# An Equation for Identifying the Candidate Set of Eigenvectors From Which Ones Are Selected when Constructing an Eigenvector Spatial Filter

M. Lee, P. Sinha, Y. Chun, D. A. Griffith

The University of Texas at Dallas, 800 W. Campbell rd., Richardson, Texas, USA
Telephone: 1-972-883-4950
Fax: 1-972-883-69676
Email: mxl120631@utdallas.edu
Email: parmanand.sinha@utdallas.edu
Email: yxc070300@utdallas.edu
Email: dagriffith@utdallas.edu

## 1. Introduction

Eigenvector spatial filtering (ESF) (Griffith 2000; Griffith 2003; Getis and Griffith, 2002) employs traditional regression techniques, while ensuring that regression residuals behave according to the traditional model assumption of their containing no spatial autocorrelation (SA). ESF uses a set of spatial proxy variables, which usually are extracted as eigenvectors from an underlying spatial relationship matrix that ties the spatial objects together, and adds these vectors as control variables in a model. These eigenvectors, which are extracted from a posited spatial relationship matrix, exhibit distinctive spatial map patterns with an associated level of SA. These control variables identify and isolate the stochastic spatial dependencies among observations, thus allowing model building to proceed as if the observations are independent.

Although the ESF model specification is flexible, and has become more popular in addressing SA latent in georeferenced data (e.g., Thany and Simanis 2012), it faces two major computational challenges. The first lies in computing the eigenvectors from an *n*-by-*n* modified spatial weights matrix, which often requires substantial computational resources as *n* increases. For example, eigenvector generation for a high resolution remotely sensed image is computationally challenging, although an algorithm for sparse matrices can improve the computation of the eigenvectors (Pace et al. 2011). The second computational challenge lies in the selection of a subset from the resulting set of *n* eigenvectors to construct a spatial filter, which becomes increasingly challenging as *n* increases. The selection of eigenvectors is conducted through two stages. First, a candidate set is identified; second, a final set is selected from the candidate set with stepwise regression techniques. In this second step, a forward stepwise regression technique can select significant eigenvectors (say at the 10% level) at each step, continually increasing $R^2$ until no significant vectors remain outside of the regression equation. Alternatively, the forward stepwise regression technique can select eigenvectos at each step that minimize residual SA, continuing to do so until the residual MC $\approx$ E(MC), the expected value of the MC.

This paper focuses on formalating an equation for identifying the candidate set of eigenvectors from which ones are selected to construct an eigenvector spatial filter. This procedure should reduce the size of an original candidate eigenvector set, making the construction of an eigenvector spatil filter less computational intensity in the stepwise regression selection stage.

## 2. Eigenvector spatial filtering

The ESF methodology utilizes the properties of eigenvectors and corresponding eigenvalues of the transformed spatial weights matrix $(\mathbf{I} - \mathbf{1}\mathbf{1}^T/n)\mathbf{C}(\mathbf{I} - \mathbf{1}\mathbf{1}^T/n)$, where $\mathbf{I}$ is an identity matrix, $\mathbf{1}$ is an $n$-by-1 vector of ones, $\mathbf{C}$ is a spatial weights matrix, and superscript T denotes the matrix transpose operator. Studies, including Tiefelsdorf and Boots (1995) and Griffith (1996), show that its $n$ mutually orthogonal and uncorrelated (Griffith 2000) eigenvectors, $\mathbf{E}=\{\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_n\}$, and $n$ corresponding eigenvalues, $\boldsymbol{\lambda}=\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, relate to SA. Important proproties of these vectors include: 1) they furnish distinct map pattern descriptions of latent spatial autocorrealtion in georeferenced variables, and 2) the eigenvalues index the level of SA of a map pattern that is generated when the corresponding eigenevecotor is mapped on the given tessellation. That is, the Moran Coefficience (MC) of the map pattern produced by $\mathbf{E}_j$ is $MC_j = \lambda_j \cdot n/\mathbf{1}^T\mathbf{C}\mathbf{1}$.

The ESF linear regression model specification can be written as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}_k\boldsymbol{\beta}_E + \boldsymbol{\varepsilon}$, where $\mathbf{E}_k$ is an $n$-by-$K$ matrix containing $K$ eigenvectors, $\boldsymbol{\beta}_E$ is the corresponding vector of regression parameters, and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma_\varepsilon^2)$ is an $n$-by-1 error vector whose elements are iid normal random variates. The selection of $K$ eigenvectors to construct a spatial filter, which is the linear combination of the eigenvectors ($\mathbf{E}_k\boldsymbol{\beta}_E$), is one geo-computational feature of ESF.

Griffith and Chun (2009) suggest a criterion based upon the level of SA in a variable, rather than using an arbitrary relative MC value, to identify a finely tuned candidate set of eigenvectors. They developed the following equation to achieve this goal:

$$MC_j \geq 2.9970 - 2.8805/(1 + e^{-0.6606 - 0.2525 z_{MC}}), \quad (1)$$

where, $z_{MC}$ denotes the Z score of the MC for residuals from a stepwise regression. This equation suggest that the size of the candidate set inreasing as the degree of SA increases. It was constructed with a simulation experiment unilizing an 20-by-20 regular square tessellation (n = 400). Accordingly, it needs to be generalized for $n$.

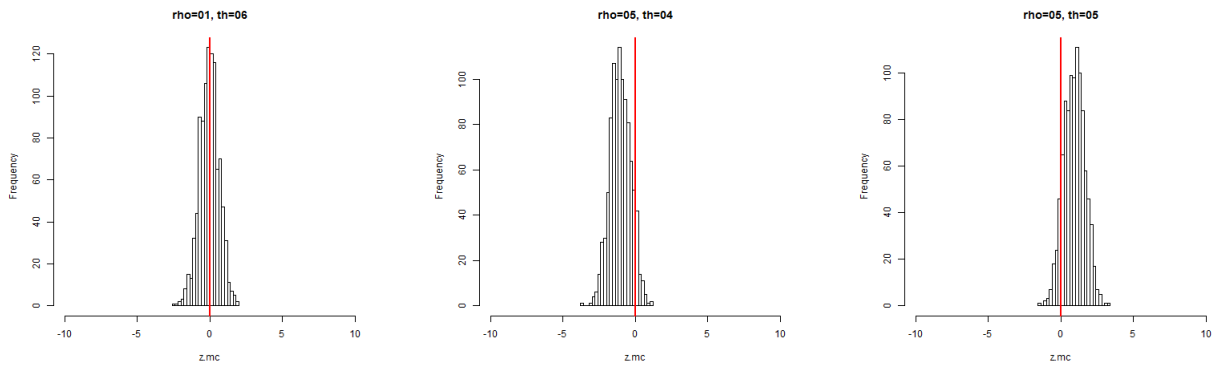
## 3. Simulation design and results

A simulation experiment was designed to uncover an equation for identifying the threshold MC value needed to identify a candidate set of eigenvectors. This involves calculating different $z_{MC}$s based on different levels of SA, and relating them to different threshold values. The steps of this experiment are: (1) generate 1,000 pseudo-random numbers with different SA level of $\rho$ values, {0.1, 0.2, … , 0.9, 0.95}, using a standard simultaneous autoregressive model (SAR) model; (2) compile candidate sets for 10 different *threshold* values {0.015, 0.1, 0.2, … , 0.9} (these values determine which threshold values have the highest $R^2$s for the model spatial filters; (3) select eigenvectors for spatial fiter constructon from the candidates sets using stepwise regression (the dependent variables are the 1,000 pseudo-random numbers in which SA is embedded with an SAR model, and the independent variables are sets of candidate eigenvectors); and, (4) calculate $z_{MC}$ for the regression residuals (a histogram of 1,000 $z_{MC}$ values establishes a suitable *threshold* value).

The result of interest is a specific threshold value that gives the most efficient and accurate candidate set of eigenvectors for each pair of $n$ and $\rho$. Table 1 reports selected simulation experimental results. *Threshold* values tend to decrease asn $\rho$ values increase and $n$ decreases. The range of *threshold values* is from 0.3 to 0.8 in this case.

| $\rho$ value | *threshold* value | | | | |
|---|---|---|---|---|---|
| | 10-by-10 | 20-by-20 | 30-by-30 | 40-by-40 | 50-by-50 |
| 0.1 | 0.6 | 0.8 | 0.8 | 0.8 | 0.8 |
| 0.2 | 0.5 | 0.7 | 0.7 | 0.7 | 0.7 |
| 0.3 | 0.4 | 0.6 | 0.6 | 0.6 | 0.6 |
| 0.4 | 0.4 | 0.5 | 0.5 | 0.5 | 0.5 |
| 0.5 | 0.3 | 0.5 | 0.5 | 0.5 | 0.5 |
| 0.6 | 0.3 | 0.4 | 0.4 | 0.4 | 0.4 |
| 0.7 | 0.3 | 0.4 | 0.4 | 0.4 | 0.4 |
| 0.8 | 0.3 | 0.4 | 0.4 | 0.4 | 0.4 |
| 0.9 | 0.3 | 0.3 | 0.4 | 0.4 | 0.4 |
| 0.95 | 0.3 | 0.3 | 0.4 | 0.4 | 0.4 |

Table 1. Simulation results for threshold values to construct an eigenvector spatial filter




(a) (b) (c)

Figure 1. Z-scores of MC for resduals. (a) threshold value = 0.4, ρ = 0.1, and a 10-by-10 regular square tessellation (b) threshold value = 0.4, ρ = 0.5, and a 50-by-50 regular square tessellation (c) resduals for value = 0.5, ρ = 0.5, and a 50-by-50 regular square tessellation

Additonal experiments are required to establish more precise *threshold* values, partly because the search grid used here is too coarse. Figure 1a shows the threshold value for a 10-by-10 regular square tessellation with $\rho = 0.1$. The mean $z_{MC}$ is close to 0; however, the *threshold* value for a 50-by-50 regular square tessellation with $\rho = 0.5$ (Figure 1b and 1c) is not clear, and lies somewhere between 0.4 and 0.5.

## 4. Implications

By identifying accurate *threshold* values, ESF will become more efficient and less numerically intensive, supporting its use with much larger datasets. This paper contributes to this goal by extending equation (1) to a wider range of *n* values.

## 5. References


Getis A and Griffith DA, 2002, Comparative spatial filtering in regression analysis. *Geographical Analysis*, 34: 130–140.

Griffith DA, 1996, Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying georeferenced data. *The Canadian Geographer*, 40: 351-367.

Griffith DA, 2000, Eigenfunction properties and approximations of selected incidence matrices employed in spatial analyses, *Linear Algebra & Its Applications*, 321:95-112.

Griffith DA, 2003, *Spatial Autocorrelation and Spatial Filtering: Gaining Understanding through Theory and Scientific Visualization*, Springer-Verlag, Berlin.

Griffith DA and Chun Y, 2009, Eigenvector selection with stepwise regression techniques to construct spatial filters. paper presented at the 105th annual Association of American Geographers meeting, Las Vegas, NV, March 25.

Pace K, LeSage J and Zhu S, 2011, Interpretation and Computation of Estimates from Regression Models using Spatial Filtering., paper presented to the Vth World Conference of the Spatial Econometrics Association, Toulouse, FR, July 6-8.

Patuelli, R., Schanne, N., Griffith, D. A., & Nijkamp, P. (2012). Persistence of Regional Unemployment: Application of a Spatial Filtering Approach to Local Labor Markets in Germany*. *Journal of regional science*, 34:253-280.

Thayn, J., and J. Simanis. 2012. Accounting for spatial autocorrelation in linear regression models using spatial filtering with eigenvectors, *Annals of the Association of American Geographers*. 103: 47-66.

Tiefelsdorf M and Boots B, 1995, The exact distribution of Moran's *I*. *Environment and Planning A*, 27:985-999.