Sciknow Publications Ltd.
Econometrics

E 2013, 1(2): 20-30
Print ISSN: 2328-5044
Online ISSN: 2328-5052
DOI: 10.12966/e.05.01.2013

Spatial Weighting Matrix Selection in Spatial Lag Econometric Model

Philip Kostov*

Lancashire Business School, University of Central Lancashire, Preston, UK

*Corresponding author (Email: pkostov@uclan.ac.uk)

Abstract – This paper investigates the choice of spatial weighting matrix in a spatial lag model framework. In the empirical literature the choice of spatial weighting matrix has been characterized by a great deal of arbitrariness. The number of possible spatial weighting matrices is large, which until recently was considered to prevent investigation into the appropriateness of the empirical choices. Recently Kostov (2010) proposed a new approach that transforms the problem into an equivalent variable selection problem. This article expands the latter transformation approach into a two-step selection procedure. The proposed approach aims at reducing the arbitrariness in the selection of spatial weighting matrix in spatial econometrics. This allows for a wide range of variable selection methods to be applied to the high dimensional problem of selection of spatial weighting matrix. The suggested approach consists of a screening step that reduces the number of candidate spatial weighting matrices followed by an estimation step selecting the final model. An empirical application of the proposed methodology is presented. In the latter a range of different combinations of screening and estimation methods are employed and found to produce similar results. The proposed methodology is shown to be able to approximate and provide indications to what the 'true' spatial weighting matrix could be even when it is not amongst the considered alternatives. The similarity in results obtained using different methods suggests that their relative computational costs could be primary reasons for their choice. Some further extensions and applications are also discussed.

Keywords – Spatial Weighting Matrix, Variable Selection, Spatial Lag, Spatial Econometrics

1. Introduction

Models of 'spatial' dependence have recently become increasingly popular in the regional science literature. In spatial econometrics the spatial dependence is typically represented via either spatial lag or spatial error specification. The potential underlying causes and reasons for these two distinct forms of spatial dependence are rather different and in many cases explicitly distinguishing between them is of major interest, particularly when a substantive understanding of the underlying processes generating the spatial dependence patterns is desired. Technically speaking however the spatial lag representation is much more interesting for two main reasons. First, ignoring spatial lag dependence has more serious implications when inference is concerned. The resulting estimates are typically inconsistent and biased. In contrast, ignoring spatial error dependence leads to consistent, though inefficient estimates, in the same manner as in any other heteroscedastic model. Second, the spatial lag representation nests within itself both spatial lag and spatial error dependence in the sense that the spatial error model can have an alternative representation that technically resembles the spatial lag representation. In a linear model the spatial error representation is a (testable) restriction on the spatial lag model.

This paper looks at the choice of spatial weighting matrix. When this is the focus of analytical attention, the question about exact nature of spatial dependence is of secondary importance and both forms can be subsumed in a spatial lag type of specification (strictly speaking one may want to use the more general so called spatial Durbin model specification, but here we will ignore such technical issues and focus on the spatial lag representation only). This is by no means restrictive, since once the precise type of spatial weighting matrix has been determined, one can go further into investigating which form of spatial dependence is present, if this is of interest. Why is the choice of spatial weighting matrix important? If existing spatial lag dependence is ignored the resulting parameter estimates will in general be biased (see e.g. Anselin, 2002). But similarly if spatial lag dependence is included when the true model does not exhibit it, it is accounting for a general model misspecification, which could also result in erroneous inference (McMillen, 2003). One can also have the case where spatial dependence is existing, but the wrong spatial weighting matrix is used. In such cases Griffith and Lagona (1998) show that the mean estimates could be consistent (although under non-trivial conditions), while the variance estimates will be typically biased and inefficient thus impeding inference.

The paper proceeds as follows. The next section briefly reviews the issues surrounding the spatial weighting matrix in terms of significance, alternative specifications etc. Then the general background of the proposal is discussed. This describes the motivation and the philosophical basis for the proposed methodology. The actual methodology is then presented clarifying the technical details and the reason behind their choices. An empirical application of the proposed methods is presented. Finally the obtained results are presented and interpreted and some conclusions and possible future extensions are briefly outlined.

2. The Spatial Weighting Matrix

The specification of spatial dependence via a spatial weighting matrix is a convenient way to describe theoretical or a priori knowledge and understanding of the underlying structure generating the 'spatial' dependence between different economic agents and units of analysis. There are different approaches to specifying a spatial weighting matrix (see Getis, 2009 for an overview). In simple words defining a spatial weighting matrix involves two choices, namely a neighbourhood scheme and spatial weights. The neighbourhood scheme involves determination of which units of analysis are linked and which are not. When units are economic agents this means the decisions of which agents are to be included in the objective functions of other agents. A social network structure could for example be used to infer the neighbourhood scheme. The weighting scheme on the other hand defines the strength of these links. The weighting scheme is based on some distance metrics, which could be spatial, economic distance, or in the case of the social network example a social distance (e.g. family, close friends, acquaintances etc.). The weighting scheme takes the distance metrics and combines it in order to derive the strength of the impact each unit has on another unit.

In some applications some of the above choices may be logically predetermined, e.g. the nature of the problem may suggest the neighbourhood scheme and/or equal weights could be a logical choice. In most cases however this choice is far from trivial. The choice of spatial weighting matrix in empirical applications has been subject to some arbitrariness. This arbitrariness presents a serious problem to the inference in such models since estimation results have been shown to often critically depend on the choice of spatial weighting matrix (Anselin, 2002; Fingleton, 2003).

For identification purposes the spatial weighting matrix needs to be exogenous (Manski, 1993). One reason for the popularity of spatial weighting matrices based on geographical distances is the fact that their exogeneity is automatically ensured. Furthermore very often spatial distances may reasonably well approximate the underlying 'true' metrics, which may be unobservable or unavailable. For example often spatial distance can approximate the strength of social relationships. Therefore in the absence of direct measurement of the underlying relationship, the spatial distances could be used. Note however that in such an approximation process even if one knows the exact form of the linkages, as expressed in the underlying unavailable metrics, translation into spatial distances (or any other alternative metrics system) changes matters. The translation may effectively break down the theoretical spillover definition. Hence the uncertainty about what the spatial distances measure introduces additional uncertainty in the process of specifying an appropriate spatial weighting matrix.

The issue of spatial weighting matrix has been outstanding for considerable amount of time. There have been a number of proposals how to alleviate the problem. A major stumbling block in identifying an appropriate spatial weighting matrix is that the number of potential alternatives is extremely large. This puts a great computational burden on any method designed to deal with it.

Kooijman (1976) proposed to choose the spatial weighting matrix by maximizing Moran's coefficient. In a more general vein this has led to the practice of choosing spatial weighting matrix maximising alternative spatial dependence statistics. Research into reducing the degree of arbitrariness in spatial weighting matrix choice has been particularly active in recent years. One could classify this strand of research into two main types. First, new and more flexible ways to specify the neighbourhood and/or the weighting schemes have been proposed. The second type of proposals deals with essentially selecting the spatial weighting matrix either implicitly or explicitly from a pre-defined set of candidates. This paper falls in this second category. Bhattacharjee and Jensen-Butler (2005) proposed estimating spatial weighting matrix consistent with the data distribution, but their approach only applies to the spatial error model. Lima and Macedo (1999) proposed an interesting procedure dealing with estimating the weights decay and thus the spatial weights matrix with a predefined 'soft' neighbourhood (soft in the sense that the weight decay can exclude some observations from the neighbourhood definition). When we have an explicit set of competing spatial weighting matrices, LeSage and Parent (2007) proposed a Bayesian model averaging procedure for spatial model which incorporates the uncertainty about the correct spatial weighting matrix while LeSage and Fischer (2008) expanded this approach to select a spatial weighting matrix. Holloway and Lapar (2007) used a Bayesian marginal likelihood approach to select a neighbourhood definition (cut-off points for the neighbourhood), but one can consider their approach as a general model selection approach, which could be applied to any other set of competing models.

Recently Kostov (2010) proposed applying a component-wise boosting algorithm to a reformulated spatial weighting matrix selection problem. Kostov's (2010) proposal is computationally efficient in that it can deal with thousands of alternatives. We build upon Kostov's (2010) proposal and extend it by applying alternative variable selection methods. The paper is organised as follows. The next section reviews the proposal of Kostov (2010). Then we outline our method and its justification. Finally we apply the proposed methods to the same data as in Kostov's (2010) original application in order to compare the results.

3. Background and Motivation

Kostov's (2010) approach implements a component-wise boosting counterpart to the spatial two-stage least squares approach of Kelejian and Prucha (1998). The latter uses the spatially lagged independent variables as instruments for the spatially lagged dependent variable. Thus one can simply project the spatially lagged dependent variable in the vector space of the instruments and use the transformed in this way variable instead of the original one. The novelty of Kostov's (2010) approach consists in applying a variable selection method in the second step. In simple words the first step in the spatial two-stage least squares approach (Kelejian and Prucha, 1998) can be viewed as instrumental variables transformation applied to a spatially lagged dependent variable. Kostov (2010) proposes applying the above transformation to all candidates for spatial weighting matrices to be considered in an empirical application. Then treating the first step as given, the spatial weighting matrix selection problem becomes equivalent to a variable selection problem, defined with regard to the transformed spatially dependent variables. In a parametric modelling framework, the latter variable selection problem can be dealt with standard tools. Kostov (2010) further proposes component-wise boosting for this particular purpose, partly motivated from the fact that the potential set of spatial weighting matrices can be very large thus requiring methods able to carry out variable selection in an ultra high dimensional case at a low computational cost. As already noted the essence of Kostov's (2010) proposal is not so much the component-wise boosting method, but rather the transformation of the spatial weighting matrix selection into a variable selection problem. Therefore any other variable selection methods could be used in the second step. A popular class of variable selection methods are penalisation methods, such as the nonnegative garrote (Breiman, 1994), LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2002), LARS (Efron, et al. 2004), the bridge estimator (Huang et al., 2007) and the Dantzig selector (Candes and Tao, 2007). See Kostov (2010) for a brief overview of these methods.

In this paper we will consider the penalisation approaches. The reasons for this are briefly outlined below. A desirable property of any variable selection method is the so called 'oracle property' (Fan and Li, 2001, 2006). In simple words an estimator is said to possess the oracle property when it is both consistent in terms of variable selection and efficient in estimation in the sense that the estimator's asymptotic variance matrix is essentially the same as this of the 'oracle' estimator (i.e. the estimator obtained by knowing which variables have to be selected). Fan and Li (2001, 2006) provide detailed technical discussion on the oracle property and we will not discuss it here in any detail. Kostov (2010) claims that the oracle property is not necessary in justifying his approach. The reason for this seems rather intuitive. The proposed method is a two-step equivalent to the spatial two-stage least squares. It is however computationally complicated to obtain covariance estimates for the overall approach. Owing to this Kostov (2010) suggested that the methodology has to be used to obtain the final model that will need to be estimated by the standard spatial two-stage least squares approach. Hence by differentiation between the consistency (in terms of variable selection) and efficiency (in the oracle sense) it looks like only consistency is required, since the results will after all be obtained by applying the 'oracle' estimator.

Unfortunately the above logic suffers from an important drawback. A variable selection method that does not possess the oracle property may fail do identify the oracle model. Being consistent in terms of variable selection means that the variables that belong to the model will be retained. Nevertheless this does not guarantee that a number of irrelevant variables would not be retained too. In order to better explain the intuition behind this, consider the following. The variable selection methods would typically need a criterion to define how to select a crucial parameter (the number of boosting iterations in boosting or the value of the penalty parameter in penalisation approaches). This is designed to avoid over-fitting. Conventional methods, such as e.g. cross-validation would typically select over-parameterised models (see e.g. James and Radchenko, 2009). The reason for this is that such methods are constructed with fixed designs in mind while in variable selection problems this is no longer the case. As a result the basic variable selection algorithms need to be modified to account for this. The SCAD method uses two penalty parameters to correct for this problems, the adaptive lasso (Zou, 2006) applies adaptive weighting to the classical lasso estimates, the relaxed lasso (Meinshausen, 2007) interpolates between two estimates to attenuate the problem and the double Dantzig (James and Radchenko, 2009) applies similar logic. Since we are interested in correctly identifying the important variables in such setting, it is desirable that our variable selection methods possess the oracle property. To be more precise, in this particular setting we are not interested in the oracle property, but in the rather weaker 'persistency' property (Greenshtein and Ritov 2004). Nonetheless the oracle property would be desirable. Another argument for it would be the fact that the set of candidate spatial weighting matrices, that needs to be constructed by the researcher is not guaranteed to contain the 'true' one. In this case the results would approximate the unknown 'true' spatial weighting matrix. When such approximations are involved, the prediction properties of the model become important and therefore the stronger oracle property could be useful in achieving efficient approximation.

4. Methodology

The discussion above does not imply that methods that do not possess the oracle property are not useful. Even without the oracle property, the variable selection consistency ensures that the relevant variables are retained within the set of predictors. Therefore any consistent variable selection methods can still be employed as screening methods to greatly reduce the set of candidate variables. When the latter set is very large, as it is in the case of spatial weighting matrix choice, this is an advantageous development. Another important point of consideration is that the rate of convergence of variable selection algorithms depends on the dimension of the problem. It is therefore advantageous if the initial problem is pre–screened in eliminating irrelevant variables to reduce its dimensionality. Such a strategy will bring two distinct types of advantages. The first is the improvement in the rate of convergence of the employed variable selection algorithm, which will improve the results. The second is more practical. Whenever the screening method is a simple and computationally fast, reducing the dimensionality of the problem will considerably speed up estimation when compared to a direct application of variable selection to the larger problem. Therefore we suggest implementing the variable selection task in two steps, namely a screening step that eliminates (most of the) irrelevant variables (in this case candidates spatial weighting matrices) followed by a variable selection procedure that obtains the final model. Below we briefly discuss what particular methods could be implemented at each of these steps.

Without entering into too much technical detail we can state that most of the variable selection algorithms mentioned in the previous section can be used as screening methods. From a practical point of view however it is advisable to use simple and computationally cheap methods. The screening is to be applied to the whole problem and more complicated methods could be computationally demanding. The general idea behind screening is rather simple. Screening methods reduce the dimensionality of the problem and then a variable selection method possessing the oracle property can be used to infer the final structure. This approach has several important advantages. First, the dimension reduction allows one to be able to use methods that would otherwise have been infeasible with the original problem. Take for example the adaptive lasso method. It is not applicable when the number of variables exceeds the sample size, but when screening that reduces the number of candidate variables so that it is lower than the sample size is carried out, it can be implemented. The other advantage is that once irrelevant variables have been filtered out, the resulting estimator will have better convergence rate compared to being applied to the original unrestricted problem. Take for example the Dantzig selector, the convergence of which is a function of the relative (with regard to the sample) size of the problem. Screening will drastically increase its converge rate and hence result in more reliable inference (see Fan and Lv, 2008 for detailed discussion). Finally, since most consistent variable selection methods possess screening power, irrespective of whether they are characterised by the oracle property or not, it would be advantageous to combine different types of such methods in a consecutive matter.

The screening idea originates from Fan and Lv (2008) who proposed and justified (by establishing its theoretical properties) the so called Sure Independence Screening method designed to reduce the dimensionality of the variable selection problem. The ISIS method of Fan and Lv (2008) which is an iterative version of the basic SIS is numerically similar to component-wise boosting, but is less greedy. Wang (2009) established the screening properties of the classical forward regression, which can be viewed as limiting greedy learning case of the boosting algorithm (achieved with the maximum updating factor of unity). Taking the above connections into considerations, as well as the general links amongst different variable selection algorithms (see e.g. Meinshausen *et al*, 2007) it would be advantageous to combine different screening and variable selection methods. A particular concern in the present application is the fact that by construction the variables created using a set of candidate spatial weighting matrices, following the proposal of Kostov (2010) will exhibit considerable correlation. The other possible complication is that we cannot be sure that the 'true' spatial weighting matrix is in the set of alternatives that is constructed to investigate the problem. This means that often our search for an appropriate spatial weighting matrix will yield an approximation. This suggests that the estimation problem we are solving is likely to be characterised by a relatively low signal to noise ratio, which will impact negatively on the performance of most screening methods. Wang (2009) presents some extensive numerical simulations comparing SIS, ISIS, LARS (least angle regression) and forward regression implemented alone or followed by a consistent variable selection method (adaptive lasso or SCAD). Their results show that no method clearly dominates the others.

Another important consideration is that by their nature screening methods have to be very simple (see the discussion on the paper by Fan and Lv, 2008 in the same issue). There is obviously some trade-off here since 'better' methods should be able to achieve a greater reduction in the dimension of the initial problem (i.e. to eliminate more irrelevant variables) and hence reduce the computational requirements for the consequent variable selection methods, as well as improve its (theoretical) convergence rate. In highly correlated designs that will typically characterise the spatial weighting matrix selection problem as reformulated in Kostov (2010) too simple methods or methods that are not 'robust' with regard to the correlated design, could be inconsistent. Hence it could be useful to compare the relative performance of different such methods. Typically such comparisons are carried out on simulated datasets. In this case we will take a slightly different approach and implement such comparison on a real dataset.

We will consider the following candidates for screening methods. First we will use the component-wise boosting method. Since this is the method that have been implemented in the original proposal of Kostov (2010) it should allow direct comparison with his results, particularly if the same dataset is employed. Following Kostov (2010) we will use the g-prior Minimum Description Length (gMDL) of Hansen and Yu (2001) as stopping criterion. Kostov (2010) shows that it compares favourably to different forms of cross-validation at a fraction of their computational costs and hence this choice allows us to obtain a fast and reliable screening method. We could have used a more traditional criteria, such as AIC resulting in larger models to be submitted to the second step in our approach, but we felt that ensuring direct correspondence with Kostov's (2010) approach which we are building upon is desirable. The next screening method is the forward regression with AIC as stopping criterion. This is the best known classical method for dimensionality reduction and as shown in Wang (2009) it possesses screening power. The other screening methods we consider include the LASSO (Least Absolute Sum of Squares Operator, see Tibshirani, 1996), forward stagewise regression and LARS (Least Angle Regression, see Efron et al., 2004)) with Mallow's Cp as stopping rule. The full regularisation path for the latter three methods can be easily computed by modifications to the computationally efficient lars algorithm (Efron et al., 2004) and therefore these are all fast and suitable for variable screening purposes. Finally mainly for comparison purposes we will also implement a more complicated screening method, mainly the relaxed lasso (Meinshausen, 2007) with cross-validation used to select the regularisation (i.e. penalty and relaxation parameter) parameters. This is obviously a more demanding method both in terms of complexity and computational requirements. Since however it is a generalisation of the lasso it can be useful to consider it in comparative perspective and see whether the simplicity in the proposed screening methods does not come at a price.

Furthermore we consider the methods to be used on the screened data. Firstly, we use two generalisations of the Dantzig selector, namely the Gauss-Dantzig (Candes and Tao, 2007) and the Double Dantzig (James and Radchenko, 2009). The other method is the adaptive lasso (Zou, 2006). Finally we implement two non-convex penalisation methods namely SCAD (smoothly clipped absolute deviation, Fan and Li, 2001) and MCP (minimax concave penalty, Zhang 2007). All the above methods possess the oracle property and therefore are suited for implementation in the second step of our approach.

5. Study Design and Implementation Details

For comparative purposes we follow closely the design outlined in Kostov's (2010) study. This involves using the same dataset, model specification as well as set of competing alternative spatial weighting matrices. Since all these are discussed in some detail in Kostov (2010) we will only briefly sketch them here.

The corrected version of the popular Boston housing dataset (Harrison and Rubinfeld, 1978) is used. It consists of 506 observations and incorporates some corrections and additional latitude and longitude information, due to Gilley and Pace (1996). This dataset contains one observation for each census tract in the Boston Standard Metropolitan Statistical Area. The variables comprise of proxies for pollution, crime, distance to employment centres, geographical features, accessibility, housing size, age, race, status, tax burden, educational quality, zoning, and industrial externalities. A detailed description of the variables, to be used in this study is presented in table 1.

Table 1. Description of variables

Variable	Description
MEDV	Median values of owner-occupied housing in thousands of USD
LON	Tract point longitude in decimal degrees
LAT	Tract point latitude in decimal degrees
CRIM	Per capita crime
ZN	Proportion of residential land zoned for lots over 25,000 sq. ft per town
INDUS	Proportion of non-retail business acres per town
CHAS	An indicator: 1 if tract borders Charles River; 0 otherwise
NOX	Nitric oxides concentration (parts per 10 million) per town
RM	Average number of rooms per dwelling
AGE	Proportions of owner-occupied units built prior to 1940
DIS	Weighted distance to five Boston employment centres
RAD	Index of accessibility to radial highways per town
TAX	Property-tax rate per USD 10,000 per town
PTRATIO	Pupil-teacher ratio per town
B	Calculated as 1000*(Bk - 0.63)^2 where Bk is the proportion of blacks
LSTAT	Percentage of lower status population

The basic model as implemented in Kostov (2010) is as follows:

log(MEDV)= f {CRIM, ZN, INDUS, CHAS, NOX^2, RM^2, AGE, log(DIS), log(RAD), TAX, PTRATIO, B, log(LSTAT)}

A linear functional form specification is used and the latter is augmented with alternative candidate spatial weighting matrices, constructed using the longitude and latitude information. The set of alternative spatial weighting matrices is constructed using inverse distance raised on a power weights specification and nearest neighbours definition of the neighbourhood scheme

We will adopt the naming conventions used in Kostov (2010) combining the codes for the neighbourhood definition and the weighting scheme to refer to the corresponding spatial weighting matrix and the resulting additional variables to be included in the boosting model. All these variables are named using the following convention: *nxwy*, where *x* is the number of neighbours and *y* is the weighting parameter (which is the inverse power of the weight decay). For example the spatial weighting matrix with the nearest 50 observations as neighbours and inverse squared distance weights as well as the resulting transformed variable will be denoted as *n*50*w*2. We employ all values for number of neighbours from 1 to 50 and evaluate *w* in the interval [0.4, 4] using increments of 0.1. In simple words this means that we are combining 50 possible neighbourhood definitions with 37 alternatives for the weighting parameter resulting in 1,850 alternative spatial weighting matrices to be considered simultaneously.

Kostov (2010) projects the spatially weighted dependent variable into the column vector space of the spatially weighted independent variables, by taking the fitted values from a least-squares regression to obtain the transformed variables, named according to the above convention. Here we built upon that strategy and instead of applying a single variable selection method in the second step we use consecutive application of two such methods. The first is to be used as a screening method while the second (which in this case would be a method possessing the 'oracle' property) will fine tune the selection results.

Table 2. Screening and estimation methods used

Code	Method
Screening step	
BS	Component-wise boosting
FR	Forward regression
LR	LARS
LS	LASSO
RL	Relaxed LASSO
FS	Forward stagewise
Estimation step	
GD	Gauss-Dantzig
DD	Double Dantzig
ALASSO	Adaptive lasso
MCP	MCP
SCAD	SCAD

To simplify discussion from here on, unless specified otherwise, under first and second step we will understand the screening and the consequent estimation step. Given the large number of combinations of different estimation methods, for labeling purposes, it is convenient to adopt the following convention. We will use short codes to denote each of the used methods. Then each combination will be referred to as X_Y, where X will be the code for the screening method and Y the code for variable selection method implemented on the dataset reduced by the corresponding screening method. The corresponding codes are shown in Table 2. The regularisation parameters for all estimation step methods and for the relaxed lasso are chosen by 5-fold cross-validation. The non-convex penalty approaches (MCP and SCAD) involve two penalty parameters. In order to reduce the computational load (particularly due to the non-convexity of the optimization problem) we follow a commonly used in empirical applications convention and fix the second penalty parameter to 3.7. See e.g. Fan and Li (2001) for discussion on the theoretical reasons for this particular choice.

6. Results

Before we proceed to the detailed results, we will briefly review the results of Kostov (2010) who's design we follow. Table 3 shows the coefficients corresponding to the spatial weighting matrices retained in the model implementing the boosting approach of Kostov (2010), which in essence is our BS screening method, with gMDL stopping rule and updating parameter of 0.3, which is in the middle of the commonly used range of [0.1, 0.5].

Table 3. Boosting estimation results for the spatial weighting matrices

Variable	Coefficient
n3w1.2	0.0374
n3w1.3	0.0061
n6w0.4	0.1877
n6w0.5	0.0109
n6w0.6	0.0100
n6w0.7	0.0091
n6w0.8	0.0099
n6w0.9	0.0113
n6w1	0.0069

Kostov (2010) only presents a list of the retained spatial weighting matrices and notes that since all spatial weighting matrices from *n*6*w*0.4 to *n*6*w*1 are selected, using a single spatial weighting matrix by centring over the range should reasonably well approximate the true underlying structure. Taking into account the actual contributions of the retained spatial weighting matrices however suggests that *n*6*w*0.4 should have been the preferred option, since on one hand it has by far the largest (in magnitude) coefficient and on the other it is actually at the centre of the 'mass' distribution for the retained spatial weighting matrices.

Another notable feature of the present analysis is that following the discussion of Kostov (2010) one could from the very outset suspect that the 'true' spatial weighting matrix is not in the set of alternatives included in the study design. This however provides a further insight into this how the proposed approach can approximate this unknown spatial weighting matrix. Kostov (2010) speculated that a spatial weighting matrix based on contiguity definition of the neighbourhood and some form of common border weighting (using the tracts) is what is probably most consistent with the obtained results. Although as we show above the original results of Kostov (2010) need to be reconsidered, the modified results (i.e. using n6w0.4) are still consistent with this conjecture.

Another important point to address is why did not we try the original screening approach, i.e. the SIS and ISIS methods of Fan and Lv (2008). We actually implemented the latter, but the results were disappointing. In simple words the resulting models excluded virtually all main variables (i.e. variables other than the transformed spatial weighting matrices) and correspondingly the results yielded an approximation to the correlation structure over transformed spatial weighting matrices. Furthermore the exclusion of the main variables occurred during the screening step and therefore the consequent estimation methods could not recover meaningful model. The implicit simplicity of the SIS and ISIS methods in this case could not deal with the highly correlated nature of the study design.

Table 4. Number of retained variables by screening method

Code	Method	Number of retained variables
BS	Component-wise boosting	21
FR	Forward regression	43
LR	LARS	332
LS	LASSO	144
RL	Relaxed LASSO	20
FS	Forward stagewise	1164

We now describe the results. Table 4 presents the size of the reduced set of covariates (i.e. counting both 'main' variables and 'transformed' spatial weighting matrices), following the implementation of a particular screening method. One should note that the degree to which different methods reduce the dimensionality of the original problem depends on the stopping rule and hence the results in table 4 should not be viewed as comparison between different screening methods in general, but rather as a setting in which to evaluate the performance of the consequent estimation step methods. Furthermore the main purpose of the screening step is not maximum reduction, but considerable reduction that avoids as much as possible the danger of falsely omitting important variables. For this reason for example the AIC is implemented to stop the forward regression, rather than e.g. the gMDL which would have yielded greater reduction in the size of the problem. The greater reduction however could have risked dropping the most appropriate spatial weighting matrix.

Both boosting (see also Kostov 2010 for an indication about the relative number of selected variables under alternative stopping rules) and forward regression have managed to considerably reduce the size of the problem. The only other screening method that achieved similar reduction is the relaxed lasso, but it is considerably more demanding in computational terms, particularly since cross-validation is needed to select the regularisation parameters. LARS and LASSO also reduce the dimensionality below the sample size (of 506) and hence can be useful as screening methods. At first sight it looks like LARS and LASSO are retaining too many variables and hence might impede the consequent estimation methods. Note however that in this particular case the cross-validated relaxed lasso chose a relaxation parameter of 1 (see Meinshausen, 2007 for details), which effectively reduces the relaxed lasso to the conventional LASSO estimates. Therefore we can view in this particular instance the relaxed lasso as LASSO, where cross-validation is used (instead of the Mallow's Cp) to select the penalty. Moreover here the cost of omitting a relevant variable in the screening step is higher that the potential advantages in speeding estimation in the next step. Furthermore avoiding costly cross-validation in the first step (when the dimensionality is considerably higher) more than offsets the additional computational cost incurred when dealing with a larger problem in the second step. Finally the forward stagewise regression only achieves moderate reduction in the size of the problem, which remains above the sample size. As above a different stopping rule could have been employed but this would have compromised the speed of the proposed methodology.

We now proceed to the actual estimation results. These are presented in Tables 5-7. To facilitate discussion we have adopted the following ordering for the results. The results from non-convex second step (i.e. estimation) methods are presented separately in Table 7 with results ordered by estimation (i.e. second step) method. Table 5 and 6 present the results from the other methods, ordered by screening method. To simplify the presentation we have omitted the intercept from all results. Each of the above tables contains the main variables in the same order followed by n6w0.4. Three of the main variables, namely ZN, INDUS and AGE are not chosen by any of the applied methods and for this reason we do not include them in the result tables. The rest of the tables contain other spatial weighting matrices retained by the corresponding method. The latter are specific to each table, for reasons to become clear during the discussion.

First of all, the results obtained by the different methods are broadly speaking comparable. The spatial weighting matrix that best fits the model is n6w0.4, conforming to the conjecture of Kostov (2010). Moreover in about half of the methods used this is the only spatial weighting matrix, while in most other cases the additional retained spatial weighting matrices have rather small contributions. There are some small differences between different methods in that some of them deselect some of the main variables. We will not explicitly comment on these differences unless they are essential in explaining what is happening with regard to the main focus of the study, namely the choice of spatial weighting matrix. Hereafter we will refer to the model with n6w0.4 as the only spatial weighting matrix as the default model.

Boosting performs very well as screening method, producing results which are consistent amongst the different second step methods. This should come as no surprise since the boosting application has resulted in a rather small set of candidate spatial weighting matrices. The only deviation from this rule is BS_MCP case which selects n6w0.9 instead of the n6w0.4 spatial weighing matrix. Interestingly the MCP algorithm selects n5w0.8 or n6w0.9 in four out of the six pre-screened sets (see table 7), which suggest that this ‘preference’ for slightly higher weighting parameter could have something to do with the algorithm itself. The non-convex nature of the algorithm which can have at least three distinct implementations as well as the issue what type of cross-validation would be most appropriate for the problem in hand are some issues that may require some additional attention. Nevertheless even with the slight difference in the BS_MCP result, the results obtained using boosting as screening method conform to the expectations.

Forward regression also performs very well. Similar results are obtained across the whole range of second step methods (see tables 5 and 7). The only two methods that deviate from the default model are FR_ALASSO which selects n16w0.7 in addition to n6w0.4, and FR_MCP, where n6w0.8 is selected instead. These deviations can be viewed as ‘spreading’ the spatial dependence in comparison to the default model because they imply in simple terms an additional effect characterized by more neighbours but also larger weight decay. In this way such effects could be consistent with additional (possibly of non-spatial origin) heteroscedasticity present in the default model. We will revisit this point later.

The application of LARS as screening method yields very similar results. The LR_GD ‘augments’ the default model with a very small contribution from n9w0.4, while LR_ALASSO drastically increase the additional ‘contributions’ by including n16w08 and n33w0.4. The SCAD and MCP replace n6w0.4 with respectively n6w0.8 and n6w0.9. Again this suggests some ‘spreading’ of the pattern of spatial dependence.

As explained earlier in this particular application the relaxed lasso reduces to ordinary lasso (with cross-validation for penalty choice rather than Mallow’s Cp). The use of cross-validation in place of simple selection criterion, does not seem to affect the results too much. LS_DD and LS_DD do not select the default spatial weighting matrix but choose n5w0.4 which is virtually the same. The difference amongst the lasso and relaxed lasso screened models are essentially due to the second step methods. Although such difference do not change the conclusions about the nature of the spatial weighting matrix, they are somewhat more substantive with regard to the main variables and this is certainly an issue that deserves more thorough investigation. Perhaps surprisingly SCAD and MCP produce identical estimates for the lasso and relaxed lasso screened model essentially coinciding with the default model, which may prompt closer look at these.

Forward stagewise regression did not manage to sufficiently reduce the problem size. This means that FS_ALASSO cannot be implemented because the number of variables retained by the FS exceeds the sample size and consequently initial weights for the adaptive lasso algorithm cannot be computed. For other second step methods however the corresponding algorithms can be implemented and the results are not substantively different from those obtained using the other methods.

It is worth mentioning that the application of SCAD and MCP in the second step produces remarkably similar results regardless of the screening method used. This could be a property of the methods themselves, but given the implicit difficulties in optimizing non-convex objective functions and the already mentioned fact that we fixed the second penalty parameter, it could also be due to the particular application.

In order to elaborate on the earlier point about ‘spreading’ of the spatial dependence, consider Table 8 that lists the estimation results from spatial two stage least squares estimation of the default model (excluding the three main variables that are not selected by any of the used methods). Standard errors produced without and with heteroscedasticity correction are shown together with their ratio. These results are indicative of considerable residual heteroscedacticity. We will not elaborate on the possible reasons for this, since it may be due to the approximation that the default spatial weighting matrix provides for the ‘true’ one. Furthermore it may also be due to the functional form assumptions employed here. The presence of such heteroscedasticity however can and as already discussed does to some extent affect the results, which is to be expected since all methods considered in this study, whether used for screening or estimation purposes are ultimately based on least squares and hence will be affected by the presence of heteroscedasticity. Note furthermore that the relative effect of the heteroscedacity is larger for the spatial dependence parameter, which in this particular case is also to be expected given that it can be viewed as an approximation.

Table 5. Estimation results (part 1)

	BS_DD	BS_GD	BS_ALASSO	FR_DD	FR_GD	FR_ALASSO	LR_DD	LR_GD	LR-ALASSO
CRIM	-0.0098	-0.0101	-0.0076	-0.0098	-0.0101	-0.0086	-0.0098	-0.0101	-0.0087
CHAS	0.0283	0.0332		0.0283	0.0332		0.0282	0.0329	
NOX^2	-0.3561	-0.4059		-0.3561	-0.4059	-0.2295	-0.3551	-0.4060	-0.2863
RM^2	0.0069	0.0065	0.0066	0.0069	0.0065	0.0067	0.0069	0.0065	0.0067
log(DIS)	-0.1495	-0.1580	-0.0940	-0.1495	-0.1580	-0.1497	-0.1493	-0.1584	-0.1564
log(RAD)	0.0307	0.0334	0.0057	0.0307	0.0334	0.0542	0.0306	0.0333	0.0663
TAX			0.0000			-0.0002			-0.0003
PTRATIO	-0.0124	-0.0152		-0.0124	-0.0152	-0.0090	-0.0123	-0.0152	-0.0115
B			0.0001			0.0003			0.0003
log(LSTAT)	-0.2643	-0.2765	-0.2733	-0.2643	-0.2765	-0.2759	-0.2641	-0.2768	-0.2725
n6w0.4	0.5003	0.4655	0.5376	0.5003	0.4655	0.3770	0.5011	0.4544	0.2873
n9w0.4								0.0120	
n16w0.7						0.1056			
n16w0.8									0.2690
n33w0.4									-0.1037

Tabale 6. Estimation Results (part 2)

	LS_DD	LS_GD	LS-ALASSO	RL_DD	RL-GD	RL_ALASSO	FS_DD	FS_GD
CRIM	-0.0093	-0.0095	-0.0082	-0.0094	-0.0096	-0.0083	-0.0100	-0.0101
CHAS	0.0283	0.0326		0.0248	0.0291		0.0317	0.0323
NOX^2	-0.2565	-0.2826	-0.0678	-0.2481	-0.2743	-0.1286	-0.3714	-0.4109
RM^2	0.0078	0.0075	0.0069	0.0074	0.0071	0.0068	0.0069	0.0065
log(DIS)	-0.1348	-0.1412	-0.1130	-0.1372	-0.1436	-0.1268	-0.1507	-0.1601
log(RAD)	0.0173	0.0175	0.0250	0.0196	0.0198	0.0389	0.0309	0.0331
TAX			-0.0001			-0.0002		
PTRATIO			-0.0040			-0.0055	-0.0132	-0.0154
B			0.0002			0.0002		
log(LSTAT)	-0.2530	-0.2655	-0.2709	-0.2570	-0.2693	-0.2739	-0.2689	-0.2793
n6w0.4			0.5107	0.5125	0.4982	0.4693	0.4789	0.4765
n1w0.4								-0.0263
n2w0.4	0.0613	0.0521		0.0454	0.0353			
n5w0.4	0.4785	0.4635						
n9w0.4							0.0009	0.0104
n16w0.6						0.0272		

Tabale 7. Estimation results (part 3)

	BS_SCAD	FR_SCAD	LR_SCAD	LS-SCAD	RL_SCAD	FS_SCAD	BS_MCP	FR_MCP	LR_MCP	LS_MCP	RL_MCP	FS_MCP
CRIM	-0.0092	-0.0092	-0.0091	-0.0092	-0.0092	-0.0092	-0.0091	-0.0091	-0.0091	-0.0092	-0.0092	-0.0091
CHAS		0.0023										
NOX^2	-0.3140	-0.3145	-0.3172	-0.3140	-0.3140	-0.3140	-0.3192	-0.3172	-0.3192	-0.3140	-0.3140	-0.3300
RM^2	0.0067	0.0067	0.0068	0.0067	0.0067	0.0067	0.0068	0.0068	0.0068	0.0067	0.0067	0.0068
log(DIS)	-0.1630	-0.1639	-0.1605	-0.1630	-0.1630	-0.1630	-0.1601	-0.1605	-0.1601	-0.1630	-0.1630	-0.1658
log(RAD)	0.0741	0.0743	0.0755	0.0741	0.0741	0.0741	0.0759	0.0755	0.0759	0.0741	0.0741	0.0766
TAX	-0.0003	-0.0003	-0.0003	-0.0003	-0.0003	-0.0003	-0.0003	-0.0003	-0.0003	-0.0003	-0.0003	-0.0003
PTRATIO	-0.0135	-0.0134	-0.0138	-0.0135	-0.0135	-0.0135	-0.0139	-0.0138	-0.0139	-0.0135	-0.0135	-0.0145
B	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
log(LSTAT)	-0.2722	-0.2720	-0.2690	-0.2722	-0.2722	-0.2722	-0.2686	-0.2690	-0.2686	-0.2722	-0.2722	-0.2750
n6w0.4	0.4426	0.4422		0.4426	0.4426	0.4426				0.4426	0.4426	
n6w0.8			0.4397					0.4397				
n6w0.9							0.4375		0.4375			0.5148
n1w2.2												-0.0931

7. Conclusions and Possible Extensions

This paper considered the choice of spatial weighting matrix in a spatial Durbin model framework. Building upon the transformation approach of Kostov (2010) we propose a two-step selection approach with a screening step reducing the number of candidate spatial weighting matrices and estimation step selecting the final model. In an empirical application of the proposed methodology a range of different combinations of screening and estimation methods are found to produce similar results. We also demonstrate the ability of the proposed methodology to approximate and provide indications to what the 'true' spatial weighting matrix could be even when it is not amongst the considered alternatives. The similarity in results obtained using different methods suggests that their relative computational costs could be primary reasons for their choice. Note however that there are some numerical and algorithmic issues still to be resolved that can affect the comparative performance of different methods, which is to be subject of further research. Another unresolved issue refers to the presence of heteroscedacticity in the estimated models, something that may prompt search of more robust alternatives of the proposed methods. Finally, another important issue that we have not discussed here is this of functional form. Since non-parametric estimators are still consistent, although inefficient under the presence of (ignored) spatial dependence, the proposed methods could still be applied to non-parametrically filtered data, although the question of potential interplay of simultaneous selection of main variables (in non-parametric setup) and spatial weighting matrices is something that would require much more careful consideration.

References

Anselin, L. (2002). Under the hood: Issues in the specification and interpretation of spatial regression models. *Agricultural Economics, 27,* 247-267. http://dx.doi.org/10.1111/j.1574-0862.2002.tb00120.x

Berry, B. J. L., D. A. Griffith and M.R. Tiefelsdorf (2008). From Spatial Analysis to Geospatial Science. *Geographical Analysis, 40,* 229–238. http://dx.doi.org/10.1111/j.1538-4632.2008.00723.x

Bhattacharjee, A. and C. Jensen-Butler (2005) Estimation of Spatial Weights Matrix in a Spatial Error Model, with an Application to Diffusion in Housing Demand, University of St. Andrews, Centre for Research into Industry, Enterprise, Finance and the Firm (CRIEFF), Discussion Paper 0519

Bühlmann, P. (2006). Boosting for high-dimensional linear models, *The Annals of Statistics, 34,* 559–583. http://dx.doi.org/10.1214/009053606000000092

Bühlmann, P. and T. Hothorn (2007a). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science, 22*(4), 477-505. http://dx.doi.org/10.1214/07-STS242

Bühlmann, P. and T. Hothorn (2007b). Rejoinder: Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science, 22*(4), 516-522. http://dx.doi.org/10.1214/07-STS242REJ

Bühlmann, P. and Yu, B. (2003). Boosting with L2 loss: Regression and classification. *Journal of the American Statistical Association, 98,* 324– 338. http://dx.doi.org/10.1198/016214503000125

Candes E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics, 35*(6), 2313-2351. http://dx.doi.org/10.1214/009053607000000532, http://dx.doi.org/10.1214/009053606000001523

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics 32*, 407–499. http://dx.doi.org/10.1214/009053604000000067

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360. http://dx.doi.org/10.1198/016214501753382273

Getis A. (2009). Spatial Weights Matrices, *Geographical Analysis*, *41*, 404–410. http://dx.doi.org/10.1111/j.1538-4632.2009.00768.x

Griffith D.A. and Lagona, F. (1998). On the quality of likelihood-based estimators in spatial autoregressive models when the data dependence structure is misspecified, *Journal of. Statistical Planning and Inference*, *69*, 153-174. http://dx.doi.org/10.1016/S0378-3758(97)00156-0

Hansen, M and Yu, B (2001). Model selection and minimum description length principle. *Journal of the American Statistical Association*, *96*, 746–774. http://dx.doi.org/10.1198/016214501753168398

Holloway G. and M. L. A. Lapar (2007). How Big is Your Neighbourhood? Spatial Implications of Market Participation Among Filipino Smallholders, *Journal of Agricultural Economics*, *58*(1): 37-60. http://dx.doi.org/10.1111/j.1477-9552.2007.00077.x

Hothorn, T. and P. Bühlmann (2006). Model-based boosting in high dimensions. *Bioinformatics*, *22*(22): 2828-2829. http://dx.doi.org/10.1093/bioinformatics/btl462

James, G M, and P. Radchenko (2009). A generalized Dantzig selector with shrinkage tuning. *Biometrika*, *96*(2), 323–337. http://dx.doi.org/10.1093/biomet/asp013

Kneib, T., T. Hothorn, and G. Tutz (2009). Variable selection and model choice in geoadditive regression models. *Biometrics*, *65*,626-634. http://dx.doi.org/10.1111/j.1541-0420.2008.01112.x

Kooijman, S. A. L. M. (1976). Some Remarks on the Statistical Analysis of Grids Especially with Respect to Ecology. *Annals of Systems Research 5*, 113-132. http://dx.doi.org/10.1007/978-1-4613-4243-4_6

Kostov, P. (2010). Model boosting for spatial weighting matrix selection in spatial lag models. *Environment and Planning B: Planning and Design*, *37*(3), 533–549. http://dx.doi.org/10.1068/b35137

Le Sage, J., and R. K. Pace. (2009). Introduction to Spatial Econometrics. *Boca Raton, FL: Taylor and Francis/CRC*. http://dx.doi.org/10.1201/9781420064254.

LeSage, J. P. and Fischer, M. M. (2008). Spatial Growth Regressions: Model Specification, Estimation and Interpretation. *Spatial Economic Analysis*, *3*(3), 275-304. http://dx.doi.org/10.1080/17421770802353758

LeSage, J. P. and O. Parent (2007). Bayesian Model Averaging for Spatial Econometric Models. *Geographical Analysis*, *39*, 241–267. http://dx.doi.org/10.1111/j.1538-4632.2007.00703.x

Lima, E.C.R. and P.B.R. Macedo (1999). Estimation of a weights matrix for determining spatial effects, IPEA Discussion paper 672, Rio de Janeiro.

Manski, C. F. (1993). Identification of endogenous social effects: The reflexion problem, *Review of Economic Studies*, *60*, 531–542. http://dx.doi.org/10.2307/2298123

Meinshausen, N, G. Rocha and Bin Yu (2007). Discussion: A tale of three cousins: Lasso, L2Boosting and Dantzig. *Annals of Statistics*,*35*(6), 2373-2384. http://dx.doi.org/10.1214/009053607000000460

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics and Data Analysis*, *52*(1), 374–93. http://dx.doi.org/10.1016/j.csda.2006.12.019

McMillen D.P. (2003). Spatial autocorrelation or model misspecification? *International Regional Science Review*, *26*, 208–217. http://dx.doi.org/10.1177/0160017602250977

Pace R. K., J. LeSage and S. Zhu (2009). Impact of Cliff and Ord on the Housing and Real Estate Literature. *Geographical Analysis*, *41*, 418–424. http://dx.doi.org/10.1111/j.1538-4632.2009.00770.x

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, *104*, 1512–1524. http://dx.doi.org/10.1198/jasa.2008.tm08516

Zhang, C.-H. (2007). Penalized linear unbiased selection. *Technical Report* No. 2007-003. Department of Statistics, Rutgers University

Zou, H. (2006). The Adaptive Lasso and its Oracle Properties. *Journal of the American Statistical Association*, *101*, 1418-1429. http://dx.doi.org/10.1198/016214506000000735