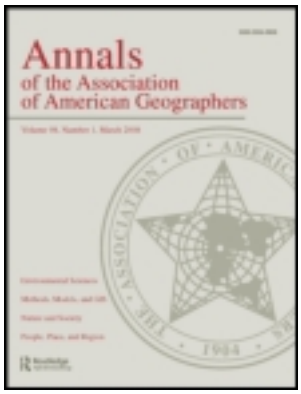

Annals of the Association of American Geographers
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/raag20

Accounting for Spatial Autocorrelation in Linear Regression Models Using Spatial Filtering with Eigenvectors

Jonathan B. Thayn [a] & Joseph M. Simanis [a]
[a] Department of Geography-Geology, Illinois State University
Published online: 20 Jun 2012.

To cite this article: Jonathan B. Thayn & Joseph M. Simanis (2013) Accounting for Spatial Autocorrelation in Linear Regression Models Using Spatial Filtering with Eigenvectors, Annals of the Association of American Geographers, 103:1, 47-66, DOI: 10.1080/00045608.2012.685048

To link to this article: http://dx.doi.org/10.1080/00045608.2012.685048

Accounting for Spatial Autocorrelation in Linear Regression Models Using Spatial Filtering with Eigenvectors

Jonathan B. Thayn and Joseph M. Simanis

Department of Geography–Geology, Illinois State University

Ordinary least squares linear regression models are frequently used to analyze and model spatial phenomena. These models are useful and easily interpreted, and the assumptions, strengths, and weaknesses of these models are well studied and understood. Regression models applied to spatial data frequently contain spatially autocorrelated residuals, however, indicating a misspecification error. This problem is limited to spatial data (although similar problems occur with time series data), so it has received less attention than more frequently encountered problems. A method called *spatial filtering with eigenvectors* has been proposed to account for this problem. We apply this method to ten real-world data sets and a series of simulated data sets to begin to understand the conditions under which the method can be most usefully applied. We find that spatial filtering with eigenvectors reduces spatial misspecification errors, increases the strength of the model fit, frequently increases the normality of model residuals, and can increase the homoscedasticity of model residuals. We provide a sample script showing how to apply the method in the R statistical environment. Spatial filtering with eigenvectors is a powerful geographic method that should be applied to many regression models that use geographic data. *Key Words: eigenvectors, linear regression, spatial filtering, spatial misspecification.*

普通最小二乘法线性回归模型经常被用于分析和模型空间现象。这些模型是有用并容易被理解的，这些模型的假设，长处，弱点已经被我们很好地研究和理解了。但是，用到空间数据的回归模型通常包含空间自相关的冗余，它表明了一个设定错误。因此该问题局限于空间数据（尽管在时间序列的数据中有类似问题的发生），所以它受到的关注少于更经常遇到的问题。本文提出一个名为特征向量空间滤波的方法来考虑这个问题。我们将这种方法用于十个真实的数据集和一系列的模拟数据集，开始理解在哪种条件下，该方法可以被最有效的应用。我们发现特征向量空间滤波法降低了空间设定的错误，提高了模型拟合的能力，频繁地增加了模型冗余的正态性，并能增加模型冗余的方差齐性。我们提供了一个示例程序来显示如何在 R 统计环境中使用该方法。特征向量空间滤波是一个功能强大的地理方法，应该被应用于使用地理数据的许多回归模型。*关键词：向量，线性回归，空间滤波，空间设定错误*。

Los modelos de regresión lineal ordinaria de cuadrados mínimos se utilizan con frecuencia para analizar y modelar fenómenos espaciales. Estos modelos son útiles y fáciles de interpretar, y sus fortalezas, debilidades y supuestos, han sido bien estudiados y entendidos. No obstante, los modelos de regresión aplicados a datos espaciales frecuentemente contienen residuos espacialmente autocorrelacionados, lo cual indica un error de especificación equivocada. Este problema se limita a datos espaciales (aunque problemas similares ocurren con los datos de series de tiempo), por lo que ha recibido menos atención de la que se concede a problemas de mayor ocurrencia. Para enfrentar este problema, se ha propuesto un método denominado *filtro espacial con eigenvectores*. Aplicamos ese método a diez conjuntos de datos del mundo real y a una serie de conjuntos de datos simulados, para empezar a entender las condiciones bajo las cuales el método puede ser aplicado con mayor utilidad. Descubrimos que el filtrado espacial con eigenvectores reduce los errores de especificación espacial equivocada, aumenta la fuerza de correspondencia del modelo, frecuentemente incrementa la normalidad de los residuos del modelo y puede incrementar la homocedasticidad [varianza de error constante] de los residuos. Suministramos instrucciones para indicar cómo aplicar el método en el entorno estadístico R. El filtrado espacial con eigenvectores es un método geográfico robusto que debería aplicarse a muchos modelos de regresión que utilicen datos geográficos. *Palabras clave: eigenvectores, regresión lineal, filtrado espacial, especificación espacial equivocada.*

Annals of the Association of American Geographers, 103(1) 2013, pp. 47–66
Initial submission, January 2011; revised submissions, June and August 2011; final acceptance, August 2011
Published by Taylor & Francis, LLC.

Ordinary least squares (OLS) regression models are among the most commonly used and best understood statistical procedures (Burt and Barber 1996). Linear regressions, for inferential purposes, rest on two assumptions regarding the errors of the model: first, homoscedasticity (constant variance) and second, normality. If these assumptions are not met, the results of the model are unreliable. An additional assumption regarding the model residuals is encountered in OLS models performed on geographic data sets—model errors must not be spatially autocorrelated. If the residuals of an OLS model are spatially autocorrelated, the model suffers from a misspecification problem and the results of the model are questionable (Anselin 1988). Typically, statistical tests become too liberal in the presence of positive spatial autocorrelation; that is, the null hypothesis is rejected more often than it should be (Clifford, Richardson, and Hémon 1989; Dray, Legendre, and Peres-Neto 2006; Dormann 2007; Dormann et al. 2007). This is a frequent problem in geographic analysis that, until recently, did not have a ready solution.

Several methods have been proposed that account for spatially autocorrelated residuals by filtering or screening the spatial component from model variables before submission to OLS regression. These methods derive a dummy spatial variable that is then included as an additional independent variable in the regression model. This removes the misspecification problem from the model. Thus, spatial data can be appropriately submitted to regression models and the concomitant diagnostic statistics that make interpretation of regression results easy and straightforward (Getis 1990).

The Getis (1990, 2010) method uses local statistics analysis that finds the spatial association among observations and then screens or removes most of the spatial dependence from the dependent variable. The spatial pattern is then introduced to the model as an independent variable. One strength of the Getis method is that the dummy spatial variable is based on the selection of a distance between observations that maximizes spatial autocorrelation, placing importance on the spatial pattern observed as distance increases from a focus. This method of measuring spatial autocorrelation is related to the G and O statistics (Ord and Getis 2001) and is a less rigid approach for determining neighbors than the adjacency method commonly used. The Getis method filters each independent variable individually, which allows for different scales of autocorrelation for each variable and is an excellent way to identify multicollinearity, when more than one independent variable share the same spatial pattern (Getis 2010). Unfortunately, the Getis method is limited to variables with a natural origin that are positive, excluding variables that are rates or percentage change (Getis 1990).

Griffith (2000b, 2003) and Tiefelsdorf and Griffith (2007) have developed a method called *spatial filtering with eigenvectors* (SFE) that creates a series of dummy spatial patterns by finding the eigenvectors associated with the independent variables of the linear model and a connectivity matrix (Bivand, Pebesma and Gómez-Rubio 2008). These patterns are eigenvectors (Griffith 2004; Bivand, Pebesma, and Gómez-Rubio 2008) that are mathematically associated with Moran's *I*, a very commonly used measure of spatial autocorrelation (Moran 1948), and they are orthogonal and uncorrelated, perfectly meeting that assumption of regression analysis. SFE discovers the latent spatial pattern in the independent variables as a body rather than filtering each pattern individually, so it does not identify multicollinearity as effectively as Getis's method, but it does eliminate the threat of multicollinearity in model specification. Dormann et al. (2007) found SFE to be the most adaptable of the seven spatial filtering methods they studied (they did not look at Getis's method). Griffith and Peres-Neto (2006) also commented on the flexibility of SFE.

Another intriguing aspect of SFE is that each of the eigenvectors can capture spatial autocorrelation at different scales (Diniz-Filho and Bini 2005; Dormann et al. 2007), relaxing the assumptions of spatial isotropy (gradients of spatial autocorrelation vary uniformly in all directions) and stationarity (all locations in the data are equally spatially autocorrelated). The conditions of isotropy and stationarity are rarely met in real data, and this is the only method, of which we are aware, that relaxes these assumptions.

These two methods have been compared by their originators, who determined that both methods work well, and the difference between them comes "down to a point of view" (Getis and Griffith 2002, 139). We have chosen SFE as the focus of this article simply because most of our data are rates and percentages. The Getis method would work for the other data sets, and we report the results of both methods for one of our examples.

Although SFE has been adequately documented elsewhere (Griffith 2000b, 2003; Getis and Griffith 2002; Dray, Legendre, and Peres-Neto 2006; Griffith and Peres-Neto 2006), it has been only since Tiefelsdorf and Griffith (2007) that the method has been programmed into a readily accessible software package, the R Project for Statistical Computing (R Development Core Team 2010). According to our literature review, most applications of SFE have been published in economic, statistics,

epidemiology, and computation journals or in working papers. This article does not contribute to the rationale for spatial filtering (Anselin 1988; Getis 1990) or to the methodology of such filtering with eigenvectors (Griffith 2010); rather, our purpose is to first demonstrate the effectiveness of SFE across a varying range of real and synthesized data sets; second, begin to assess the conditions under which this method can be most fruitfully applied; and third, introduce SFE to a broader geographic audience.

We use the Columbus, Ohio, crime rate data set of Anselin (1988, Table 12.1, 189) for the initial discussion. These data are included with the "spdep" R package (Bivand et al. 2010), so they are available for researchers interested in replicating our results. These data were discussed in several other works dealing with the spatial autocorrelation (Getis 1990, 2010; Griffith and Layne 1999), allowing us to make comparisons to earlier work. Anselin modeled the incidence of crime, defined as the total of residential burglaries and vehicle thefts per thousand households by census tracts in Columbus, Ohio (Figure 1A), using mean home values and per capita income as independent variables. Following the example of Bivand and Brunstad (2006) and the suggestion of Leisch and Rossini (2003) and Gentleman (2005), we provide the R script necessary to reproduce our analysis of the Ohio crime data. Throughout this text, we refer to lines of script presented in the Appendix. Our hope is that this will help researchers who are not familiar with the R statistical environment learn how to apply SFE to regression analysis.

The R statistical environment is a powerful and adaptive, open-source, and freely distributed software package for statistical computing. R has a scripting user interface that grants freedom and dexterity when manipulating data. R is extensible because new functions that add to the capabilities of the software are generously contributed by users. The SpatialFiltering function (Tiefelsdorf and Griffith 2007) is housed in the "spdep" package (Bivand et al. 2010). The "spdep" package adds to R the ability to manipulate spatial data and assess spatial dependency. The other packages used in this analysis and in the Appendix are "classInt" (Bivand, Ono, and Dunlap 2009), "lmtest" (Zeileis 2002), "maptools" (Lewin-Koh et al. 2010), "RColorBrewer" (Neuwirth 2007), and "sm" (Bowman and Azzalini 2010).

It should be common practice, when analyzing spatial data, to map the residuals of the model (Figure 1C, Line 34) and to assess them for spatial autocorrelation. We agree with Kühn (2007, 68): "If spatial autocorrelation is ignored we simply do not know if we can trust the [regression model] results at all. Therefore, . . . the presence of residual spatial autocorrelation should always be tested for . . . and appropriate methods should be used if there is shown to be significant spatial autocorrelation." Diniz-Filho, Bini, and Hawkins (2003) make a similar point. It should also be common practice to test the conditions of the Gauss–Markov theorem (Upton and Fingleton 1985; Anselin 1988), which are that the residuals are normally distributed and are homoscedastic. In this analysis we use Moran's I (Line 41) to assess spatial autocorrelation, the Shapiro–Wilks test to assess normality (Line 42), and the Breusch–Pagan test to assess homoscedasticity (Line 44).

The OLS model for the Ohio crime data was significant ($R^2 = 0.552$, $p < 0.000$). The coefficients were –1.597 and –0.274 for income and home values, respectively. The model residuals were normally distributed ($SW = 0.977$, $p = 0.450$) but unfortunately they were heteroscedastic ($BP = 7.217$, $p = 0.027$) and moderately spatially autocorrelated ($MI = 0.251$, $|z| \approx 2.9$, $p = 0.002$). This indicates a spatial autocorrelation misspecification error in the model. This needs to be corrected

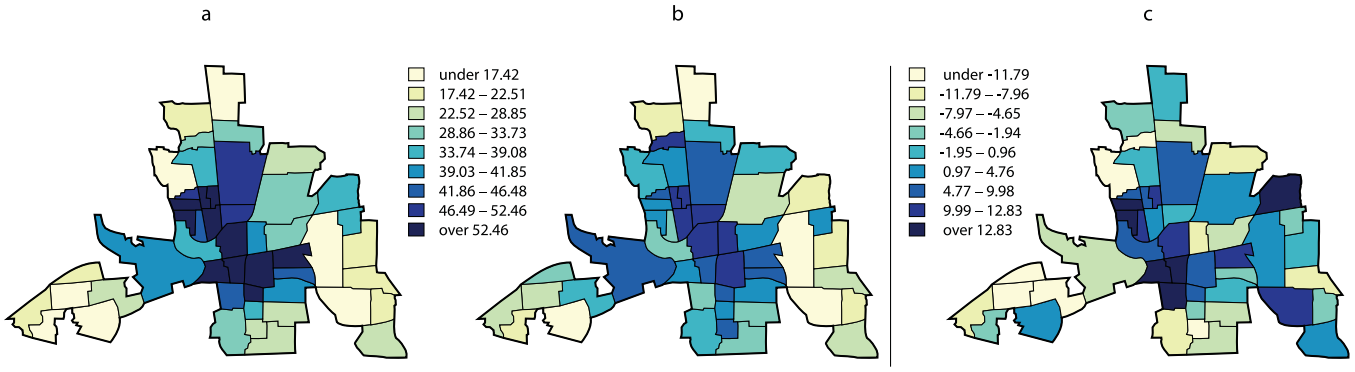

Figure 1. (A) Crime rates in Columbus, Ohio, in 1980. (B) Predicted crime rates using a linear regression model with income level and home value as independent variables. (C) The residuals of the linear model, which are strongly spatially autocorrelated ($MI = 0.251$, $|z| \approx 2.9$). (Color figure available online.)

before the results of the OLS model can be considered reliable.

Review of Spatial Filtering with the Eigenvector Approach

The first step in accounting for the spatial autocorrelation inherent in the OLS model is to establish a list of neighbors of each observation in the data set. A connectivity matrix has as many rows and columns as there are observations or polygons in the spatial pattern. Each row and each column is associated with a location. The matrix contains zeros, except at the intersection of neighboring observations, which contain ones. In other words, the value at row i and column j is one if areal unit i and areal unit j are neighbors; the value is zero otherwise.

Connectivity matrices are often weighted. The binary scheme discussed earlier is the *B*-scheme. The *W*-scheme is row standardized so that the rows in the connectivity matrix sum to one. The *C*-scheme is globally standardized, by multiplying each element of the *B*-scheme matrix by $n/\mathbf{1}^{\mathrm{T}}\mathbf{B1}$ (where $\mathbf{1}$ is a vector of length n containing ones and $\mathbf{B}$ is the *B*-scheme connectivity matrix), so that all the links in the connectivity matrix sum to n. The *U*-scheme is equal to the *C*-scheme divided by the number of neighbors so that all the links sum to one. The *S*-scheme is the variance-stabilizing scheme proposed by Tiefelsdorf, Griffith, and Boots (1999) where all links sum to n. The *W*-scheme is frequently used in spatial econometrics because it makes interpreting the underlying model easier (the value at location i is a function of the average of its neighboring values). The *C*-scheme is generally used for spatial statistics and to test for spatial dependence, although this is not a mathematical requirement (Anselin 1988; Tiefelsdorf, Griffith, and Boots 1999). We chose to use the *C*-scheme to enable comparisons with earlier work.

The R function poly2nb creates a neighborhood object (Line 20) or a list of bordering polygons (Figure 2, Lines 22–24). Weights can be assigned to the neighborhood object using any of the schemes discussed earlier to create a list weights object (Line 21). The list weights object can then be converted to a matrix (Line 47). This is the n-by-n weighted connectivity matrix, $\mathbf{C}$. This matrix is the same as the spatial link matrix used in calculating Moran's *I* (Moran 1948, 1950).

The spatial neighborhoods defined by $\mathbf{C}$ now need to be tied to the data set through the matrix $\mathbf{M}$. There are two ways of calculating $\mathbf{M}$. The first is based on a set of dummy variables created through the equation (Line 48):

Figure 2. The neighborhood adjacencies found in the Columbus, Ohio, crime rates data set using rook connectivity; that is, polygons are considered neighbors if they share a length of border, not just a single common node.

$$\mathbf{M} = \mathbf{I} - \mathbf{1}\mathbf{1}^{\mathrm{T}}/n \tag{1}$$

where $\mathbf{I}$ is an n-by-n identity matrix (a square matrix filled with zeros except along the diagonal that runs from the top left to the bottom right, which contains ones) and where $\mathbf{1}$ is a vector of length n containing ones. Multiplying matrix $\mathbf{M}$ by matrix $\mathbf{C}$ and then by matrix $\mathbf{M}$ results in the matrix $\mathbf{MCM}$ (Line 49). The eigenvectors of matrix $\mathbf{MCM}$ (Line 50) are the possible spatial patterns associated with the connectivity matrix $\mathbf{C}$. $\mathbf{MCM}$ is an n-by-n matrix, so there are as many eigenvector spatial patterns as there are observations in the data set. Because $\mathbf{M}$ is based on a nonreal variable, the spatial patterns derived by calculating the eigenvectors of $\mathbf{MCM}$ are the generic patterns that might occur in the neighborhood defined by $\mathbf{C}$.

These spatial patterns are uncorrelated map patterns of possible spatial autocorrelation (Griffith 2000a, 2000b). The first eigenvector is the set of real numbers that has the largest *MI* possible for the given connectivity matrix $\mathbf{C}$. The second eigenvector is the set of numbers with the largest *MI* possible for $\mathbf{C}$ that is uncorrelated with the first eigenvector. The *MI* value of the eigenvectors continues to decrease until the last eigenvector, which has the most negative *MI* possible for the matrix $\mathbf{C}$ that is uncorrelated with all preceding eigenvectors (Figure 3). The *MI* for each eigenvector can be found by multiplying the corresponding eigenvalue by $n/\mathbf{1}^{\mathrm{T}}\mathbf{C1}$ (Line 60, Griffith 2003).

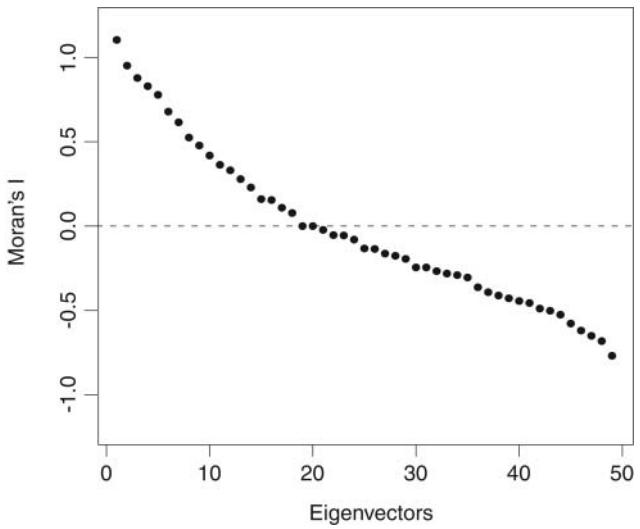

Figure 3. The Moran's *I* values of the spatial patterns derived by taking the eigenvectors of matrix **MCM** based on Equation 1. The patterns begin with strong positive spatial autocorrelation, move through random patterns, and end with strong negative spatial autocorrelation.

Using Equation 1 for **M** generates a series of possible spatial patterns that can be used to separate the underlying spatial pattern from the noise of a variable (Getis and Griffith 2002; Griffith 2003; Getis 2010). Figure 4 shows an example of the underlying spatial pattern associated with crime rates in Columbus, Ohio. Eigenvectors were calculated using Equation 1 and a subset of them was found by submitting them to a stepwise regression model with the crime rate as the dependent variable (Griffith 2000b; Tiefelsdorf and Griffith 2007; Griffith and Chun 2009). Eigenvectors 4, 1, and 3 (Figure 4) were selected. These three eigenvector patterns were then combined linearly using the coefficients derived by regressing them against the crime rate. The result represents the underlying spatial pattern of the data or the spatially filtered crime rate data (Figure 4B). In other words, this is the cleaned or filtered crime rate after the noise of the data pattern has been removed. The *MI* of this spatial filter is 0.885 ($|z| \approx 9.5$, $p < 0.000$), and it accounts for the bulk of the variability in the crime rate data ($R^2 = 0.594$, $p < 0.000$). The residuals of this model are the deviance of the actual data from the underlying spatial pattern (Figure 4C). This could be random noise or it could represent important outliers in the pattern. For example, in Figure 4C, the gray neighborhoods are places where the crime rate is lower than the underlying spatial pattern suggests, and the red neighborhoods are places where the crime rate is higher than suggested by the underlying spatial pattern. If law enforcement officials were able to determine why crime rates are lower than expected in some neighborhoods, they might be able to reduce crime in other areas as well.

The eigenvectors based on Equation 1 can be used to generate random, spatially autocorrelated variables. They were used to create the patterns displayed in Figure 5. These patterns are not directly tied to a data set, however, so they represent the generic patterns that might occur in the neighborhood defined by **C**. A series of spatial patterns that are directly tied to the variables of an OLS regression can be derived using the following equation for **M** (Line 53):

$$\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}} \tag{2}$$

Recall that **X** is a matrix with an initial column of ones followed by columns containing the independent variables. This is the same **X** that appears in the standard OLS regression equation, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. This definition of **M** is tied to the error term of the OLS model, in that (Line 54):

$$\mathbf{M}\mathbf{y} = \boldsymbol{\varepsilon} \tag{3}$$

The eigenvectors of **MCM** derived from **M** as defined by Equation 2 are not randomly generated spatial patterns. They are derived from and are orthogonal to the independent variables, **X**. They are based on the spatial arrangement of the observations through **C** (Griffith 2004). They are mathematically tied to the residuals of the model (Equation 3). Thus, the series of potential spatial patterns returned by the eigenvector approach are specific to the independent variables, their spatial distribution, and their relationship to the dependent variable. The *n* hypothetical spatial patterns generated by calculating the eigenvectors of **MCM** can be seen using Line 65.

A subset of these patterns is judiciously selected as representative of the spatial component of the error term. Because the eigenvectors are orthogonal and uncorrelated, selecting the subset of patterns is frequently done using a stepwise regression (Griffith 2003, 2010; Griffith and Chun 2009). This subset of patterns is then included in the linear regression as additional independent variables. This increases the number of independent variables and the number of coefficients. It also boosts the significance of the estimated regression coefficients by reducing the mean square error. The standard regression equation can be written to include the

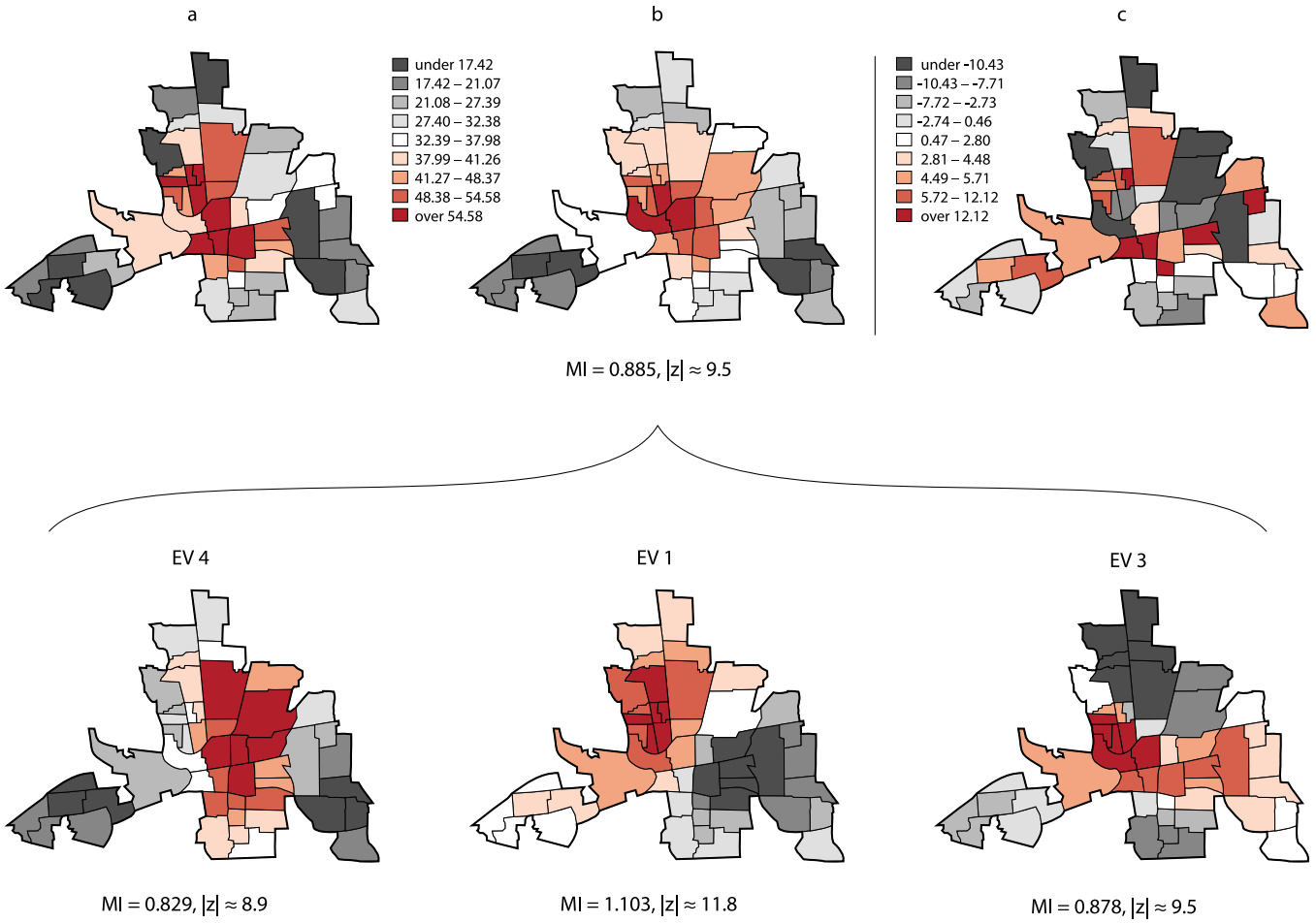

Figure 4. An example of using spatial filtering with eigenvectors (SFE) to find the underlying spatial pattern associated with a variable. (A) The actual crime rates of Columbus, Ohio. (B) The spatial filter, or the underlying spatial pattern, of the crime rate data. This pattern is a linear combination of the eigenvectors 4, 1, and 3 using the coefficients $\mathbf{b} = (35.129, -69.987, -36.278, -42.050)$. (C) The difference between the crime rate and its underlying spatial pattern. Gray areas have lower crime rates than the pattern suggests, whereas red areas have more crime than the pattern suggests. (Color figure available online.)

misspecification term (Tiefelsdorf and Griffith 2007):

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{E\gamma} + \boldsymbol{\eta} \tag{4}$$

where $\mathbf{E\gamma}$ is the misspecification term. Note that $\boldsymbol{\varepsilon}$ from Equation 3 (and from the standard regression equation) is equal to $\mathbf{E\gamma} + \boldsymbol{\eta}$ from Equation 4.

The SpatialFiltering function in R uses Equation 2 to derive **M**. It then uses an iterative brute force process to select a parsimonious subset of eigenvectors (Lines 67–68) that can be added to the OLS model as independent variables to account for and remove the spatial autocorrelation in the model residuals (Line 74; Tiefelsdorf and Griffith 2007).

In the Columbus, Ohio, crime model, eigenvectors 3, 5, 10, and 4 are selected and added to the model. These eigenvectors can be visualized using Lines 70 through 72. Because the eigenvectors are uncorrelated, they can be combined linearly and an *MI* statistic representing the filter can be found according to the following (Getis and Griffith 2002):

$$MI_{filter} = \mathbf{b}^{\mathrm{T}}\mathbf{vb}/\mathbf{b}^{\mathrm{T}}\mathbf{b} \tag{5}$$

where vector **b** is the regression coefficients that correspond to the selected eigenvectors and vector **v** is the eigenvalues of the selected eigenvectors. The value MI_{filter} returned by Equation 5 is equal to the *MI* of the spatial pattern generated by combining the selected eigenvectors using their corresponding coefficients.

The MI_{filter} for these four eigenvectors is 0.676 ($|z| \approx 7.3$, $p < 0.000$). The strong, positively spatially autocorrelated pattern encompassed in this spatial filter accounts for the spatial autocorrelation in the residuals of

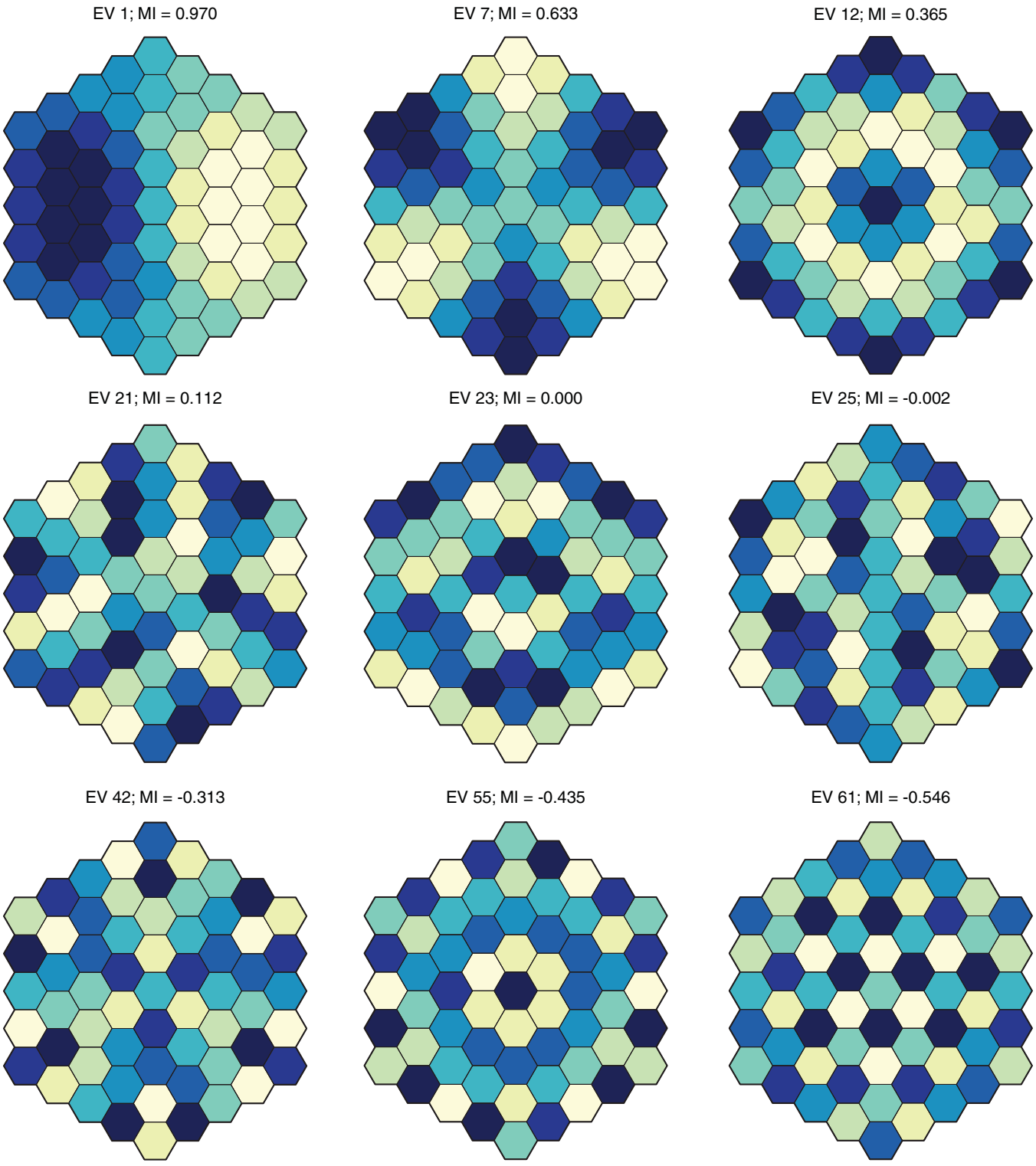

Figure 5. Examples of derived spatially autocorrelated patterns. These were created by taking the eigenvectors of matrix **MCM** based on Equation 1. (Color figure available online.)

the OLS model. The eigenvectors are synthetic variates that function as surrogates for missing variables. They are similar to the spatially structured random effects in a mixed linear model. Using these eigenvectors as additional independent variables will remove the spatial misspecification from the model. The results of the SFE model are not spatially autocorrelated ($MI = –0.013$, $|z| \approx 0.1$, $p = 0.469$), are normally distributed ($SW = 0.974$, $p = 0.358$), and are homoscedastic ($BP = 9.470$, $p = 0.149$). The spatially filtered linear regression model now meets the assumptions of normality, homoscedasticity, and nonspatial autocorrelation of the residuals.

Removing the spatial patterns inherent in the residuals and using them as independent variables also increases the predictive power of the model (Dormann 2007). The adjusted R^2 value has increased from 0.53 ($p < 0.000$) to 0.68 ($p < 0.000$), which, according to the Williams–Steiger test, is a statistically significant increase ($WS = –2.748$, $p = 0.004$). The coefficients for the income and home value variables did not change. The mean squared error (MSE) of the regression dropped from 130.759 (OLS) to 88.343 (SFE).

When Getis (2010) applied his method to the Columbus data, the results were very similar: his adjusted R^2 increased to 0.72. The residuals of the SFE approach contain less spatial autocorrelation ($|z| \approx 0.08$) than those of the Getis approach, although those of the Getis approach were successfully filtered and clearly not statistically autocorrelated ($|z| \approx 1.16$). One advantage of the Getis approach is that the researcher is able to see the effects on the regression model of each filtered, now aspatial, variable separate from and along with each variable's spatial component. The R SpatialFiltering function returns a set of spatial filters for the set of predictor variables, not for each variable individually.

Applications of Spatial Filtering with Eigenvectors in the Literature

Economists make frequent use of SFE to study economic convergence because rates of convergence depend strongly on the assumed underlying spatial patterns of the data. Cuaresma and Feldkircher (2010) examined the rate of income convergence in Europe and found a convergence rate of 1 percent, about half of the value typically reported in nonspatially filtered studies. Le Gallo and Dall'erba (2008) and Badinger, Möller, and Tondl (2004) examined economic convergence in the European Union and found that omitting spatial effects (i.e., not using SFE or a similar technique) can result in biased measurements of convergence. Pecci and Pontarollo (2010) used SFE to account for spatial interactions and structural differences in their model of economic convergence. They were able to improve the R^2 of their model from 0.519 to 0.961. Montresor, Pecci, and Pontarollo (2010) examined European policies, more specifically European Union Structural Funds, to determine their affect economic convergence. They were able to improve the R^2 of their models from approximately 0.60 to over 0.90. Chen and Rura (2008) studied the Jiangsu province in China, looking to see whether the wave of annexation of cities resulted in greater economic integration between the peripheral areas and the cities. For 1999, the R^2 of their geographically weighted regression model increased from 0.88 to 0.95 when SFE was applied.

Spatially filtered models have been used to analyze employment and production data sets as well. Mayor and López (2008) used Getis's spatial filtering technique (Getis and Griffith 2002) to analyze the evolution of regional employment in Spain. They were able to measure both the spatial and nonspatial relationships between regions. Patuelli et al. (2011) studied German unemployment using SFE and uncovered spatial patterns that were consistently significant over time. The R^2 measurements of their models improved from approximately 0.75 to 0.95. Möller and Soltwedel (2007, 99) ended their guest editorial by stating:

> Spatial econometrics is able to compensate for the lack of data for functionally defined regional labor markets. Hence, as tests of economic theories may have to rely more and more on regionally disaggregated time series, then it is easy to predict that spatial econometrics in general will play an even more important role for labor market analysis in the future.

Grimpe and Patuelli (2009) used SFE to measure the effects of research and development activities on nanomaterials patents in German districts. They examined public and private research and development and found significant positive effects of both. Further, their analysis hints that the colocation of both private and public efforts results in a positive interaction that increases productivity. Fischer and Griffith (2008) compared SFE to a spatial econometric model by analyzing patent citation data across European regions. Both methods increased the fit of the models.

Health geography and epidemiology is another area in which SFE has been used. Fabbri and Robone (2010) examined the in- and outflow of patients in 171 local health authorities in Italy. They tested the patient flows

for spatial autocorrelation and, when present, they used SFE to account for it. They found that patients who go to hospitals outside of their region tend to come from poorer regions and that neighboring hospitals compete with one another less than with hospitals that are more removed. Tiefelsdorf (2007) modeled prostate cancer in the 508 U.S. State Economic Areas using exposure to risk factors as independent variables. A common problem with disease modeling is that the long latency period associated with some diseases allows people to contact the risk factors, move to another region, and then be diagnosed. Tiefelsdorf used the 1965 to 1970 interregional migration census as the underlying spatial structure used in calculating the eigenvectors. By including SFE to account for patient migration, the R^2 improved from 0.146 to 0.538. Jacob et al. (2008) used SFE to model the incidence of arboviral and protozoan disease vectors in Gulu, Uganda, based on hydrological and geophysical factors. Their SFE model selected both positively and negatively spatially autocorrelated eigenvectors, demonstrating the complexity of disease vector distribution. They used this spatial filter to determine that 12 to 28 percent of the information in the count samples was redundant.

Migration flows have also been studied using spatial filtering techniques. Chun (2008) modeled U.S. interstate migration using population, unemployment, income per capita, and mean January temperature as independent variables in an SFE Poisson model. The filtered model had a lower standard error, a lower z score of the T statistic, and generally more significant p values associated with the independent variables. This made interpreting the parameters of the spatially filtered model easier and more intuitive. Griffith and Chun (2011) extended that work by including Poisson, linear mixed models, and generalized linear mixed models regression variants. Their results showed that network autocorrelation can be successfully accounted for in each of these models using SFE.

Spatial filtering has also been used in ecology and biogeography. Diniz-Filho and Bini (2005) used SFE methods to evaluate spatial patterns in bird species richness in South America. They achieved an $R^2 = 0.917$ and found the SFE method a simple and suitable method to measure species richness while taking into account spatial autocorrelation. Dray, Legendre, and Peres-Neto (2006) adapted the principal coordinate analysis of neighbor matrices technique by including SFE for modeling ecological distributions. Ficetola and Padoa-Schioppa (2009) used SFE to determine the effects of human activity on the extinction and colonization rates of island biogeography. Kühn (2007) studied the relationship between plant species richness and environmental correlates and discovered that large-scale environmental gradients can be inverted at the local or regional level. He concluded that these patterns would not have been recognized or included in the analysis without SFE. De Marco, Diniz-Filho, and Bini (2008) studied species distribution modeling during range expansion and determined that mechanisms that generate range cohesion and determine species' ranges under climate change can be captured using SFE.

Method

To begin to understand the conditions under which the spatial filtering with the eigenvector approach can be most fruitfully applied, we have collected a set of geo-referenced data sets and submitted them to OLS and to SFE analyses. We also generated a data set of variables wherein we controlled the amount of spatial autocorrelation in the residuals. Comparing OLS and SFE models for various data sets provides a sense of how and when SFE can be used. We tested the residuals of each model for normality, homoscedasticity, and spatial autocorrelation. We used the Williams–Steiger test (Williams 1959; Steiger 1980) to determine whether the fit of the SFE model is statistically different from that of the OLS model.

For the simulated data set, we used a normal distribution generator to create two independent variables, x_1 and x_2. These data contained 100 observations and were associated with a hexagonal tessellation with ten rows and ten columns. The x variables were scaled so that they ranged from zero to one. A series of 100 eigenvectors was generated using Equation 1 and these were also scaled to range from zero to one. The dependent variable was created by linearly combining the two x variables and one of the eigenvectors using the coefficients $\mathbf{c} = (1, 1, 2)$. The two x variables and the dependent variable were included in the OLS and the SFE models, thus the spatial autocorrelation of the residuals of the models ranged from a strong positively correlated pattern, through random patterns, to a strong negatively correlated pattern. The model was then run for each of the 100 eigenvectors. Varying the last coefficient of $\mathbf{c}$ allowed us to alter the fit of the model. Increasing the last coefficient of $\mathbf{c}$ added weight to the derived spatial pattern—and therefore the residuals—and decreased the R^2 of the OLS model. Decreasing the last of the coefficients reduced the weight of the spatial pattern and increased the R^2 of the model. This was done

repeatedly and the results were compared to ensure that the emergent patterns were consistent.

Tiefelsdorf and Griffith (2007) used a simulated data set to determine how much power is lost by using SFE. They found that SFE does not lead to biased results and that it is able to recover the pattern of the data satisfactorily. Like the results of their simulated model, ours are limited because they are tied to a specific ten-by-ten tessellation of hexagons.

Results

Simulated Data Sets

The coefficients used to create the model results presented in Figure 6 were $\boldsymbol{c} = (1, 1, 1)$, which resulted in OLS models with R^2 values that were centered near 0.6 (Figure 6B). The *MI* values for the residuals of the models began at 0.98 and ended at –0.38 (Figure 6A).

The spatial filtering technique lowered the *MI* of the model residuals to near zero for all models, effectively removing the misspecification error (Figure 6A). OLS models with residuals that are randomly distributed (*MI* near zero), and therefore are not spatially misspecified, do not need to be spatially filtered. Processing time for models with randomly distributed residuals was longer than for positively or negatively spatially autocorrelated models. During several iterations of the simulated data sets, the R SpatialFiltering function struggled to correct for negatively spatially autocorrelated misspecification errors. This appeared to be an issue with the default parameters associated with the eigenvector selection process. A trial-and-error process was used to determine the most appropriate parameters for these models and the problem disappeared (these are the data shown in Figure 6A). We discuss the parameters of the selection process in more detail in the Discussion section of this article. The effects of negative spatial autocorrelation on regression inferences are not well understood, although our understanding is increasing (Griffith and Arbia 2010). Negative spatial autocorrelation is very rare in empirical data (Griffith and Arbia 2010), so it is unlikely that negatively spatially autocorrelated residuals will be a concern for most researchers.

As the spatial patterns are removed from the residual term and are included as independent variables, the R^2 will increase (Dormann et al. 2007; Dutilleul, Pelletier, and Alpargu 2008). The magnitude of the increase in R^2 values increases with increasing misspecification in the models. The most dramatic increase occurred at the extremes of the *MI* range. The largest increase in R^2 was from 0.555 to 0.917. The increase in R^2 of the models at these extremes was statistically significant. Figure 6B includes the *p* values of the Williams–Steiger test. The increase in R^2 was not statistically significant when the *MI* of the model residuals was near zero. Obtaining a high R^2 for a spatially filtered model indicates that the model contains strongly spatially autocorrelated data.

A stronger model fit is generally desirable; however, we suggest that a dramatic increase in R^2 values could be problematic. This might indicate that important independent variables are missing from the model. When this is the case, we suggest that researchers identify and use the missing independent variable rather than continue with the SFE model. The derived spatial filter will mirror the distribution of the missing variable, which should aid in its discovery. Awareness of the missing variable, and a knowledge of its pattern, would not have been possible without SFE. The *p* values of the

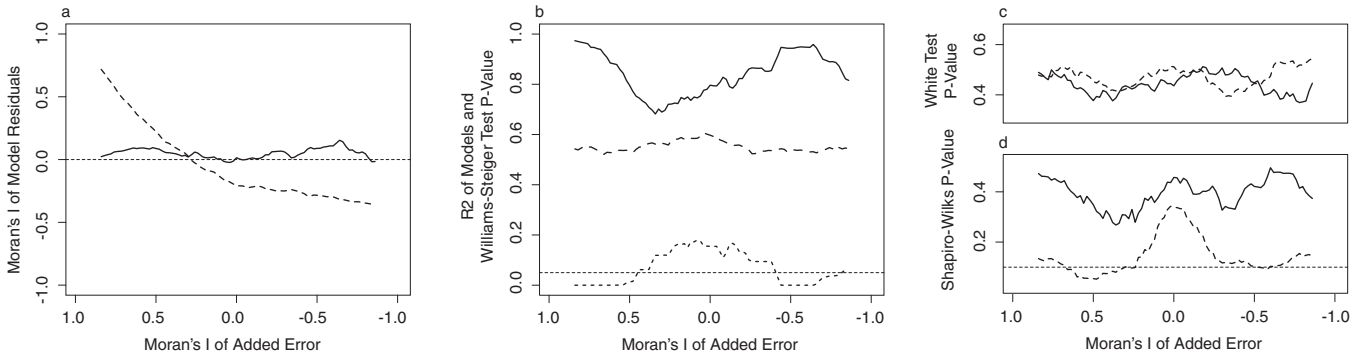

Figure 6. Results of spatial filtering applied to the simulated data sets. The solid line is the results of the spatially filtered linear models, and the dashed line is the results of the nonfiltered models. (A) The Moran's *I* of model residuals. (B) The R^2 of the models and the *p* values of the Williams–Steiger tests used to determine whether the difference between the R^2 values was statistically significant (the dotted line); an α of 0.05 is indicated with a horizontal dotted line. (C) The *p* values of the White tests used to assess the homoscedasticity of the models' residuals. (D) The *p* values of the Shapiro–Wilks tests used to assess the normality of the models' residuals; an α of 0.1 is indicated with a horizontal dotted line.

Williams–Steiger tests performed on our simulated data sets indicate that the increase in R^2 values is statistically significant when the OLS model residuals have an $MI \geq 0.25$ (Figures 6A, 6B), although this threshold will likely change in empirical studies and each model should be thoughtfully evaluated.

An additional benefit of the SFE models is that the residuals can become more normally distributed, which is another assumption of regression models. The residuals of the OLS models of our simulated data sets were all nonnormal, or close to nonnormal ($\alpha = 0.1$, Shapiro–Wilks test), except for when the MI of the added error centered on zero (Figure 6D). In all cases, SFE resulted in normally distributed model residuals. Even the normality of nonspatially misspecified models was improved (larger p values), although this was not a serious problem in these models.

We also tested for changes in the homoscedasticity of model residuals using the Breusch–Pagan test and the White test, but there was no significant change between the OLS and SFE models. The White test is more conservative but it is currently a little more difficult to implement in R, so the Breusch–Pagan test is used in the Appendix (Line 44). Interested researchers can learn about the White test in R using the searchable archive of the R users' mailing list (http://tolstoy.newcastle.edu.au/R/). The results of the White test are shown in Figure 6C. Both the OLS and the SFE models' residuals were homoscedastic. We suspect this is because of the way the data were simulated, not necessarily because spatial filtering with eigenvectors is incapable of improving the homoscedasticity of model residuals. Several of the real data sets we examined had OLS models with heteroscedastic residuals that became homoscedastic under the SFE models.

Real Data Sets

The results of these models are presented in Table 1. The Columbus, Ohio, crime data have already been discussed. The other real data examples demonstrate the advantages and potential problems associated with SFE.

Bladder Cancer by State Economic Areas. Tiefelsdorf and Griffith (2007) modeled the occurrence of bladder cancer by state economic areas using exposure to risk factors as predictor variables. They used lung cancer rates as a surrogate for smoking rates and they used the population density as a surrogate for environmental and occupational risks, as well as behavioral differences in urban and rural lifestyles. Indoor radon concentrations were also included as independent variables (Tiefelsdorf 2007). The data were obtained from the *Atlas of Cancer Mortality in the United States: 1950–94* (Devesa et al. 1999). The connectivity matrix for this model was derived by Tiefelsdorf and Griffith (1997) and is based on the 1965 to 1970 interregional migration census rather than geographic adjacency. The SFE model selected nineteen eigenvectors as the spatial filter, more eigenvectors than any of the other examples. This might be due to the increased spatial complexity of this model—these data have 508 observations and a $\mathbf{C}$ matrix based on the migration census. The R^2 increased from 0.26 to 0.50 and the increase was statistically significant ($WS = -8.201$, $p < 0.000$). The residuals for both the OLS and the SFE models were normally distributed. The OLS residuals were heteroscedastic and moderately spatially autocorrelated. The SFE model residuals were homoscedastic and not spatially autocorrelated. This is a successful implementation of SFE.

Note that our results on the bladder cancer example do not match those of Tiefelsdorf and Griffith (2007) because we used an additional independent variable, indoor radon concentrations, which was added to the data set later (Tiefelsdorf 2007). This increased the R^2 of our model. Also, the inclusion of an additional independent variable meant that our spatial filtering model generated a slightly different collection of eigenvectors. Our SFE R^2 value was lower than that reported by Tiefelsdorf and Griffith (2007) because we selected fewer and different eigenvectors as our spatial filter.

Per Capita Income at Various Scales. To begin to assess the effects of scale on spatial filtering with eigenvectors, we modeled per capita income using mean household size, percentage of population in urban areas, and percentage of population that is foreign-born as predictor variables. These data came from the 2000 U.S. Census. We built models for the forty-eight conterminous U.S. states plus the District of Columbia, for the 102 counties of Illinois, and for the forty-one census tracts of McLean County, Illinois. These data were accessed from the National Historical Geographic Information System (NHGIS) database.

The per capita income for U.S. states example is a successful application of SFE. Residuals were spatially autocorrelated under the OLS model and not spatially autocorrelated under the SFE model. Residuals were normally distributed and homoscedastic under both models. The per capita income for Illinois counties example did not need SFE. The residuals of both models were normally distributed, homoscedastic, and not spatially autocorrelated. The R^2 increased from

Table 1. The results of ten real data sets analyzed using spatial filtering with eigenvectors

	Columbus, Ohio crime		Bladder cancer rates		Per capita income: U.S. states		Per capita income: Counties of Illinois		Per capita income: Census tracts of McLean Co., Illinois	
	OLS	SFE	OLS	SFE	OLS	SFE	OLS	SFE	OLS	SFE
n	49	—	508	—	49	—	102	—	41	—
MI (**Y**)	0.519 ($\|z\| \approx 5.7$)	—	0.446 ($\|z\| \approx 16.7$)	—	0.333 ($\|z\| \approx 3.9$)	—	0.529 ($\|z\| \approx 9.2$)	—	0.375 ($\|z\| \approx 4.4$)	—
SE	130.759	88.343	0.030	0.021	4,182,633	3,043,687	3,187,803	2,369,256	43,353,637	22,566,841
R^2	0.552 ($p < 0.000$)	0.724 ($p < 0.000$)	0.261 ($p < 0.000$)	0.496 ($p < 0.000$)	0.587 ($p < 0.000$)	0.719 ($p < 0.000$)	0.670 ($p < 0.000$)	0.690 ($p < 0.000$)	0.095 ($p = 0.292$)	0.580 ($p < 0.000$)
WS	—	–2.748 ($p = 0.004$)	—	–8.201 ($p < 0.000$)	—	–2.383 ($p = 0.011$)	—	–1.270 ($p = 0.104$)	—	–3.800 ($p < 0.000$)
Filter										
m	—	4	—	19	—	3	—	1	—	4
MI_{filter}	—	0.676 ($\|z\| \approx 7.3$)	—	0.928 ($\|z\| \approx 34.7$)	—	0.571 ($\|z\| \approx 6.6$)	—	1.046 ($\|z\| \approx 17.5$)	—	0.663 ($\|z\| \approx 7.5$)
Results										
MI ($\hat{\mathbf{Y}}$)	0.397 ($\|z\| \approx 4.5$)	0.561 ($\|z\| \approx 6.1$)	0.309 ($\|z\| \approx 11.6$)	0.794 ($\|z\| \approx 29.7$)	0.207 ($\|z\| \approx 2.5$)	0.428 ($\|z\| \approx 5.0$)	0.674 ($\|z\| \approx 11.6$)	0.711 ($\|z\| \approx 12.2$)	0.286 ($\|z\| \approx 3.5$)	0.659 ($\|z\| \approx 7.5$)
Residuals										
SW	0.977 ($p = 0.450$)	0.974 ($p = 0.358$)	0.998 ($p = 0.827$)	0.996 ($p = 0.158$)	0.966 ($p = 0.165$)	0.953 ($p = 0.047$)	0.966 ($p = 0.011$)	0.950 ($p = 0.001$)	0.978 ($p = 0.583$)	0.971 ($p = 0.363$)
BP	7.217 ($p = 0.027$)	9.470 ($p = 0.149$)	14.560 ($p = 0.002$)	26.729 ($p = 0.222$)	1.981 ($p = 0.576$)	3.914 ($p = 0.688$)	31.173 ($p < 0.000$)	30.938 ($p < 0.000$)	4.297 ($p = 0.231$)	8.353 ($p = 0.303$)
MI	0.251 ($\|z\| \approx 2.9$)	–0.013 ($\|z\| \approx 0.1$)	0.298 ($\|z\| \approx 11.2$)	0.004 ($\|z\| \approx 0.2$)	0.193 ($\|z\| \approx 2.4$)	0.014 ($\|z\| \approx 0.4$)	0.065 ($\|z\| \approx 1.2$)	0.001 ($\|z\| \approx 0.2$)	0.349 ($\|z\| \approx 4.2$)	–0.013 ($\|z\| \approx 0.1$)
	1940 school years		1980 telephone		Texas mortgages		Access to health care: Mexico states		Latin American immigration	
	OLS	SFE	OLS	SFE	OLS	SFE	OLS	SFE	OLS	SFE
n	49	—	49	—	254	—	32	—	25	—
MI (**Y**)	0.567 ($\|z\| \approx 6.5$)	—	0.292 ($\|z\| \approx 3.4$)	—	0.526 ($\|z\| \approx 13.9$)	—	0.514 ($\|z\| \approx 4.9$)	—	0.151 ($\|z\| \approx 1.4$)	—
SE	0.303	0.196	0.0005	0.0002	7,590.883	6,837.204	32.803	23.294	0.002	0.002
R^2	0.591 ($p < 0.000$)	0.759 ($p < 0.000$)	0.748 ($p < 0.000$)	0.915 ($p < 0.000$)	0.768 ($p < 0.000$)	0.795 ($p < .000$)	0.847 ($p < 0.000$)	0.899 ($p < 0.000$)	0.593 ($p = 0.001$)	0.804 ($p = 0.002$)
WS	—	–2.913 ($p = 0.003$)	—	–4.841 ($p < 0.000$)	—	–2.902 ($p = 0.002$)	—	–1.948 ($p = 0.031$)	—	–2.499 ($p = 0.010$)
Filter										
m	—	4	—	4	—	5	—	2	—	6
MI_{filter}	—	0.695 ($\|z\| \approx 7.8$)	—	0.841 ($\|z\| \approx 9.3$)	—	0.528 ($\|z\| \approx 13.8$)	—	0.708 ($\|z\| \approx 6.7$)	—	–0.290 ($\|z\| \approx 1.8$)
Results										
MI ($\hat{\mathbf{Y}}$)	0.646 ($\|z\| \approx 7.3$)	0.715 ($\|z\| \approx 8.1$)	0.097 ($\|z\| \approx 1.3$)	0.290 ($\|z\| \approx 3.4$)	0.616 ($\|z\| \approx 16.2$)	0.603 ($\|z\| \approx 15.9$)	0.623 ($\|z\| \approx 05.9$)	0.571 ($\|z\| \approx 5.4$)	0.329 ($\|z\| \approx 2.5$)	0.246 ($\|z\| \approx 2.0$)
Residuals										
SW	0.971 ($p = 0.266$)	0.972 ($p = 0.297$)	0.955 ($p = 0.058$)	0.975 ($p = 0.363$)	0.692 ($p < 0.000$)	0.731 ($p < 0.000$)	0.983 ($p = 0.983$)	0.988 ($p = 0.968$)	0.871 ($p = 0.004$)	0.972 ($p = 0.701$)
BP	21.954 ($p = 0.000$)	17.104 ($p = 0.129$)	5.864 ($p = 0.118$)	9.355 ($p = 0.228$)	5.632 ($p = 0.344$)	20.573 ($p = 0.024$)	1.837 ($p = 0.607$)	8.448 ($p = 0.133$)	9.313 ($p = 0.054$)	18.060 ($p = 0.054$)
MI	0.276 ($\|z\| \approx 3.3$)	–0.018 ($\|z\| \approx 0.0$)	0.561 ($\|z\| \approx 6.3$)	0.004 ($\|z\| \approx 0.3$)	0.055 ($\|z\| \approx 1.7$)	–0.008 ($\|z\| \approx 0.1$)	0.207 ($\|z\| \approx 2.2$)	–0.053 ($\|z\| \approx 0.2$)	–0.162 ($\|z\| \approx 0.9$)	–0.024 ($\|z\| \approx 0.1$)

Note: OLS = ordinary least squares; SFE = spatial filtering with eigenvectors; n = number of observations; $MI(\mathbf{Y})$ = Moran's I of the dependent variable; SE = standard error of the model; R^2 = coefficient of determination of the model; WS = the statistic from the Williams–Steiger test used to determine whether the increase in R^2 is statistically significant; m = number of eigenvectors selected as the spatial filter; MI_{filter} = Moran's I of the spatial filter; $MI(\hat{\mathbf{Y}})$ = Moran's I of the predicted dependent variable; SW = Shapiro–Wilks statistic of normality; BP = Breusch–Pagan test statistic for homoscedasticity; MI = Moran's I of the model residuals.

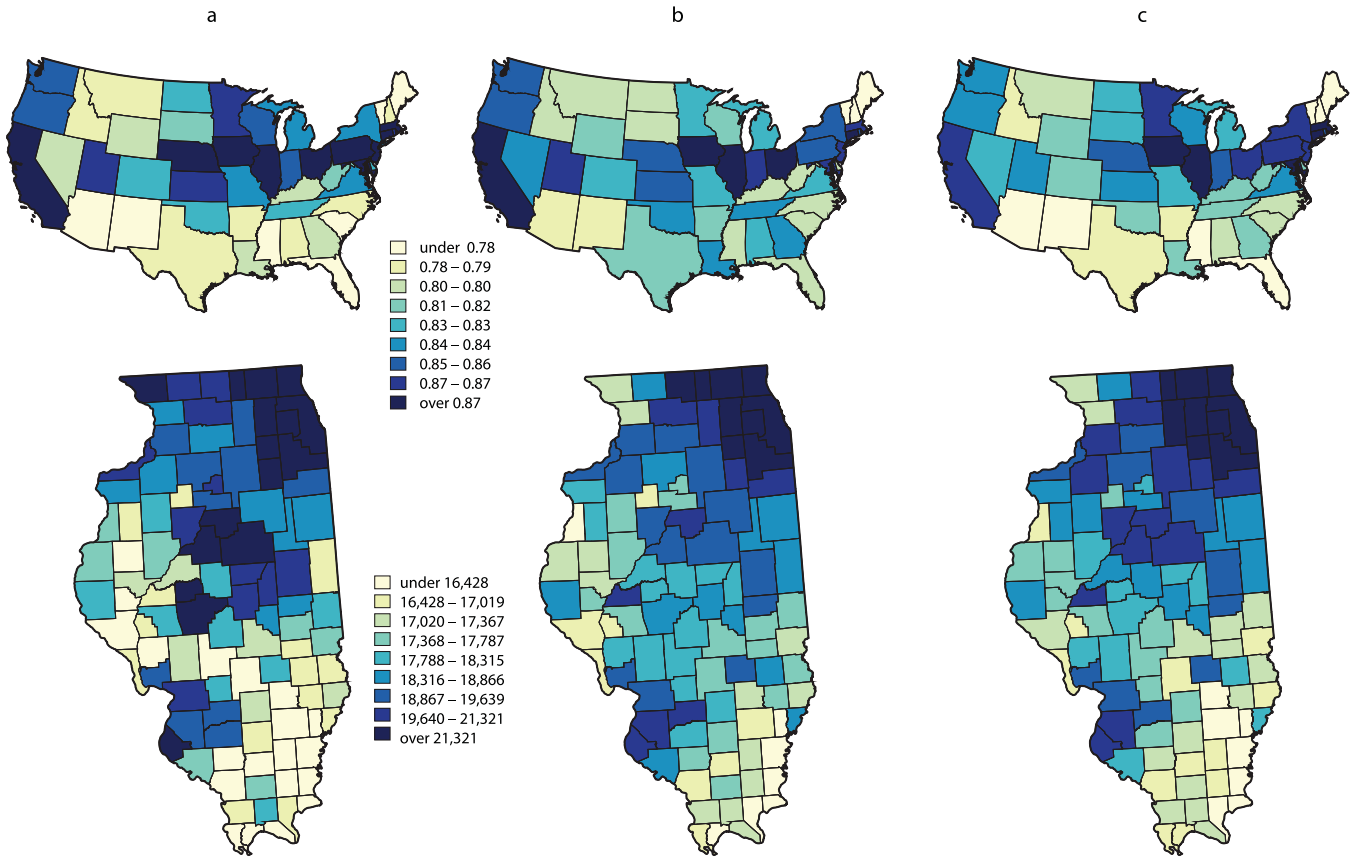

Figure 7. Two examples of spatially filtered linear models. Column A displays the original data, column B shows the predicted dependent variable of the nonspatially filtered model, and column C shows the predicted dependent variable of the model spatially filtered with eigenvectors. The first example shows the percentage of people by state who had telephones in their homes in 1980. Notice that the spatially filtered model does a better job of predicting the spatial pattern of telephone ownership. The second example models per capita income by county in Illinois and is an example of a model that does not suffer from spatial misspecification problems and therefore does not need to be spatially filtered. (Color figure available online.)

0.67 to 0.69, an increase that was not statistically significant at $\alpha = 0.05$. The OLS model was not misspecified and did not need correction. The original data and the predicted results of the OLS and SFE models are presented in Figure 7. Residuals for the per capita income for McLean County, Illinois, example were normally distributed and homoscedastic under both OLS and SFE. They were strongly spatially autocorrelated under the OLS; this was corrected under the SFE. The fit of the OLS model was weak and insignificant ($R^2 = 0.095$, $p = 0.292$) but moderately strong and significant in the SFE model ($R^2 = 0.580$, $p < 0.000$). Obviously, this model is missing at least one dependent variable. Before analysis of this model can continue, these missing data need to be found.

1940 Median School Years Attended. We modeled the median terminal year of school achievement by U.S. conterminous states and the District of Columbus using data from the 1940 census. Independent variables were mean household value, percentage of homes with refrigerators, percentage of homes with running water, and the percentage of the population living in urban areas. These data were accessed from the NHGIS. This model benefited from SFE. The residuals of both the OLS and SFE models for this example were normally distributed. They were heteroscedastic under the OLS model ($BP = 21.94$, $p < 0.000$) and homoscedastic under the SFE model ($BP = 17.10$, $p = 0.129$). The *MI* of the residuals decreased from 0.276 ($|z| \approx 3.3$) under OLS to –0.018 ($|z| \approx 0.0$) under the SFE. The increase in model fit was statistically significant.

1980 Percentage of U.S. Homes with Telephones. The percentage of homes with telephones in 1980 by U.S. state was predicted using the percentage of

homes with indoor plumbing, the percentage of the population living in urban areas, and the median value of households as independent variables. These data originated with the 1980 U.S. Census and were gathered from the NHGIS. The OLS model residuals were not normally distributed but were strongly spatially autocorrelated ($MI = 0.561$, $|z| \approx 6.3$); they were normally distributed and not spatially autocorrelated ($MI = 0.004$, $|z| \approx 0.3$) under the SFE model. Residuals were homoscedastic under both models. The increase in R^2 was statistically significant ($WS = -4.841$, $p < 0.000$). The *MI* of the results of the SFE model was nearly identical to that of the dependent variable, which strongly suggests that the SFE model is doing a better job of predicting the dependent variable. The original data and the predicted results of the OLS and the SFE models are presented in Figure 7.

Texas Mortgage Payments by Housing Units. Another data set was prepared by Hubenig, Beckstead, and Tiefelsdorf and distributed as part of the 2008 Spatial Filtering Workshop held in Dallas, Texas, on 16–20 June. The independent variable was the median monthly mortgage payment by county. The independent variables were population density, the percentage of the population aged twenty-five and older with at least a high school education, the median household income, the percentage of housing units that were built since 1980, and the median age. The residuals of the OLS model were not spatially autocorrelated ($MI = 0.055$, $|z| \approx 1.7$), indicating that there was no spatial misspecification in the model. Not only was SFE unnecessary to correct for spatial autocorrelation but it seems to have introduced a new problem: The residuals of the OLS model were homoscedastic, whereas those of the SFE model were heteroscedastic.

Access to Health Care in Mexico. We modeled the percentage of the population with access to health care by states of Mexico in 2009. The independent variables were the percentages of the population with running water, with sewer service, and with a refrigerator at home. The data were accessed through the Web site of the Institute Nacional de Estadística y Geographía of Mexico (INEGI 2010). The residuals of both models were normal and homoscedastic. Those of the OLS model were significantly spatially autocorrelated ($MI = 0.207$, $|z| \approx 2.2$), but the spatial autocorrelation was successfully removed in the SFE model ($MI = -0.053$, $|z| \approx 0.2$). The R^2 of SFE model was slightly larger than that of the OLS model, but the increase was statistically significant ($WS = -1.948$, $p = 0.031$). The SFE model is an improvement over the OLS model.

Immigration from Latin America to the United States. We also examined immigration to the United States from Latin America. The dependent variable was the percentage of the populations of Latin American countries living in the United States. These data were collected from the 2000 U.S. Census. The predictor variables were the gross domestic product–purchasing power parity (GDP–PPP), the human development index (HDI), the GINI coefficient, and the population density. These data are available from the United Nations Development Programme (2010) and the Central Intelligence Agency (CIA) World Factbook (CIA 2010). This was the only model with negatively spatially autocorrelated residuals in the OLS model ($MI = -0.162$, $|z| \approx 0.9$). SFE successfully removed the autocorrelation from the residuals ($MI = -0.024$, $|z| \approx 0.1$). The residuals of the OLS model were not normal, but those of the SFE model were. Both models' residuals were heteroscedastic. The increase in the fit of the model ($R^2 = 0.593$ to $R^2 = 0.804$) was dramatic and significant. One might be tempted to look for a missing dependent variable; however, the rarity of negatively spatially autocorrelated data might make this difficult.

Several trends were observed in the models' results. First, the residuals of the SFE models were less spatially autocorrelated than those of the OLS models. Also, most of the SFE results were more spatially autocorrelated than those of the corresponding OLS models. This is expected, of course, because the spatial patterns in the residuals have been moved to the independent variables and thus to the dependent variable. Of the three examples where the SFE results were less spatially autocorrelated than those of the OLS, one was not spatially misspecified and so did not need to be spatially filtered (Texas mortgages), and another had negatively spatially autocorrelated OLS residuals and a negatively spatially autocorrelated spatial filter (Latin American immigration). The third, access to health care in Mexico, had positively spatially autocorrelated OLS model residuals and a positively spatially autocorrelated SFE spatial filter; nonetheless, the predicted dependent variable of the SFE model was less spatially autocorrelated than those of the OLS model.

Second, the increase of the R^2 of the SFE models over those of the OLS models was statistically significant as determined by the Williams–Steiger test ($\alpha = 0.05$) for all examples except the per capita income by Illinois counties, which was not spatially misspecified.

As in the simulated data, the significance of the increase in R^2 values increases as the OLS model residuals become more spatially autocorrelated. Several of the real data sets had OLS model residuals with *MI* values less than 0.25 ($|z| \approx 5.0$), however, which is the threshold suggested by the simulated data.

Third, the *MI* of the SFE-predicted dependent variable tended to be closer to the *MI* of the original dependent variable than was the *MI* of the predicted variable of the OLS model. This was true for the following examples: Columbus, Ohio, crime; per capita income of U.S. states; per capita income of McLean County, Illinois; 1980 telephones; Texas mortgages; access to health care in Mexico; and Latin American immigration. We take this as a sign, along with the others, that the SFE model is superior to the OLS model in these cases.

Fourth, unlike the simulated data sets whose residuals were more normally distributed using the SFE model, most of the real data sets' residuals became somewhat less normally distributed (the p value of the Shapiro–Wilks test decreased); however, the decrease in normality was never sufficient to make the residuals statistically nonnormal at $\alpha = 0.1$. The residuals of the 1980 telephones and the Latin American immigration examples were not normally distributed under the OLS model but became normally distributed under the SFE model.

Fifth, homoscedasticity of the real data set examples varied, with eight of the ten examples showing improvement or no change after spatial filtering. The Texas mortgages example, which was not spatially misspecified, was the only data set with residuals that went from being statistically homoscedastic to statistically heteroscedastic. All other data sets had homoscedastic residuals after spatial filtering.

Discussion

One potential difficulty with applying the SFE model is determining which of the n eigenvectors should be included in the spatial filter and therefore in subsequent analysis. The simple answer is that the most parsimonious collection of eigenvectors is the most desirable; in other words, as few eigenvectors as possible as long as they adequately remove the spatial misspecification from the model. The R SpatialFiltering function uses an iterative process that searches through the eigenvectors for the one that reduces the Moran's *I* of the regression residuals most and continues until no additional eigenvectors reduce the residuals' *MI* by more than a provided threshold. Eigenvector selection is based on one of two parameters (Tiefelsdorf and Griffith 2007). The first is a convergence tolerance parameter. The *MI* of the model residuals is estimated and compared to the tolerance threshold (the default is 0.1). When the estimated *MI* of the residuals is less than this threshold, the SpatialFiltering function terminates and the selected eigenvectors are returned (Tiefelsdorf and Griffith 2007). Increasing the tolerance threshold will reduce the number of eigenvectors chosen. The second parameter that can be used to choose eigenvectors is a threshold for the alpha value of each eigenvector in the model. Eigenvectors with an alpha that is less than or equal to the supplied alpha threshold are chosen. The alpha parameter default is null and the tolerance threshold is used unless an alpha threshold is provided.

While processing the data sets used in this analysis, we noticed that the default parameter for eigenvector selection in the R SpatialFiltering function tended to be too liberal and inclusive, especially for negatively spatially autocorrelated models. Tiefelsdorf and Griffith (2007) noticed a similar situation in their simulation experiment. When the selected eigenvectors were included in the regression model, some of the models were overcorrected (Griffith 2003) and the residuals went from displaying significant positive spatial autocorrelation to significant negative spatial autocorrelation, with little change in the magnitude of the spatial autocorrelation. This does not remove the spatial misspecification from the model. Many of our simulated data that had strong negative spatially autocorrelated OLS residuals did not improve when the defaults of the R SpatialFiltering function were used. Once appropriate parameters were used, the negative misspecification of the models was removed. Only one of our real data sets displayed negative misspecification (the Latin American immigration example); nonetheless, we did not rely on the default parameters when processing any of these data.

We chose to manipulate the alpha parameter to create SFE models that used fewer eigenvectors and were better able to eliminate the spatial autocorrelation of the model residuals. Determining the appropriate alpha for each model was a trial-and-error process. We wrote an R script that ran the SpatialFiltering function for a sequence of alpha values with 0.05 increments. The alpha parameter that returned the SFE model with residuals that had the lowest *MI* was used for subsequent analysis. Most of the alpha values selected in this manner ranged from 0.05 to 0.3.

The issue of determining how many spatial filters to include is unique to the eigenvector method of spatial filtering. The advantage of using multiple filters is that each can account for spatial autocorrelation at a different scale (Diniz-Filho and Bini 2005; Dormann et al. 2007), and each could be a surrogate for a different missing independent variable. With a little patience, researchers can determine the most appropriate parameters for the R SpatialFiltering function.

We urge caution when the fit of the model increases dramatically when spatial filtering is used. It is likely that the residuals display such strong spatial autocorrelation not just because of a spatial misspecification error but because the independent variables cannot account for all the spatial patterns displayed by the dependent variable. Our suggestion in this situation is not to rely unquestioningly on the results of the SFE model (which will likely exaggerate the predictive ability of the model); rather, we suggest that researchers look for and use additional independent variables that can explain more of the remaining pattern of the dependent variable before continuing with SFE analysis (Griffith 2004). The decision between when to rely on the SFE model and when to search for additional independent variables needs to be made on a case-by-case basis and should be carefully weighed by researchers. The quality of the analysis cannot exceed that of the data (Wakefield 2003; Tiefelsdorf and Griffith 2007).

This work, and the earlier work of others cited herein, has shown that spatial filtering with eigenvectors is a powerful, flexible, and useful method for reducing spatial misspecification errors in linear regression models that use spatial data. The spatial autocorrelation of model residuals can be reduced to random, spatially nonautocorrelated patterns. We demonstrate here that spatial filtering with eigenvectors also tends to increase the normality of model residuals and increase the homoscedasticity of model residuals (although one of our real data examples was an exception to this trend). We suggest that spatial filtering with eigenvectors be applied to any linear regression model using spatial data, except in those cases, like the Texas mortgages and the per capita income by Illinois counties examples, when the residuals of the OLS model have near-zero *MI* values and therefore do not have significant spatial misspecification problems.

References

Anselin, L. 1988. *Spatial econometrics: Methods and models*. Dordrecht, The Netherlands: Kluwer Academic.

Badinger, H., W. G. Möller, and G. Tondl. 2004. Regional convergence in the European Union (1985–1999): A spatial dynamic panel analysis. *Regional Studies* 38:241–53.

Bivand, R., M. Altman, L. Anselin, R. Assunção, O. Berke, A. Bernat, E. Blankmeyer, et al. 2010. spdep: Spatial dependence: Weighting schemes, statistics and models. R package version 0.5–24. http://cran.r-project.org/web/packages/spdep/index.html (last accessed 4 August 2011).

Bivand, R., and R. Brunstad. 2006. Regional growth in Western Europe: Detecting spatial misspecification using the R environment. *Papers in Regional Science* 85:277–97.

Bivand, R., H. Ono, and R. Dunlap. 2009. classInt: Choose univariate class intervals. R package version 0.1–14. http://cran.r-project.org/web/packages/classInt/index.html (last accessed 4 August 2011).

Bivand, R., E. J. Pebesma, and V. Gómez-Rubio. 2008. *Applied spatial data analysis with R*. New York: Springer.

Bowman, A. W., and A. Azzalini. 2010. R package 'sm': Nonparametric smoothing methods. R package version 2.2–4. http://cran.r-project.org/web/packages/sm/index.html (last accessed 4 August 2011).

Burt, J. E., and G. M. Barber. 1996. *Elementary statistics for geographers*. New York: Guilford.

Central Intellience Agency (CIA). 2010. *The world factbook 2010*. Washington, DC: U.S. Central Intelligence Agency.

Chen, S., and M. Rura. 2008. The economic impact of administrative annexation in Jiangsu Province, China—Spatial filtering perspective. Paper presented at the annual meeting of the Association of American Geographers, Boston.

Chun, Y. 2008. Modeling network autocorrelation within migration flows by eigenvector spatial filtering. *Journal of Geographical Systems* 10:317–44.

Clifford, P., S. Richardson, and D. Hémon. 1989. Assessing the significance of the correlation between two spatial processes. *Biometrics* 45:123–34.

Cuaresma, J. C., and M. Feldkircher. 2010. Spatial filtering, model uncertainty and the speed of income convergence in Europe. Working Paper 160, Oesterreichische Nationalbank (Austrian Central Bank), Vienna, Austria.

De Marco, P., Jr., J. A. F. Diniz-Filho, and L. M. Bini. 2008. Spatial analysis improves species distribution modelling during range expansion. *Biology Letters* 4:577–80.

Devesa, S. S., D. J. Brauman, W. J. Blot, G. A. Pennello, R. Hover, and J. F. Fraumeni. 1999. *Atlas of cancer mortality in the United States: 1950–94*. Bethesda, MD: National Cancer Institute.

Diniz-Filho, J. A. F., and L. M. Bini. 2005. Modelling geographical patterns in species richness using eigenvector-based spatial filters. *Global Ecology and Biogeography* 17:177–85.

Diniz-Filho, J. A. F., L. M. Bini, and B. A. Hawkins. 2003. Spatial autocorrelation and red herrings in geographical ecology. *Global Ecology and Biogeography* 12:53–64.

Dormann, C. F. 2007. Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global Ecology and Biogeography* 16:129–38.

Dormann, C. F., J. M. McPherson, M. G. Araújo, R. Bivand, J. Bolliger, G. Carl, R. G. Davies, et al. 2007. Methods

to account for spatial autocorrelation in the analysis of species data: A review. *Ecography* 30:609–28.

Dray, S., P. Legendre, and P. Peres-Neto. 2006. Spatial modeling: A comprehensive framework for principal coordinate analysis of neighborhood matrices (PCNM). *Ecological Modeling* 196:483–93.

Dutilleul, P., B. Pelletier, and G. Alpargu. 2008. Modified F tests for assessing the multiple correlation between one spatial process and several others. *Journal of Statistical Planning and Inference* 138:1402–15.

Fabbri, D., and S. Robone. 2010. The geography of hospital admission in a national health service with patient choice. *Health Economics* 19:1029–47.

Ficetola, G. F., and E. Padoa-Schioppa. 2009. Human activities alter biogeographical patterns of reptiles on Mediterranean islands. *Global Ecology and Biogeography* 18:214–22.

Fischer, M. M., and D. A. Griffith. 2008. Modeling spatial autocorrelation in spatial interaction data: An application to patent citation data in the European Unions. *Journal of Regional Science* 48:969–89.

Gentleman, R. 2005. Reproducible research: A bioinformatics case study. *Statistical Applications in Genetics and Molecular Biology* 4:Article 1.

Getis, A. 1990 Screening for spatial dependence in regression analysis. *Papers of the Regional Science Association* 69:69–81.

———. 2010. Spatial filtering in a regression framework: Examples using data on urban crime, regional inequality, and government expenditures. In *Perspectives on spatial data analysis*, ed. L. Anselin and S. J. Rey, 191–202. Heidelberg, Germany: Springer.

Getis, A., and D. A. Griffith. 2002. Comparative spatial filtering in regression analysis. *Geographical Analysis* 34:130–40.

Griffith, D. A. 2000a. Eigenfunction properties and approximations of selected incidence matrices employed in spatial analyses. *Linear Algebra and Its Applications* 321:95–112.

———. 2000b. A linear regression solution to the spatial autocorrelation problem. *Journal of Geographical Systems* 2:141–56.

———. 2003. *Spatial autocorrelation and spatial filtering: Gaining understanding through theory and scientific visualization*. Berlin: Springer Verlag.

———. 2004. Distributional properties of georeferenced random variables based on the eigenfunction spatial filter. *Journal of Geographical Systems* 6:263–88.

———. 2010. Spatial filtering. In *Handbook of applied spatial analysis: Software tools, methods, and applications*, ed. M. M. Fischer and A. Getis, 301–18. Berlin, Germany: Springer Verlag.

Griffith, D. A., and G. Arbia. 2010. Detecting negative spatial autocorrelation in georeferenced random variables. *International Journal of Geographical Information Science* 24:417–37.

Griffith, D. A., and Y. Chun. 2009. Eigenvector selection with stepwise regression techniques to construct spatial filters. Paper presented at the annual meeting of the Association of American Geographers, Las Vegas, NV.

———. 2011. Modeling network autocorrelation in space–time migration flow data: An eigenvector spatial filtering approach. *Annals of the Association of American Geographers* 101 (3): 523–36.

Griffith, D. A., and L. R. Layne. 1999. *A casebook for spatial statistical data analysis*. New York: Oxford University Press.

Griffith, D. A., and P. R. Peres-Neto. 2006. Spatial modeling in ecology: The flexibility of eigenfunction spatial analyses. *Ecology* 87:2603–13.

Grimpe, C., and R. Patuelli. 2009. Regional knowledge production in nanomaterials: A spatial filtering approach. *The Annals of Regional Science* 46:519–41.

Institute Nacional de Estadística y Geographía (INEGI). 2010. Institute Nacional de Estadística y Geographía, Mexico City, Mexico. http://www.inegi.org.mx (last accessed 4 August 2011).

Jacob, B. G., E. J. Muturi, E. X. Caamano, J. T. Gunter, E. Mpanga, R. Ayine, J. Okelloonen, et al. 2008. Hydrological modeling of geophysical parameters of arboviral and protozoan disease vectors in internally displaced people camps in Gulu, Uganda. *International Journal of Health Geographics* 7:11.

Kühn, I. 2007. Incorporating spatial autocorrelation may invert observed patterns. *Diversity and Distributions* 13:66–69.

Le Gallo, J., and S. Dall'erba. 2008. Spatial and sectoral productivity convergence between European regions, 1975–2000. *Papers in Regional Science* 87:505–25.

Leisch, F., and A. Rossini. 2003. Reproducible statistical research. *Chance* 16:46–50.

Lewin-Koh, N. J., R. Bivand, E. J. Pebesma, E. Archer, A. Baddeley, H. Bibiko, S. Dray, et al. 2010. maptools: Tools for reading and handling spatial objects. R package version 0.7–38. http://cran.r-project.org/web/packages/maptools/index.html (last accessed 4 August 2011).

Mayor, M., and A. J. López. 2008. Spatial shift-share analysis versus spatial filtering: An application to Spanish employment data. *Empirical Economics* 34:123–42.

Möller, J., and R. Soltwedel. 2007. Recent developments of regional labor market analysis using spatial econometrics: Introduction. *International Regional Science Review* 30:95–99.

Montresor, E., F. Pecci, and N. Pontarollo. 2010. The evaluation of European structural funds on economic convergence with the application of spatial filtering technique. Working paper No. 07/2010. Università di Verona, Dipartimento di Scienze economiche, Verona, Italy. http://dse.univr.it//workingpapers/Montresor_Pecci_Pontarollo_Madeira.pdf (last accessed 4 August 2011).

Moran, P. A. P. 1948. The interpretation of statistical maps. *Journal of the Royal Statistical Society Series B* 10:245–51.

———. 1950. Notes on continuous stochastic phenomena. *Biometrika* 37:17–23.

Neuwirth, E. 2007. RColorBrewer: ColorBrewer palettes. R package version 1.0–2. http://cran.r-project.org/web/packages/RColorBrewer/index.html (last accessed 8 May 2012).

Ord, J. K., and A. Getis. 2001. Testing for local spatial autocorrelation in the presence of global autocorrelation. *Journal of Regional Science* 41:411–32.

Patuelli, R., D. A. Griffith, M. Tiefelsdorf, and P. Nijkamp. 2011. Spatial filtering and eigenvector stability:

Space–time models for German unemployment data. *International Regional Science Review* 34 (2): 253–380.

Pecci, F., and N. Pontarollo. 2010. The application of spatial filtering technique to the economic convergence of the European regions between 1995 and 2007. In *Computational science and its applications—ICCSA 2010, Lecture notes in computer science*, ed. D. Taniar, O. Gervasi, B. Murgante, E. Pardede, and B. Apduhan, 46–61. Berlin, Germany: Springer Verlag.

R Development Core Team. 2010. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.r-project.org (last accessed 4 August 2011).

Steiger, J. H. 1980. Tests for comparing elements of a correlation matrix. *Psychological Bulletin* 87:245–51.

Tiefelsdorf, M. 2007. Controlling for migration effects in ecological disease mapping of prostate cancer. *Stochastic Environmental Research and Risk Assessment* 21:615–24.

Tiefelsdorf, M., and D. A. Griffith. 2007. Semiparametric filtering of spatial autocorrelation: The eigenvector approach. *Environment and Planning A* 39:1193–1221.

Tiefelsdorf, M., D. A. Griffith, and B. Boots. 1999. A variance-stabilizing coding scheme for spatial link matrices. *Environment and Planning A* 31:165–80.

United Nations Development Programme. 2010. *Human development reports*. New York: United Nations Development Programme.

Upton, G., and B. Fingleton. 1985. *Spatial data analysis by example: Point pattern and quantitative data*. Chippenham, UK: Wiley.

Wakefield, J. C. 2003. Sensitivity analyses for ecological regression. *Biometrics* 59:9–17.

Williams, E. J. 1959. The comparison of regression variables. *Journal of the Royal Statistical Society* 21:396–99.

Zeileis, A. 2002. Diagnostic checking in regression relationships. *R News* 2 (3): 7–10. http://cran.r-project.org/web/packages/lmtest/index.html (last accessed 4 August 2011).

Correspondence: Department of Geography–Geology, Illinois State University, Normal, IL 61790, e-mail: jthayn@ilstu.edu (Thayn); jmsiman@ilstu.edu (Simanis).

Appendix

```
# This creates a function called plot.map() that will display values in the ESRI shapefile
plot.map <- function(theme, poly, color = NULL, ncl = 9,main = NULL){
require(RColorBrewer); require(maptools); require(classInt)
if(is.null(color)) color <- "YlGnBu"
int <- classIntervals(theme,n = ncl,style = "quantile")$brks
pal <- brewer.pal(ncl,color)
cols <- pal[findInterval(theme,int,rightmost.closed = T)]
plot(poly,col = cols)
title(main = main)
}
# opens the ESRI shapefile and prepares the data
library(spdep) #this opens the spatial dependency package
col.poly <- readShapePoly(system.file("etc/shapes/columbus.shp", package = "spdep")[1])
# the readShapePoly("") function opens polygon shapefiles, just type the file directory path between the quotation marks
col.data <- slot(col.poly, "data")
attach(col.data)
colnames(col.data)
plot.map(CRIME,col.poly,main="CRIME")
# creates the neighborhood and neighborhood weights objects
col.nb <- poly2nb(col.poly,queen = F)
col.listw <- nb2listw(col.nb,style = "C")
plot(col.poly,col = "gray",border = "white")
coords <- coordinates(col.poly)
plot(col.nb,coords,add = T)
# performs the ordinary least squares regression and view the results
lm.ols <- lm(CRIME~INC+HOVAL)
summary(lm.ols)
coef(lm.ols)
lm.ols.pred <- predict(lm.ols)
lm.ols.res <- residuals(lm.ols)
par(mfrow = c(2,2))
plot.map(CRIME,col.poly,main="CRIME")
plot.map(lm.ols.pred,col.poly,main = "Predicted Crime")
plot.map(lm.ols.res,col.poly,main = "Residuals")
plot(CRIME,lm.ols.pred,pch = 20)
abline(coef(lm(lm.ols.pred~CRIME)))
par(mfrow = c(1,1))
# calculates Moran's I for the data and performs a Shapiro-Wilks test for normality on the residuals of the model
moran.test(CRIME,col.listw)
moran.test(lm.ols.pred,col.listw)
moran.test(lm.ols.res,col.listw)
shapiro.test(lm.ols.res)
library(lmtest)
bptest(lm.ols)
# generates a series of hypothetical potential spatial patterns
n <- length(col.nb)
C <- listw2mat(col.listw)
M <- diag(1,n)-1/n
MCM <- M%*%C%*%M
E <- eigen(MCM)$vectors
# generates the spatial patterns tied to this model
X <- cbind(1,INC,HOVAL)
M <- diag(1,n)-tcrossprod(X%*%qr.solve(crossprod(X)),X)
cbind(M%*%CRIME,lm.ols.res)
MCM <- M%*%C%*%M
eig <- eigen(MCM)
E <- eig$vector
# calculates and plots the Moran's I values for the spatial patterns tied to this model
ones <- rep(1,n)
mi <- eig$values*n/crossprod(ones,C%*%ones)
plot(mi,ylim = c(-1,1),pch = 20,xlab = "Eigenvector Spatial Pattern",ylab = "Moran's I")
abline(0,0,lty = 3)
# displays the spatial patterns tied to this model
library(sm)
for(i in 1:n){plot.map(E[,i],col.poly,main = paste("EV",i,": Moran's I = ",round(mi[i],3))); pause()}
# creates the spatial filter – the function SpatialFiltering() generates MCM and calculates its eigenvectors and then uses a simultaneous autoregressive model to judiciously select a subset of the spatial patterns
```

```
67. sf.err <- SpatialFiltering(lm.ols,nb = col.nb,
    style = "C",alpha = 0.25,ExactEV = T)
68. E.sel <- fitted(sf.err)
69. dim(E.sel)
70. par(mfrow = c(2,2))
71. for(i in 1:4) plot.map(E.sel[,i], col.poly, main
    = colnames(E.sel)[i])
72. par(mfrow = c(1,1))
73. # performs the spatially filtered ordinary
    least squares regression and displays the re-
    sults
74. lm.sf <- lm(CRIME~INC+HOVAL+E.sel)
75. summary(lm.sf)
76. lm.sf.pred <- predict(lm.sf)
77. lm.sf.res <- residuals(lm.sf)
78. par(mfrow = c(2,2))
79. plot.map(CRIME,col.poly,main="CRIME")
80. plot.map(lm.sf.pred,col.poly,main="SF Pre-
    dicted Crime")
81. plot.map(lm.sf.res,col.poly,main="SF Resid-
    uals")
82. plot(CRIME,lm.sf.pred,pch = 20)
83. abline(coef(lm(lm.sf.pred~CRIME)))
84. par(mfrow = c(1,1))
85. # calculates Moran's I for the data and per-
    forms a Shapiro-Wilks test for normality on
    the residuals of the spatially filtered model
86. moran.test(CRIME,col.listw)
87. moran.test(lm.sf.pred,col.listw)
88. moran.test(lm.sf.res,col.listw)
89. shapiro.test(lm.sf.res)
90. bptest(lm.sf)
```