.048	0.058	0.044	0.046	1.000	1.000	0.703	0.849
0.8	0.8	0.054	0.059	0.054	0.056	1.000	1.000	0.762	0.845

explanatory variables, which are randomly drawn from a uniform $[-10, 10]$ distribution, $\boldsymbol{\beta}$ is a 4×1 vector of ones and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector which is randomly drawn from a normal distribution with mean zero and variance equal to one. Data are generated for each possible pair of $\rho = \{0.2, 0.5, 0.8\}$ and $\lambda = \{0.2, 0.5, 0.8\}$. These values for the autoregressive parameters are chosen in order to represent small, medium and large positive spatial autocorrelation. The explanatory variables remain the same in all replications, but the coordinates and the vector $\boldsymbol{\varepsilon}$ are drawn 1000 times for each simulated size and power realization reported in Table 1 and 1000 times for each simulated rejection probability in the sequence of spatial J-tests reported in Tables 2-4.

We begin with simulating the size and power of the spatial J-test. In order to get the size of the test, model M_8, which is the data generating process, is individually tested against the other models M_k, where $k = \{4, 6, 10, 12\}$. To get the power of the test each one of the other models are tested against the model M_8. The results are shown in Table 1.

The results show that the size of the test is close to the nominal size, which is 0.050. However, for smaller sample sizes than $n = 900$, which is used here, the test is found to be slightly oversized (Ahlgren and Gerkman 2010). The power of the test is high when the models with underspecified weights matrices, M_4 and M_6, are tested against M_8 and further, if ρ is at

least moderate the power is found to be one. By underspecified we mean that the weighs matrix is based on a smaller number of nearest neighbours than the one in the data generating process. When the models with overspecified weights matrices, M_{10} and M_{12}, are tested against the data generating M_8, and ρ is small, the power of the test is low. By overspecified we mean that the weighs matrix is based on a larger number of nearest neighbours than the one in the data generating process. When ρ is increased, the power of the test increases, but for all cases it remains lower when the null model is overspecified than when it is underspecified. No clear impact of the value of λ on the power can be seen in these results.

The results indicate that when the weights matrix is overspecified the test has lower power than when the weights matrix is underspecified. An overspecified weights matrix contains the true neighbours and some additional observations. These are all given the same weight, $1/k$. The influence of the additional observations, which also are weighted into the spatial lag, will diminish the power of the test. In the case of an underspecified weights matrix the number of neighbours are less than in the correct specified one, but they are all true neighbours and they are all given the same weight. Even if there are neighbours missing, there are no additional, false neighbours weighted into the spatial lag and therefore the power of the test is less influenced in comparison to the overspecified case. The finding is in favour of the increasing neighbours approach.

Next we simulate the rejection probabilities for the null hypothesis of the spatial J-test at a 5% significance level. Every model M_k, including the data generating M_8, is tested against every other model. The number of neighbours are the same as above, $k = \{4, 6, 10, 12\}$ and data are generated for each possible pair of $\rho = \{0.2, 0.5, 0.8\}$ and $\lambda = \{0.2, 0.5, 0.8\}$ by the general spatial model (9).

The rejection probabilities of the sequences of tests conducted by applying the increasing neighbours approach are the bold faced cases above the diagonals, beginning from upper left to lower right in Tables 2-4. The results of the decreasing neighbours approach are the rejection probabilities below the diagonals, beginning from lower right to upper left. In these simulations the tests in the sequences are not conditioned on the outcomes of the previous tests. Each rejection probability is therefore the probability of rejecting a specific null hypothesis when tested against a specific alternative hypothesis.

Table 2 shows the outcomes when the spatial autocorrelation in the dependent variable of the data generating process is small, $\rho = 0.2$. We can

see that when the spatial autocorrelation in the error term is small ($\lambda = 0.2$) or intermediate ($\lambda = 0.5$) the results are similar. If we apply the increasing number of neighbours approach we will begin with testing M_4 against M_6. In this first step of the sequence the null model will always be rejected, since the probability is 1.00. Then we test M_6 against M_8. If $\lambda = 0.2$, M_6 will be rejected with probability 0.99 and if $\lambda = 0.5$, M_6 will be rejected with probability 0.92. When we further increase the number of neighbours and test M_8 against M_{10} the rejection probabilities are 0.06 and 0.05, respectively, which are the sizes of the tests. However, if ρ is small and λ is large, as in the third panel of the table, the individual rejection probabilities for the tests in the sequence are 0.99, 0.71 and 0.06. Here the rejection probability in the second step, i.e. the power, which is obtained when M_6 tested against the data generating M_8, is 71%. The consequence may be that the sequence of tests is stopped too early and the outcome is an underspecified weights matrix.

Now assume that we apply the decreasing number of neighbours approach. We start with M_{12}, which is the model that has a weights matrix with the largest number of neighbours, and test it against M_{10}. According to the results in Table 2, where ρ is small and λ is small, intermediate or large, there is a low probability of rejecting the null hypothesis in all three cases of the first step of the sequence, 58%, 55% and 34%, respectively. In the second step, when we test M_{10} against M_8 the probability of rejecting the null model is even lower, only 32%, 28% and 18%, respectively. The size of the last test in the sequence is still correct, 0.04, 0.05 and 0.05, respectively. However, the problem is that when the rejection probabilities in the sequence are low, the sequence might be stopped too early, and we never reach the stage of the last test. When that happens the outcome of the specification search by this approach will be an overspecified weights matrix.

Table 3 shows the result when ρ is intermediate and Table 4 shows the result when ρ is large. The results are very similar. When the increasing neighbours approach is applied, in both cases a false null hypothesis is always rejected, since the rejection probability is one in all steps. Since the number of neighbours in the weights matrix is step-wise increased and the sequence of test is continued until the null model is not rejected, this means that the sequence ends only when the data generating model is the null model, i.e. when M_8 is tested against M_{10}. When $\lambda = 0.5$ or $\lambda = 0.8$, the significance level of the last test in the sequence is 0.05, which is equal to the nominal size. When $\lambda = 0.2$, the size is almost correct, 0.06.

Table 2: Rejection probabilities of the null hypotesis for the spatial J-test when ρ is small, $n = 900$ and $MC = 1000$.

dgp: M_8		$\rho = 0.2$		$\lambda = 0.2$		
		H_1				
		M_4	M_6	M_8	M_{10}	M_{12}
	M_4	-	**1.00**	1.00	1.00	1.00
	M_6	0.98	-	**0.99**	0.98	0.98
H_0	M_8	0.05	**0.04**	-	**0.06**	0.05
	M_{10}	0.29	0.30	**0.32**	-	0.26
	M_{12}	0.56	0.58	0.62	**0.58**	-

dgp: M_8		$\rho = 0.2$		$\lambda = 0.5$		
		H_1				
		M_4	M_6	M_8	M_{10}	M_{12}
	M_4	-	**1.00**	1.00	1.00	1.00
	M_6	0.92	-	**0.92**	0.92	0.93
H_0	M_8	0.05	**0.05**	-	**0.05**	0.06
	M_{10}	0.26	0.26	**0.28**	-	0.25
	M_{12}	0.53	0.55	0.56	**0.55**	-

dgp: M_8		$\rho = 0.2$		$\lambda = 0.8$		
		H_1				
		M_4	M_6	M_8	M_{10}	M_{12}
	M_4	-	**0.99**	0.99	0.99	0.99
	M_6	0.71	-	**0.71**	0.72	0.71
H_0	M_8	0.07	**0.05**	-	**0.06**	0.05
	M_{10}	0.17	0.18	**0.18**	-	0.18
	M_{12}	0.37	0.37	0.36	**0.34**	-

Table 3 and Table 4 show that when the decreasing neighbours approach is applied and when neither the null model M_{12} nor the alternative model M_{10} is the data generating model, the rejection probability is high, although it is not one. But in the next step of the sequence, when M_{10} is tested against the data generating model M_8 the rejection probability is much lower. For instance, when $\rho = 0.5$ and $\lambda = 0.2$ there is only a 77% probability of rejecting the null model, despite that the alternative model is the data generating process. When $\rho = 0.8$ and $\lambda = 0.2$ the rejection probability of the test in this step of the sequence is 69%. The implication is that when this approach to specification search is applied, the sequence might be stopped too early and we never reach the step where the null model reflects the data generating process. When this happens k will be specified too high in comparison to the data generating model. The outcome of the specification search by the decreasing neighbours approach is then an overspecified weights matrix.

When conducting tests in a sequence, ideally all rejection probabilities for false null hypothesis should be one, otherwise there is no guarantee that the sequence will come to the test where the null model corresponds to the data generating process. The results show that the increasing number of neighbours approach to specification search clearly performs better than the decreasing number of neighbours approach. According to the simulations, applying the increasing number of neighbours approach, all false null models are rejected and the significance level of the last test is unaffected by the sequence of test, as long as the spatial autocorrelation in the dependent variable is at least intermediate. Therefore, if the outcomes of the approaches are not the same, we suggest that the outcome of the increasing neighbours approach to specification search should be the choice for the number of neighbours in a k-nearest weights matrix. In the next section the strategy is illustrated by an empirical example of house price data.

5 An Empirical Illustration

A k-nearest neighbours specification of the weights matrix is often suitable in house price models. Assume that the buyer has chosen a certain neighbourhood where he or she wants to purchase a house. The buyer will then compare the available houses in that area. The price that he or she is willing to pay for a particular house is affected by the price level for houses in the neighbourhood. Also the seller's expectation of the price is affected by the

Table 3: Rejection probabilities of the null hypotesis of the spatial J-test when ρ is intermediate, $n = 900$ and $MC = 1000$.

dgp: M_8		$\rho = 0.5$		$\lambda = 0.2$		
		H_1				
		M_4	M_6	M_8	M_{10}	M_{12}
	M_4	-	**1.00**	1.00	1.00	1.00
	M_6	1.00	-	**1.00**	1.00	1.00
H_0	M_8	0.06	**0.06**	-	**0.06**	0.06
	M_{10}	0.72	0.75	**0.77**	-	0.70
	M_{12}	0.95	0.95	0.96	**0.95**	-

dgp: M_8		$\rho = 0.5$		$\lambda = 0.5$		
		H_1				
		M_4	M_6	M_8	M_{10}	M_{12}
	M_4	-	**1.00**	1.00	1.00	1.00
	M_6	1.00	-	**1.00**	1.00	1.00
H_0	M_8	0.05	**0.05**	-	**0.05**	0.05
	M_{10}	0.74	0.76	**0.78**	-	0.73
	M_{12}	0.95	0.96	0.97	**0.95**	-

dgp: M_8		$\rho = 0.5$		$\lambda = 0.8$		
		H_1				
		M_4	M_6	M_8	M_{10}	M_{12}
	M_4	-	**1.00**	1.00	1.00	1.00
	M_6	1.00	-	**1.00**	1.00	1.00
H_0	M_8	0.05	**0.05**	-	**0.05**	0.06
	M_{10}	0.70	0.70	**0.71**	-	0.69
	M_{12}	0.93	0.93	0.93	**0.93**	-

Table 4: Rejection probabilities of the null hypotesis for the spatial J-test when ρ is large, $n = 900$ and $MC = 1000$.

dgp: M_8		$\rho = 0.8$		$\lambda = 0.2$		
		H_1				
		M_4	M_6	M_8	M_{10}	M_{12}
	M_4	-	**1.00**	1.00	1.00	1.00
	M_6	1.00	-	**1.00**	1.00	1.00
H_0	M_8	0.04	**0.04**	-	**0.05**	0.05
	M_{10}	0.54	0.61	**0.69**	-	0.50
	M_{12}	0.69	0.77	0.83	**0.78**	-

dgp: M_8		$\rho = 0.8$		$\lambda = 0.5$		
		H_1				
		M_4	M_6	M_8	M_{10}	M_{12}
	M_4	-	**1.00**	1.00	1.00	1.00
	M_6	1.00	-	**1.00**	1.00	1.00
H_0	M_8	0.06	**0.06**	-	**0.06**	0.06
	M_{10}	0.58	0.65	**0.74**	-	0.58
	M_{12}	0.72	0.80	0.86	**0.81**	-

dgp: M_8		$\rho = 0.8$		$\lambda = 0.8$		
		H_1				
		M_4	M_6	M_8	M_{10}	M_{12}
	M_4	-	**1.00**	1.00	1.00	1.00
	M_6	1.00	-	**1.00**	1.00	1.00
H_0	M_8	0.05	**0.05**	-	**0.05**	0.05
	M_{10}	0.62	0.69	**0.74**	-	0.66
	M_{12}	0.73	0.81	0.86	**0.84**	-

selling prices of the other houses in the neighbourhood. The seller will be satisfied with a price that is at least as good as the average in the neighbourhood at the moment of the trade. The average price in the neighbourhood is on the other hand affected by the attributes of the neighbourhood itself, like available service, safety, or simply good reputation. Naturally, also supply and demand in a particular neighbourhood affect the price. A practical issue is how many of the nearby observations should be considered as nearest neighbours when the weights matrix is specified. In this section we illustrate how the proposed strategy for specifying the number of nearest neighbours can be of assistance when this choice has to be made and formally motivated.

Here we conduct a specification search for the number of nearest neighbours in real house price data. The data consist of 1377 transactions of single-family houses between January 2000 and May 2001 in the county of Stockholm, Sweden. The data were analyzed by Wilhelmsson (2002), and include the selling price, spatial coordinates and, in addition, information about the size of the house in square metres as well as other characteristics. See Wilhelmsson (2002) for a detailed description of the data.

We assume the same model as in the simulations and demonstrate how the suggested strategy is applied in practice. We then compare the outcomes of the approaches described in the two previous sections, and we verify that the results illustrate the findings of the simulations.

In the study of Wilhelmsson the results are in favour of a spatial error model. We therefore continue by applying the proposed strategy to specification search to Wilhelmsson's model. The outcome of this specification search is compared to the result of a Bayesian model comparison. For the Bayesian model comparison we use the routines in the Spatial econometric toolbox of LeSage (version 2010).

We begin with assuming that the model is a general spatial econometric model (8), as in the previous sections. We apply the increasing neighbours approach and start with $k = 2$. The model M_2 is tested against M_3, and M_2 is rejected. Then M_3 is tested against M_4. Again the null model is rejected. The procedure is continued until the null model is not rejected. As Table 5 shows this happens when M_{16} is tested against M_{17}, since for this test the J-statistic is 4.27, which is smaller than the critical value of $\chi^2_{0.95}(2) = 5.99$. Thereby the outcome is that $k = 16$.

When the decreasing neighbours approach is applied, we choose a large value of k, say $k = 24$. The model M_{24} is then tested against M_{23}. We find that M_{24} is rejected. The number of nearest neighbours is decreased and

Table 5: J-statistics of spatial J-tests in a sequence on house price data from Stockholm.

M_k = the general spatial econometric model				
Increasing neighbours approach				
$\mathbf{H}_0$	M_{13}	M_{14}	M_{15}	M_{16}
$\mathbf{H}_1$	M_{14}	M_{15}	M_{16}	M_{17}
	70	55	30	**4.27**
Decreasing neighbours approach				
$\mathbf{H}_0$	M_{24}	M_{23}	M_{22}	M_{21}
$\mathbf{H}_1$	M_{23}	M_{22}	M_{21}	M_{20}
	8	15	13	**4.35**

the testing procedure is continued. Table 5 shows that in the sequence of tests the null hypothesis is repeatedly rejected until we test M_{21} against M_{20}. Hence, by this approach we find that $k = 21$.

In this particular illustration the outcomes of the two approaches are not the same. Table 5 shows that the test values in the increasing neighbours approach are all large until the last test, where it is small. The null hypothesis is not rejected and the sequence is therefore stopped. In comparison to this, when the decreasing neighbours approach is applied there is no dramatic difference between the test value of the test stopping the sequence, and the test values of the previous tests in the sequence. Here all the J-statistics are small in comparison to the test values of the increasing neighbours approach. This illustrates the findings of the simulations in the previous section where we saw that the rejection probabilities of the tests in the sequence of the decreasing neighbours approach were lower than the ones in the increasing neighbours approach. If the rejection probability is low the implication is that the sequence of tests might stop too early. For these data we would therefore select the weights matrix that has 16 nearest neighbours.

In the study of Wilhelmsson the general spatial econometric model is not among the considered models for these data. His results are in favour of the spatial error model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, where $\mathbf{u} = \rho\mathbf{W}_k\mathbf{u} + \varepsilon$. To this model he applies a k-nearest neighbours weights matrix where $k = 4$. No motivation is given for this specific choice of k. We therefore apply the increasing neighbours approach on these data in order to formally specify k in the weights matrix of this model. The results are in Table 6. We begin the sequence at

Table 6: J-statistics of spatial J-tests in a sequence on house price data from Stockholm.

M_k = the spatial error model

Increasing neighbours approach										
$\mathbf{H}_0$	M_{11}	M_{12}	M_{13}	M_{14}						
$\mathbf{H}_1$	M_{12}	M_{13}	M_{14}	M_{15}						
	21.7	9.9	7.6	**5.2**						
Decreasing neighbours approach										
$\mathbf{H}_0$	M_{24}	M_{23}	M_{22}	M_{21}	M_{20}	M_{19}	M_{18}	M_{17}	M_{16}	M_{15}
$\mathbf{H}_1$	M_{23}	M_{22}	M_{21}	M_{20}	M_{19}	M_{18}	M_{17}	M_{16}	M_{15}	M_{14}
	3.4	**0.5**	**0.5**	**0.7**	**5.6**	**3.8**	**0.5**	**4.1**	**2.5**	**5.1**

M_2 against M_3 (not reported) and continue the testing until the null hypothesis is not rejected. The outcome is $k = 14$. As a comparison, Table 6 also contains the result of a decreasing number of neighbours approach. We find that this approach does not work on these data, since the null hypothesis is not rejected at any stage. The sequence of tests stops immediately at the first step when M_{24} is tested against M_{23}. The spatial J-test looses power when this strategy is applied, which again is in line with the findings in the simulations. The outcome of the suggested strategy is thereby 14 nearest neighbours, where as Wilhelmsson uses 4.

We conclude this empirical example by including a Bayesian model comparison. All spatial error models, M_4 to M_{24}, are estimated with uninformative priors. According to LeSage and Pace (2009) this results in estimates similar to maximum likelihood estimates. In Table 7 we can see that the preferred model entertains a nearest neighbours weights matrix with $k = 13$. This result is almost the same as the outcome of the increasing neighbours approach to specification search and we can conclude that the weights matrix with $k = 4$ is underspecified.

We note that when estimates of spatial error models with different k-nearest neighbours weights matrices (not reported) are compared, $k = 14$ and $k = 13$ give the same parameter estimates, but when $k = 4$ is applied the estimates are different. According to Pace et al. (2009), the exact specification of the weights matrix in large samples is not so important, since different weights matrices do not give materially different coefficient estimates. We argue though, that the specification is important in small samples.

Table 7: Model probabilities for spatial error models on house price data from Stockholm.

M_k = the spatial error model
Bayesian model comparison

M_k	M_4	...	M_{10}	M_{11}	M_{12}	M_{13}	M_{14}	...	M_{24}
Model Prob	0	...	0	0.056	0.195	**0.749**	0	...	0

6 Conclusions

This paper focuses on specification search for the number of neighbours in a k-nearest neighbours weights matrix of a general spatial econometric model. As the means of the specification search we use the spatial J-test proposed by Kelejian (2008). We suggest two approaches for finding the number of nearest neighbours k, the increasing and the decreasing number of neighbours approaches. When the increasing neighbours approach is applied we begin with a null model where k is small. Then k is gradually increased. In the resulting sequence of tests k is always smaller in the null model than it is in the alternative model. In the other approach we start with a large k, which is gradually decreased. Hence, in the resulting sequence of tests k is always larger in the null model than it is in the alternative model.

The results of the simulation study show that the spatial J-test can be used for distinguishing between general spatial models with different k-nearest neighbours weights matrices. The size of the test is acceptable and the power of the test is very high when the weights matrix of the null model is underspecified, especially if the parameter of the spatially lagged dependent variable ρ is at least moderate. If the weights matrix of the null model is overspecified, the power of the test is not equally high. The power of the test is low when the amount of spatial autocorrelation ρ in the dependent variable is small and the weights matrix of the null model is overspecified. If ρ is increased, the power of the test increases, although it never gets as high as when the weights matrix of the null model is underspecified. The results are therefore clearly in favour of the increasing neighbours approach.

When we examine the rejection probabilities of the tests in the sequences of the two considered approaches, we find a clear difference between the approaches. When ρ is at least moderate and the increasing neighbours approach is applied, all false models are rejected and the significance level of

the final test is unaffected by the sequential testing. However, if we conduct the specification search by the decreasing neighbours approach, all false models might not be rejected, since the sequence contains tests whose rejection probabilities are less than one. The implication is that when applying the decreasing neighbours approach we might stop the sequence of tests too early and the outcome is too large a k. Ideally the two approaches find the same k. But if this is not the case and we have to choose between the outcomes of the two approaches, the results support that the increasing neighbours approach should be used for specification search for the number of neighbours in a k-nearest weights matrix.

The paper also includes a specification search for the number of nearest neighbours on real data, which illustrate how the suggested strategy is applied in practice. For the general spatial econometric model we see that the outcomes of the approaches are not the same and that they illustrate the findings in the simulations. The suggested strategy is also applied to a spatial error model. The outcome of the strategy is different from what Wilhelmsson (2002) applied in his spatial error model for these data. However, the outcome of a Bayesian model comparison is close to the outcome of the increasing neighbours approach to specification search. The conclusion is that the suggested strategy can be applied for specification search when the objective is to specify k in a k-nearest neighbours weight matrix. It gives a formal justification for the choice of k which has been missing in the classical spatial econometric literature.

References

[1] Anselin, L. (1988) *Spatial Econometrics: Methods and Models,* Kluwer Academic Publishers, Dordrecht.

[2] Burridge, P. and Fingleton, B. (2010) Bootstrap Inference in Spatial Econometrics: The J Test. *Spatial Econometric Analysis* 5, 1, p.93-119.

[3] Davidson, R. and MacKinnon, J. (1981) Several tests for model specification in the presence of alternative hypothesis. *Econometrica* 49, p. 781-794

[4] Florax, R.J.G.M., Folmer, H. and Rey, S.J. (2006) A comment on specification searches in spatial econometrics: The relevance of Hendry's

methodology: A reply. *Regional Science and Urban Economics* 36, Issue 2, March 2006, p. 300-308.

[5] Florax, R.J.G.M., Folmer, H. and Rey, S.J. (2003) Specification searches in spatial econometrics: The relevance of Hendry's methodology. *Regional Science and Urban Economics*, 33, p. 557–579.

[6] Hendry, D. (2006) A comment on "Specification searches in spatial econometrics: The relevance of Hendry's methodology" *Regional Science and Urban Economics* Volume 36, Issue 2, March 2006, p. 309-312.

[7] Kelejian, H.H. (2008) A Spatial J-Test for Model Specification Against a Single or a Set of Non-Nested Alternatives. *Letters in Spatial and Resource Sciences* 1, 1, p. 3-11.

[8] Kelejian, H.H. and Prucha, I.R.(1998) A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Parameter in a Spatial Model with Autoregressive Disturbances. *Journal of Real Estate Finance and Economics* 17, p.99-121.

[9] Kelejian, H.H. and Prucha, I.R.(1999) A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model. *International Economic Review* 40 (1999), p. 509-533.

[10] LeSage, J.P. (2005, 2010) Spatial Econometric Toolbox for Matlab, http://www.spatial-econometrics.com.

[11] LeSage, J.P. and Pace, R. K. (2009) *Introduction to Spatial Econometrics.*, CRC Press, Taylor and Francis Group.

[12] Maddala, G.S. (2001) *Introduction to Econometrics.,* 3rd Ed., John Wiley & Sons Ltd, England.

[13] Pace, R. K., LeSage, J. P. and Zhu, S. (2009) Impact of Cliff and Ord on the Housing and Real Estate Literature, *Geographical Analysis* 41, 418-424.

[14] Piras, G. and Lozano-Garcia N. (2008) Spatial J-test: Some Monte Carlo Evidence. Paper presented at the Annual Meeting of the RSAI in New York November 2008.

[15] Wilhelmsson, M. (2002), Spatial Models in Real Estate Economics, *Housing, Theory and Society*, 19, p. 92-101.

Bootstrap Spatial J-Tests for k-Nearest Neighbours

Niklas Ahlgren Linda Gerkman
Hanken School of Economics,
Department of Finance and Statistics

Abstract

The weights matrix plays an important role in the specification of spatial econometric models. It contains the assumed spatial structure of the data. The problem is that there are usually several alternative formulations. An important practical issue is therefore how to choose the weights matrix. This paper studies the properties of the spatial J-test when it is applied to discriminate between spatial models with different k-nearest neighbours weights matrices. We find that the asymptotic test is oversized in small samples. The bootstrap is found to correct the size of the asymptotic test in small samples. Regarding the power of the test, we find that if the null model entertains an underspecified weights matrix, the power of the test is high. If the null model entertains an overspecified weights matrix, the test has low power.
Keywords: Bootstrap, k-nearest neighbours; Spatial J-test; Weights matrix.

1 Introduction

The weights matrix plays an important role in the specification of spatial econometric models. It contains the assumed spatial structure of the variables in the model, and is assumed to be exogenous to the model (Anselin 1988). Ideally, the weights matrix is based on theory rather than on spatial patterns found in the data. The problem is that the economic theory may be subjectively interpreted, and may result in several different but equally realistic weights matrices. An important practical issue is therefore how to choose the weights matrix.

In the general spatial autoregressive model

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{W}_1\mathbf{y} + \mathbf{u}, \\ \mathbf{u} &= \lambda\mathbf{W}_2\mathbf{u} + \boldsymbol{\varepsilon}, \end{aligned} \tag{1}$$

the dependent variable $\mathbf{y}$ is explained by the exogenous variables in the matrix $\mathbf{X}$ and a spatial lag $\mathbf{W}_1\mathbf{y}$, where $\mathbf{W}_1$ is a weights matrix. The error term $\mathbf{u}$ is spatially autocorrelated with spatial lag $\mathbf{W}_2\mathbf{u}$, where $\mathbf{W}_2$ is a weights matrix, and $\boldsymbol{\varepsilon}$ is a vector of IID$(0, \sigma_\varepsilon^2)$ error terms. Here $\boldsymbol{\beta}$ is a vector of regression coefficients, and ρ and λ are scalar spatial autoregressive parameters. The spatial dependence is formulated in the weights matrices $\mathbf{W}_1$ and $\mathbf{W}_2$. In practice it often is assumed that $\mathbf{W}_1 = \mathbf{W}_2 = \mathbf{W}$, say.

In general, if the sample size is n and $\mathbf{y}$ is an $n \times 1$ vector, then the weights matrix is given by an $n \times n$ matrix

$$\mathbf{W} = \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{pmatrix}. \tag{2}$$

The weights matrix specifies which of the other units in the system that affect the observed value at some particular location. In this paper we focus on weights matrices based on the number of nearest neighbours, say k. Every unit is assigned k nearest neighbours according to the shortest distance in space. Equal weights are given to all k neighbours, and the weights matrix is row normalized. Therefore, every non-zero element of the weights matrix will be equal to $1/k$. In the spatial autoregressive model, the spatial lag $\mathbf{W}_1\mathbf{y}$, or $\mathbf{W}_2\mathbf{u}$, becomes an averages of the k-nearest observations on the dependent

variable, or the disturbances. Even if the average of, say, the four nearest neighbours contains exactly the same two neighbours as the average of the two nearest neighbours and two additional, they are weighted differently by a weights matrix with $k = 4$ than by a weights matrix with $k = 2$. Therefore, spatial models are non-nested if they entertain different weights matrices. The specification of the weights matrix may also be based on distance or contiguity. See for instance LeSage and Pace (2009) for different alternatives.

Kelejian (2008) suggests a spatial J-test in order to discriminate between spatial econometric models entertaining different weights matrices. The test is a Wald test. Piras and Lozano-Garcia (2008), and Burridge and Fingleton (2010) show that the spatial J-test can be used to discriminate between different types of weights matrices, for instance, contiguity, inverse distance or k-nearest neighbours weights matrices. Burridge and Fingleton (2010) study the performance of a bootstrap spatial J-tests in this context. They find that the bootstrap test is superior in most cases.

This paper focuses on the spatial J-test when it is used in order to distinguish between spatial models with different k-nearest neighbours weights matrices. If both the null and the alternative model entertain k-nearest neighbours weights matrices, the models are closer to each other than they are in the case with different types of weights matrices. In a set-up where the null and the alternative model are close, the power of the test may be low. We study the properties of the asymptotic test and compare it with a bootstrap version of the test in order to examine whether the properties of the asymptotic spatial J-test test can be improved by the bootstrap. The motivation for the bootstrap is that the asymptotic J-test may suffer from size distortion in small samples (Piras and Lozano-Garcia 2008, Gerkman 2010).

The paper is organised as follows. The next section presents the spatial J-test, and the bootstrap spatial J-test is introduced in Section 3. Section 4 contains a simulation study. The last section concludes.

2 Spatial J-Test

The J-test, introduced in a seminal paper by Davidson and MacKinnon (1981), is a test for non-nested hypothesis. The main idea of the J-test is to estimate a model which consists of the null model and the predicted value of the alternative model. Then the significance of the additional term

is tested.

Spatial econometric models are non-nested if they entertain different weights matrices. Kelejian (2008) suggests a spatial J-test to test a spatial model against one or more non-nested alternative models. Following Kelejian, the spatial model of the null hypothesis is

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{W}_1\mathbf{y} + \mathbf{u} = \mathbf{Z}\boldsymbol{\gamma} + \mathbf{u}, \quad (3) \\ \mathbf{u} &= \lambda\mathbf{W}_2\mathbf{u} + \boldsymbol{\varepsilon}, \quad (4) \end{aligned}$$

where $\mathbf{Z} = (\mathbf{X}, \mathbf{W}_1\mathbf{y})$, $\boldsymbol{\gamma} = (\boldsymbol{\beta}', \rho)'$. The vectors $\mathbf{y}, \mathbf{u}$ and $\boldsymbol{\varepsilon}$, the matrices $\mathbf{X}, \mathbf{W}_1$ and $\mathbf{W}_2$, the parameter vector $\boldsymbol{\beta}$, and the scalar parameters ρ and λ are as before. The sample is size n.

The alternative models are of the form

$$\begin{aligned} \mathbf{y} &= \mathbf{X_i}\boldsymbol{\beta}_i + \rho_i\mathbf{W}_{1i}\mathbf{y} + \mathbf{u}_i = \mathbf{Z}_i\boldsymbol{\gamma}_i + \mathbf{u}_i, \quad (5) \\ \mathbf{u}_i &= \lambda_i\mathbf{W}_{2i}\mathbf{u}_i + \boldsymbol{\varepsilon}_i, \qquad i = 1, ..., G, \end{aligned}$$

where G is a finite constant equal to the number of possible alternatives to the null model.

The J-test and its properties are described in detail in the appendix of Kelejian (2008). Our discussion is based on Kelejian. To derive the J-statistic, first solve (4) and to obtain $\mathbf{u} = (\mathbf{I} - \lambda\mathbf{W}_2)^{-1}\boldsymbol{\varepsilon}$. Insert the solution into (3), and pre-multiply the expression by $(\mathbf{I} - \lambda\mathbf{W}_2)$ to get

$$(\mathbf{I} - \lambda\mathbf{W}_2)\mathbf{y} = (\mathbf{I} - \lambda\mathbf{W}_2)\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}.$$

Define $\mathbf{y}(\lambda) = (\mathbf{I} - \lambda\mathbf{W}_2)\mathbf{y}$ and $\mathbf{Z}(\lambda) = (\mathbf{I} - \lambda\mathbf{W}_2)\mathbf{Z}$, and write

$$\mathbf{y}(\lambda) = \mathbf{Z}(\lambda)\boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \quad (6)$$

Likewise for the alternative models (5), define $\mathbf{Z}_i(\lambda_i) = (\mathbf{I} - \lambda_i\mathbf{W}_{2i})\mathbf{Z}_i$. Then for all $i = 1, ..., G$, estimate $\boldsymbol{\gamma}_i$ in the ith alternative model (5) by generalized spatial two stage least squares (Kelejian and Prucha 1998, 1999), and get $\widehat{\boldsymbol{\gamma}}_i$. Add the predictive power $\mathbf{Z}_i(\lambda_i)\widehat{\boldsymbol{\gamma}}$ of all $i = 1, ..., G$ alternative models to the null model (6) to get:

$$\mathbf{y}(\lambda) = \mathbf{Z}(\lambda)\boldsymbol{\gamma} + \sum_{i=1}^{G} \alpha_i[\mathbf{Z}_i(\lambda_i)\widehat{\boldsymbol{\gamma}}_i] + \boldsymbol{\varepsilon}, \quad (7)$$

where $\widehat{\boldsymbol{\gamma}}_i$ is the 2SLS estimate of $\boldsymbol{\gamma}_i$. The parameter α_i is a scalar. Given that the null model is true, $\alpha_i = 0$ for all $i = 1, ..., G$. Insert $\mathbf{Z}_i(\lambda_i) = (\mathbf{I} - \lambda_i \mathbf{W}_{2i})\mathbf{Z}_i$ into (7), and define $\phi_i = -\alpha_i \lambda_i$, to get

$$\mathbf{y}(\lambda) = \mathbf{Z}(\lambda)\boldsymbol{\gamma} + \sum_{i=1}^{G} \alpha_i [\mathbf{Z}_i \widehat{\boldsymbol{\gamma}}_i] + \sum_{i=1}^{G} \phi_i [\mathbf{W}_{2,i} Z_i \widehat{\boldsymbol{\gamma}}_i] + \boldsymbol{\varepsilon}.$$

Then let $\boldsymbol{\delta} = (\alpha_1, ..., \alpha_G, \phi_1, ..., \phi_G)'$. The test of the null model (3) against the alternative models (5) is simply a Wald test of $\boldsymbol{\delta} = \mathbf{0}$. At the significance level α, reject the null hypothesis if

$$J = \widehat{\boldsymbol{\delta}}' \widehat{\mathbf{V}}_{\widehat{\boldsymbol{\delta}}}^{-1} \widehat{\boldsymbol{\delta}} > \chi^2_{1-a}(2G),$$

where $\widehat{\mathbf{V}}_{\widehat{\boldsymbol{\delta}}}$ is the estimated small-sample variance-covariance matrix of $\widehat{\boldsymbol{\delta}}$.

3 Bootstrap

We denote bootstrap quantities by a '$*$'. For generating samples of bootstrap observations $\mathbf{y}^*$, we use a parametric bootstrap algorithm.

Algorithm 1 (Bootstrap J-test)

1. *Estimate the spatial autoregressive model (1) by the generalized method of moments (GMM) procedure suggested in Kelejian and Prucha (1999) to obtain the estimates $\widehat{\rho}$, $\widehat{\lambda}$ and $\widehat{\boldsymbol{\beta}}$.*

2. *Check whether $|\widehat{\rho}| < 1$ and $|\widehat{\lambda}| < 1$.*

3. *If both conditions are satisfied, generate bootstrap observations $\mathbf{y}^*$ from*

$$\begin{aligned} \mathbf{y}^* &= \mathbf{X}\widehat{\boldsymbol{\beta}} + \widehat{\rho}\mathbf{W}\mathbf{y}^* + \mathbf{u}^*, \\ \mathbf{u}^* &= \widehat{\lambda}\mathbf{W}\mathbf{u}^* + \boldsymbol{\varepsilon}^*, \end{aligned}$$

where $\boldsymbol{\varepsilon}^* \sim N_n(0, \mathbf{I})$.

If the number of bootstrap replications is B, repeated application of Algorithm 1 gives B bootstrap samples $\mathbf{y}_j^*$, which are used to compute bootstrap J-statistics J_j^*. The empirical distribution of the J_j^* is used to approximate

the distribution of J under the null hypothesis. The bootstrap critical value at the significance level α is given by the $1-\alpha$ quantile of the J_j^*.

Let $\widehat{J}$ denote the realized value of the J-statistic. For a test at significance level α we reject the null hypothesis if $\widehat{J}$ is larger than the bootstrap critical value.

In many cases in practice it is more convenient to translate the test statistics into p-values. The bootstrap p-value is defined as

$$p^* = \frac{1}{B}\sum_{j=1}^{B} I(J_j^* > \widehat{J}), \tag{8}$$

i.e. the fraction of the bootstrap samples for which J_j^* is larger than $\widehat{J}$. For a test at significance level α we reject the null hypothesis if $p^* < \alpha$.

4 Simulation Study

This section contains a simulation study of the properties of the asymptotic and bootstrap spatial J-tests in a k-nearest neighbours setting. The simulations are performed using Matlab. The function `randn` in Matlab, which uses Margasalia's ziggurat algorithm, is used to generate pseudo-random numbers. The Spatial Econometric Toolbox for Matlab by LeSage (version 2010) is used in the estimations.

The model of interest is the general spatial autoregressive model with the same k-nearest neighbours weights matrix $\mathbf{W}_k$ in the spatial lag and in the spatially correlated error term

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{W}_k\mathbf{y} + \mathbf{u}, \\ \mathbf{u} &= \lambda\mathbf{W}_k\mathbf{u} + \boldsymbol{\varepsilon}. \end{aligned} \tag{9}$$

We consider non-nested models M_k, which entertain different k-nearest neighbours weights matrices $\mathbf{W}_k$, where $k = \{4, 6, 8, 10, 12\}$.

By solving (9) and setting $\mathbf{W}_k = \mathbf{W}_8$, we get the data generating process (DGP)

$$\mathbf{y} = (\mathbf{I} - \rho\mathbf{W}_8)^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \rho\mathbf{W}_8)^{-1}(\mathbf{I} - \lambda\mathbf{W}_8)^{-1}\boldsymbol{\varepsilon}. \tag{10}$$

The matrix $\mathbf{W}_8$ in the DGP (10) is a weights matrix where every observation has $k = 8$ nearest neighbours. All weights matrices are based on coordinates on the two-dimensional plane. The coordinates are randomly drawn

integers from a uniform $[1, 10000]^2$ distribution. The studied sample size is $n = \{100, 400, 900, 2500\}$. When the sample size is growing the domain is not expanding, instead it becomes more dense. Commonly used terminology is 'fill-in asymptotics' as opposed to 'increasing domain asymptotics'. This means that the k nearest neighbours of a particular point are usually not the same when the sample size is different.

The matrix $\mathbf{X}$ contains a column of ones and three columns of explanatory variables, which are randomly drawn from a uniform $[-10, 10]$ distribution, $\boldsymbol{\beta}$ is a 4×1 vector of ones and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector which is randomly drawn from a normal distribution with mean zero and variance equal to one. Data are generated for all possible pairs of $\rho = \{0.2, 0.5, 0.8\}$ and $\lambda = \{0.2, 0.5, 0.8\}$. These values for the autoregressive parameters are chosen in order to represent small, medium and large positive spatial autocorrelation. The explanatory variables remain the same in all replications. The coordinates and the vector of errors $\boldsymbol{\varepsilon}$ are drawn 1000 times for each sample of size n.

To estimate the size of the spatial J-test, the null model M_8 which corresponds to the data generating process is tested against the other models. The power of the test is estimated as each one of the other models are tested against M_8. The nominal significance level of the tests in the simulations is 5%.

From the DGP (10) we generate samples $\mathbf{y}_m$, indexed by $m = 1, .., MC$. In each replication we compute the J-statistic $\widehat{J}_m$. The rejection probability of the test is then estimated by

$$\frac{1}{MC} \sum_{m=1}^{MC} I(\widehat{J}_m > \chi^2_{1-a}(2G)), \tag{11}$$

i.e. the fraction of the MC replications for which the value of the J-statistic $\widehat{J}_m$ is larger than the critical value $\chi^2_{1-a}(2G)$.

The size and power of the bootstrap spatial J-test can be estimated in an analogous way (Horowitz 1994, and Davidson and MacKinnon 2006). For each replication $m = 1, \ldots, MC$, generate B bootstrap J-statistics J^*_{mj} using Algorithm 1, and indexed by $j = 1, \ldots, B$. Compute $J^*_m(1-\alpha)$, the $1-\alpha$ quantile of the J^*_{mj}. The rejection probability of the bootstrap test is then estimated by

$$\frac{1}{MC} \sum_{m=1}^{MC} I(\widehat{J}_m > J^*_m(1-\alpha)), \tag{12}$$

i.e. the fraction of the MC replications for which the value of the J statistic $\widehat{J}_m$ is larger than the bootstrap critical value $J^*_m(1-\alpha)$.

The procedure above involves computing $MC \times (B+1)$ J-statistics. The computational time for estimating the rejection probability becomes prohibitive when MC and B are large. Commonly used values of MC and B are $MC = 10000$ and $B = 1000$, resulting in computing approximately 10 million J-statistics.

We use a procedure proposed by Davidson and MacKinnon (2006) for estimating the size and power of a bootstrap test which substantially reduces the computational burden. For each replication $m = 1, \ldots, MC$, compute the J-statistic $\widehat{J}_m$, and generate one bootstrap J-statistic J^*_m. Compute $J^*(1-\alpha)$, the $1-\alpha$ quantile of the J^*_m. The approximate rejection probability of the bootstrap test is then estimated by

$$\frac{1}{MC}\sum_{m=1}^{MC} I(\widehat{J}_m > J^*(1-\alpha)). \tag{13}$$

See Davidson and MacKinnon (2006) for more details.

Both procedures estimate the nominal rejection probability of the bootstrap test, i.e. the size and power of the bootstrap test at the nominal level α. The difference between (12) and (13) is in the estimation of the $1-\alpha$ quantile of the bootstrap replications. In (12) the $\widehat{J}_m$ are compared to different estimated critical values, whereas in (13) the $\widehat{J}_m$ are compared to the same estimated critical value. The amount of computation required to estimate (13) is $2 \times MC$ J-statistics, which is much smaller than $MC \times (B+1)$. Davidson and MacKinnon (2006) establish the validity of (13) under the assumption of independence between the bootstrap DGP and the test statistic. For the J-test, the bootstrap DGP and the J-statistic are asymptotically independent.

We begin by examining the size and power of the asymptotic and bootstrap spatial J-tests for the small sample size $n = 100$. The results are reported in Table 1. We find that the asymptotic spatial J-test is oversized, in particular when the spatial autocorrelation in the error term is large and the spatial autocorrelation in the dependent variable is small or intermediate. The bootstrap corrects the size for all combinations of ρ and λ. For instance, for $\rho = 0.2$ and $\lambda = 0.8$, when M_8 is tested against M_4, the bootstrap corrects the size from 11.4% to 5.7%.

Turning to power, it is interesting to note that for all combinations of ρ and λ, the power of both tests are prominently lower when a model entertaining an overspecified weights matrix, like M_{10} or M_{12}, is tested against the data generating alternative model, than when the null model entertains an underspecified weights matrix. For instance, when $\rho = 0.5$ and $\lambda = 0.5$, the power of the asymptotic spatial J-test is 81.8% when M_6 is tested against M_8, and 29.6% when M_{10} is tested against M_8. The powers of the corresponding bootstrap tests are 73.4% and 27.0%, respectively. Regarding the impact of ρ and λ, we find that the power of the test is low when the amount of spatial autocorrelation ρ in the dependent variable is small, and in particular when the weights matrix of the null model is overspecified. The power of the test is high when ρ is at least intermediate and the weights matrix of the null model is underspecified.

Table 2 shows the results when the sample size is $n = 400$. Also for this sample size the asymptotic test is slightly oversized, in particular when ρ is small or intermediate, and when λ is large. For instance, for $\rho = 0.2$ and $\lambda = 0.8$, when M_8 is tested against M_4, the size is 6.3%. The bootstrap corrects the size to 5.4%. For most values of ρ and λ, the size of the asymptotic and bootstrap tests are close to each other. Table 2 shows that when the underspecified null model M_4, or M_6 is tested against the model corresponding to the data generating process M_8, the power of the test is higher for $n = 400$ than for $n = 100$. This is what we would expect from the usual asymptotics. In contrast, when the null model entertains the overspecified weights matrix M_{10}, or M_{12}, we find that the power mostly decreases when the sample size is increased from $n = 100$ to $n = 400$.

The results for $n = 900$ are reported in Table 3. The size of both tests are very close to nominal significance level. As before, we find that when the null model is overspecified and ρ is small, the power is low. For all combinations of ρ and λ, the power increases for underspecified as well as for overspecified weights matrices, as the sample is increased from $n = 400$ to $n = 900$.

Table 4 presents the results for the largest sample size $n = 2500$. We note that the size and power of the asymptotic and bootstrap tests are very close to each other. The power of the test is now very high when the null model entertains an underspecified weights matrix. The power is in general high, even when the overspecified model M_{12} is tested against M_8. But when we test the model M_{10}, which entertains an overspecified weights matrix, against M_8, we find that the power when low if ρ is small.

Table 1: Size and power of the spatial J-test and the bootstrapped spatial J-test for $n = 100$, $MC = 1000$ and $B = 1$.

			Size				Power			
	$H_0:$		M_8	M_8	M_8	M_8	M_4	M_6	M_{10}	M_{12}
	$H_1:$		M_4	M_6	M_{10}	M_{12}	M_8	M_8	M_8	M_8
ρ	λ									
0.2	0.2	J	0.070	0.079	0.087	0.071	0.568	0.282	0.127	0.195
		J^*	0.053	0.063	0.062	0.052	0.507	0.235	0.100	0.169
0.2	0.5	J	0.078	0.063	0.070	0.070	0.519	0.272	0.134	0.200
		J^*	0.053	0.046	0.065	0.068	0.488	0.242	0.105	0.174
0.2	0.8	J	0.114	0.099	0.118	0.121	0.429	0.271	0.161	0.239
		J^*	0.057	0.056	0.054	0.051	0.358	0.206	0.102	0.162
0.5	0.2	J	0.056	0.068	0.064	0.065	0.971	0.833	0.300	0.474
		J^*	0.046	0.043	0.049	0.045	0.965	0.811	0.271	0.433
0.5	0.5	J	0.082	0.071	0.079	0.068	0.957	0.818	0.296	0.465
		J^*	0.073	0.050	0.048	0.038	0.942	0.734	0.270	0.354
0.5	0.8	J	0.081	0.077	0.097	0.088	0.886	0.655	0.273	0.380
		J^*	0.044	0.031	0.043	0.034	0.855	0.627	0.220	0.293
0.8	0.2	J	0.065	0.059	0.064	0.062	0.995	0.997	0.445	0.541
		J^*	0.036	0.038	0.027	0.029	0.993	0.968	0.395	0.517
0.8	0.5	J	0.088	0.075	0.059	0.066	0.994	0.955	0.388	0.533
		J^*	0.063	0.054	0.043	0.051	0.990	0.955	0.378	0.463
0.8	0.8	J	0.082	0.079	0.082	0.080	0.991	0.956	0.412	0.530
		J^*	0.071	0.060	0.070	0.072	0.987	0.946	0.351	0.429

Table 2: Size and power of the spatial J-test and the bootstrapped spatial J-test for $n = 400$, $MC = 1000$ and $B = 1$.

			Size				Power			
	$H_0:$		M_8	M_8	M_8	M_8	M_4	M_6	M_{10}	M_{12}
	$H_1:$		M_4	M_6	M_{10}	M_{12}	M_8	M_8	M_8	M_8
ρ	λ									
0.2	0.2	J	0.069	0.067	0.080	0.068	0.989	0.741	0.159	0.311
		J^*	0.067	0.066	0.080	0.070	0.989	0.706	0.140	0.259
0.2	0.5	J	0.041	0.048	0.040	0.048	0.961	0.625	0.164	0.317
		J^*	0.046	0.054	0.034	0.053	0.972	0.653	0.134	0.327
0.2	0.8	J	0.063	0.053	0.064	0.054	0.811	0.419	0.136	0.163
		J^*	0.054	0.058	0.054	0.052	0.817	0.378	0.116	0.192
0.5	0.2	J	0.072	0.056	0.062	0.068	1.000	1.000	0.516	0.773
		J^*	0.059	0.046	0.067	0.080	1.000	1.000	0.531	0.776
0.5	0.5	J	0.056	0.049	0.045	0.045	1.000	1.000	0.511	0.752
		J^*	0.054	0.065	0.042	0.039	1.000	1.000	0.518	0.762
0.5	0.8	J	0.058	0.071	0.071	0.069	1.000	0.979	0.433	0.669
		J^*	0.047	0.047	0.074	0.070	1.000	0.975	0.422	0.599
0.8	0.2	J	0.047	0.046	0.053	0.050	1.000	1.000	0.497	0.648
		J^*	0.046	0.041	0.048	0.047	1.000	1.000	0.451	0.651
0.8	0.5	J	0.053	0.044	0.047	0.052	1.000	1.000	0.507	0.654
		J^*	0.041	0.037	0.047	0.053	1.000	1.000	0.465	0.642
0.8	0.8	J	0.065	0.057	0.056	0.063	1.000	1.000	0.509	0.642
		J^*	0.054	0.042	0.045	0.059	1.000	1.000	0.448	0.647

Table 3: Size and power of the spatial J-test and the bootstrapped spatial J-test for $n = 900$, $MC = 1000$ and $B = 1$. The bold faced cases are the ones referred to in the text.

			Size				Power			
	$H_0:$		$\mathbf{M}_8$	$\mathbf{M}_8$	$\mathbf{M}_8$	$\mathbf{M}_8$	$\mathbf{M}_4$	$\mathbf{M}_6$	$\mathbf{M}_{10}$	$\mathbf{M}_{12}$
	$H_1:$		$\mathbf{M}_4$	$\mathbf{M}_6$	$\mathbf{M}_{10}$	$\mathbf{M}_{12}$	$\mathbf{M}_8$	$\mathbf{M}_8$	$\mathbf{M}_8$	$\mathbf{M}_8$
ρ	λ									
0.2	0.2	J	0.041	0.061	0.057	0.068	1.000	1.000	**0.317**	**0.606**
		J^*	0.040	0.058	0.056	0.050	1.000	0.970	**0.310**	**0.582**
0.2	0.5	J	0.046	0.050	0.047	0.048	0.999	0.933	**0.281**	**0.560**
		J^*	0.036	0.038	0.036	0.036	0.999	0.935	**0.277**	**0.532**
0.2	0.8	J	0.043	0.048	0.048	0.048	0.985	0.730	**0.208**	**0.376**
		J^*	0.038	0.041	0.031	0.036	0.985	0.728	**0.181**	**0.359**
0.5	0.2	J	0.043	**0.035**	0.055	0.053	1.000	1.000	0.762	0.953
		J^*	0.040	**0.029**	0.048	0.044	1.000	1.000	0.762	0.941
0.5	0.5	J	0.059	0.047	0.048	0.046	1.000	1.000	0.811	0.965
		J^*	0.072	0.051	0.046	0.047	1.000	1.000	0.831	0.965
0.5	0.8	J	0.053	0.058	0.049	0.049	1.000	1.000	0.728	0.941
		J^*	0.049	0.058	0.045	0.047	1.000	1.000	0.731	0.957
0.8	0.2	J	0.051	0.053	0.061	0.059	1.000	1.000	0.717	0.834
		J^*	0.050	0.043	0.059	0.058	1.000	1.000	0.673	0.797
0.8	0.5	J	0.048	0.058	0.044	0.046	1.000	1.000	0.703	0.849
		J^*	0.043	0.040	0.036	0.035	1.000	1.000	0.684	0.825
0.8	0.8	J	0.054	0.059	0.054	0.056	1.000	1.000	0.762	0.845
		J^*	0.057	0.059	0.057	0.061	1.000	1.000	0.772	0.858

Table 4: Size and power of the spatial J-test and the bootstrapped spatial J-test for $n = 2500$, $MC = 1000$ and $B = 1$.

			Size				Power			
	$H_0:$		M_8	M_8	M_8	M_8	M_4	M_6	M_{10}	M_{12}
	$H_1:$		M_4	M_6	M_{10}	M_{12}	M_8	M_8	M_8	M_8
ρ	λ									
0.2	0.2	J	0.057	0.055	0.062	0.054	1.000	1.000	0.657	0.949
		J^*	0.051	0.046	0.053	0.050	1.000	1.000	0.650	0.954
0.2	0.5	J	0.046	0.054	0.047	0.056	1.000	1.000	0.627	0.940
		J^*	0.049	0.058	0.051	0.050	1.000	1.000	0.648	0.940
0.2	0.8	J	0.053	0.044	0.052	0.052	1.000	0.981	0.447	0.799
		J^*	0.044	0.037	0.037	0.043	1.000	0.978	0.422	0.794
0.5	0.2	J	0.052	0.048	0.055	0.046	1.000	1.000	0.999	1.000
		J^*	0.056	0.062	0.077	0.057	1.000	1.000	0.999	1.000
0.5	0.5	J	0.045	0.054	0.051	0.044	1.000	1.000	0.995	1.000
		J^*	0.045	0.039	0.057	0.043	1.000	1.000	0.993	1.000
0.5	0.8	J	0.044	0.037	0.047	0.040	1.000	1.000	0.982	1.000
		J^*	0.044	0.031	0.041	0.039	1.000	1.000	0.982	1.000
0.8	0.2	J	0.059	0.064	0.063	0.064	1.000	1.000	0.950	0.991
		J^*	0.050	0.064	0.051	0.053	1.000	1.000	0.950	0.990
0.8	0.5	J	0.045	0.050	0.056	0.053	1.000	1.000	0.967	0.998
		J^*	0.039	0.041	0.050	0.046	1.000	1.000	0.959	0.998
0.8	0.8	J	0.049	0.048	0.045	0.048	1.000	1.000	0.989	0.999
		J^*	0.046	0.037	0.036	0.035	1.000	1.000	0.989	0.998

5 Conclusions

This paper studies the properties of the asymptotic spatial J-test proposed by Kelejian (2008). We suggest and study a bootstrap spatial J-test in order to see whether the properties of the test in finite samples can be improved. The simulation study shows that the spatial J-test can be used for distinguishing between general spatial models with different k-nearest neighbours weights matrices. We find that the asymptotic test is oversized in small samples. The bootstrap is useful for correcting the size of the asymptotic test. Turning to power, we find that when the sample size is small, the amount of spatial autocorrelation in the dependent variable is small, and the weights matrix of the null model is overspecified, the power of the test is low. In contrast, the power of the test is high when the spatial autocorrelation in the dependent variable is at least intermediate, and the weights matrix of the null model is underspecified. When the sample size is increased and the null model entertains an underspecified weights matrix, the power increases. For large samples we find that the power of the test is very high when the null model entertains an underspecified weights matrix. The power of the asymptotic test is very close to the power of the bootstrapped version.

References

[1] Anselin, L. (1988) *Spatial Econometrics: Methods and Models.* Kluwer Academic Publishers, Dordrecht.

[2] Burridge, P. and Fingleton, B. (2010) Bootstrap Inference in Spatial Econometrics: The J Test, *Spatial Econometric Analysis*, 5, 1, 93–119.

[3] Davidson, R. and MacKinnon, J. G. (1981) Several Tests for Model Specification in the Presence of Alternative Hypothesis, *Econometrica,* 49, 781–794

[4] Davidson, R. and MacKinnon, J. G. (2006) The power of bootstrap and asymptotic tests. *Journal of Econometrics,* 133: 421–441.

[5] Gerkman, L. (2010) A Practical Proposal to Specification Search of a k-Nearest Neighbours Weights Matrix, Manuscript. Hanken School of Economics, Helsinki.

[6] Horowitz (1994) Bootstrap-Based Critical Values for the Information Matrix Test, *Journal of Econometrics*, 61: 395–411.

[7] Kelejian, H.H. and Prucha, I.R.(1999) A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model. *International Economic Review,* 40 (1999), 509–533.

[8] Kelejian, H.H. (2008) A Spatial J-Test for Model Specification Against a Single or a Set of Non-Nested Alternatives. *Letters in Spatial and Resource Sciences,* 1, 1, 3–11.

[9] LeSage, J.P. and Pace, R. K. (2009) *Introduction to Spatial Econometrics.*. CRC Press, Taylor and Francis Group.

[10] Piras, G. and Lozano-Garcia N. (2008) Spatial J-test: Some Monte Carlo Evidence, paper presented at the Annual Meeting of the RSAI in New York November 2008.

EKONOMI OCH SAMHÄLLE
Skrifter utgivna vid Svenska handelshögskolan

ECONOMICS AND SOCIETY
Publications of the Hanken School of Economics

185. MARIA SUOKANNAS: Den anonyma seniorkonsumenten identifieras. Om identitetsskapande processer i en marknadsföringskontext. Helsingfors 2008.

186. RIIKKA SARALA: The Impact of Cultural Factors on Post-Acquisition Integration. Domestic and Foreign Acquisitions of Finnish Companies in 1993-2004. Helsinki 2008.

187. INGMAR BJÖRKMAN et al. (Eds.): Innovation, Leadership, and Entrepreneurship. A Festschrift in Honour of Professor Martin Lindell on his 60th Birthday. Helsinki 2008.

188. JOACIM TÅG: Essays on Platforms. Business Strategies, Regulation and Policy in Telecommunications, Media and Technology Industries. Helsinki 2008.

189. HENRIK TÖTTERMAN: From Creative Ideas to New Emerging Ventures. Entrepreneurial Processes Among Finnish Design Entrepreneurs. Helsinki 2008.

190. ANNIKA RAVALD: Hur uppkommer värde för kunden? Helsingfors 2008.

191. TOM LAHTI: Angel Investing in Finland: An Analysis Based on Agency Theory and the Incomplete Contracting Theory. Helsinki 2008.

192. SYED MUJAHID HUSSAIN: Intraday Dynamics of International Equity Markets. Helsinki 2009.

193. TEEMU TALLBERG: The Gendered Social Organisation of Defence. Two Ethnographic Case Studies in the Finnish Defence Forces. Helsinki 2009.

194. JONAS HOLMQVIST: Language Influence in Services. Perceived Importance of Native Language Use in Service Encounters. Helsinki 2009.

195. ENSIO ERÄ-ESKO: Beskattningsrätt och skattskyldighet för kyrkan i Finland. Steuerrecht und Versteuerung der Kirche in Finnland. Mit einer deutschen Zusammenfassung. Helsingfors 2009.

196. PIA BJÖRKWALL: Nyttighetsmodeller - ett ändamålsenligt innovationsskydd? Helsingfors 2009.

197. ARTO THURLIN: Essays on Market Microstructure. Price Discovery and Informed Trading. Helsinki 2009.

198. PETER NYBERG: Essays on Risk and Return. Helsinki 2009.

199. YANQING JIANG: Growth and Convergence: The Case of China. Helsinki 2009.

200. HANNA WESTMAN: Corporate Governance in European Banks. Essays on Bank Ownership. Helsinki 2009.

201. CATHARINA von KOSKULL: Use of Customer Information. An Ethnography in Service Development. Helsinki 2009.

202. RITVA HÖYKINPURO: Service Firms' Action upon Negative Incidents in High Touch Services: A Narrative Study. Helsinki 2009.

203. SUVI NENONEN: Customer Asset Management in Action. Using Customer Portfolios for Allocating Resources Across Business-to-Business Relationships for Improved Shareholder Value. Helsinki 2009.

204. CAMILLA STEINBY: Multidimensionality of Actors in Business Networks. The Influence of Social Action in Pharmacy Networks in Finland. Helsinki 2009.

205. JENNIE SUMELIUS: Developing and Integrating HRM Practices in MNC Subsidiaries in China. Helsinki 2009.

206. SHERAZ AHMED: Essays on Corporate Governance and the Quality of Disclosed Earnings – Across Transitional Europe. Helsinki 2009.

207. ANNE HOLMA: Adaptation in Triadic Business Relationship Settings. A Study in Corporate Travel Management. Helsinki 2009.

208. MICHAL KEMPA: Monetary Policy Implementation in the Interbank Market. Helsinki 2009.

209. SUSANNA SLOTTE-KOCK: Multiple Perspectives on Networks. Conceptual Development, Application and Integration in an Entrepreneurial Context. Helsinki 2009.

210. ANNA TALASMÄKI: The Evolving Roles of the Human Resource Function. Understanding Role Changes in the Context of Large-Scale Mergers. Helsinki 2009.

211. MIKAEL M. VAINIONPÄÄ: Tiering Effects in Third Party Logistics: A First-Tier Buyer Perspective. Helsinki 2010.

212. ABDIRASHID A. ISMAIL: Somali State Failure. Players, Incentives and Institutions. Helsinki 2010.

213. ANU HELKKULA: Service Experience in an Innovation Context. Helsinki 2010.

214. OLLE SAMUELSON: IT-innovationer i svenska bygg- och fastighetssektorn. En studie av förekomst och utveckling av IT under ett decennium. Helsingfors 2010.

215. JOANNA BETH SINCLAIR: A Story about a Message That was a Story. Message Form and Its Implications to Knowledge Flow. Helsinki 2010.

216. TANJA VILÉN: Being in Between. An Ethnographic Study of Opera and Dialogical Identity Construction. Helsinki 2010.

217. ASHIM KUMAR KAR: Sustainability and Mission Drift in Microfinance. Empirical Studies on Mutual Exclusion of Double Bottom Lines. Helsinki 2010.

218. HERTTA NIEMI: Managing in the "Golden Cage". An Ethnographic Study of Work, Management and Gender in Parliamentary Administration. Helsinki 2010.