ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/321490869

A Local Indicator of Multivariate Spatial Association

Working Paper · November 2017

DOI: 10.13140/RG.2.2.18101.58084

CITATIONS
0

READS
540

1 author:

Luc Anselin
University of Chicago

229 PUBLICATIONS **29,513** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project NIH/NCI R01 Grant 2-R01CA126858 “Geospatial Factors & Impacts II” View project

Project Violence(s)as public health, criminological, and sociological phenomena View project

All content following this page was uploaded by Luc Anselin on 04 December 2017.

The user has requested enhancement of the downloaded file.

A Local Indicator of Multivariate Spatial Association: Extending Geary's c.[*]

Luc Anselin

Center for Spatial Data Science

University of Chicago

anselin@uchicago.edu

November 9, 2017

[*] This research was funded in part by Award 1R01HS021752-01A1 from the Agency for Healthcare Research and Quality (AHRQ), "Advancing spatial evaluation methods to improve healthcare efficiency and quality." Comments by Sergio Rey, Julia Koschinsky and the referees on an earlier draft are greatly appreciated.

Abstract

This paper extends the application of the Local Geary c statistic to a multivariate context. The statistic is conceptualized as a weighted distance in multivariate attribute space between an observation and its geographical neighbors. Inference is based on a conditional permutation approach. The interpretation of significant univariate Local Geary statistics is clarified and the differences with a multivariate case outlined. An empirical illustration uses Guerry's classic data on moral statistics in 1830s France.

Key words: spatial clusters, LISA, multivariate spatial association, Local Geary c, spatial data science

1 Introduction

An important and growing component of geographical analysis is a focus on the *local*, reflected in new methods to deal with local spatial autocorrelation and local spatial heterogeneity (for general overviews, see, e.g., Unwin 1996, Fotheringham 1997, Unwin and Unwin 1998, Fotheringham and Brunsdon 1999, Boots 2002, Boots and Okabe 2007, Lloyd 2010). Specifically, considerable interest has been devoted to local indicators of spatial association (LISA) since the original LISA framework was outlined in Anselin (1995, 1996), building upon the initial work by Getis and Ord (1992, 1996), and Ord and Getis (1995). Since the local statistics form the basis for a hypothesis test for (local) spatial randomness, they are strictly speaking outside the scope of exploratory data analysis (EDA) as outlined by, among others, Tukey (1977) and Good (1983), and its spatial counterparts (ESDA). Narrowly defined, EDA and ESDA are focused on generating hypotheses, not testing them. Nevertheless, the local statistics are often considered to be an important part of an exploratory strategy (see, e.g. Sokal et al. 1998b). This is the spirit in which they are considered here.

The idea of a local test for spatial autocorrelation has been extended in multiple directions, such as applications to categorical data (Boots 2003, 2006), points on networks (Yamada and Thill 2007), the construction of optimal spatial weights (Getis and Aldstadt 2004, Aldstadt and Getis 2006), as well as space-time and income mobility (Rey 2016). Considerable attention has been paid to problems of statistical inference, both exact and asymptotic, as well as more fundamental issues of multiple comparisons and correlated tests. For example, Sokal et al. (1998a) examined the properties of asymptotic approximations based on analytical moments, whereas Tiefelsdorf (2002) developed a saddlepoint approximation to exact inference. The multiple comparison problem was discussed in general by de Castro and Singer (2006), and investigations into inference in the presence of global spatial autocorrelation are reported by Ord and Getis (2001) and Rogerson (2015). More technical issues have been considered as well, such as power calculations (Bivand et al. 2009), the design of optimal spatial weights (Rogerson 2010, Rogerson and Kedron 2012), and conceptual and computational issues pertaining to randomization inference (Lee 2009, Hardisty and Klippel 2010). In addition, the Local Moran test and the Getis-Ord local G statistics have been implemented in both commercial and open source spatial analytical software, such as GeoDa (Anselin et al. 2006), spdep and other packages in

R (Bivand 2006, Bivand et al. 2013), the PySAL Python library for spatial analysis (Rey and Anselin 2007), ESRI's ArGIS Spatial Analyst, and the online spatial analytical functionality in Carto (`https://carto.com/blog/cluster-outlier-intro`).

Most of the discussion of local spatial autocorrelation has been situated in a univariate context. The treatment of spatial autocorrelation in a multivariate setting has focused on global statistics, specifically Moran's I. This started with the work by Wartenberg (1985) that extended the notion of principal components to include spatial autocorrelation. This line of thinking was further generalized by Dray et al. (2008) into the concept of MULTISPATI, which adds a matrix of spatially lagged variables to the statistical triplet used in co-inertia analysis (see also Dray and Jombart 2011). A different approach was taken in Lee (2001) specifically for a bivariate case, where a distinction is made between the correlative and the spatial association between two variables.

The current paper has two objectives. One is a closer examination of the univariate Local Geary statistic. This test was also proposed as part of the general LISA framework in Anselin (1995), but it has received less attention to date than its counterpart the Local Moran statistic, or the Getis-Ord local statistics. Nevertheless, it forms an interesting alternative to these statistics, since it is not limited to linear associations, as it is based on a squared difference. Specifically, the interpretation and visualization of this statistic are discussed in some detail, with the emphasis on its use as a data exploratory tool in the spirit of unsupervised (machine) learning and spatial data mining, rather than as a statistical test in a strict sense. The second and main goal is to extend the univariate case to a multivariate setting, and to introduce a Local Geary statistic for multivariate spatial autocorrelation. The statistic is outlined and its inference and interpretation are discussed in detail, again with an emphasis on its use in data exploration, rather than as a strict test statistic. The new statistics are illustrated with a local take on the analysis by Dray and Jombart (2011) of global multivariate spatial autocorrelation based on the classic data set with "moral statistics of France," attributed to an 1833 essay by André-Michel Guerry. The paper closes with some concluding remarks.

2 Local Geary c Revisited

As is well-known in the spatial analysis literature, Geary (1954) introduced a global measure of spatial autocorrelation as:

$$c = \frac{(n-1)\sum_i \sum_j w_{ij}(x_i - x_j)^2}{2S_0 \sum_i (x_i - \bar{x})^2}, \tag{1}$$

where x_i is an observation on the variable of interest at location i, $\bar{x}$ is its mean, n is the total number of observations, and w_{ij} are the elements of the familiar spatial weights matrix, which embodies a prior notion of the neighbor structure of the observations.[1] The term S_0 corresponds to the sum of all the weights ($\sum_i \sum_j w_{ij}$). The Geary c statistic can equivalently be expressed as a ratio of two sums of squares, i.e., the squared difference between observations at i and j in the numerator, and the sum of squared deviations from the mean in the denominator:

$$c = \frac{\sum_i \sum_j w_{ij}(x_i - x_j)^2 / 2S_0}{\sum_i (x_i - \bar{x})^2 / (n-1)}. \tag{2}$$

Clearly, the denominator is an unbiased estimator for the variance. The numerator on the other hand is a rescaled sum of weighted squared differences. The factor 2 is included to center the expected value of the statistic under the null hypothesis of no spatial autocorrelation to the value of 1 (not zero). Statistics smaller than one, indicating a small difference between an observation and its neighbors, suggest positive spatial autocorrelation. Statistics larger than one suggest negative spatial autocorrelation (large differences between an observation and its neighbors).[2]

Geary's c statistic is reminiscent of the pairwise squared deviation measure that underlies the empirical semi-variogram in geostatistics (for example, see the overarching framework that includes a range of cross-product statistics outlined in Getis 1991). However, there are two important differences. First, in the semi-variogram, all pairwise differences are considered, which results in $n(n-1)/2$ estimates. In Geary's c statistic, the difference between an observation and its neighbors is summarized as a weighted average for each location (roughly $n\bar{k}$ comparisons, with $\bar{k}$ as the average number of neighbors) and yields a single statistic. Second, whereas in the semi-variogram the squared difference measure is sorted by the distance that separates the

[1] By convention, $w_{ii} = 0$, so that self-neighbors are excluded.

[2] While this may seem somewhat counterintuitive at first, this is easily remedied by subtracting 1 from the statistic and changing its sign.

observation pairs, in Geary's c, the neighbors are pre-defined through the spatial weights. In sum, the semi-variogram focuses on all pairs of observations, but Geary's c provides a single measure for each individual observation. Both approaches take the same perspective in the sense that small values of the statistic suggest similarity, or positive spatial autocorrelation, with large values of the statistic suggesting the reverse. In addition, since the statistic is based on a squared difference, it is not constrained to linear forms of association.

A local version of Geary's c was outlined in Anselin (1995) as:

$$c_i = \sum_j w_{ij}(x_i - x_j)^2. \tag{3}$$

Since the squared deviations cancel out the mean, it is irrelevant whether the variable is expressed on the original scale, or in standardized form, although in a multivariate setting, the latter is the preferred practice. Also, a number of variants of this statistic can be defined, depending on which of the scaling constants are included. For example, an alternative form, also given in Anselin (1995) and further investigated by Sokal et al. (1998a) includes a consistent estimate for the variance as a scaling factor:

$$c_i = (1/m_2) \sum_j w_{ij}(x_i - x_j)^2, \tag{4}$$

where $m_2 = \sum_i (x_i - \bar{x})^2/n$.[3]

The inclusion of the scaling factor only results in a monotone transformation of the value in Equation 3, so it is easier to keep the simplest formulation. This expression is also the only aspect of the global Geary c that changes with each observation i, since both the denominator (the variance) and S_0 are constants.

The analytical moments for the Local Geary c_i were given in Sokal et al. (1998a, p. 353), using the expression in Equation 4 (see also the extensive discussion in Boots 2002). More specifically, the expected value for the Local Geary under a randomization approach is shown to be:

$$\mathrm{E}[c_i] = 2nw_i/(n-1), \tag{5}$$

where w_i is the sum of the weights in row i, i.e., $\sum_j w_{ij}$. Some straightforward algebraic

[3] Note how this is a consistent estimator for the variance, but not an unbiased one. The unbiased estimator used in the expression for the global Geary c divides the sum of squared deviations by $n - 1$.

manipulations yield the expected value of the expression in Equation 3 as:

$$\mathrm{E}[c_i] = 2nw_i m_2/(n-1). \tag{6}$$

In the case of row-standardized weights, $w_i = 1$. Futhermore, with standardized z_i, $\mathrm{E}[m_2] = (n-1)/n$. Substituting these results in the expression for the expected value yields:

$$\begin{aligned} \mathrm{E}[c_i] &= 2n(n-1)/[(n-1)n] \qquad (7) \\ &= 2 \qquad (8) \end{aligned}$$

With expressions for the expected value and the variance in hand an asymptotic approximation can be developed, as shown in Sokal et al. (1998a). However, these same authors also cautioned that asymptotic inference based on these moments tends to fail. Instead, the approach taken in practice is to use conditional permutation, as outlined in Anselin (1995). This consists of creating a reference distribution for each individual location by randomly permuting the remaining values (i.e., all observations except the value at location i) and recomputing the statistic each time. Inference can then be based on a pseudo p-value of a one-sided test computed from the number of replicated statistics that are more *extreme* (either larger or smaller) than the observed local statistic. As is well known, the resulting pseudo p-values should be interpreted with caution, since they suffer from multiple comparisons, the potential biasing effect of global autocorrelation, and other such complicating factors (see, among others, the reviews in Sokal et al. 1998b, Ord and Getis 2001, de Castro and Singer 2006, Rogerson 2015, as well as the discussion below).

Finally, note that, similar to all local statistics, the hypothesis test associated with the Local Geary statistic is a *diffuse* test. The null hypothesis is that of spatial randomness. More precisely, this means that locally, i.e., focused on a given observation, any organization of values in the surrounding neighbors is equally likely. Differences from such randomness are detected by using the weighted squared difference as a criterion. However, unlike what is the case for a *focused* test (such as the Likelihood Ratio test used in a regression context), there is no specified alternative. Different local statistics use different criteria to detect deviations from the null, such as a squared difference for the Local Geary, a cross-product for the Local Moran, or a sum in the Getis-Ord statistics. These different criteria will have more power against specific alternatives, but also have power against all others. In other words, while the rejection of the null suggests the absence of (local) spatial randomness, it cannot suggest the presence of a particular form of association. Hence, the main purpose of these statistics is in a data exploration sense (see also the argument in Sokal et al. 1998b).

2.1 Inference and Interpretation

The pseudo p-value obtained from the conditional permutation procedure has to be interpreted with caution, since it is unlikely to correctly reflect the actual Type I error. In addition to being only an approximation to the actual p-value, it is also affected by the multiple comparisons inherent in any local analysis that considers many (all) locations in the data set in turn. The textbook correction for multiple comparison consists of a Bonferroni or Sidak bound (see, for example, the discussion in Boots 2002). Both consider a target overall p-value, sometimes referred to as the family wide error rate, or FWER, i.e., the probability of making even one false rejection (see, for example Efron and Hastie 2016, Chapter 15, for a detailed technical discussion).

With a target p-value of α and k comparisons, the Bonferroni bound for each individual test would be α/k. The corresponding Sidak bound would be $1 - (1 - \alpha)^{1/k}$. In the context of LISA statistics, the practical question is what value k should take. The total number of observations is likely too conservative, and some measure of overlap between the locations and their neighbors could be considered, as suggested in Getis and Ord (2000), although its implementation in practice is not straightforward. An alternative approach was suggested in de Castro and Singer (2006), based on the false discovery rate (FDR) proposed by Benjamini and Hochberg (1995) (see also Efron 2010, Efron and Hastie 2016, for an extensive technical treatment).

The FDR is obtained in two steps. First, the pseudo p-values $p(i)$ are ranked from smallest to largest. The observation i_{max} is selected as the largest value of i for which $p(i) \leq (i/N)\alpha$ (with N as the total number of observations). All observations with $i \leq i_{max}$ are taken to reject the null hypothesis. However, as Efron and Hastie (2016, Footnote 6, on p. 276) caution, "the classic term *significant* for a non-null identification doesn't seem quite right for FDR control ... and we will use *interesting* instead."

Even with these caveats in mind, the interpretation of a "significant" (or, rather, "interesting") Local Geary statistic is arguably not as intuitive as that of the other commonly used local statistics, such as the Getis-Ord G_i and G_i^* statistics and the Local Moran. For the former, a positive and significant value indicates a *hot spot* or cluster of high values, whereas a negative and significant value suggests a *cold spot*, or cluster of low values (Getis and Ord 1992, 1996, Ord and Getis 1995). The interpretation of the Local Moran is facilitated in conjunction with the quadrants of the Moran scatter plot and suggests spatial clusters (high-high, and low-low) as well as spatial outliers (high-low, and low-high) (Anselin 1995, 1996)

A significant c_i statistic that is less than its expected value under the null hypothesis (either the analytical value, or the average of the empirical reference distribution in a permutation approach) suggests a clustering of *similar* values. Unlike what is the case for the Moran scatter plot, there is no unambiguous differentiation of the type of association. This follows from the fact that the local Moran is a cross-product statistic, which naturally aligns with a linear fit (regression line) of points in a Moran scatter plot. The c_i statistic is based on squared differences, irrespective of whether these are differences between high values or low values. So, while the Local Moran focuses on *linear* associations, the Local Geary is not constrained by this, and can detect associations of a non-linear form as well. In sum, a small squared difference suggests *similarity*, but cannot divulge the type of similarity.

Similar neighbors could thus have either similar high values (the counterpart of high-high in the Local Moran case), or similar low values (the counterpart of low-low in the Local Moran case). However, they could also result from two data points that span the mean (e.g., one above the mean and one below), but that are very close together in value. So, unlike the Local Moran case where there is a clear categorization of the results, this is not the case for the Local Geary statistic.

Nevertheless, it is still possible to categorize the type of association in some instances. This is accomplished by locating the pairs $x_i, \sum_j w_{ij} x_j$ in the Moran scatterplot. Those pairs that correspond with a significant *small* value of c_i and that fall clearly in the high-high or low-low quadrants can be classified as such. For those pairs where that is not the case (e.g., falling in a low-high quadrant), there is no corresponding classification, and this case has to be referred to as *other*.

There is no counterpart to this classification for negative spatial autocorrelation as indicated by a significant value for c_i. In this instance, the Local Geary statistic simply indicates a large (larger than under spatial randomness) difference between neighboring values, without suggesting a particular high-low or low-high pattern. Due to the use of a squared difference as the criterion for attribute similarity, such a distinction is not possible.

3 A Multivariate Extension

3.1 A Multivariate Local Geary Statistic

The Local Geary c_i is a univariate statistic. In essence, it measures the squared distance in attribute space (i.e., along a line for the univariate case) between the value at a geographic location and that at each neighboring location (in geographic space), and summarizes this in the form of a weighted sum.[4] In practice, since the spatial weights are typically row-standardized, this boils down to a weighted average of the squared distances in attribute space between an observation and its geographic neighbors (as defined by a spatial weights matrix).

This concept can be extended in a straightforward manner to a multivariate context. For example, consider two variables, z_1 and z_2. Following standard practice in multivariate clustering analysis, these variables have been standardized such that the mean of the transformed variable is zero and its variance is one. The squared distance d_{ij}^2 in two-dimensional attribute space between the values at observation i and its geographic neighbor j is:

$$d_{ij}^2 = (z_{1,i} - z_{1,j})^2 + (z_{2,i} - z_{2,j})^2 \tag{9}$$

A weighted average of this expression incorporating the squared distance in two-dimensional attribute space between location i and all its geographic neighbors is then:

$$\begin{aligned}\sum_j w_{ij} d_{ij}^2 &= \sum_j w_{ij} [(z_{1,i} - z_{1,j})^2 + (z_{2,i} - z_{2,j})^2] \\ &= \sum_j w_{ij} (z_{1,i} - z_{1,j})^2 + \sum_j w_{ij} (z_{2,i} - z_{2,j})^2 \\ &= c_{1,i} + c_{2,i} \end{aligned} \tag{10}$$

[4] Note that the squared distance is used to keep the similarity with the original formulation of Geary's c, but instead the distance, i.e., the square root of this expression, could be used equivalently.

In other words, the concept of a Local Geary statistic is additive in the attribute dimension. In general then, for k attributes, a multivariate Local Geary can be defined as:

$$c_{k,i} = \sum_{v=1}^{k} c_{v,i}, \tag{11}$$

with $c_{v,i}$ as the Local Geary statistic for variable v. This measure corresponds to a weighted average of the squared distances in multidimensional attribute space between the values observed at a given geographic location i and those at its geographic neighbors. As an alternative to the simple sum in Equation (11), the average could be taken, which would keep the scale of the multivariate measure in line with the univariate measures:

$$c_{k,i} = \sum_{v=1}^{k} c_{v,i}/k. \tag{12}$$

The expected value of the multivariate statistic under the randomization null hypothesis follows as a direct extension of the univariate case given above. For the expression in Equation (12), this remains $\mathrm{E}[c_{k,i}] = 2$, and for the unscaled version $\mathrm{E}[c_{k,i}] = 2k$. However, unlike the univariate case, the derivation of the variance of the multivariate counterpart is quite complex, and not analytically tractactable, since the general variance-covariance among the variables needs to be accounted for (in addition to their spatial correlation). Given the poor results of an asymptotic approximation reported in the literature for the univariate case, the more practical way to obtain inference should again be based on conditional permutation.

The expression in Equation (10) can be generalized in a number of ways. As mentioned, other distance measures can be applied, such as a Manhattan distance (absolute differences), or, in general, any Minkowski distance metric. In addition, different weights could be applied to each individual variable, for example, by means of using the inverse variance matrix as a weight. However, since the main objective of such weighting is to compensate for different variances among the variables, this becomes unnecessary when the variables have been standardized, which is best practice in multivariate analysis.

In a univariate LISA analysis applied to several variables (in turn), it is quite common to employ a range of different spatial weights, each appropriate for one or more variables (e.g., with weights based on geographical distance bands). In a multivariate setting, the core concept is to compare geographical neighbors with neighbors in multi-attribute space, so the former requires a single definition of spatial weights, just as the notion of neighbors in multi-attribute space requires a single distance metric. Nevertheless, a sensitivity analysis of the results with respect to the choice of the spatial weights remains recommended practice in any empirical application.

3.2 Inference and Interpretation

A conditional permutation approach consists of holding the *tuple* of values observed at i fixed, and computing the statistic for m permutations of the remaining tuples over the other locations. This results in an empirical reference distribution that represents a computational approach at obtaining the distribution of the statistic under the null. The resulting pseudo p-value corresponds to the fraction of statistics in the empirical reference distribution that are equal to or more extreme than the observed statistic.

Such an approach suffers from the same problem of multiple comparisons mentioned for the univariate case (see Section 2.1). In addition, there is a further complication. When comparing the results for k univariate Local Geary statistics, the multiple comparisons need to be accounted for. For example, for each univariate test, the target p-value of α would typically be adjusted to α/k (with k variables, each with a univariate test), as a Bonferroni bound. Since the multivariate statistic is in essence a sum of the statistics for the univariate cases, this would suggest a similar approach by dividing the target p-value by the number of variables (k). Alternatively, and preferable, a FDR strategy can be pursued. The extent to which this actually compensates for the two dimensions of multiple comparison (multiple variables and multiple observations) remains to be further investigated.

As in the univariate case, the resulting pseudo p-values should only be taken as providing some indication of *interesting* locations in a data exploration exercise, and they should not be interpreted in a strict sense. In practice, some sensitivity analysis is therefore in order.

The interpretation of a location with a “significant” statistic (in the limited sense outlined above) is more complex than in the univariate case. Since multiple variables are involved, the notion of a *hot spot* or *cold spot* is not necessarily meaningful. In low-dimensional comparisons, such as in a bivariate case, it is possible to construct cross-classification of whether each individual variable is above or below the mean relative to its neighbors, but for higher dimensions, this quickly becomes unwieldy, resulting in many cells with zero elements. In an interactive exploration environment such as GeoDa (Anselin et al. 2006), it is possible to combine a cluster map of the locations of significant multivariate Geary with a brushing and linking operation on the respective quadrants of the univariate Moran scatter plots, yielding some insight into the combinations involved. Overall, however, the statistic indicates a combination of the notion of distance in multi-attribute space with that of geographic neighbors. This is the essence of any spatial autocorrelation statistic. It is also the trade-off encountered in spatially constrained multivariate clustering methods (for a recent discussion, see, for example Grubesic et al. 2014).

Finally, it is important to keep in mind that, even though the multivariate statistic is the sum of the univariate statistics, it is not so that significant locations for the univariate case necessarily translate into significant locations for the multivariate case. In fact, an important motivation for the use of the multivariate statistic is that it focuses on a combination of the distances along the different variable dimensions, rather than taking each as being orthogonal. This may provide additional insight in a spatial data exploration exercise.

4 Empirical Illustration

Recently, Dray and Jombart (2011) illustrated new concepts of multivariate *global* spatial autocorrelation, based on the inclusion of spatially lagged variables, in what they refer to as co-inertial analysis (see also Dray et al. 2008). Their empirical examples used the classic data set with "moral statistics of France," attributed to an 1833 essay by André-Michel Guerry. The data set consists of a collection of observations for 86 French départements on a range of social indicators, including crime, literacy, suicides, etc.[5]

To highlight the different insights gained between the global analyses in Dray and Jombart (2011) and the Local Geary statistics, the same data set and the same six variables are considered here: Crimes Against Persons, Crimes Against Property, Literacy, Donations, Infants Born out of Wedlock, and Suicides. All variables are expressed such that larger values are "better." For example, rather than expressing Crimes Against Persons as the usual crime rate consisting of the ratio of crimes over population, the reverse is used, i.e., the ratio of population

[5] The data is contained in the R package *Guerry*, developed by Michael Friendly and Stéphane Dray, available at `https://CRAN.R-project.org/package=Guerry`. It is also part of the sample data sets provided with the GeoDa software, available at `https://geodacenter.github.io/data-and-lab//Guerry/`.

Table 1: Correlations between the six variables

	Cr. pers.	Cr. prop.	Lit.	Don.	Inf.	Suic.	Moran's I
Crime persons	1.000	0.255	-0.021	0.134	-0.027	-0.134	0.412
Crime property		1.000	-0.363	-0.082	0.278	0.523	0.264
Literacy			1.000	-0.196	-0.412	-0.374	0.718
Donations				1.000	0.159	-0.035	0.353
Infants					1.000	0.289	0.229
Suicides						1.000	0.402

over crime. This operation is applied to the two Crime variables, to Infants Born out of Wedlock and to Suicides. In the analysis that follows, all variables are also standardized, such that their mean equals zero and their variance equals one. Finally, as in Dray and Jombart (2011), Corsica (an island) is removed from the data, which results in a final set of 85 observations. All the analyses were carried out using the latest version of the GeoDa software (Version 1.12), and can be replicated using the sample data set available with the software.[6] Note that the analysis is intended as a straightforward illustration of the different patterns identified by each of the local autocorrelation methods, and not as a substantive study of the moral statistics in 1830 France.

As shown in Dray and Jombart (2011), the six variables are characterized by a high degree of positive spatial autocorrelation, indicated by a positive and highly significant global Moran's I statistic (see the last column in Table 1).[7] However, this spatial correlation is not matched by a similarly strong bivariate correlation between the variables, as shown in Table 1. In fact, none of the correlations are particularly high, with the largest value 0.523, between Property Crime and Suicides. Literacy turns out to be negatively correlated with all the other variables. Several of the bivariate relationships are very weak, such as the correlation between Crime Against Persons and Literacy (-0.021) as well as with Infants Born out of Wedlock (-0.027), and between Donations and Suicides (-0.035).

As a first step, following the approach taken in Dray and Jombart (2011), the six variables are converted into principal components. The results are only moderately successful at cap-

[6] The Guerry data set is installed with the software as one of the built-in sample data sets. The Local Geary statistic and its multivariate generalization are implemented in GeoDa since version 1.10.

[7] All statistics are computed using queen contiguity spatial weights and are significant at $p < 0.001$, based on 999 permutations.

Table 2: Squared correlations: six variables and PC1, PC2

	PC1	PC2
Crime persons	0.009	0.419
Crime property	0.562	0.009
Literacy	0.561	0.020
Donations	0.024	0.587
Infants	0.436	0.013
Suicides	0.549	0.153

turing the overall variance, with the first two components explaining only slightly over half the variance (0.56). The squared correlations in Table 2 show how the first component is primarily constructed from Crime against Property, Literacy, Infants and Suicides, whereas the second component is highly correlated with Crime against persons and Donations.[8] The spatial distribution of the first two components is illustrated in the choropleth maps shown in Figures 1 and 2 (computed as natural breaks maps with 6 intervals).

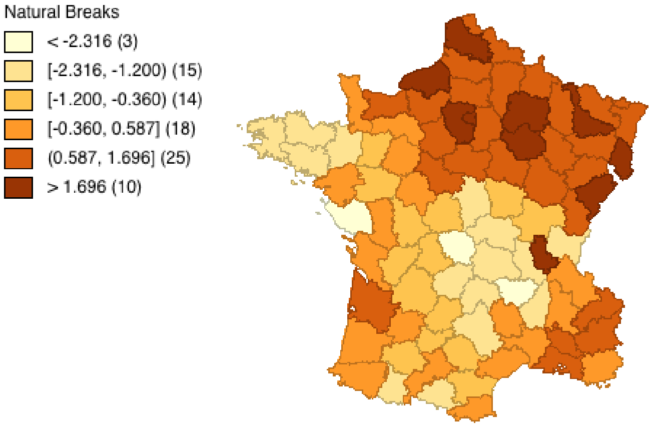

Figure 1: PC1 (natural breaks)

Considering the univariate case first, the cluster centers identified for the local Moran, Getis-Ord G_i and the Local Geary statistics are shown for the first principal component (PC1)

[8] These results match the findings reported in Dray and Jombart (2011). In order to explain more than 90% of the variance, five components (out of six variables!) are needed. This is a direct consequence of the low bivariate correlations between the variables.

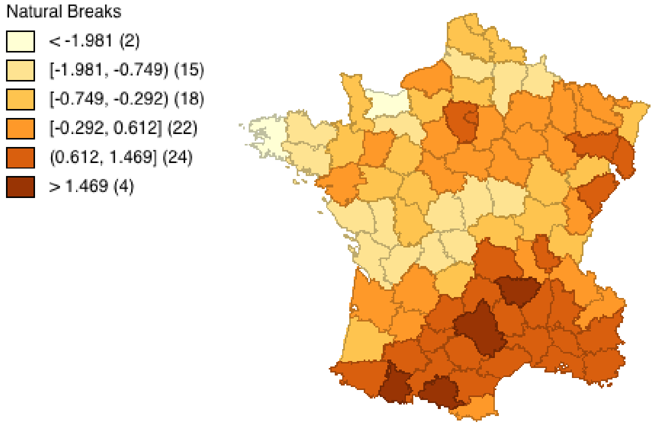

Figure 2: PC2 (natural breaks)

in Figures 3 to 5.[9] Note that, even though these visualizations are typically referred to as cluster maps, they in fact only identify the core or center of a cluster. The cluster itself also includes the neighboring locations (as defined by the spatial weights).

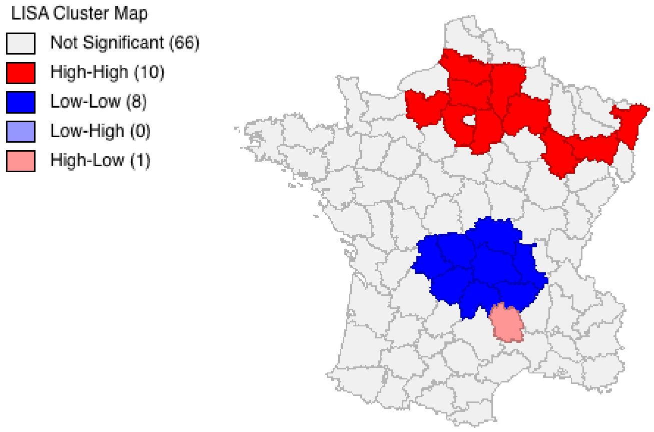

Figure 3: Local Moran Clusters, PC1 (0.01)

[9] PC1 also shows strong positive *global* spatial autocorrelation, as indicated by a Moran's I coefficient of 0.551 for queen contiguity, significant at 0.001 for 999 permutations.

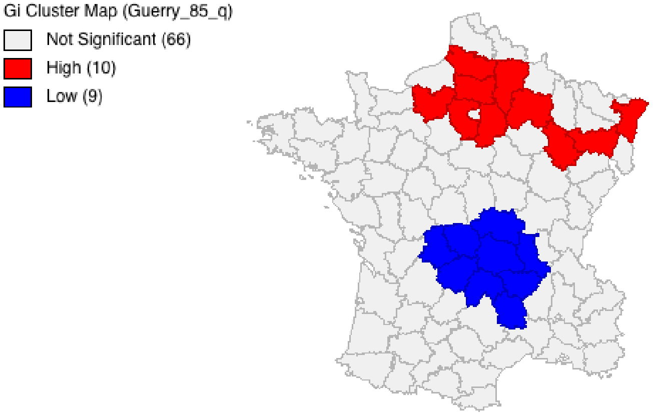

Figure 4: Local G_i Clusters, PC1 (0.01)

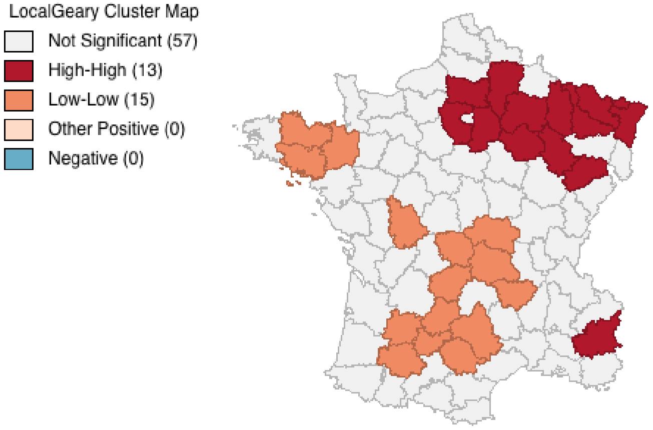

Figure 5: Local Geary Clusters, PC1 (0.01)

For comparison purposes, all three maps are shown for a pseudo p-value of 0.01, based on 99999 permutations.[10] As is typically the case, the results for the Local Moran and the Local G_i statistics match almost exactly, except for the high-low spatial outlier identified in the Local Moran cluster map. The Local G_i focuses only on hot spots and cold spots and classifies this location as part of a low-low cluster. The Local Geary seems to be more liberal

[10] All random permutations are based on the same random seed so that the analysis can be replicated exactly.

in identifying cluster centers, with 28 such locations (for a pseudo p-value of 0.01), compared to 19 for the other local statistics. Figure 6 shows the overlap between locations identified by the Local Moran and Local G_i on the one hand, and those indicated by the Local Geary (the matching locations are highlighted in the Local Geary map). Of the 19 locations from Local Moran/G_i, 12 are also significant for the Local Geary, but seven are not (the grey shaded areas in the map). The 16 other locations deemed significant in the Local Geary cluster map are not identified as such for the two other local measures. This is not surprising, since the statistic use different approaches to quantify *attribute similarity*. Both the Local Moran and the Local G_i statistic are heavily influenced by the weighted average of the neighbors, whereas the Local Geary measures squared difference. There is no a priori reason why the two would give the same results.

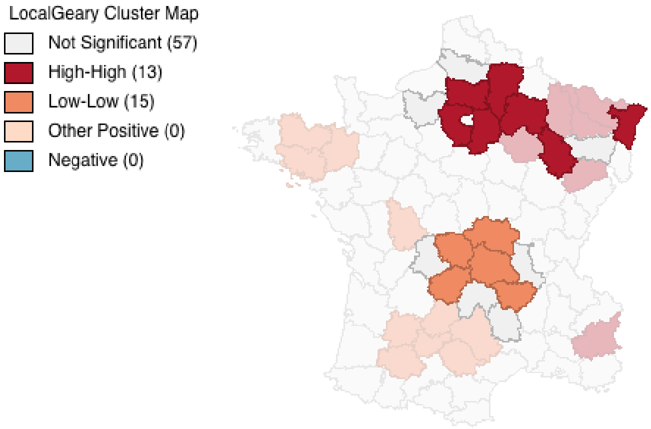

Figure 6: Matching Local Clusters, PC1 (0.01)

So far, cluster centers have only been reported for a pseudo p-value of 0.01. Given the issue of multiple comparisons and the presence of strong global spatial autocorrelation, this is arguably too liberal. To assess the effect of different target p-values, Table 3 shows the number of significant locations for p-values of 0.01, 0.005 (the *new* standard as argued in Benjamin and 72 others 2017), and 0.001, as well as for the FDR suggested rates (the FDR for Local Moran and Local G_i with a target α of 0.01 is 0.00035; for Local Geary it is the same as the Bonferroni bound, 0.00012). In addition to the number of significant cluster centers, the table

Table 3: Number of significant locations by target p-value

	Local Moran	Local G_i	Local Geary	Overlap
0.01	19	19	28	12
FDR	3	3	0	0
0.005	10	10	18	6
0.001	3	3	4	0

also lists the number of overlapping locations. As is to be expected, the main effect is to shrink the clusters to only those locations where the similarity with neighbors is extreme. At 0.001, there are only three and four locations identified for respectively Local Moran/G_i and Local Geary, without any overlap between the two sets.

To further illustrate the effect of changing p-values, the Local Geary clusters are shown in Figures 7 and 8 for p-values of 0.005 and 0.001, which reduces the number of significant locations to respectively 18 and four.

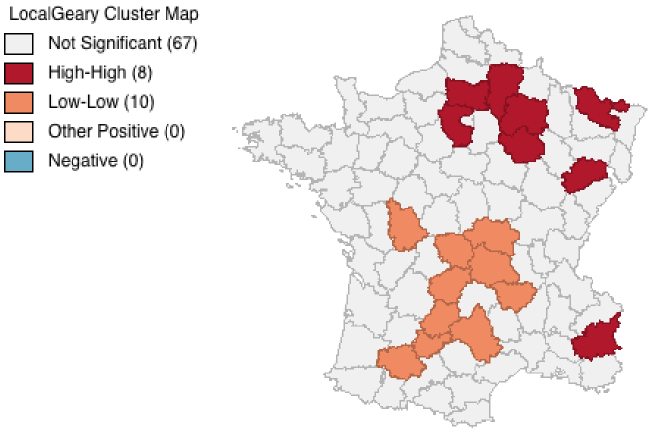

Figure 7: Local Geary Clusters, PC1 (0.005)

Moving on to the bivariate case, the Local Geary map for the second principal component (PC2) is shown in Figure 9, again using a p-value of 0.01 for comparison purposes (based on 99999 permutations). The pattern is quite distinct from PC1, as is to be expected from a cursory examination of the maps in Figures 1 and 2. At this p-value, 19 “significant” cluster centers are identified.

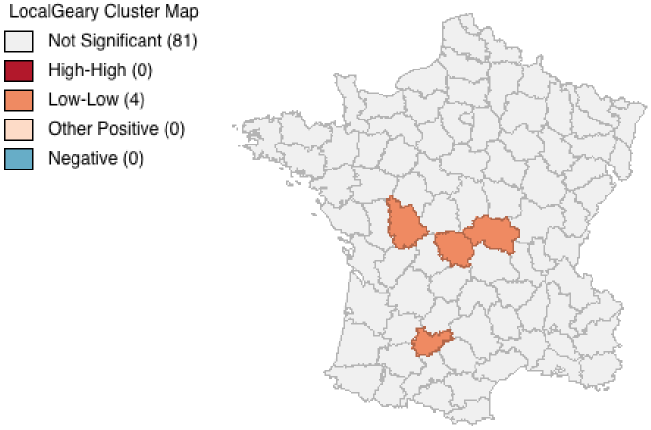

Figure 8: Local Geary Clusters, PC1 (0.001)

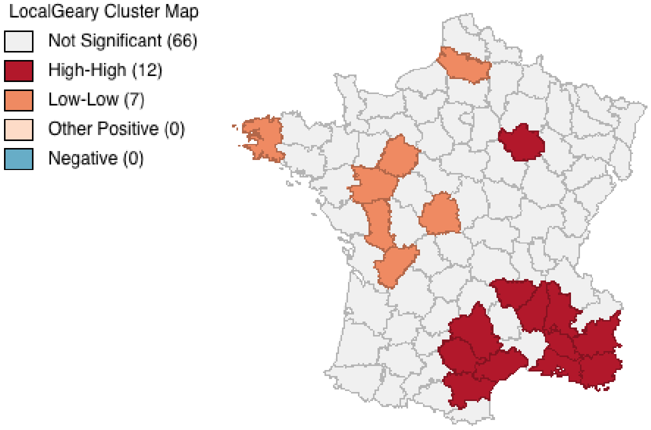

Figure 9: Local Geary Clusters, PC2 (0.01)

Comparing the results for PC1 and PC2, it turns out that only five locations are identified as significant for both variables. These five locations show a significant similarity with their neighbors (based on queen contiguity) for each of the two variables taken individually.

The bivariate Local Geary cluster map focuses on a different aspect of the local spatial association between the two variables. Instead of considering each variable separately, the *joint* location in the attribute space is the focus of attention, i.e., points in attribute space for the two dimensions that are close to their geographical neighbors. Unlike the case for a bivariate Moran's I, there is no confusion of the inter-variable correlation with the bivariate association.[11]

Figure 10 shows the cluster centers for the bivariate Local Geary applied to the first two principal components, using the FDR with an α of 0.01 as the indicator of significance (the resulting FDR is 0.00247 based on 99999 permutations). Twenty-one such locations are identified. Compared to the individual Local Geary cluster maps for each variable (using the same p-value of 0.0025) indicates that of the 11 significant locations for PC1, six are also included as cluster centers based on the bivariate analysis. For PC2, of the two cluster locations, one is in common with the bivariate map (maps not shown). Note, that as pointed out earlier, there was no overlap between the cluster centers for PC1 and PC2 when considered individually. This suggests that the bivariate approach is able to indicate *interesting* locations that go beyond what is found in the univariate analyses. These are locations where the joint profile of PC1 and PC2 is similar to that of the neighbors, but not necessarily for each individual variable.

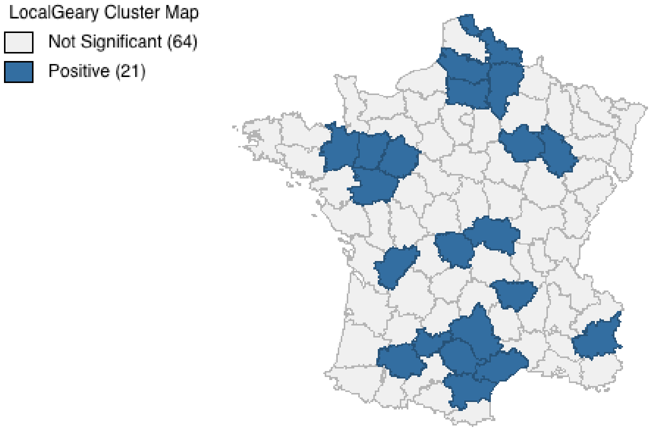

Figure 10: Bivariate Local Geary, PC1 and PC2 (FDR)

As a final comparison, the clusters obtained from a multivariate Local Geary analysis for all six variables are considered (i.e., the original variables, not the principal components). This is illustrated in Figure 11. Again using the FDR (which is the same as in the bivariate case),

[11] In any case, the principal components are uncorrelated by construction, so this aspect can be safely ignored.

this yields 21 cluster centers. While this is the same number as for the bivariate analysis on the principal components, the locations are not the same, although there is substantial overlap. Of the 21 cluster centers, 13 are shared in both analyses. In some sense, this suggests that the analysis based on the first two principal components captures a lot of the same multivariate spatial association as the analysis using all original variables. Much of the same patterns are identified, with some differences at the margins. A further sensitivity analysis can be carried out by changing the critical p-values and assessing the differences and similarity between the two sets of maps, to "discover potentially explicable patterns" (Good 1983), or "to detect the expected and discover the unexpected" (Thomas and Cook 2005). The interactive visualization implemented in the GeoDa software is designed to facilitate such a data exploration.

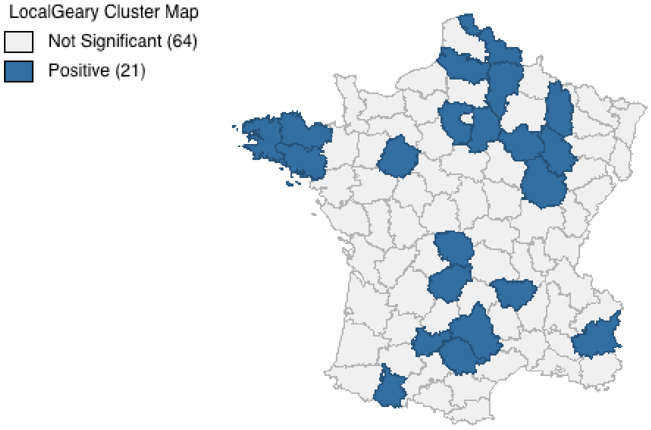

Figure 11: Multivariate Local Geary, six variables (FDR)

Overall, the multivariate measure brings out patterns that are not obvious in its univariate counterparts. There is little overlap in the local patterns of the six individual variables (not shown here), but there seems to be some evidence of grouping along a multivariate dimension. Again, this emphasizes the point that the multivariate measure of attribute similarity is not a simple extrapolation of the univariate measures, but it involves complex trade-offs in all attribute dimensions considered.

5 Conclusion

The transition from a univariate setting to a multivariate context brings interesting challenges to the construction of measures of local spatial autocorrelation. Such statistics represent a mathematical compromise between a measure of attribute similarity and an indication of locational similarity. In one dimension, there are several candidates for attribute similarity, the most commonly used one being the cross product (e.g., in Moran's I). In a multivariate setting, it seems more intuitive to use a concept related to the distance in attribute space (i.e., squared differences) between a point (an observation) and the points that correspond to its geographic neighbors. As is well known from the literature on contiguity constrained spatial clustering, neighbors in attribute space are not necessarily also neighbors in geographic space.

The generalization of the Local Geary c statistic to multiple variables is a way to formalize the combination of attribute similarity and locational similarity. It turns out that the statistic is simply the sum of the individual local statistics for each variable. This corresponds with a notion of the average squared distance in multivariate attribute space to the observations that are neighbors in geographic space, as formalized in a spatial weights matrix. The combination of the different dimensions introduces trade-offs so that the resulting clusters provide insights that differ from the simple overlay of univariate statistics.

The problem of inference in this situation is complex, and may not have a satisfactory solution in the traditional sense. This aspect of computation-driven statistical analysis is also highlighted in the recent work by Efron and Hastie (2016), who suggest that the focus instead should be on identifying "interesting" features. In this spirit, the bivariate extension of the Local Geary in particular may provide insights that do not necessarily follow from multiple univariate analyses. With more that two variables, the curse of dimensionality may affect the usefulness of the approach, and it seems that resorting instead to an analysis of the main principal components may be more fruitful. This remains to be further investigated.

References

Aldstadt, J. and Getis, A. (2006). Using AMOEBA to create a spatial weights matrix and identify spatial clusters. *Geographical Analysis*, 38:327–343.

Anselin, L. (1995). Local indicators of spatial association — LISA. *Geographical Analysis*, 27:93–115.

Anselin, L. (1996). The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In Fischer, M., Scholten, H., and Unwin, D., editors, *Spatial Analytical Perspectives on GIS in Environmental and Socio-Economic Sciences*, pages 111–125. Taylor and Francis, London.

Anselin, L., Syabri, I., and Kho, Y. (2006). GeoDa, an introduction to spatial data analysis. *Geographical Analysis*, 38:5–22.

Benjamin, D. and 72 others (2017). Redefine statistical significance. *Nature Human Behaviour*, 1. Sept. 1, DOI 10.1038/s41562-017-0189-z.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57:289–300.

Bivand, R. S. (2006). Implementing spatial data analysis software in R. *Geographical Analysis*, 38:23–40.

Bivand, R. S., Muller, W., and Reder, M. (2009). Power calculations for global and local Moran's I. *Computational Statistics and Data Analysis*, 53:2859–2872.

Bivand, R. S., Pebesma, E. J., and Gómez-Rubio, V. (2013). *Applied Spatial Data Analysis with R, Second Edition*. Springer, New York, NY.

Boots, B. (2002). Local measures of spatial association. *Ecoscience*, 9:168–176.

Boots, B. (2003). Developing local measures of spatial association for categorical data. *Journal of Geographical Systems*, 5:139–160.

Boots, B. (2006). Local configuration measures for categorical spatial data: Binary regular lattices. *Journal of Geographical Systems*, 8:1–24.

Boots, B. and Okabe, A. (2007). Local statistical spatial analysis: Inventory and prospect. *International Journal of Geographical Information Science*, 21:355–375.

de Castro, M. C. and Singer, B. H. (2006). Controlling the false discovery rate: An application to account for multiple and dependent tests in local statistics of spatial association. *Geographical Analysis*, 38:180–208.

Dray, S. and Jombart, T. (2011). Revisiting Guerry's data: Introducing spatial constraints in multivariate analysis. *The Annals of Applied Statistics*, 5(4):2278–2299.

Dray, S., Saïd, S., and Débias, F. (2008). Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. *Journal of Vegetation Science*, 19:45–56.

Efron, B. (2010). *Large-Scale Inference. Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, Cambridge, UK.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference. Algorithms, Evidence, and Data Science*. Cambridge University Press, Cambridge, UK.

Fotheringham, A. S. (1997). Trends in quantitative methods I: Stressing the local. *Progress in Human Geography*, 21:88–96.

Fotheringham, A. S. and Brunsdon, C. (1999). Local forms of spatial analysis. *Geographical Analysis*, 31:340–358.

Geary, R. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician*, 5:115–145.

Getis, A. (1991). Spatial interaction and spatial autocorrelation: A cross-product approach. *Environment and Planning A*, 23:1269–1277.

Getis, A. and Aldstadt, J. (2004). Constructing the spatial weighs matrix using a local statistic. *Geographical Analysis*, 36:90–104.

Getis, A. and Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24:189–206.

Getis, A. and Ord, J. K. (1996). Local spatial statistics: an overview. In Longley, P. and Batty, M., editors, *Spatial Analysis: Modeling in a GIS Environment*, pages 261–277. GeoInformation International.

Getis, A. and Ord, J. K. (2000). Seemingly independent tests: Addressing the problem of multiple simultaneous and dependent tests. Paper presented, 39th Annual Meeting of the Western Regional Science Association, Kauai, HI.

Good, I. (1983). The philosophy of exploratory data analysis. *Philosophy of Science*, 50:283–295.

Grubesic, T. H., Wei, R., and Murray, A. T. (2014). Spatial clustering overview and comparison: Accuracy, sensitivity, and computational expense. *Annals of the Association of American Geographers*, 104:1134–1156.

Hardisty, F. and Klippel, A. (2010). Analysing spatio-temporal autocorrelation with LISTA-Viz. *International Journal of Geographical Information Science*, 24:1515–1526.

Lee, S.-I. (2001). Developing a bivariate spatial association measure: An integration of Pearson's r and Moran's I. *Journal of Geographical Systems*, 3:369–385.

Lee, S.-I. (2009). A generalized randomization approach to local measures of spatial association. *Geographical Analysis*, 41:221–248.

Lloyd, C. D. (2010). *Local Models for Spatial Analysis, Second Edition.* CRC Press, Boca Raton, FL.

Ord, J. K. and Getis, A. (1995). Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis*, 27:286–306.

Ord, J. K. and Getis, A. (2001). Testing for local spatial autocorrelation in the presence of global autocorrelation. *Journal of Regional Science*, 41:411–432.

Rey, S. J. (2016). Space-time patterns of rank concordance: Local indicators of mobility association with application to spatial income inequality dynamics. *Annals of the American Association of Geographers*, 106:788–803.

Rey, S. J. and Anselin, L. (2007). PySAL, a Python library of spatial analytical methods. *The Review of Regional Studies*, 37(1):5–27.

Rogerson, P. A. (2010). Optimal geograpahic scales for local spatial statistics. *Statistical Methods in Medical Research*, 20:119–129.

Rogerson, P. A. (2015). Maximum Getis-Ord statistic adjusted for spatially autocorrelated data. *Geographical Analysis*, 47:20–33.

Rogerson, P. A. and Kedron, P. (2012). Optimal weights for focused tests of clustering using the local moran statistic. *Geographical Analysis*, 44:121–133.

Sokal, R. R., Oden, N. L., and Thompson, B. A. (1998a). Local spatial autocorrelation in a biological model. *Geographical Analysis*, 30:331–354.

Sokal, R. R., Oden, N. L., and Thompson, B. A. (1998b). Local spatial autocorrelation in biological variables. *Biological Journal of the Linnean Society*, 65:41–62.

Thomas, J. and Cook, K. (2005). *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE Computer Society Press, Los Alamitos, CA.

Tiefelsdorf, M. (2002). The saddlepoint approximation of Moran's I and local Moran's I_i's reference distribution and their numerical evaluation. *Geographical Analysis*, 34:187–206.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison Wesley, Reading, MA.

Unwin, A. (1996). Exploratory spatial analysis and local statistics. *Computational Statistics*, 11:387–400.

Unwin, A. and Unwin, D. (1998). Exploratory spatial data analysis with local statistics. *The Statistician*, 47(3):415–421.

Wartenberg, D. (1985). Multivariate spatial correlation: A method for exploratory geographical analysis. *Geographical Analysis*, 17:263–283.

Yamada, I. and Thill, J.-C. (2007). Local indicators of network-constrained clusters in spatial point patterns. *Geographical Analysis*, 39:268–292.

View publication stats