Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER Linear Algebra and its Applications 388 (2004) 201–219

LINEAR ALGEBRA AND ITS APPLICATIONS

www.elsevier.com/locate/laa

Extreme eigenfunctions of adjacency matrices for planar graphs employed in spatial analyses[☆]

Daniel A. Griffith

Department of Geography, 144 Eggers Hall, Syracuse University, Syracuse, NY 13244-1020, USA

Received 22 June 2001; accepted 27 December 2002

Submitted by H.J. Werner

Abstract

Mathematical properties of extreme eigenfunctions of popular geographic weights matrices used in spatial statistics are explored, and applications of these properties are presented. Three theorems are proposed and proved. These theorems pertain to the popular binary geographic weights matrix—an adjacency matrix—based upon a planar graph. They uncover relationships between the determinant of this matrix and its extreme eigenvalues, regression and the minimum eigenvalue of this matrix, and the eigenvectors of a row-standardized asymmetric version of this matrix and its symmetric similarity matrix counterpart. In addition, a conjecture is posited pertaining to estimation of the largest eigenvalue of the binary geographic weights matrix when the estimate obtained with the oldest and well-known method of matrix powering begins to oscillate between two trajectories in its convergence. An algorithm is outlined for calculating the extreme eigenvalues of geographic weights matrices based upon planar graphs. And, applications results for selected very large adjacency matrices are reported.

Keywords: Eigenfunction; Incidence matrix; Spatial analysis; Irregular tessellation; Stochastic matrix; Extreme eigenvalues; Jacobian

1. Introduction

Let G be an irreducible, undirected planar graph with $n \geqslant 2$ vertices. Let $\mathbf{C}$ be the binary 0-1 n-by-n adjacency matrix constructed from G, where $c_{ij} = 1$ if vertices i and j are adjacent, and $c_{ij} = 0$ otherwise. G is commonly employed in spatial

[☆] This research was supported by the National Science Foundation, research grant #BCS-9905213.

E-mail address: griffith@maxwell.syr.edu (D.A. Griffith).

0024-3795/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/S0024-3795(03)00368-9

analyses, where it is constructed for surface tessellations (e.g., pixels of a remotely sensed image, counties of a state). Its planar property means that $c_{ij} = 1$ if tessellation cells i and j share a nonzero length boundary, and $c_{ij} = 0$ otherwise. Especially in spatial statistics, matrix **C** may be converted to its row-standardized, stochastic version, say **W**; hence, $w_{ij} = c_{ij}/(\sum_{j=1}^{n} c_{ij})$. Suppose an attribute variable has n numerical values, where each of these values is associated with one and only one of the vertices of G. Spatial autoregressive model estimation involves one of the following matrix determinant Jacobian terms, say $e^{J(\rho)}$, based upon these matrices: $\det(\mathbf{I} - \rho\mathbf{C})$ and $\det(\mathbf{I} - \rho\mathbf{W})$, where **I** is an n-by-n identity matrix, and the scalar ρ is a parameter denoting the nature and degree of spatial autocorrelation, the correlation depicting tendencies for similar or dissimilar numerical values of a given attribute variable to cluster within a tessellation (i.e., on a map)—in other words, similar or dissimilar numerical values to be associated with adjacent nodes of G. Following Ord [25], the logarithmic counterpart to this Jacobian term may be rewritten as $J(\rho) = -(1/n)\sum_{i=1}^{n} \mathrm{LN}(1 - \rho\lambda_i)$, where λ_i are the i eigenvalues of matrix **C** or **W**, depending upon which matrix is employed in a spatial statistical analysis. Suppose these eigenvalues are arranged in descending order; hence λ_1 and λ_n respectively denote the largest and smallest eigenvalues. The objective of this paper is to give analytically derived eigenfunction results that allow $J(\rho)$ to be approximated by $\widehat{J}(\rho)$ for very large, unpatterned versions of matrices **C** and **W**. The need to address this problem of computing log-determinants of large sparse matrices is emphasized by, among others, Barry and Pace [3]. Advantages of approaches outlined in this paper over these others include an ability to accurately approximate asymptotic standard errors (see [17]), and avoidance of complications arising from selecting poor simulation samples. The current underdeveloped state of efficient eigenvalue algorithms, and a need to rectify this situation, is stressed by Luk and Qiao [23].

2. Deriving $\widehat{J}(\rho)$, an approximation for the log-determinant of selected matrices

The log-summation for $J(\rho)$ is easily calculated for values of n as large as several thousand; the upper limit of efficient and feasible calculation of sets of eigenfunctions has been increasing with the advancement of computer technology. This log-summation term also can be efficiently calculated for matrix **C** constructed for either the linear (i.e., the minimally connected G) or the regular square tessellation surface partitioning, whose eigenvalues are analytically known. Griffith [15, p. 102] presents an extremely accurate approximation for the eigenvalues of the corresponding matrix **W** for a regular square tessellation, and reports the analytical eigenvalues of matrix **W** for the linear arrangement. Theorem 2.1 extends the calculation simplification for $J(\rho)$ to the entire class of planar graphs G.

Theorem 2.1. *If* **M** *is an n-by-n irreducible adjacency matrix—either a binary* 0-1 *matrix or its row-standardized counterpart—based upon an undirected planar* *graph, and λ_1 and λ_n respectively are its extreme eigenvalues, then the affiliated spatial autoregressive log-Jacobian term, $J(\rho)$, may be approximated with $\widehat{J_1}(\rho)$, whose expression is given by*

$$\rho\frac{\alpha_{2,n}\lambda_n^2-\alpha_{1,n}\lambda_1^2}{2}+\alpha_{2,n}|\lambda_n^2|\mathrm{LN}\left(\frac{\delta_{2,n}}{|\lambda_n|}\right)+\alpha_{1,n}\lambda_1^2\mathrm{LN}\left(\frac{\delta_{1,n}}{\lambda_1}\right)$$
$$-\alpha_{2,n}\lambda_n^2\mathrm{LN}\left(\frac{\delta_{2,n}}{|\lambda_n|}+\rho\right)-\alpha_{1,n}\lambda_1^2\mathrm{LN}\left(\frac{\delta_{1,n}}{\lambda_1}-\rho\right), \tag{1}$$

where the coefficients $\alpha_{1,n}$ and $\alpha_{2,n}$ are half of the average distance between consecutive eigenvalues in, respectively, the negative and the positive ranges of a set of n eigenvalues, and coefficients $\delta_{1,n}$ and $\delta_{2,n}$ help compensate for use of a truncated log-series expansion.

Proof

$$\begin{aligned}J(\rho)&=-\frac{1}{n}\sum_{i=1}^{n}\mathrm{LN}(1-\rho\lambda_i)\\&=-\frac{1}{n}\left[\sum_{\lambda_i<0}\mathrm{LN}(1-\rho\lambda_i)+0+\sum_{\lambda_i>0}\mathrm{LN}(1-\rho\lambda_i)\right].\end{aligned}$$

Since the eigenvalues are not necessarily uniformly spaced across their range, guided by both the trapezoidal rule and the Gaussian formula from calculus,

$$\begin{aligned}\sum_{\lambda_i<0}\mathrm{LN}(1-\rho\lambda_i)&\approx\alpha_1\int_{\lambda_n}^{0}\mathrm{LN}(1-\rho\lambda)\mathrm{d}\lambda\\&=-\alpha_1\frac{-\mathrm{LN}(1+\rho|\lambda_n|)(1+\rho|\lambda_n|)+\rho|\lambda_n|}{\rho}\end{aligned}$$

and

$$\begin{aligned}\sum_{\lambda_i>0}\mathrm{LN}(1-\rho\lambda_i)&\approx\alpha_2\int_{0}^{\lambda_n}\mathrm{LN}(1-\rho\lambda)\mathrm{d}\lambda\\&=-\alpha_2\frac{-\mathrm{LN}(1-\rho\lambda_1)(1-\rho\lambda_1)-\rho\lambda_1}{\rho}.\end{aligned}$$

Substituting these two results into the equation for $J(\rho)$ after replacing $(\mathrm{LN}(1-\rho\lambda))/\rho$ with the first two terms of its log-series expansion, namely $(-\rho\lambda-\rho^2\lambda^2/2)/\rho$, yields expression (1), where the coefficients $\alpha_{1,n}$ and $\alpha_{2,n}$ are subscripted with n because, respectively, they are computed from α_1/n and α_2/n, and 1 is replaced by $\delta_{1,n}$ and $\delta_{2,n}$. □

If $\alpha_{1,n}$ and $\alpha_{2,n}$ respectively have $1/|\lambda_n|$ and $1/\lambda_1$ factored out of them, expression (1) becomes the approximation expression reported and evaluated by Griffith and Sone [20].

Three sources of error are affiliated with Eq. (1). The first arises from using the trapezoid rule of integration, and is of order $\mathrm{O}(\lambda_k^3/n_k^2)$, where λ_k is either λ_1 or λ_n, and n_k denotes the number of positive ($\lambda_k = \lambda_1$) or negative ($\lambda_k = \lambda_n$) eigenvalues. This quantity is rather modest because λ_1 at most tends to be proportional to $\sqrt{n}$ (see Eq. (3), for example). The second arises from truncating a series expansion of a logarithm expression, and is of order $\mathrm{O}(\rho^3\lambda_k^3)$. This quantity is rather modest for most values of ρ, since the autocorrelation parameter value is restricted such that $1/\lambda_n < \rho < 1/\lambda_1$. The third source of error arises from the nonuniform spacing of the eigenvalues, which diminishes with increasing n because the eigenvalue range is divided into increasingly smaller intervals—n increases faster than the interval $[\lambda_n, \lambda_1]$, which in some cases is constant over n. A database containing 130 graph adjacency matrices (see Appendix A for a description of it) has been complied and used here to both evaluate mathematical specifications and supply model-based inferential support for generalizing findings to other graphs not contained in the database. Calculations obtained with Eq. (1) suggest that the combination of these three sources of error results in modest total error in practice; based on dividing the feasible spatial autocorrelation parameter space into 20 intervals, the relative error sum of squares (RESS; the residual sum of squares divided by the corrected total sum of squares) has a mean of 0.00057, a standard deviation of 0.00096, a minimum of 0.00000, and a maximum of 0.00485. Roughly 44% of the RESS is accounted for by the following trend in these data:

$$\mathrm{RESS} \approx \frac{0.00268}{1 + 8.65638e^{-0.00485n}}.$$

This trend suggests that the total error should be negligible across n.

The log-summation version of $J(\rho)$ focuses attention on both the eigenfunctions associated with G and the importance of the pairs of extreme eigenvalues. In order to calibrate expression (1), then, these extreme eigenvalues need to be computable.

3. The principal eigenvalues of matrices C and W

Because matrix $\mathbf{W}$ is stochastic, its largest eigenvalue, say $\lambda_1(\mathbf{W})$, theoretically is known to be 1 (all rows sum to 1). The largest eigenvalue of matrix $\mathbf{C}$, say $\lambda_1(\mathbf{C})$, is easily calculated by using one of the oldest and the well-known method of $\lim_{k\to\infty}(\mathbf{1}'\mathbf{C}^{k+1}\mathbf{1})/(\mathbf{1}'\mathbf{C}^{k}\mathbf{1}) = \lambda_1$ for matrix $\mathbf{C}$ [7, p. 213]; this Rayleigh quotient usually converges to $\lambda_1(\mathbf{C})$ as $k \to \infty$. Because matrix $\mathbf{C}$ is sparse—since G is planar, the maximum number of 1s is $6(n-2)$—the calculation of powers of matrix $\mathbf{C}$ can be expedited by restricting attention to only those $c_{ij} = 1$; Anselin and Smirnov [2] discuss efficient procedures for constructing these types of powered matrices, too.

And, the irreducibility of G can be checked quite easily by tracing a path through the graph with a numerical algorithm to see whether or not it passes through all nodes. For purposes of this paper, this approach has allowed a quick calculation of $\lambda_1(\mathbf{C})$ for n as large as 45,974.

Friedman [14] discusses error bounds for the calculation of $\lambda_1(\mathbf{C})$. Another approach to accuracy assessment is based upon the trajectory traced by $\lambda_1(\mathbf{C})_\tau$, the estimate of $\lambda_1(\mathbf{C})$ at iteration τ, which often asymptotically approaches $\lambda_1(\mathbf{C})$ from below. Inspection of a number of these trajectories has suggested the following conjecture.

Conjecture 3.1. *If* $\mathbf{C}$ *is an n-by-n irreducible binary* 0-1 *adjacency matrix based upon an undirected planar graph, and* λ_1 *is computed with* $\lim_{\tau\to\infty}(\mathbf{1}'\mathbf{C}^{\tau+1}\mathbf{1})/(\mathbf{1}'\mathbf{C}^{\tau}\mathbf{1})$, *then*

$$\lambda_1(\mathbf{C})_\tau \approx \widehat{\lambda}_1(\mathbf{C}) + \widehat{\alpha} I_{\text{even/odd}} + \widehat{\beta} e^{-\widehat{\gamma}\tau}, \tag{2}$$

where $I_{\text{even/odd}} = 1$ *if* τ *is even*, *and* -1 *otherwise*.

The ideal trajectory sketched by $\lambda_1(\mathbf{C})_\tau$ is a concave curve that converges upon asymptote $\widehat{\lambda}_1(\mathbf{C})$. If $\mathbf{C}$ is a periodic matrix, then the trajectory $\lambda_1(\mathbf{C})_\tau$ oscillates between two concave curves, one traced by even powers and the other traced by odd powers of matrix $\mathbf{C}$. This trajectory pair is detected by $\widehat{\alpha} \neq 0$, which renders $\widehat{\lambda}_1(\mathbf{C})$ as a weighted average of the two trajectories; $\widehat{\alpha} \approx 0$ when a single trajectory exists. Because all trajectories converge upon asymptotes, $\widehat{\beta} e^{-\widehat{\gamma}\tau}$ describes this convergence; this specification of error disappearance is suggested by the matrix exponentiation involved in constructing the trajectories.

In practice, the first $\tau = 1, 2, \ldots, L$ of $\lambda_1(\mathbf{C})_\tau$ need to be discarded in order for the intercept estimate, $\widehat{\lambda}_1(\mathbf{C})$, to be approximately equal to $\lambda_1(\mathbf{C})$; this action is similar to that taken with simulation work, where a simulated process requires a burn-in time. Experience suggests that L should be at least $0.10\tau_{\max}$, where $\tau_{\max}$ denotes the maximum number of iterations performed when convergence is attained. The parameter α accounts for the possible presence of both an upper and a lower bound trajectory for $\lambda_1(\mathbf{C})$, rather than the trajectory itself that converges on $\lambda_1(\mathbf{C})$, a situation that arises when adjacency matrices constructed for G are periodic. If these two separate trajectories do not exist, then $\alpha = 0$; if they do, then $\widehat{\lambda}_1(\mathbf{C})$ is an average of the asymptotic upper and lower bound values, and may only be approximately equal to $\lambda_1(\mathbf{C})$ in the limit. For example, the triangular–hexagonal Archimedean tiling of a plane (e.g., tiling 3.6.3.6 in the notation of Ahuja and Schachter [1, p. 7]), for $n = 36$, has $\lambda_1(\mathbf{C}) = 3.75130$, $\widehat{\lambda}_1(\mathbf{C}) = 3.75706$, $\widehat{\alpha} = -0.20813$, $\widehat{\beta} = -0.17181$, $\widehat{\gamma} = 0.48740$, $L = 7$ (roughly 10% of the $\tau_{\max} = 74$ iterations resulting from using a convergence criterion of sequential estimate change less than 10^{-12}), and asymptotic upper and lower bounds of 3.96519 and 3.54893.

A statistical description of $\lambda_1(\mathbf{C})$ has been obtained from the database containing 130 matrices. This description exploits the following three known properties of the principal eigenvalue:

(1) $\lambda_1(\mathbf{C}) \leqslant \mathrm{MAX}\left(\sum_{j=1}^{n} c_{ij}\right)$, the maximum row sum of the binary adjacency matrix **C**, an upper bound [5];

(2) $$\sqrt{\frac{\sum_{i=1}^{n}\left(\sum\limits_{j=1}^{n} c_{ij}\right)^2}{n}} \leqslant \lambda_1(\mathbf{C}),$$

a lower bound [9]; and,

(3) $$\lambda_1(\mathbf{C}) = \sqrt[p]{\frac{\sum_{i=1}^{n}\left(\sum\limits_{j=1}^{n} c_{ij}\right)^p}{n}},$$

for some real number p [21], a central tendency measure when p is estimated by pooling a collection of surface partitionings.

A value of p for this third property was calculated for each of the 130 matrices in the database. The average value is approximately 4; estimation of p for the 130 matrices simultaneously also yielded a value for p of roughly 4. A boxplot of the 130 values suggests the presence of a half-dozen outliers, all values of p between 6 and 7. These six graphs display no conspicuous differences from the remaining 124. The remaining values of p are symmetrically distributed, conforming closely to a bell-shaped curve, and are contained in the interval (2.50, 5.75). The Box–Cox transformation $(p + 3.25)^{-1.50}$, applied to the full set of p values, conforms very well to a normal distribution.

A very good statistical description of the principal eigenvalue, obtained with linear regression, and casting $\lambda_1(\mathbf{C})$ as a weighted average of its bounds and the Hofmeister-based measure of central tendency, is given by

$$\widehat{\lambda}_1(\mathbf{C}) = -0.94606 + 0.54806\sqrt{\frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{n} c_{ij}\right)^2}{n}} + 1.32521\sqrt{\sqrt[4]{\frac{\sum_{i=1}^{n}\left(\sum\limits_{j=1}^{n} c_{ij}\right)^4}{n}}} + 0.05962\,\mathrm{MAX}\left(\sum_{j=1}^{n} c_{ij}\right). \quad (3)$$

Diagnostic statistics for Eq. (3) include 97.04% of the variance in $\lambda_1(\mathbf{C})$ being accounted for, implying a very good fit (which is confirmed by a $\lambda_1(\mathbf{C})$-versus-$\widehat{\lambda}_1(\mathbf{C})$ plot), a Shapiro–Wilk statistic of 0.99013 for the regression residuals, implying very close conformity with a bell-shaped curve, and a residuals-versus-predicted plot that exhibits no apparent variance heterogeneity. In sum, the rather small residual values associated with Eq. (3) are statistically well behaved, furnishing sound model-based inferential support for generalizing its results to the wider population of graphs used in empirical spatial statistical analyses. Furthermore, 34 of the $\widehat{\lambda}_1(\mathbf{C})$ values deviate from their corresponding $\lambda_1(\mathbf{C})$ values by less than 1%, 81 deviate by between 1% and 5%, and 34 deviate by between 6% and 9%. The only apparent graph type feature detectable here is that Archimedean tilings other than the square and hexagon do not fall into the "less than 1% deviation" category. Finally, based upon a stepwise regression procedure, the central tendency variable,

$$\sqrt{\sqrt[4]{\frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{n} c_{ij}\right)^4}{n}}},$$

accounts for 91.87% of the variance in $\lambda_1(\mathbf{C})$, the upper bound accounts for 3.76% of this variance, and the lower bound accounts for 1.41% of this variance. Of note is that for the case of $n = 2$ (a graph not in the database), namely

$$\mathbf{C} = \mathbf{W} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

for which $\lambda_1(\mathbf{C}) = 1$ and $\lambda_n(\mathbf{C}) = -1$ by inspection,[1] Eq. (3) predicts a value of 0.98683. And for the previously mentioned case of the triangular–hexagonal–square Archimedean tiling, Eq. (3) predicts a value of 3.99660 (a 6.1% error).

4. The smallest eigenvalues of matrices C and W

In large part, spatial statistics is concerned with the spatial autocorrelation latent in attribute variables distributed across geographic space; most often, similar values tend to cluster on a map, a feature labeled spatial autocorrelation. Spatial autocorrelation is characterized in terms of either matrix **C** or **W**; these matrices capture the arrangement of numerical values on a map. One index of spatial autocorrelation is the Moran Coefficient, which looks and behaves very much like a Pearson product–moment correlation coefficient. But its extreme values are not ±1; rather, de Jong et al. [11] show that the Moran Coefficient extremes are determined by a pair of extreme eigenvalues. The matrix from which these extreme eigenvalues are extracted is the following modified version of binary matrix **C**:

$$(\mathbf{I} - \mathbf{11}'/n)\mathbf{C}(\mathbf{I} - \mathbf{11}'/n), \tag{4}$$

where **1** is an n-by-1 vector of ones; this matrix expression appears in the numerator of the Moran Coefficient. Tiefelsdorf and Boots [26] show that all of the eigenvalues of matrix expression (4) relate to specific Moran Coefficients. Griffith [16] shows that the corresponding eigenvectors relate to distinct types of numerical map

[1] The maximum eigenvalue is contained in the interval defined by the minimum and maximum row sums, and the sum of the eigenvalues equals the trace of the matrix.

patterns. Hence, the eigenvector associated with the principal eigenvalue of expression (4) depicts that geographic distribution of numerical values having the highest level of positive spatial autocorrelation that is possible. Similarly, the eigenvector associated with the smallest eigenvalue of expression (4) depicts that geographic distribution of values having the highest level of negative spatial autocorrelation that is possible. Finally, Griffith and Amrhein [18] show that a Moran Coefficient can be computed for some attribute variable Y by regressing vector $\mathbf{C}(\mathbf{I}-\mathbf{11}'/n)\mathbf{Y}$ on vector $(\mathbf{I}-\mathbf{11}'/n)\mathbf{Y}$.

This link between the Moran Coefficient and regression indicates that regressing some eigenvector $\mathbf{E}_k$ of expression (4) on vector $\mathbf{C}(\mathbf{I}-\mathbf{11}'/n)\mathbf{E}_k$ should display appealing properties. Accordingly,

Theorem 4.1. *If* $\mathbf{E}_k$ *is an eigenvector of an n-by-n real symmetric matrix* $\mathbf{M}$, *and its corresponding eigenvalue is* $\lambda_k \neq 0$, *then the simple linear regression ordinary least squares estimates of the intercept and slope coefficients obtained by regressing* $\mathbf{E}_k$ *on* $\mathbf{ME}_k$ *respectively are* 0 *and* $1/\lambda_k$.

Proof

$$\begin{aligned}
\mathbf{b} &= \left(<\mathbf{1} \vdots \mathbf{ME}_k>'<\mathbf{1} \vdots \mathbf{ME}_k>\right)^{-1} <\mathbf{1} \vdots \mathbf{ME}_k> \mathbf{E}_k \\
&= \begin{pmatrix} n & \mathbf{1}'\mathbf{ME}_k \\ \mathbf{E}_k\mathbf{M1} & \mathbf{E}_k'\mathbf{M}'\mathbf{ME}_k \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{1}'\mathbf{E}_k \\ \mathbf{E}_k'\mathbf{ME}_k \end{pmatrix} \\
&= \left(n\lambda_k^2 - \lambda_k^2\mathbf{E}_k'\mathbf{11}'\mathbf{E}_k\right)^{-1} \begin{pmatrix} \lambda_k^2 & -\lambda_k\mathbf{1}'\mathbf{E}_k \\ -\lambda_k\mathbf{E}_k\mathbf{1} & n \end{pmatrix} \begin{pmatrix} \mathbf{1}'\mathbf{E}_k \\ \lambda_k \end{pmatrix} \\
&= \left(n\lambda_k^2 - \lambda_k^2\mathbf{E}_k'\mathbf{11}'\mathbf{E}_k\right)^{-1} \begin{pmatrix} 0 \\ -\lambda_k\mathbf{E}'\mathbf{11}'\mathbf{E} + n\lambda_k \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{1}{\lambda_k} \end{pmatrix}. \qquad \square
\end{aligned}$$

Besides knowing the regression equation coefficients analytically, establishing the goodness-of-fit of the associated regression line also is desirable. Hence,

Theorem 4.2. *If* $\mathbf{E}_k$ *is an eigenvector of an n-by-n real symmetric matrix* $\mathbf{M}$, *and its corresponding eigenvalue is* $\lambda_k \neq 0$, *then the simple linear regression equation obtained by regressing* $\mathbf{E}_k$ *on* $\mathbf{ME}_k$ *has a sum of squared errors* (SSE) *term equal to* 0.

Proof

$$\begin{aligned}
\mathrm{SSE} = \mathbf{E}_k'\mathbf{E}_k - b' <\mathbf{1} \vdots \mathbf{ME}_k> \mathbf{E}_k &= 1 - \begin{pmatrix} 0 & \frac{1}{\lambda_k} \end{pmatrix} \begin{pmatrix} \mathbf{1}'\mathbf{E}_k \\ \lambda_k \end{pmatrix} \\
&= 1 - 1 = 0. \qquad \square
\end{aligned}$$

These two results allow efficient algorithms to be developed, ones that quickly converge upon minimum eigenvalues.

4.1. The smallest eigenvalue of matrix **C**

Griffith [15] shows that, for the regular square tessellation, as n goes to infinity, ordered pairings of eigenvalues of matrix **C** and expression (4) converge, once $\lambda_1(\mathbf{C})$ is replaced with 0 and $\mathbf{E}_1(\mathbf{C})$ is replaced with $(1/\sqrt{n})\mathbf{1}$—this modification is attributable to the multiplicative presence of the projection matrix $(\mathbf{I}-\mathbf{1}\mathbf{1}'/n)$ in expression (4). Because the matrix expression (4) is symmetric, its eigenvectors are orthogonal. This property can be fruitfully combined with the two presented in Theorems 4.1 and 4.2, as well as the negative spatial autocorrelation result that the eigenvector associated with the smallest eigenvalue of expression (4) will be such that $(\mathbf{I}-\mathbf{1}\mathbf{1}'/n)\mathbf{E}_k \equiv \mathbf{C}(\mathbf{I}-\mathbf{1}\mathbf{1}'/n)\mathbf{E}_k$. Therefore, $\lambda_n(\mathbf{C})$ can be approximated iteratively as follows:

Phase 1: initialize the eigenvector corresponding to the minimum eigenvalue.

Step 1: compute $\mathbf{E}_1$ from the results of $\lim_{\tau\to\infty}(\mathbf{1}'\mathbf{C}^{\tau+1}\mathbf{1})/(\mathbf{1}'\mathbf{C}^{\tau}\mathbf{1})$; the normalized vector $\mathbf{C}^{\tau_{\max}}\mathbf{1}$ converges on $\mathbf{E}_1$,

Step 2: let $\mathbf{E}_{n,\tau=0} = \mathbf{E}_1$.

Phase 2: sequentially move $\mathbf{E}_{n,\tau=0}$ toward maximum negative spatial autocorrelation.

Step 1: let $e_{i,n,\tau=1} = -\sum_{j=1}^{n} c_{ij}\, e_{j,n,\tau=1}$,

$$\text{if} \left| -\sum_{j=1}^{n} c_{ij}e_{j,n,\tau=1} - \sum_{j=1}^{n} c_{ij}e_{j,n,\tau=1} \right| + \sum_{j=1}^{n} c_{ij} \left| e_{j,n,\tau=1} - \left[\sum_{k=1}^{n} c_{jk}e_{k,n,\tau=1} - e_{i,n,\tau=1} + \left(-\sum_{j=1}^{n} c_{ij}e_{j,n,\tau=1} \right) \right] \right| < \left| e_{i,n,\tau=1} - \sum_{j=1}^{n} c_{ij}e_{j,n,\tau=1} \right| + \left| \sum_{j=1}^{n} c_{ij} \left(e_{j,n,\tau=1} - \sum_{k=1}^{n} c_{jk}e_{k,n,\tau=1} \right) \right|, \quad i = 1, 2, \ldots, n,$$

Step 2: center the vector by subtracting its mean from it:

$$e_{i,n,\tau=2} = e_{i,n,\tau=1} - \frac{\sum_{i=1}^{n} e_{i,n,\tau=1}}{n}, \quad i = 1, 2, \ldots, n,$$

Step 3: normalize the vector by dividing it by the square-root of its sum of squares:

$$e_{i,n,\tau=3} = \frac{e_{i,n,\tau=2}}{\sqrt{\sum_{i=1}^{n} e_{i,n,\tau=2}^2}}, \quad i = 1, 2, \ldots, n,$$

Step 4: Repeat Steps 1–3 until the SSE from regressing $\mathbf{E}_{n,\tau}$ on $\mathbf{CE}_{n,\tau}$ stops decreasing.
Step 5: let $\lambda_{n,r} = \mathbf{E}'_{n,\tau}\mathbf{CE}_{n,\tau}/\mathbf{E}'_{n,\tau}\mathbf{E}_{n,\tau}$.

Phase 3: update eigenvector values with the regression coefficients—where a denotes the intercept term and b denotes the slope coefficient—obtained with a simple linear regression of $\mathbf{E}_{n,\tau}$ on $\mathbf{CE}_{n,\tau}$.
Step 1: let $e_{i,n,\tau+1} = a + b\sum_{j=1}^{n} c_{ij}e_{j,n,\tau}$,

$$\begin{aligned}\text{if}\ &\left|\left(a + b\sum_{j=1}^{n} c_{ij}e_{j,n,\tau}\right) - \sum_{j=1}^{n} c_{ij}e_{j,n,\tau}\right|\\ &+ \sum_{j=1}^{n} c_{ij}\left|e_{j,n,\tau} - \left[\sum_{k=1}^{n} c_{jk}e_{k,n,\tau} - e_{i,n,\tau} + \left(a + b\sum_{j=1}^{n} c_{ij}e_{j,n,\tau}\right)\right]\right|\\ &< \left|e_{i,n,\tau} - \sum_{j=1}^{n} c_{ij}e_{j,n,\tau}\right| + \left|\sum_{j=1}^{n} c_{ij}\left(e_{j,n,\tau} - \sum_{k=1}^{n} c_{jk}e_{k,n,\tau}\right)\right|,\\ &\quad i = 1, 2, \ldots, n,\end{aligned}$$

Step 2: center the vector by subtracting its mean from it (see Phase 2, Step 2),
Step 3: normalize the vector (see Phase 2, Step 3),
Step 4: let $\lambda_{n,\tau+1} = \mathbf{E}'_{n,\tau+1}\mathbf{CE}_{n,\tau+1}/\mathbf{E}'_{n,\tau+1}\mathbf{E}_{n,\tau+1}$,
Step 5: repeat Steps 1-4 until $\left|\lambda_{n,\tau+1} - \lambda_{n,\tau}\right| < c$, where c is a very small constant,[2] or the SSE from regressing $\mathbf{E}_{n,\tau}$ on $\mathbf{CE}_{n,\tau}$ stops decreasing.

Phase 4: fine tune the eigenfunction estimation.
Step 1: obtain estimates of coefficients a and b by regressing $\mathbf{E}_{n,\tau}$ on $\mathbf{CE}_{n,\tau}$,
Step 2: let $\mathbf{E}_{n,\tau+1} = a + b\mathbf{CE}_{n,\tau}$,
Step 3: center the vector by subtracting its mean from it (see Phase 2, Step 2),
Step 4: normalize the vector (see Phase 2, Step 3),
Step 5: repeat Steps 1–4 until the SSE $< 10^{-15}$,
Step 6: let $\widehat{\lambda}_n = \mathbf{E}'_{n,\tau+1}\mathbf{CE}_{n,\tau+1}/\mathbf{E}'_{n,\tau+1}\mathbf{E}_{n,\tau+1}$.

The resulting $\widehat{\lambda}_n$ estimates the minimum eigenvalue of matrix expression (4), say $\lambda_n^*(\mathbf{C})$, and serves as a good estimate for the minimum eigenvalue of matrix $\mathbf{C}$, as it is a tight upper bound for $\lambda_n(\mathbf{C})$. In fact, regressing $\lambda_n(\mathbf{C})$ of $\widehat{\lambda}_n$ yields

$$\lambda_n(\mathbf{C}) = -0.03861 + 0.92207\,\widehat{\lambda}_n + e,$$

[2] A reasonable value for c appears to be 10^{-12}, while a maximum value of 10^{-6} allows a reasonably accurate but very quick calculation.

where e is the regression residual error term, and $R^2 = 0.990$. This equation has only a few $\widehat{\lambda}_n$ values that are conspicuously less than their $\lambda_n(\mathbf{C})$ counterparts; the worst is for the $n = 39$ Archimedean tiling comprising hexagons, squares and dodecagons and labeled 4.6.12 in the notation of Ahuja and Schachter [1, p. 7]. And, the intercept and slope estimates respectively are consistent with statistical null hypothesis values of 0 and 1. Mäkeläinen [24] gives the necessary and sufficient conditions for these two quantities to be equal.

A statistical description of $\lambda_n(\mathbf{C})$ also has been obtained from the previously mentioned database containing 130 matrices. This description exploits the following three known properties of the minimum eigenvalue:

(1) $\lambda_n(\mathbf{C}) \leqslant -\frac{1}{2} - \frac{1}{2}\sqrt{1 + 4(n-3)/(n-1)}$, an upper bound [28];
(2) $-\lambda_1(\mathbf{C}) \leqslant \lambda_n(\mathbf{C})$, a lower bound [5]; and,
(3) $\lambda_1(\mathbf{C})/(1 - (1/e_{1,\max})) \leqslant \lambda_n(\mathbf{C})$, after some algebraic manipulations of a result reported by Fiol [13], where $e_{1,\max}$ is the maximum value of the normalized principal eigenvector, a positive quantity by the Perron–Frobinius theorem.

A reasonably good statistical description of the minimum eigenvalue of matrix $\mathbf{C}$, obtained with nonlinear regression, and casting $\lambda_n(\mathbf{C})$ as a weighted average of its bounds and the Fiol-based measure, is given by

$$\begin{aligned}\widehat{\lambda}_n(\mathbf{C}) = {} & 20.50694 - \frac{0.19973}{5.7 - e^{-1.05\left(-0.5-0.5\sqrt{1+4\frac{n-3}{n-1}}\right)}} \\ & + 0.03992[\lambda_1(\mathbf{C}) - 1]^2 - 0.28036\frac{\lambda_1(\mathbf{C})}{1 - \frac{1}{e_{1,\max}}} \\ & - 6.55956\,\mathrm{LN}\left[\mathrm{MAX}\left(\sum_{j=1}^{n} c_{ij}\right) + 29\right]. \qquad (5)\end{aligned}$$

Diagnostic statistics for Eq. (5) include 84.54% of the variance in $\lambda_n(\mathbf{C})$ being accounted for, implying a respectably good fit (which is confirmed by a $\lambda_n(\mathbf{C})$-versus-$\widehat{\lambda}_n(\mathbf{C})$ plot), a Shapiro–Wilk statistic of 0.94157 for the regression residuals, implying a failure for conformity with a bell-shaped curve, and a residuals-versus-predicted plot that is less than ideal. Eq. (5) furnishes model-based inferential support weaker than that associated with Eq. (3) for generalizing its results to the wider population of graphs used in empirical spatial statistical analyses. In addition, the transformed version of variable $\mathrm{MAX}\left(\sum_{j=1}^{n} c_{ij}\right)$ accounts for 73.46% of the variance in $\lambda_n(\mathbf{C})$, the transformed version of the upper bound term,

$$e^{\frac{1}{2}+\frac{1}{2}\sqrt{1+4\frac{n-3}{n-1}}},$$

accounts for 7.30% of this variance, the lower bound term, $[\lambda_1(\mathbf{C}) - 1]^2$, accounts for 2.64% of this variance, and $\lambda_1(\mathbf{C})/(1 - (1/e_{1,\max}))$ accounts for 1.14% of this variance. Of note is that for the case of a minimally connected G with $n = 4$ (a graph not contained in the database), where

$$\mathbf{C} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

for which $\lambda_n(\mathbf{C}) = 2\,\mathrm{COS}(4\pi/5) = -1.61803$ and $e_{1,\max} = (\sqrt{2/5})\,\mathrm{SIN}(2\pi/5) = 0.60150$ are known analytically [4], Eq. (5) predicts a value of -1.84957, which contains a 14.3% error. Thus, while Eq. (5) helps identify important descriptors of the variation in $\lambda_n(\mathbf{C})$, it should not be considered a reliable predictor of this quantity. Of note is that the preceding algorithm iteratively renders $\hat{\lambda}_n^*(\mathbf{C}) = -1.61803$, and that the lower bound of $-\sqrt{2n-4}$ presented by Hong and Shu [22], which needed to be transformed to $-\mathrm{LN}(2n-13)$ to optimize its linear relationship with the $\lambda_n(\mathbf{C})$ in the database, failed to improve the statistical description once the other terms were entered into Eq. (5).

As an aside, the maximum value of the normalized principal eigenvector, $e_{1,\max}$, tends to be exactly estimated by the maximum value, $\widehat{e}_{1,\max}$, of the normalized vector $\mathbf{C}^k\mathbf{1}$, where k denotes the power of matrix $\mathbf{C}$ for which convergence of the Rayleigh quotient $(\mathbf{1}'\mathbf{C}^{\tau+1}\mathbf{1})/(\mathbf{1}'\mathbf{C}^{\tau}\mathbf{1})$ has occurred. When $\mathbf{C}$ is a periodic matrix, then this value can be well-approximated by the maximum value of vector

$$\frac{\left[\frac{\mathbf{C}^k\mathbf{1}}{\sqrt{\mathbf{1}'\mathbf{C}^k\mathbf{1}}} + \frac{\mathbf{C}^{k+1}\mathbf{1}}{\sqrt{\mathbf{1}'\mathbf{C}^{k+1}\mathbf{1}}}\right]}{2},$$

using the same principle as mentioned before for estimating $\lambda_1(\mathbf{C})$ for a periodic matrix $\mathbf{C}$. The two cases in the database for which matrix $\mathbf{C}$ is periodic, namely the Archimedean tilings labeled 3.4.6.4 (triangles, squares and hexagons, with $n = 39$) and 3.6.3.6 (triangles and hexagons, with $n = 36$) in the notation of Ahuja and Schachter [1, p. 7], respectively have $\widehat{e}_{1,\max} = 0.32062$ for $e_{1,\max} = 0.32222$, and $\widehat{e}_{1,\max} = 0.36604$ for $e_{1,\max} = 0.37186$. Nevertheless, the intercept and slope regression coefficient estimates obtained by regressing $e_{1,\max}$ on $\widehat{e}_{1,\max}$ respectively are statistically indistinguishable from 0 and 1.

4.2. *The smallest eigenvalue of matrix* W

While $\lambda_1(\mathbf{W}) \equiv 1$ for all planar tessellations (all rows sum to 1), and $\lambda_n(\mathbf{W})$ almost always lies in the interval $[-1, -0.5]$, except for the regular square tessellation, for which $\lambda_n(\mathbf{W}) = -1$, this quantity is unknown. But while matrix $\mathbf{W}$ virtually always is asymmetric, it has a symmetric counterpart that is an algebraically similar matrix. A special case of Theorem 2.3 in Griffith [15] identifies a property of this similarity matrix that can be used to modify the preceding algorithm in order to estimate $\lambda_n(\mathbf{W})$, namely that the normalized version of vector $\mathbf{D}^{1/2}\mathbf{1}$ is the principal eigenvector of matrix $\mathbf{W}$. Exploiting this result to modify the previous minimum eigenvalue estimation algorithm enables it iteratively to converge on $\lambda_n(\mathbf{W})$ rather than $\lambda_n(\mathbf{C})$.

Two modifications are necessary in order to convert the preceding algorithm to one that estimates $\lambda_n(\mathbf{W})$. First, the $c_{ij} = 1$ entries in matrix $\mathbf{C}$ need to be replaced by $w_{ij} = 1\Big/\left(\sqrt{\left(\sum_{i=1}^{n} c_{ij}\right)\left(\sum_{j=1}^{n} c_{ij}\right)}\right)$. Second, in computing $\widehat{\lambda}_n(\mathbf{W})$, the estimated eigenvector undergoes centering in order to have a mean of zero (i.e., Phase 2, Step 2; Phase 3, Step 2; and Phase 4, Step 3). This step needs to be replaced with one that orthogonalizes $\mathbf{E}_{n,\tau}$, as follows:

$$\text{(1) compute } o = \mathbf{E}'_{n,\tau}\frac{\mathbf{D}^{1/2}\mathbf{1}}{\sqrt{\mathbf{1}'\mathbf{D}^{1/2}\mathbf{1}}},$$

$$\text{(2) } e_{i,n,\tau+1} = e_{i,n,\tau} - \frac{o}{n\dfrac{\sum_{j=1}^{n} c_{ij}}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}}}}.$$

This second modification reduces to subtracting the mean of the vector when orthogonalization is done with respect to vector $\mathbf{1}$, which characterizes matrix expression (4). These two modifications of the algorithm render a set of estimates $\widehat{\lambda}_n(\mathbf{W})$ for $\lambda_n(\mathbf{W})$ that never deviate more than 0.00001 for the 130 cases in the database. In addition, the algorithm quickly calculates $\widehat{\lambda}_n(\mathbf{W}) = 0.92485$ ($\lambda_n = 0.92506$) in 32.4 seconds (using a FORTRAN 77 program) for the case of $n = 45{,}974$; accuracy can be improved by changing the stopping rule (10^{-6} here), but with a cost of increased execution time.

A moderately good statistical description of the minimum eigenvalue of matrix $\mathbf{W}$, obtained with nonlinear regression, is given by

$$\widehat{\lambda}_n(\mathbf{W}) = -0.4000 - \frac{0.5800}{1 + 7.5644e^{0.5502L_p + 7.1387\frac{\lambda_n}{\lambda_1} + 2.2809e_{1,\max}}}, \tag{6}$$

where $L_p = \sqrt[p]{\sum_{i=1}^{n}\left(\sum_{j=1}^{n} c_{ij}\right)^p \Big/ n}$ with $p = -9$. Diagnostic statistics for this estimated equation include 66.73% of the variance in $\lambda_n(\mathbf{W})$ being accounted for, implying a modestly good fit (which is confirmed by a $\lambda_n(\mathbf{W})$-versus-$\widehat{\lambda}_n(\mathbf{W})$ plot), a Shapiro–Wilk statistic of 0.95572 for the regression residuals, implying a failure for conformity with a bell-shaped curve, and a residuals-versus-predicted plot that is less than ideal. Eq. (6) furnishes model-based inferential support far weaker than that associated with Eq. (3) for generalizing its results to the wider population of graphs used in empirical spatial statistical analyses. Furthermore, eight of the $\widehat{\lambda}_n(\mathbf{W})$ values deviate from their corresponding $\lambda_n(\mathbf{W})$ values by less than 1%, 45 deviate by between 1% and 5%, 44 deviate by between 5% and 10%, and 32 deviate by between 10% and 34%. Of note is that for the case of a minimally connected G with $n = 4$,

$\lambda_n(\mathbf{W}) = -1$, Eq. (6) yields an estimate of -0.95613, and the algorithm renders an estimate of -1. As with Eq. (5), Eq. (6) reveals important covariates of $\lambda_n(\mathbf{W})$ but may not necessarily furnish reliable predictions of this quantity.

Eq. (6) reflects a substantially different specification than the one reported by Diaconis and Stroock [12].

5. Selected evaluations of small, medium and large G

Analysis results for the 20 possible irreducible adjacency matrices based upon $n = 5$—a set of graphs not contained in the database—are presented here; a complete enumeration of the planar graphs appears in [10, pp. 236–238]. Of note is that figures #26, #28, #29, and #30 are periodic, resulting in imprecise estimates for $\lambda_1(\mathbf{C})$ and $e_{1,\max}$. The quantities yielded by the algorithm for $\lambda_n(\mathbf{C})$ and $\lambda_n(\mathbf{W})$ are correct for all Gs but #13. Both of these problems appear to relate to graphs that have a pronounced star structure [27]. A new complication arising in this set of adjacency matrices is associated with the torus tessellation, graph #27. Because $\mathbf{E}_1$ for matrix $\mathbf{C}$ of this G is a constant, centering it can result in division by zero. This problem was circumvented by adding a pseudo-random error term (a perturbation) to each element of the resulting vector of zeroes that is normally distributed with a mean of zero and a variance of $(1/100)^2$. In all cases, the values for $\lambda_n(\mathbf{C})$ and $\lambda_n^*(\mathbf{C})$ continue to be close. Finally, the value of p for the set of L_p values ranges from 1 to 3.40462. As Eq. (7) demonstrates, of note is that a number of these complications may well be due to how small n is, namely 5, and may disappear for larger values of n.

Analysis results for two medium size tessellations (not contained in the database) are presented next. The first is from the incomplete regular square tessellation for the coal ash data presented in [8], for which $n = 209$. The second is from a Thiessen polygon surface partitioning of Pennsylvania toxic release inventory sites, for which $n = 1040$. The coal ash tessellation has the anomalous feature of a dominant negative eigenvalue for matrix expression (4). The preceding algorithm is able to compute correctly $\lambda_1(\mathbf{C}) = 3.87977$, $\lambda_n^*(\mathbf{C}) = -3.87977$—which exactly equals $\lambda_n(\mathbf{C})$—and $\lambda_n(\mathbf{W}) = -1$; the estimate of $\hat{e}_{1,\max} = 0.13968$ it produces differs slightly from $e_{1,\max} = 0.13951$. Meanwhile, the toxic release inventory tessellation has no anomalies. The preceding algorithm is able to compute all quantities correctly for it: $\lambda_1(\mathbf{C}) = 6.43152$, $\lambda_n^*(\mathbf{C}) = -3.42483$—which differs slightly from $\lambda_n(\mathbf{C}) = -3.42499$—$\lambda_n(\mathbf{W}) = -0.56192$, and $e_{1,\max} = 0.18982$.

The principal eigenvalue, $\lambda_1(\mathbf{C})$, for selected patterned matrices can be used to evaluate Eq. (3) for very large G. The first evaluation is for the star graph, which often is denoted by $K_{1,n-1}$, and for which $\lambda_1(\mathbf{C}) = \sqrt{n-1}$ [27, p. 15]. Letting n span the interval [2, 50,000] and pooling these cases results in the estimate for L_p of $p = 2$, the correct value. The second evaluation is for the minimally connected G, for which $\lambda_1(\mathbf{C}) = 2\,\mathrm{COS}(\pi/(n+1))$. Letting n span the interval [3, 9800], and again pooling these cases, results in the estimate

$$p = 1.58092[1 + 0.28924(n - 2)],$$

which yields estimates that never deviate from the actual values by more than 0.2%. The third evaluation is for G constructed from a triangular tessellation covering a U-by-V rectangular region, where U denotes the horizontal axis and V denotes the vertical axis, and $n = UV$. Once more, letting n span the interval [3, 9800] and pooling these cases results in the estimate

$$p = 27.6686\left[1 - \frac{14.4229}{U + 12} - \frac{14.4229}{V + 12} + \frac{236.016}{(U + 12)(V + 12)}\right],$$

which yields 326 estimates that deviate from their actual value counterparts by no more than 1%, and 28 that deviate between 1% and 5%. The fourth evaluation is for G constructed from a regular square tessellation covering a U-by-V rectangular region, for which

$$\lambda_1(\mathbf{C}) = 2\left[\mathrm{COS}\left(\frac{\pi}{U + 1}\right) + \mathrm{COS}\left(\frac{\pi}{V + 1}\right)\right].$$

Letting n span the interval [2-by-2, 1000-by-1000] and again pooling these cases results in the estimate

$$p = 65.6175\left\{1 - \frac{4.72644}{\mathrm{LN}(U + 10)} - \frac{4.72644}{\mathrm{LN}(V + 10)} + \frac{13.8800}{\mathrm{LN}[(U + 12)(V + 12)]}\right\},$$

which yields estimates that never deviate from the actual values by more than 0.7%. The fifth evaluation is for G constructed from a regular hexagonal tessellation covering a U-by-V rectangular region, for which Griffith [15] reports a very good approximation equation. Letting n span the interval [3-by-3, 99-by-93] and, as before, pooling these cases results in the estimate

$$p = 39.7119\left\{1 - \frac{15.6857}{U + 14} - \frac{15.6857}{V + 14} + \frac{271.592}{[(U + 14)(V + 14)]}\right\},$$

which yields 710 estimates that do not deviate from their actual value counterparts by more than 1%, and 11 estimates that deviate between 1% and 5%. The sixth, and final, evaluation is for matrix $\mathbf{C}$ constructed from a maximally connected G [6], for which Griffith and Sone [20] report a very good approximation equation. Letting n span the interval [4, 46,000] and once more pooling these cases results in the estimate

$$p = 2\left[1 + \frac{0.61265}{(n - 1)^{0.60163}}\right],$$

which yields estimates that never deviate from the actual values by more than 0.2%.

Table 1
The principal eigenvalue for selected very large adjacency matrices

Graph G configuration	n	$\lambda_1(\mathbf{C})$	Estimate for L_p	Estimate from Eq. (3)
Star	50,000	223.61	223.61	3179.17
Minimally connected	9800	2.00000	2.00000	2.14333
Triangular	9800	2.99842	2.99565	3.15150
Square	1,000,000	3.99998	3.99962	4.13226
Hexagon	9207	5.99623	5.99138	5.89586
Maximally connected	46,000	304.809	304.798	2988.75

Comparisons of results for these estimates of p and Eq. (3) appear in Table 1, for selected graphs not in the database; these comparisons indicate that those p values for the special cases reported here lead to very good estimates of $\lambda_1(\mathbf{C})$. These comparisons also indicate that Eq. (3) performs best for surface partitionings resembling a mixture of the square and hexagonal tessellations, which for the most part is what is found in practice. In addition, Eq. (3) performs very poorly for tessellations where MAX $\left(\sum_{j=1}^{n} c_{ij}\right)$ becomes too large. Moreover, while Eq. (3) serves to identify prominent features of G that help to describe variation in $\lambda_1(\mathbf{C})$, in general it should not be relied upon for accurate estimates of this principal eigenvalue. Of interest here is the potential that formulae for p can be established for patterned matrices.

Comparative results for other quantities appear in Table 2, again for graphs not in the database. Of note is that the actual values for the hexagon and maximally connected examples were computed with MATLAB sparse matrix routines. Of note is that the equations in [15,20] yield results very close to their respective actual values: 5.99647 and -2.99733 for the hexagonal case, and 46.17941 and -43.20280 for the maximally connected case. Occasionally the algorithm converges upon $\lambda_{n-1}(\mathbf{C})$ rather than $\lambda_n(\mathbf{C})$; but the algorithm usually correctly converges on $\lambda_n(\mathbf{C})$. Finally, for the commonly encountered irregular surface partitions that resemble mixtures of square and hexagonal tessellations, the algorithm appears to work well.

Table 2
Selected eigenvalues for particular very large adjacency matrices

Graph G	n	$\lambda_1(\mathbf{C})$		$\lambda_n(\mathbf{C})$		$\lambda_n(\mathbf{W})$	
configuration		Actual	Algorithm	Actual	Algorithm	Actual	Algorithm
Minimally connected	1000	1.99999	1.99999	-1.99999	-1.99998	-1	-1
Square	10,000	3.99807	3.99807	-3.99807	-3.99807	-1	-1
Hexagon	10,000	5.99661	5.99661	-2.99831	-2.99831	-0.57309	-0.57309
Maximally connected	1001	46.20080	46.20080	-43.20280	-1.99931	-0.5	-0.5

6. Conclusion

Findings summarized in this paper demonstrate that efficient algorithms for computing extreme eigenvalues of adjacency matrices based on planar graphs can be built with the standard regression procedure. They also furnish good statistical descriptions for the extreme eigenvalues studied in this paper, revealing and assigning relative degrees of importance to prominent covariates. In addition, Hofmeister's [21] specification of L_p shows considerable promise as a basis for developing formulae that render accurate statistical estimates of principal eigenvalues of patterned matrices.

Numerous out-of-sample evaluations of the statistical results are presented, which when coupled with model-based inference allow generalizations to be made about the extreme eigenvalues of the population of graphs used in empirical spatial statistical analyses, which appear to be some mixture of square and hexagonal tessellations. The specific out-of-sample cases evaluated are: the $n = 2$ graph, a selected $n = 4$ graph, the entire set of $n = 5$ planar graphs, two moderate sized graphs used in empirical spatial statistical analyses ($n = 209$ and $n = 1040$), and 10 very large, patterned planar graphs. These assessments suggest that formulae and the algorithm reported here furnish quick and accurate approximations of extreme eigenvalues for graphs used in spatial statistical analyses.

The extreme eigenvalues studied here are of particular importance because they govern calculation of the determinant of large, sparse matrices. With reference to spatial statistical analyses, they help allow Eq. (1) to be implemented as an approximation for the spatial statistical Jacobian term based upon this determinant. This approximation is particularly appealing when analyzing massively large geographic data sets, whose associated graphs and adjacency matrices prevent the full set of n eigenvalues from being calculated. For example, procedures described in this paper have been used to successfully complete a spatial statistical analysis based upon the nearly 225,000 coterminous US 1990 census blockgroups.

Acknowledgements

I am very grateful to Ms. Michele R. Griffith, Ms. Claire Saint-Rossy, and Ms. Reneé Reinertsen for compiling most of the database. I am also grateful to the numerous geographers and regional scientists who contributed copies of their surface partitionings so that the database could be constructed.

Appendix A

The database comprises 130 graph duals of planar surface partitionings used by spatial scientists in empirical analyses, converted to binary adjacency matrices, and whose n ranges from 7 to 45,974. These graphs are included strictly because of their availability. Forty-seven were extracted from Canadian census sources: 27 urban census tracts for 1971, 18 urban census tracts for 1986, and two sets of national enumeration areas for 1991. Twelve were extracted from Puerto Rico: island-wide municipios, and municipios for five agricultural regions. Seven were extracted from US census sources: 1990 census tracts, blockgroups, and blocks for Syracuse, NY; 1980 census tracts for Houston; 1990 census tracts for Chicago and for Washington, DC; and counties for the coterminous US. Nine were constructed for the Archimedean tilings appearing in [1]. Twenty were taken from Griffith and Layne [19]. Six came from a spatial statistical project undertaken in Peru. Four came from a soil pollution project and three came from an archaeology project, each being a Thiessen polygon surface partitioning respectively based upon soil sample and archaeological site locations. Fifteen were gleaned from the quantitative geography literature. Three were obtained from conference presenters. And, four were constructed from special publications.

The Box–Cox transformation $(n - 3.40)^{-0.28}$, for the size distribution of these graphs, is approximately normally distributed. The average of the mean number of graph connections is 4.6, ranging from 2.5 to 5.9. The average of the maximum number of graph connections is 9.6, ranging from 3 to 52. The average percentage of nodes with 4, 5 or 6 links is 62.5, ranging from 0% to 94%. The average percentage of maximum planar connections is 82, ranging from 44% to 100%. Only two graphs are periodic. This entire set of graphs appears to be representative of those typically found in geography and regional science work to date. As such they should furnish a solid foundation for model-based inferences.

References

[1] M. Ahuja, B. Schachter, Pattern Models, Wiley, New York, 1983.

[2] L. Anselin, O. Smirnov, Efficient algorithms for constructing proper higher order spatial lag operators, J. Reg. Sci. 36 (1996) 67–89.

[3] R. Barry, R. Pace, Monte Carlo estimates of the log determinant of large sparse matrices, Linear Algebra Appl. 289 (1999) 41–54.

[4] A. Basilevsky, Applied Matrix Algebra in the Statistical Sciences, North-Holland, New York, 1983.

[5] A. Berman, R. Plemmons, Nonnegative Matrices in the Mathematical Sciences, SIAM, Philadelphia, 1994.

[6] B. Boots, G. Royal, A conjecture on the maximum value of the principal eigenvalue of a planar graph, Geograph. Anal. 23 (1991) 276–282.

[7] F. Chatelin, Eigenvalues of Matrices (translated by W. Ledermann), New York, Wiley, 1993.

[8] N. Cressie, Statistics for Spatial Data, Wiley, New York, 1991.

[9] D. Cvetković, P. Rowlinson, The largest eigenvalue of a graph: a survey, Linear and Multilinear Algebra 28 (1990) 3–33.

[10] D. Cvetković, P. Rowlinson, S. Simić, Eigenspaces of Graphs, Cambridge University Press, New York, 1997.

[11] P. de Jong, C. Sprenger, F. van Veen, On extreme values of Moran's I and Geary's c, Geograph. Anal. 16 (1984) 17–24.

[12] P. Diaconis, D. Stroock, Geometric bounds for eigenvalues of Markov chains, Ann. Appl. Probab. 1 (1991) 36–61.

[13] M. Fiol, Eigenvalue interlacing and weight parameters of graphs, Linear Algebra Appl. 290 (1999) 275–301.

[14] J. Friedman, Error bounds on the power method for determining the largest eigenvalue of symmetric, positive definite matrix, Linear Algebra Appl. 280 (1998) 199–216.

[15] D. Griffith, Eigenfunction properties and approximations of selected incidence matrices employed in spatial analyses, Linear Algebra Appl. 321 (2000) 95–112.

[16] D. Griffith, A linear regression solution to the spatial autocorrelation problem, J. Geograph. Systems 2 (2000) 141–156.

[17] D. Griffith, Quick but not so dirty ML estimation of spatial autoregressive models, unpublished manuscript, Department of Geography, Syracuse University.

[18] D. Griffith, C. Amrhein, Multivariate Statistical Analysis for Geographers, Prentice-Hall, Englewood Cliffs, NJ, 1997.

[19] D. Griffith, L. Layne, A Casebook for Spatial Statistical Data Analysis, Oxford University Press, New York, 1999.

[20] D. Griffith, A. Sone, Trade-offs associated with normalizing constant computational simplifications for estimating spatial statistical models, J. Statist. Comput. Simulation 51 (1995) 165–183.

[21] M. Hofmeister, Spectral radius and degree sequence, Math. Nachr. 139 (1988) 37–44.

[22] Y. Hong, J. Shu, Sharp lower bounds of the least eigenvalue of planar graphs, Linear Algebra Appl. 296 (1999) 227–232.

[23] F. Luk, S. Qiao, A fast eigenvalue algorithm for Hankel matrices, Linear Algebra Appl. 316 (2000) 171–182.

[24] T. Mäkeläinen, Extrema for characteristic roots of product matrices, Comment. Physico-Math. 38 (1970) 28–53.

[25] J. Ord, Estimation methods for models of spatial interaction, J. Amer. Statist. Assoc. 70 (1975) 120–126.

[26] M. Tiefelsdorf, B. Boots, The exact distribution of Moran's I, Environ. Plann. A 27 (1995) 985–999.

[27] R. Wilson, Introduction to Graph Theory, second ed., Longman, Harlow, 1979.

[28] X. Yong, On the distribution of eigenvalues of a simple undirected graph, Linear Algebra Appl. 295 (1999) 73–80.