Geographical Analysis ISSN 0016-7363

Controlling the False Discovery Rate: A New Application to Account for Multiple and Dependent Tests in Local Statistics of Spatial Association

Marcia Caldas de Castro[1], Burton H. Singer[2]

[1]Department of Geography, University of South Carolina, Columbia, SC, [2]Office of Population Research, Princeton University, Princeton, NJ

Assessing the significance of multiple and dependent comparisons is an important, and often ignored, issue that becomes more critical as the size of data sets increases. If not accounted for, false-positive differences are very likely to be identified. The need to address this issue has led to the development of a myriad of procedures to account for multiple testing. The simplest and most widely used technique is the Bonferroni method, which controls the probability that a true null hypothesis is incorrectly rejected. However, it is a very conservative procedure. As a result, the larger the data set the greater the chances that truly significant differences will be missed. In 1995, a new criterion, the false discovery rate (FDR), was proposed to control the proportion of false declarations of significance among those individual deviations from null hypotheses considered to be significant. It is more powerful than all previously proposed methods. Multiple and dependent comparisons are also fundamental in spatial analysis. As the number of locations increases, assessing the significance of local statistics of spatial association becomes a complex matter. In this article we use empirical and simulated data to evaluate the use of the FDR approach in appraising the occurrence of clusters detected by local indicators of spatial association. Results show a significant gain in identification of meaningful clusters when controlling the FDR, in comparison to more conservative approaches. When no control is adopted, false clusters are likely to be identified. If a conservative approach is used, clusters are only partially identified and true clusters are largely missed. In contrast, when the FDR approach is adopted, clusters are fully identified. Incorporating a correction for spatial dependence to conservative methods improves the results, but not enough to match those obtained by the FDR approach.

Correspondence: Marcia Caldas de Castro, Department of Geography, University of South Carolina, 125 Callcott Hall, Columbia, SC 29208
e-mail: mcaldas@sc.edu

Submitted: July 1, 2004. Revised version accepted: August 18, 2005.

Introduction

Tobler's First Law of Geography says that "everything is related to everything else, but near things are more related than distant things" (Tobler 1979). This law applies to any phenomena that have a spatial nature, with considerable implications for studies in disciplines such as sociology, demography, economics, epidemiology, urban planning, ecology, biology, archeology, and, of course, geography. The statistical investigation of these phenomena has been called spatial data analysis (Bailey and Gatrell 1995). The objectives are identification of the spatial distribution of the data, spatial patterns, and the occurrence of outliers (Anselin 1996). A spatial arrangement can be clustered, dispersed, or random depending on the observed spatial dependence (also referred to as spatial autocorrelation or spatial association). Measures of spatial association can be global or local. Global measures consider all available locations simultaneously, utilizing a single statistic that summarizes the spatial pattern. However, the larger the number of locations, the less will be the interpretability of the statistic, as a spatial pattern can vary substantially by location. Local measures represent the association between each location and its neighbors based on defined distances. One statistic is provided for each location, facilitating the identification of clusters, testing of stationarity assumptions, and inference about distances over which spatial association occurs (Getis and Ord 1996). Anselin (1995) proposed criteria to classify a statistic within a class of local indicators of spatial association (LISA).

Local statistics rely on tests of spatial association for each location in the data, and the issue of multiple comparisons is a concern when assessing their significance (Kurtz et al. 1965; Miller 1981; Tukey 1991). In other words, if multiple inferences (tests) are drawn from a given data set, the selection of statistically significant effects/differences is carried out utilizing formal multiple comparison methods. As a result, the probability that some differences will be declared significant by chance alone cannot be neglected and needs to be controlled. Multiplicity arises in situations where more than one comparison/test is to be evaluated. In this setting, the Type I error rate is the probability of rejecting one *or more* null hypotheses when each one is, in fact, true. This overall simultaneous error rate will frequently exceed, often substantially, the nominal Type I error rate, α, for a single comparison/test. Historically, a standard criterion for significance when multiple tests are carried out is the demand that the probability of any single false positive among all tests carried out is at most 0.05. This strict criterion has been used primarily in studies where only a few comparisons are expected to yield meaningful differences, and the Bonferroni adjustment is a simple and trustworthy procedure for assuring *simultaneously* that the probability of any single Type I error is no greater than α.

In the context of spatial analyses in geography, where hundreds, or even thousands, of comparisons are to be carried out, using a procedure that guards against any single false positive occurring is often going to be much too strict and will lead to many missed meaningful findings. This issue is not unique to geography. Indeed, it has already led to extensive investigation and use of alternative criteria and multiple comparison adjustment strategies in genomewide studies in molecular genetics (Storey and Tibshirani 2003), functional magnetic resonance imaging in neurobiology, complex plant breeding studies (Basford and Tukey 1999), and analyses of state-to-state differences in educational achievement (Williams, Jones, and Tukey 1999).

In addition to the multiple comparisons issue, local statistics are calculated on the basis of a defined neighborhood determined by a selected distance, and the results for locations containing common neighbors are likely to be correlated (Anselin 1995; Getis and Ord 1996; Rogerson 2001). Therefore, although largely ignored in spatial data analysis research, a correction procedure to account for both multiple and dependent tests is recommended. Up to now, the most common approach has been the use of classical—and overly conservative—multiple comparison procedures (MCPs) such as Bonferroni and Sidak corrections (Anselin 1995; Getis and Ord 1996, 2000). A recent application adopted a sharper Bonferroni correction based on a stepwise procedure (Paez, Uchida, and Miyamoto 2002), which provides slightly less conservative results (Hochberg 1988; Hommel 1988; Liu 1996). Regarding spatial dependency, Getis and Ord (2000) proposed a method to account for common neighbors shared by nearby locations.

In this article we evaluate for the first time the use of a multiple testing procedure introduced by Benjamini and Hochberg (1995) for the purposes of assessing the significance of statistics of local association. The procedure is a useful compromise between the rigidity of assuring simultaneously—via, for example, the Bonferroni approach—that the probability of any single type I error is no greater than α, and the lack of control associated with comparisons unadjusted for multiplicity. It is called the false discovery rate (FDR), and controls the average rate that declarations of significance are truly nonsignificant (Benjamini and Hochberg 2000). Recent applications of FDR as a controlling procedure are in state-of-the-art genome research (including DNA microarray analysis) and neuroimaging, both entailing extremely large data sets (Storey and Tibshirani 2001, 2003; Efron and Tibshirani 2002; Genovese, Lazar, and Nichols 2002). We also extend the Getis and Ord (2000) spatial dependence method to conservative procedures to assess if their performance improves significantly so that they would provide results similar to those obtained with the FDR approach.

The FDR procedure naturally leads to more ''significant'' findings than would be the case using a conservative approach such as Bonferroni. Whether or not these are meaningful findings—or discoveries —depends upon the particular scientific context. To acquire some insight about the use of the FDR in geography, we apply it to two kinds of problems. First, we study a real data situation—malaria on the Amazon frontier—involving many comparisons and where we understand enough about the science to know whether or not we are getting meaningful results. We also compare our FDR-based analyses with what is learned, or not, using more conventional criteria. Second, we simulate a spatial data set, where we obviously know, in advance, what is the signal and what is the noise, and compare the performance of the different MCP approaches.

Local statistics: multiple and dependent tests

The class of local statistics was first proposed by Getis and Ord (1992). Anselin (1995), while further developing these statistics, stated that they had two main purposes: (i) detection of local pockets of nonstationarity, which may be a result of atypical observations (outliers) or the presence of different spatial regimes, and (ii) identification of significant local clusters. LISA statistics (Anselin 1995), are a powerful tool for exploratory spatial data analysis (ESDA). In general matrix notation, they can be expressed as

$$\Gamma_i = \sum_{j}^{n} w_{ij} y_{ij} \tag{1}$$

where $\mathbf{\Gamma}$ is a particular measure of spatial association, n is the total number of locations, w_{ij} are the elements of a weight matrix $\mathbf{W}$ that characterizes the relationship between location i and the other locations j, and y_{ij} are the elements of a matrix $\mathbf{Y}$ representing the interactions between location i and the other locations j (Getis and Ord 1996). The possible interactions represented in the $\mathbf{Y}$ matrix can be expressed as an addition, subtraction, multiplication, division, covariance, or combinations of the first four types. Each one will generate a different kind of local statistic. Getis and Ord (1996) highlight six statistics to measure local association: Moran's I_i—based on covariance; Geary's c_i, K_{1i}, and K_{2i}—based on differences; and $G_i(d)$ and $G_i^*(d)$—based on additive interactions. The first two are considered as part of the LISA statistics as defined by Anselin (1995), while the remaining are generally accepted as an overall class of LISA indicators (Getis and Ord 1996). Inference regarding the significance of Moran's I_i, Geary's c_i, and $G_i(d)$ and $G_i^*(d)$ are made based on a null hypothesis of no spatial association between the realization of a variable at location i and its neighbors, and tests for each location are compared with critical values of the normal distribution.

In this article we focus on Moran's I_i, $G_i(d)$, and $G_i^*(d)$. All these statistics facilitate the identification of a clustering pattern. The $G_i^*(d)$ statistic, in particular, indicates the presence of clusters of high or low values surrounding a particular location i within a radius of distance d from i. Considering an area divided into n locations, each identified with a point i and associated with a value x_i (a realization of a random variable X), the $G_i^*(d)$ statistic is defined as (Getis and Ord 1992)

$$G_i^*(d) = \frac{\sum_{j=1}^{n} w_{ij}(d) x_j}{\sum_{j=1}^{n} x_j} \tag{2}$$

where w_{ij} are the elements of a weight matrix. Ord and Getis (1995) generalized Equation (2) so that $G_i^*(d)$ values are given as standard normal variates ($Z[G_i^*(d)]$). Under the null hypothesis ($Z[G_i^*(d)]$) are asymptotically normally distributed,

$N(0,1)$, as $n \to \infty$ (Getis and Ord 1992). Therefore, significant negative $Z[G_i^*(d)])$ reveals spatial clustering of low values of X within distance d, while significant positive $Z[G_i^*(d)])$ are indicative of spatial clustering of high values of X within distance d. Choosing the right distance is a key factor. When d is very small or very large (covering the total area) normality is lost. Getis and Ord (1996) suggest that the maximum distance should never exceed 1/2 of the shorter side of the study area, while the number of neighbors should be at least 30 for large samples and 8 for small ones. However, there is no definite rule to guide the decision. It is also important to mention that edge cells are not a cause of concern, unless the distance chosen results in a very small number of neighbors for those cells, which could compromise the convergence of the statistic to normality.

The $G_i(d)$ statistic is analogous to the $G_i^*(d)$, with the difference that the value of location i is not included in the sum. In other words, $G_i(d)$ is defined as Equation (2) for $j \neq i$. It is particularly useful for spread and diffusion studies. Moran's I_i is defined as (Anselin 1995)

$$I_i = (z_i/s^2)\sum_{j=1}^{n} w_{ij}z_j, \quad j \neq i \tag{3}$$

where w_{ij} are the elements of a weight matrix; z_i and z_j are deviations from the mean, $(x_i - \bar{x})$ and $(x_j - \bar{x})$, respectively; and $s^2 = \sum_{i=1}^{n} z_i^2/n$. Standard normal variates for the statistic, $Z(I_i)$, are also computed (Anselin 1995) and the assessment of significance is based on a normal distribution. Significant high values of $Z(I_i)$ indicate a cluster of similar values (either high or low), while significant low values of $Z(I_i)$ indicate a cluster of dissimilar values.

Assessing the significance of any of these local statistics requires careful attention. First, each location i is assigned a test and the rejection or retention of the null hypothesis raises questions of multiplicity. Second, local statistics have two

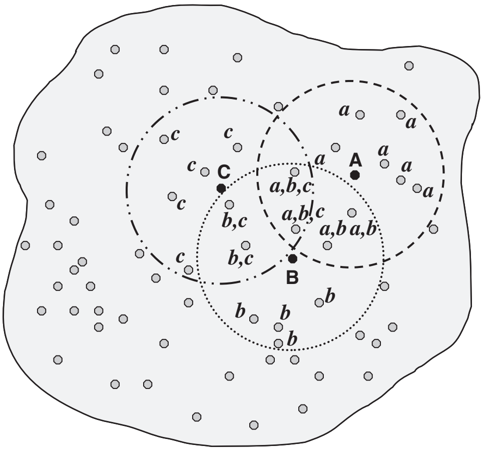

Figure 1. Selected points and neighborhood at a distance d for an irregular grid.

possible sources of spatial dependence: (i) geometric, which is caused by the fact that nearby locations share common elements in the neighborhood defined by the weight matrix, and (ii) true dependence that might exist between the values of nearby locations. Fig. 1 illustrates the geometric dependence (overlap) for an irregular gridded area. Three points, *A*, *B*, and *C*, were selected, and a circular neighborhood of radius *d* was drawn around each of those points. The observations within this distance from points *A*, *B*, and *C* are labeled *a*, *b*, and *c*, respectively. Observations labeled with more than one letter are within the neighborhood of multiple points. In this scenario, the greater the number of common neighbors (overlap), the greater will be the dependence or correlation between the tests. Therefore, procedures to account for multiplicity and spatial dependency are required to properly assess the significance of local statistics.

Traditional procedures to adjust for multiplicity

Testing multiple hypotheses is a problem encountered by scientists in diverse fields. As Dunnett and Tamhane (1992) noted, there are at least three schools of thought regarding the need for correcting for multiplicity. The first, called the comparison-wise error rate approach, postulates that no correction is needed, and each test should be assessed for significance separately. The second, known as experiment-wise or family-wise error rate (FWER—the term family stands for the set of statistical tests whose error rate needs to be controlled) approach, suggests that the probability that a true null hypothesis is incorrectly rejected (called type I error in hypothesis testing; type II error is when a false null hypothesis fails to be rejected) is controlled at a specific significance level α. The third, Bayesian decision–theoretic approach, evaluates type I and type II errors using assumptions about the probability distribution of unknown parameters. Most recently, a fourth school of thought emerged, as we will discuss in the next section of this article. It is similar to the experiment-wise approach, but seeks to control the FDR, instead of the type I error (Benjamini and Hochberg 1995).

MCPs seek to control the FWER, which can be understood as the probability that a type I error does occur among all hypotheses being tested. Considering F as the number of tests for which a type I error occurred, FWER is simply expressed as FWER $= Pr(F \geq 1)$. MCP procedures control the FWER for all n tests, at significance level α (for a two-sided test the appropriate critical value should be evaluated at $\alpha/2n$) (Jones, Lewis, and Tukey 2001). The simplest procedure, but also most conservative (especially when the tests are highly correlated), is the Bonferroni method. It evaluates the significance of the test statistics at a critical probability value (p_{critical}) set equal to α/n, where α is the overall type I error rate for the data. All test statistics whose probability values (p) satisfy the condition $p \leq p_{\text{critical}} = \alpha/n = p_{\text{BON}}$ are considered significant (null hypothesis is rejected) (Sankoh, Huque, and Dubey 1997). Different stepwise modifications to the Bonferroni method were proposed, such as those of Holm (1979), Simes (1986), Hommel (1988), and Hochberg

(1988). All provided more powerful results than the Bonferroni method and proved to be more appropriate under certain distributions and correlation profiles (Hommel 1989; Brown and Russell 1997; Sankoh, Huque, and Dubey 1997).

Sidak (1967, 1968, 1971) proposed a procedure that proved to be more powerful than the original Bonferroni when the test statistics are independent and uniformly distributed. The Sidak correction also controls for the overall probability of type I error, but with critical values appraised at a level $1 - (1 - \alpha)^{1/n}$. Therefore, a test is considered significant when $p \leq p_{\text{critical}} = 1 - (1 - \alpha)^{1/n}$.

Tukey, Ciminera, and Heyse (1985) proposed another modification that specifically accounts for high correlation among the tests. Significance is evaluated at $p \leq p_{\text{critical}} = 1 - (1 - \alpha)^{1/\sqrt{n}}$. A similar approach was suggested by Dubey and Armitage–Parmar (Sankoh, Huque, and Dubey 1997) in which the significance of the test statistics is evaluated at $p \leq p_{\text{critical}} = 1 - (1 - \alpha)^{1/w_k}$, where $w_k = n^{(1-r_{.k})}$, $r_{.k} = (n-1)^{-1}\sum_{i \neq k}^{n} r_{ik}$, and r_{ik} is the correlation between tests i and k. If the tests are not correlated, the procedure is equal to the Sidak correction. Instead, if the tests are perfectly correlated, the procedure reduces to the unadjusted comparison-wise error rate approach. Finally, if the correlation is equal to 0.5 the procedure is equivalent to the Tukey, Ciminera, and Heyse (1985) method.

These procedures are not an exhaustive list of all approaches ever proposed to control the FWER. However, some lessons can be taken from these generic methods. First, the Bonferroni method is the most conservative approach in current use. Second, if one wishes to control the FWER there is not a single best procedure that can be used for any kind of data set, and the recommended approach will depend on the distribution and level of correlation observed in the test statistics. Finally, although some of the procedures described in this section do show more power than the Bonferroni method, family-wise approaches usually produce conservative results.

An alternative procedure to adjust for multiplicity

Instead of controlling for the probability that a true null hypothesis is incorrectly rejected among all possible hypotheses, one may want to control the proportion of false declarations of significance among those individual deviations from null hypotheses considered to be significant (FDR). This approach is particularly useful in exploratory data analysis. Often in these situations, the main goal is to identify as many significant cases as possible, such as testing different treatments for a disease, evaluating multiple chemical components for drug development, and selecting some key genes (among thousands) for further investigation during microarray experiments. Similarly, in the particular case of local statistics, one wants to be able to identify as many locations with a significant local spatial association as possible. For example, assume that a local statistic is applied to disease data as a surveillance tool. The main goal is to check for the presence of clusters and indicate the areas where interventions should be immediately put into place. In this case, it is crucial to be able to identify as many significant clusters as possible, so that a disease outbreak can be effectively curtailed. Another example is the use of local statistics to identify clusters of crime in a region. Public officials interested in reducing crime rates want to be able to identify as many potential critical areas as possible.

Benjamini and Hochberg (1995) proposed a procedure that addresses this issue. Assume that there are m hypotheses to be tested. Among those, R will be declared significant (null hypothesis rejected), F are false positives (null hypothesis incorrectly rejected, or type I error), and S are the true positives (null hypothesis correctly rejected). Moreover, consider a variable Q defined as the proportion of null hypotheses incorrectly rejected among all those that were rejected. This variable can be expressed as $Q = F/(F+S)$ or $Q = F/R$, and, by definition, when $R = 0$ the variable Q is set to equal zero. Benjamini and Hochberg (1995) define the FDR as the expected value of variable Q. For independent tests it can be controlled for each test at a level α by the following stepwise procedure: (i) order the test statistics p-values (p_i) in ascending order ($p_1 \leq p_2 \leq \ldots \leq p_m$); (ii) starting from p_m find the first p_i for which $p_i \leq (i/m)\alpha$; and (iii) regard all tests as significant for which $p_i \leq p_{\text{critical}} = (i/m)\alpha = p_{\text{FDR}}$. Therefore, if p_{critical} equals 5% it means that, on average, among the rejected null hypotheses, 5% were truly null (Storey and Tibshirani 2003). The gain in power provided by this procedure becomes larger as m increases. As one would have expected, the compromise between Bonferroni critical values and unadjusted values is exemplified by the inequalities $p_{\text{BON}} \leq p_{\text{FDR}} \leq p_{\text{UNA}}$. Using simulated data, Williams, Jones, and Tukey (1999) concluded that the Benjamini and Hochberg (1995) procedure is the best available choice to correct for multiplicity. A similar positive view was reached for the analysis of neuroimaging data (Genovese, Lazar, and Nichols 2002).

An adaptive procedure was also proposed to improve the control of the FDR (also assuming that the tests are independent) (Benjamini and Hochberg 2000). The innovative feature of this adaptive procedure is the estimation of the number of true null hypotheses. The nonadaptive and the adaptive procedures show very similar results for cases when the number of true null hypotheses is large. Additionally, the issue of dependence among tests was addressed, and the Benjamini and Hochberg (1995) procedure was shown to control the FDR when the test statistics present positive regression dependence (Benjamini and Yekutieli 2001). A further development, the positive FDR ($_p$FDR) was proposed to account for the fact that an error rate will be estimated only when at least one hypothesis is declared significant, or $_p\text{FDR} = E[(F/R)|R > 0]$. The main difference between this approach and the Benjamini and Hochberg (1995) procedure is that the rejection region determined by P_{critical} is fixed, and then the significance level α is estimated (Storey 2002). The procedure was also shown to perform well under certain types of dependence in the test statistics (Storey and Tibshirani 2001). Additionally, a Bayesian approach to FDR and $_p$FDR control was introduced by Efron and Tibshirani (2002) and Storey (2003), respectively. Finally, Storey, Taylor, and Siegmund (2004) provide evidence that the FDR and the $_\text{P}$FDR are asymptotically equivalent.

Procedures to account for spatial dependence

Most procedures proposed to adjust for multiple comparisons assume that the test statistics are independent. The Tukey, Ciminera, and Heyse (1985) procedure, as previously detailed, was specifically formulated for cases of multiplicity when the tests were highly correlated. Yekutieli and Benjamini (1999) proposed a modified procedure to control the FDR for highly correlated test statistics, although it was proven that, in certain types of dependence, such as ergodic dependence, some mixing distributions, and positive regression dependence, it is still possible to control the FDR (Benjamini and Yekutieli 2001; Storey and Tibshirani 2001; Storey, Taylor, and Siegmund 2004).

In the context of local statistics, however, a more specific assessment is required. As previously highlighted, there are two sources of spatial dependence for tests on local statistics. The true dependence between the values for various nearby locations is expressed by the correlation structure of the local statistic. Ord and Getis (1995) described the correlation structure of the $G_i^*(d)$ statistic for a regular grid scenario. The correlation between the values of two locations can be negative or positive, depending on the number of observations and on the number of common neighbors that the locations share. As Anselin (1995) points out, the use of MCPs becomes too conservative as the number of tests, n, increases, as tests for locations further apart, which share no neighbors, are independent. Considering the statistical properties of the FDR, we do expect that this source of dependence can be controlled, although this will not necessarily always be true. Further research is still needed on this issue.

Regarding the geometric source of dependency (overlap), Getis and Ord (2000) proposed a method to address this issue based on the number of seemingly independent tests. Assume that among the n tests there are ν independent spatial clusters, each containing u observations perfectly correlated within the cluster, then $n=\nu u$. Additionally, as in the case of dependent tests the correlation, r, can be expressed by the overlap between nearby tests, then $\nu = n - r(n-1)$. Therefore, if there is perfect correlation ($r=1$) then there is only one independent spatial cluster ($\nu = 1$). Additionally, the lower bound for the correlation r is $-1/(n-1)$.

Using this rationale, Getis and Ord (2000) suggested a modification of the Sidak and Bonferroni procedures to account for the overlap, where the number of tests n is replaced by the number of independent spatial tests ν. Therefore, the critical probability value for the Sidak correction is appraised at level $1-(1-\alpha)^{1/\nu}$, and the Bonferroni correction at level α/ν. The implementation of this strategy is straightforward, as the proportion of overlap between tests given a certain distance can be easily assessed from the data. In fact, if the same weight matrix is used for different LISA, then the same overlap pattern, and therefore the same spatial dependence, is expected. Although the Getis and Ord (2000) suggestion can be easily incorporated in conservative MCPs, its extension to FDR procedures may result in controlling the FDR at levels above α. Further research is needed to address the impact of the overlap on highly dependent points (placed in either regular or irregular grids).

An example with empirical data: malaria transmission in the Brazilian Amazon

The main purpose of presenting this example is to emphasize the implications of adopting extreme decisions regarding the assessment of significance. On the one hand, the issue of multiple comparisons can be ignored and all tests that fail to conform to the null hypothesis are accepted as significant. On the other hand, extremely conservative methods can be used to control the error rate and result in very few significant tests. In a scenario of disease surveillance and control, the former would imply spending a large amount of human and financial resources unnecessarily and inefficiently, while the latter would result in a major failure to curb the spread of the disease. As we show below, the FDR procedure is a better and efficient alternative to those extreme decisions.

Data on malaria rates were collected in a settlement project located in the state of Rondônia (western part of the Brazilian Amazon). The project, called Machadinho, is physically divided by a river into two tracts: tract one located to the south of the river and tract two to the north. These two tracts comprise a total of 1742 plots, which are the sections of land assigned to settlers. Additionally, protected forest reserves are placed across the area of the project. Data collection started as soon as settlers moved to Machadinho in 1985, with follow-up surveys carried out in 1986, 1987, and 1995. All settlers who were effectively occupying their plots were interviewed. That means that those plots whose owners did not clear any forest area or did not live at least part-time in Machadinho were not included in the survey (Sawyer 1985). Therefore, the number of observations for each year is not the same, as shown in Table 1.

Due to a multitude of factors, detailed in Castro (2002) and Singer and Castro (2001), the area soon became high risk for malaria (Table 1). In 1985, 65.7% of the population had malaria at least once, and this number jumped to 90.1% in the next year. Also in 1986, 55.9% of people had malaria episodes in more than 5 months of the year (Sydenstricker 1992), and almost 40% of cases registered in Rondônia were

Table 1 Number of Plots and People Surveyed, Person-Months Exposed to the Disease, Malaria Cases, and Exposure Weighted Malaria Illness Rate—Machadinho (1985/95)

Year	Total number of plots surveyed	Total number of people surveyed	Number of person-months	Malaria cases	Malaria rate (%)
1985	267	1366	4587	1041	22.69
1986	545	2736	24,938	8006	32.10
1987	740	3982	38,121	9012	23.64
1995	954	5278	59,437	3939	6.63

observed in Machadinho (Sawyer and Sawyer 1987). Transmission, however, was very focal, and in certain areas some people reportedly suffered considerably more severely than others. Malaria risk is measured by the exposure-weighted malaria illness rate (referred to simply as malaria rate). Its numerator records the number of months each settler had malaria during a year, and the denominator records the exposure time—number of person-months exposed to the disease, as shown in Table 1 (Sawyer 1988).

We use LISA to check for the presence of clusters of high or low malaria rates in Machadinho. For that purpose, plots were considered as the spatial unit of analysis, and malaria rates assigned to their centroids. The definition of the neighborhood around each plot, however, needs special attention. The spread of diseases does not respect or follow any predefined borders. The most appropriate neighborhood of potential exposure is a function of a dynamic process between men, mosquitoes, and the local environment, which determines a lower or higher risk of malaria transmission. Failure to account for this process will most likely result in neighborhoods that have little use, if any, for controlling the disease. This problem is not specific to disease studies. For example, modeling the dynamics of the lion population in the Serengeti National Park depends on key distances that represent the interaction of lions and wildebeests, and considering exclusively the behavior of lions would result in a poor model (Packer et al. 2005).

In our empirical example, the distance *d* that defines the neighborhood around each plot was selected based on three factors associated with both the area and phenomena under study. The first factor is the flying behavior of mosquitoes, which ranges from 500 to 3000 m without the aid of the wind, and up to 5000 with the aid of the wind (Deane 1947; Cova-Garcia 1961; Van Thiel 1962). The second factor is the size of the plots: an average front of 400–500 m and an average depth of 700–900 m. An extensive analysis of buffers sized between 500 and 8000 m suggests that distances lower than 2000 m would result in a very small number of neighbors for each plot, with a large number of them being left as "islands" with no neighbors (Castro 2002). The third factor is related to the implementation of control measures by the local health agencies, which is done by sectors. An analysis of the distribution of sectors suggests that distances larger than 2000 m would be more appropriate (Castro 2002). Additionally, 3500 m is the minimum distance necessary to guarantee that all plots have at least one neighbor in all 4 years.

It is important to mention that different approaches have been used to choose the most appropriate neighborhood, such as a *k*-nearest neighbors method in which the number of neighbors is fixed (Baumont, Ertur, and Le Gallo 2004); using the average distance between nearest neighbors as a reference (Paez, Uchida, and Miyamoto 2001); comparison between the number of neighbors in a first-order contiguity weights matrix and a matrix defined by a certain distance *d* (Paez, Uchida, and Miyamoto 2001); and evaluating the most effective distance to remove the spatial autocorrelation from the data through the use of a filtering procedure (Getis 1995; Paez, Uchida, and Miyamoto 2001). The first three approaches are not appropriate for our data as information is not available for all 1742 plots. Some plots have no contiguous neighbor, and in those cases a fixed number of neighbors could result in a very large neighborhood with no useful meaning for purposes of malaria transmission. The last approach would require adoption of different distances for each of the 4 years of data we have. However, to guarantee temporal comparability of the outcomes of the dynamic process of malaria transmission we opted to use a unique distance.

Using 3500 m as the critical distance, the proportion of overlap between each pair of plots does not vary dramatically over time. As shown in Table 2, an overlap of 45% was observed on average each year, despite the differences in the total number of plots included in the data set. The associated standard deviation was small in magnitude and roughly the same over time. The last row of Table 2 shows the number of independent spatial clusters (ν), as proposed by Getis and Ord (2000). The difference between the total number of plots and ν is directly reflected in the critical p value to test for significance, highlighting the importance of taking into account the spatial dependence.

The calculations of $G_i(d)$ and $G_i^*(d)$ were done in Point Pattern Analysis (PPA; Aldstadt, Chen, and Getis 1998), a program developed at San Diego State University (http://www.nku.edu/~longa/cgi-bin/cgi-tcl-examples/generic/ppa/ppa.cgi), while the calculations of Moran's I_i were performed in GeoData Analysis Software (GeoDA^{TM}; http://sal.agecon.uiuc.edu/geoda_main.php), developed by Luc Anselin at the University of Illinois. All calculations to assess the significance of the local statistics were implemented in a spreadsheet, and the results are summarized in Table 3.

Initially, a comparison-wise error rate was used, with no correction for multiple testing. Independent of the size of the data set, the critical p value and the cutoff (z_{critical}) for significance are the same for all years. Although one can say that this is the least conservative method, it is also a procedure that necessarily will result in false positives (type I errors), and therefore should be avoided.

Two results shown in Table 3 highlight the problems of not correcting for multiple testing. For example, no clusters of low rates were identified by the $G_i(d)$ and $G_i^*(d)$ statistics in 1985 after multiplicity was controlled, which indicates that all

Table 2 Average and Standard Deviation of the Proportion of Overlap Between $G_i^*(d)$ Statistical Tests on Malaria Rates, and Number of Independent Spatial Clusters—Machadinho (1985/95)

Summary statistics of overlap	Year			
	1985	1986	1987	1995
Average	0.46572	0.44768	0.45237	0.45943
Standard deviation	0.03192	0.02748	0.02605	0.02663
Total # of locations	267	545	740	954
ν	143.12	301.46	405.70	516.16

Table 3 Significance Testing of the $G_i(d)$, $G_i^*(d)$, and Moran's I_i Statistics Under Different Multiple Comparison Procedures—Machadinho (1985/95)

Year	Unadjusted	Correcting for multiplicity*			Correcting for multiplicity and spatial dependence†		Recovery ratio‡
		Bonferroni	Sidak	FDR	Bonferroni	Sidak	
1985 (N = 267)							
$G_i^*(d)$							
Accept null	206	267	267	262	266	266	
Reject null	61	0	0	5	1	1	
Cluster high rates	27	0	0	5	1	1	0.082
Cluster low rates	34	0	0	0	0	0	
p_{critical}	0.025	0.0000936	0.0000948	0.0004657	0.0001747	0.0001769	
z_{critical}	± 1.95996	± 3.73560	± 3.73243	± 3.31045	± 3.57565	± 3.57236	
$G_i(d)$							
Accept null	209	266	266	265	265	265	
Reject null	58	1	1	2	2	2	
Cluster high rates	28	1	1	2	2	2	0.018
Cluster low rates	30	0	0	0	0	0	
p_{critical}	0.025	0.0000936	0.0000948	0.0001518	0.0001747	0.0001769	
z_{critical}	± 1.95996	± 3.73560	± 3.73243	± 3.61229	± 3.57565	± 3.57236	
Moran's I_i							
Accept null	206	267	267	247	267	267	
Reject null	61	0	0	20	0	0	
Cluster high rates	17	0	0	3	0	0	0.328
Cluster low rates	44	0	0	17	0	0	
p_{critical}	0.05	0.0001873	0.0001921	0.0040000	0.0003494	0.0003583	
z_{critical}	1.64485	3.55741	3.55072	2.65207	3.39008	3.38312	

1986 (N = 545)							
$G_i^*(d)$							
Accept null	323	523	523	398	516	516	
Reject null	222	22	22	147	29	29	
Cluster high rates	90	17	17	51	19	19	0.625
Cluster low rates	132	5	5	96	10	10	
p_{critical}	0.025	0.0000459	0.0000465	0.0064861	0.0000829	0.0000840	
z_{critical}	± 1.95996	± 3.91145	± 3.90840	± 2.48453	± 3.76604	± 3.76289	
$G_i(d)$							
Accept null	335	530	530	404	522	522	
Reject null	210	15	15	141	23	23	
Cluster high rates	83	11	11	52	15	15	0.646
Cluster low rates	127	4	4	89	8	8	
p_{critical}	0.025	0.0000459	0.0000465	0.0063041	0.0000829	0.0000840	
z_{critical}	± 1.95996	± 3.91145	± 3.90840	± 2.49465	± 3.76604	± 3.76289	
Moran's I_i							
Accept null	347	545	545	405	545	545	
Reject null	198	0	0	140	0	0	
Cluster high rates	64	0	0	40	0	0	0.707
Cluster low rates	134	0	0	100	0	0	
p_{critical}	0.05	0.0000917	0.0000941	0.0160000	0.0001659	0.0001701	
z_{critical}	1.64485	3.74073	3.73432	2.14441	3.58918	3.58254	

Table 3 Continued

Year	Unadjusted	Correcting for multiplicity*			Correcting for multiplicity and spatial dependence†		Recovery ratio‡
		Bonferroni	Sidak	FDR	Bonferroni	Sidak	
1987 (N = 740)							
$G_i^*(d)$							
Accept null	387	665	664	436	650	649	
Reject null	353	75	76	304	90	91	
Cluster high rates	142	20	21	116	28	28	0.824
Cluster low rates	211	55	55	188	62	63	
p_{critical}	0.025	0.0000338	0.0000342	0.0100272	0.0000616	0.0000624	
z_{critical}	± 1.95996	± 3.98469	± 3.98169	± 2.32533	± 3.83958	± 3.83648	
$G_i(d)$							
Accept null	392	675	675	437	659	659	
Reject null	348	65	65	303	81	81	
Cluster high rates	136	16	16	116	21	21	0.841
Cluster low rates	212	49	49	187	60	60	
p_{critical}	0.025	0.0000338	0.0000342	0.0101780	0.0000616	0.0000624	
z_{critical}	± 1.95996	± 3.98469	± 3.98169	± 2.31972	± 3.83958	± 3.83648	
Moran's I_i							
Accept null	439	740	740	479	740	740	
Reject null	301	0	0	261	0	0	
Cluster high rates	100	0	0	83	0	0	0.867
Cluster low rates	201	0	0	178	0	0	
p_{critical}	0.05	0.0000676	0.0000693	0.0220000	0.0001232	0.0001264	
z_{critical}	1.64485	3.81691	3.81061	2.01409	3.66588	3.65936	

1995 (N = 954)							
$G_i^*(d)$							
Accept null	452	904	904	568	900	900	
Reject null	502	50	50	386	54	54	
Cluster high rates	209	50	50	178	54	54	0.743
Cluster low rates	293	0	0	208	0	0	
p_{critical}	0.025	0.0000262	0.0000265	0.0101057	0.0000484	0.0000490	
z_{critical}	± 1.95996	± 4.04461	± 4.04165	± 2.32240	± 3.89830	± 3.89525	
$G_i(d)$							
Accept null	466	906	906	582	900	900	
Reject null	488	48	48	372	54	54	
Cluster high rates	203	48	48	174	54	54	0.736
Cluster low rates	285	0	0	198	0	0	
p_{critical}	0.025	0.0000262	0.0000265	0.0097056	0.0000484	0.0000490	
z_{critical}	± 1.95996	± 4.04461	± 4.04165	± 2.33754	± 3.89830	± 3.89525	
Moran's I_i							
Accept null	452	954	954	499	954	954	
Reject null	502	0	0	455	0	0	
Cluster high rates	119	0	0	102	0	0	0.906
Cluster low rates	383	0	0	353	0	0	
p_{critical}	0.05	0.0000524	0.0000538	0.0280000	0.0000969	0.0000994	
z_{critical}	1.64485	3.87915	3.87294	1.91104	3.72705	3.72061	

*Original formulations. †Using the proportion of overlap between tests proposed by Getis and Ord (2000). ‡To specify the recovery ratio, let S_{BON}, number of significant tests using the Bonferroni correction, S_{FDR}, number of significant tests obtained by controlling the FDR, and S_{UNA}, number of significant tests using no adjustment. The recovery ratio is defined as $RR = (S_{\text{FDR}} – S_{\text{BON}})/(S_{\text{UNA}} – S_{\text{BON}})$.

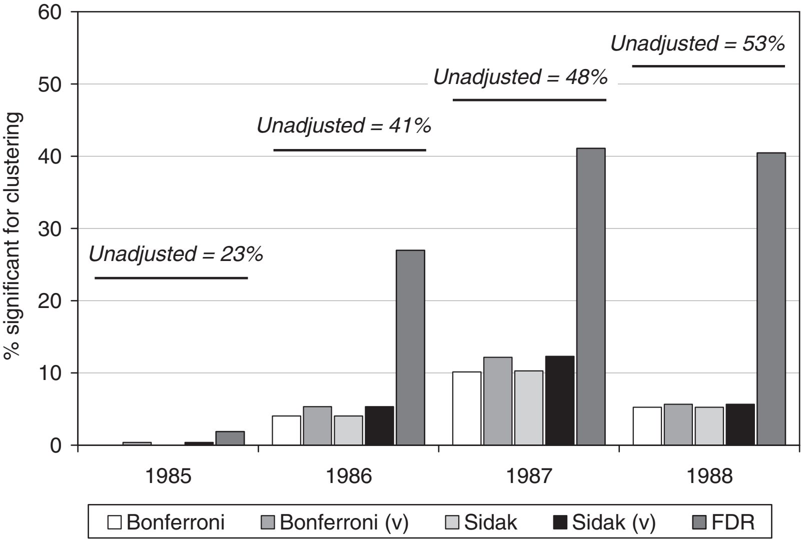

Figure 2. Percentage of plots that tested significant for clustering, according to the $G_i^*(d)$ statistic and different control procedures—Machadinho (1985/95).

those identified by the unadjusted procedure are false positives. Next, the original formulations of Bonferroni, Sidak, and FDR control procedures for accounting for multiple testing were applied. Bonferroni and Sidak led to very similar critical values, resulting in a comparable number of plots considered as significant for clustering of malaria rates, as shown in Fig. 2. The Benjamini and Hochberg (1995) procedure of controlling the FDR revealed more significant plots for clustering, as expected. Also, this effect becomes larger as the size of the data set increases, corroborating the fact that the larger the data set, the more the conventional MCPs are conservative, and the more powerful is the FDR approach.

Taking $G_i^*(d)$ as an example, the FDR p_{critical} values shown in Table 3 reveal that among the plots that were considered significant for clustering, 0.05%, 0.65%, 1.00%, and 1.01%, on average, are truly nonsignificant in 1985, 1986, 1987, and 1995, respectively. That error is acceptable for the purposes of disease surveillance in the present context. Additionally, the improvement provided by the FDR approach is captured by the recovery ratio (RR) (Williams, Jones, and Tukey 1999), which expresses the gain in the number of significant tests obtained by the FDR procedure when one moves from Bonferroni (the most conservative method) to the unadjusted approach. Table 3 shows the RR calculated for the procedures that correct for multiplicity. The ratio shows an increasing trend over the years due to the larger number of observations, and the gains are similar among the different local statistics computed for the same year. As an example, controlling the FDR for the Moran's I_i in 1985 increased the number of significant plots by 87% from the most conservative method.

The Bonferroni and Sidak methods were also applied with the correction for spatial dependence (overlap) proposed by Getis and Ord (2000). There is a slight gain in adjusting for both multiplicity and spatial dependence, as shown in Fig. 2 and Table 3. However, two issues need to be raised. First, correcting conservative MCPs for spatial dependence does not lead to results with the same power as those provided by the FDR procedure. Second, it is important to investigate if the additional plots that tested significant for a clustering pattern in this empirical example can be justified on the basis of the characteristics observed in the plots. It is not the purpose of this article to offer an exhaustive interpretation of each identified cluster, but some examples are detailed to illustrate the importance of properly assessing the significance of LISA .

All LISA applied to malaria rates in 1995 indicate no significant cluster of low rates when the Bonferroni approach is applied. However, when we control for the FDR, large clusters of low rates are observed in tract two, and their locations conform to prior expectations. For example, one cluster was identified in an area of intense and often successful agricultural production. Settlers occupying plots in that area have, among other characteristics better economic conditions, which facilitate the adoption of adequate health care, and the construction of good quality houses that provide adequate protection against mosquitoes. Intense agricultural production in the area does not leave the soil exposed, which decreases the number of potential mosquito breeding sites. Several plots in the area had only one owner during the 10-year period analyzed. This characterizes a stable area, where the population had prolonged exposure and, most likely, have developed acquired immunity against malaria.

Also in 1995, an important cluster of high malaria rates was not identified by the most conservative approaches. It is an area that registered remarkably high rates of malaria, mainly as a result of illegal deforestation that cleared, in less than 14 months, approximately 33.5 km^2 of forest in the vicinity of the area. The environmental transformations (such as slashing and burning trees, and leaving the bare soil exposed for a long period of time) contributed to an increase in the number of breeding sites. The scenario was aggravated by the fact that the labor force hired to work on the deforestation effort came from a very malarious region, resulting in ideal conditions for an intense malaria transmission (a reservoir of infected people—the hired labor force, a reservoir of people who could become infected—the settlers living in the plots close to the deforested area, and a large number of disease vectors—the mosquitoes).

A comprehensive analysis of the results for all 4 years show that important clusters of low and high malaria rates, such as those described above, were not detected by conservative control procedures, such as Bonferroni and Sidak. However, the use of the FDR resulted in a more comprehensive and defensible picture of malaria transmission in the particular area analyzed in this article.

Finally, although in this article we concentrate on the FDR approach proposed by Benjamini and Hochberg (1995), we tested the potential gains of using an adaptive FDR procedure (Benjamini and Hochberg 2000). All calculations were performed using a routine written for S-Plus® program and available for download at http://www.math.tau.ac.il/%7Eroee/FDR_Downloads2.htm. In 1985 no further improvement was obtained by the adaptive procedure (it resulted in the same p_{critical} value obtained for nonadaptive FDR). For the remaining years, however, the adaptive FDR approach identifies a larger number of significant plots, and the gains increase with the size of the data set. Considering the tests of the $G_i^*(d)$ statistic, the adaptive FDR identifies 32, 46, and 89 additional plots as significant in 1986, 1987, and 1995, respectively.

An example with simulated data: 50 × 50 grid sampled from a *N*(4,1)

Although the empirical example highlighted the costs of not accounting for multiple testing or of using highly conservative approaches, the true distribution of the data is unknown, and we cannot affirm with absolute confidence how many clusters were missed. Understanding the performance of local measures of spatial association, linked to multiple comparison methodology, requires a range of simulation studies. Ideally, we would like the simulated spatial patterns to, in some sense, be representative of what is observed across a diversity of scientific disciplines. They would include complex forms of spatial dependence, as it appears empirically in domains such as polymer chemistry, soil science, population genetics, and epidemiology of infectious diseases, to name only a few contexts. At the present time, a catalogue of simulation algorithms that could purport to represent this broad array of subject matter does not exist.

In the present article we simulate a diversity of cluster sizes, shapes, and locations that have the same features as malaria risk zones in rural settings of many endemic countries in Asia, Africa, and Latin America. The simulated arrangements are designed to challenge the $G_i^*(d)$ statistic and the FDR methodology in terms of their capacity to identify/discover actual clusters. A more elaborate comparative study across varieties of spatial arrangements from many fields would require clear delineation of the catalogue indicated in the above paragraph. Further, a rigorous mathematical understanding of the performance of $G_i^*(d)$ and FDR-based adjustments for multiple comparisons at the level of complex spatial dependencies included in the present simulations—and certainly in the more broad-based study we would all like to see—lies in the future.

Independent samples from a normal distribution with mean 4 and standard deviation 1—*N*(4,1)—were simulated and placed in a regular 50 × 50 grid. The size of the grid guarantees a big enough sample, while the mean and standard deviation produce only positive values (imitating a real-life experiment that records disease rates). We specified a significance level of 5% and defined extreme low values to be all points located below the 2.5th percentile. Extreme high values consisted of those points above the 97.5th percentile of the distribution. As simulating truly local patterns is not a trivial exercise, we rearranged the pixels in the

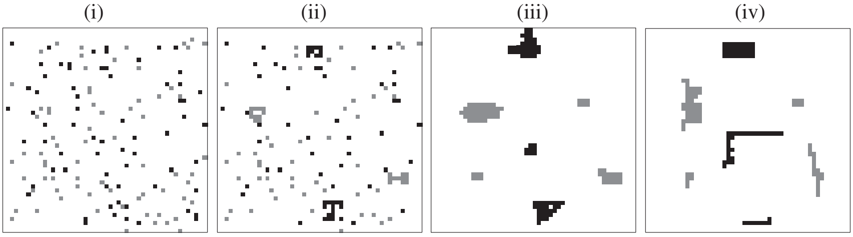

Figure 3. Simulated scenarios based on a normal distribution with mean 4 and standard deviation 1 – $N(4,1)$. Black pixels represent extreme high values (above the 97.5th percentile of the distribution), while gray pixels indicate extreme low values (below the 2.5th percentile of the distribution).

grid to create different and irregular clustering patterns of extreme values. The scenarios are shown in Fig. 3. Scenario (i) is the original data sampled from $N(4,1)$, with extreme values randomly arranged in the 50×50 grid. Scenario (ii) is a subtle alteration of the original $N(4,1)$ data in which a few pixels with extreme values were rearranged to produce four small clusters. Scenario (iii) is a more extensive rearrangement of the extreme values yielding seven compact clusters (four large and three small). Finally, scenario (iv) is a further modification of the previous scenario giving seven clusters of varied shape (compact and elongated) and size.

It is important to mention that as all LISA here addressed are expressed as normal variates, the results of this simulation can be generalized to all indicators. For the purpose of presenting the results, we chose to apply the $G_i^*(d)$ statistic to each scenario using distances equal to 2, 3, 5, and 8 units. Two remarks must be made at this point. First, any local indicator of spatial association could have been chosen, as for the purposes of the simulation what matters is the choice of the type of error rate control, not the indicator. Second, there is no real-life underlying process that could guide the researcher on the best distance d to use in this simulated grid. Unless all clusters had the same size and shape (which would have no application to real-life experiments), the choice of the critical distance faces the same challenges previously described.

The significance of the tests was assessed at a 5% level, considering no correction for multiple testing, and five different control procedures: Bonferroni, Bonferroni accounting for spatial dependence (with ν), Sidak, Sidak accounting for spatial dependence (with ν), and FDR. As Bonferroni and Sidak led to very similar outcomes, we only report the results for the former. It is expected that the $G_i^*(d)$ statistic will identify the clusters enforced in scenarios (ii)–(iv). Moreover, it is expected that pixels with nonextreme values, located outside a buffer zone of size d placed around the enforced clusters, will not test significant for a clustering pattern. Finally, we do not expect any significant clustering pattern to be identified in scenario (i).

In fact, when applied to scenario (i), the $G_i^*(d)$ statistic only reveals significant pixels if a cutoff unadjusted for multiple testing is used ($z_{\text{critical}} = 1.96$). When any correction for multiple comparisons is adopted, none of the 2500 statistics are significant. This result emphasizes the importance of accounting for multiple comparisons, as highlighted previously in the analysis of the empirical data. Although one does not want to miss real clusters, one does not want to identify false clusters either. Therefore, when local measures of spatial association are used, the issue of multiple testing must be considered. We provide some guidance on this issue using the simulated data.

Figs. 4–6 show some results for the modified scenarios. For ease of visualization, buffers of size *d* are shown around the enforced clusters. A black ''+'' indicates extreme high values of the simulated *N*(4,1) distribution that were not identified as significant for a clustering pattern, while a black ''Δ'' shows the location of extreme low values not identified as significant. Black pixels represent pixels that tested significant for a local cluster of high values, while gray pixels indicate those that tested significant for a local cluster of low values.

Fig. 4 shows the results for scenario (ii) using a distance equal to 2 and addressing multiple testing by the use of Bonferroni and FDR, respectively in grids (a) and (c). In the case of Bonferroni we also correct for spatial dependence (overlap) using ν (Getis and Ord 2000), shown in grid (b). The Bonferroni correction results in the identification of 45% of the pixels with extreme values clustered purposely, increasing to 49% when accounting for spatial dependence. The FDR approach increases this number to 81%, proving to be a more powerful procedure. None of the extreme values that were not rearranged into new clusters tested as significant, unless an unadjusted cutoff of 1.96 is used.

Fig. 5 shows results for scenario (iii). Grids (a) and (b) use a distance equal to 2, correcting for multiple testing using Bonferroni and FDR, respectively. The advantage in using FDR is considerable, especially for the small-enforced clusters. While the FDR approach identifies 95% and 65% of the extreme values placed in large and small clusters, respectively, the Bonferroni approach captures only 78% and 10% of those pixels, and two of the small-enforced clusters would be completely missed. Increasing the distance to 3 improves the identification of the large clusters, but aggravates the problem of the small clusters, as shown in grids (c) and (d), which indicates how critical the choice of *d* is especially when the clustering pattern is irregular, as one would expect to find in real-life experiments. Without correcting for spatial dependence, the Bonferroni approach identifies 88% of the extreme values constituting large enforced clusters, but misses all the small clusters. Adopting FDR changes these numbers to 97% and 40%. Correcting for spatial dependence resulted in a slight improve in Bonferroni (90% match of the extreme values in large clusters, but still none in the smaller ones).

Finally, Fig. 6 shows the application of the FDR approach for scenario (iv), using distances equal to 2, 3, and 8. As distance increases, all the pixels comprising the compact clusters are likely to be identified as significant. The elongated clusters, however, are much harder to identify. Increasing the distance, in fact, only adds pixels located inside the buffer zone, but does not improve the matching of clustered extreme values. For $d = 2$, 86% of the extreme values constituting the

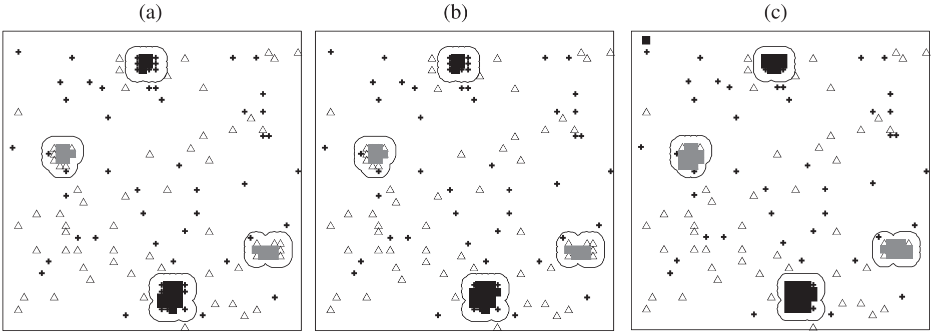

Figure 4. Results of the $G_i^*(d)$ statistic applied to modified scenario (ii) at distance $d = 2$: (a) Bonferroni, (b) Bonferroni with ν, and (c) false discovery rate.

clusters were identified as significant, while for $d = 3$ and 8 this number decreases to 76% and 74%, respectively.

These findings can be summarized by a characterization of the clusters identified under the different combinations of scenarios and MCP approaches. Consider that a true cluster is fully identified when all pixels that comprise an enforced cluster have a significant $G_i^*(d)$. Similarly, it is characterized as partially identified when only some of the pixels that encompass an enforced cluster had a significant $G_i^*(d)$. If none of the pixels had a significant $G_i^*(d)$ then the enforced cluster is classified as missed. Finally, false clusters are defined by a group of five or more pixels, located outside a buffer zone of size d placed around the enforced clusters, which had significant $G_i^*(d)$. This characterization is shown in Fig. 7 for all modified scenarios, assuming $d = 3$ for scenario (ii), and $d = 2$ for scenarios (iii) and (iv). Each ring of the graphs shows the proportion of types of clusters among all those that tested significant.

The unadjusted approach stands out as the extreme alternative. On the one hand, it fully identifies real clusters in larger numbers than any other procedure. On the other hand, it invariably reveals a significant number of false clusters. This trade-off can come at a higher cost for particular applications, for example, public health. Considering the approaches that address multiple testing, the FDR method is the only one able to fully identify real clusters. Bonferroni never does this. The Bonferroni approach is also the only one that misses real clusters, which never happens with FDR. These facts constitute a strong argument in favor of the FDR method as a much more powerful procedure than conservative MCPs, such as Bonferroni.

Regarding the need to account for spatial dependence, the answer is not so straightforward. Considering that the three modified scenarios are analyzed for four different distances and two MCP approaches, Bonferroni and FDR (the former applied with and without correcting for spatial dependence), we obtain 36 different results. Out of these, three did not show any improvement in the number of extreme values identified as significant for a clustering pattern when a correction for spatial

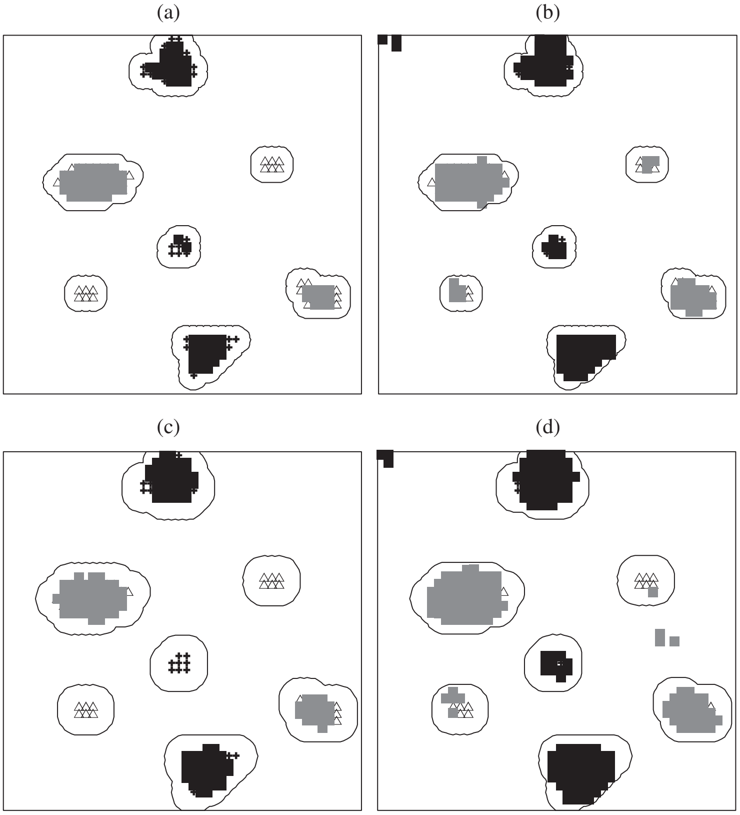

Figure 5. Results of the $G_i^*(d)$ statistic applied to modified scenario (iii): (a) using $d = 2$ and Bonferroni, (b) using $d = 2$ and false discovery rate (FDR), (c) using $d = 3$ and Bonferroni, and (d) using $d = 3$ and FDR.

dependence was included (Fig. 7a and c). Also, in two cases the number of clusters missed decreased, one of them is exposed in Fig. 7b.

While the choice of using ν, and therefore accounting for the geometric source of spatial dependence, stands out as a reasonable strategy, it appears that at least two additional issues, which are beyond the scope of this article, must also be addressed to facilitate a clear understanding of the problem. First, we require more extensive guidance on the specification of a critical distance tuned to the idiosyncrasies of a given spatial arrangement. Second, we require more nuanced understanding of how well noncompact clusters can be identified. Further research is needed on both points.

Conclusion

In light of the empirical example presented in this article, the failure to identify local spatial clusters may have multiple implications. First, local statistics of spatial association can be used as a surveillance tool for monitoring and control of dis-

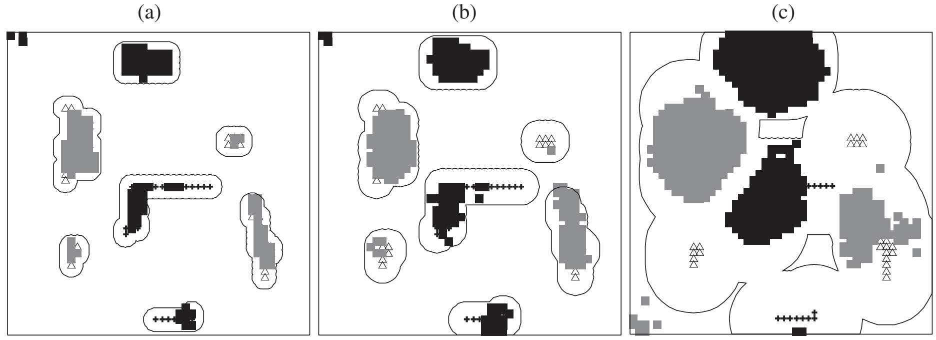

Figure 6. Results of the $G_i^*(d)$ statistic applied to modified scenario (iv): (a) using $d = 2$ and false discovery rate (FDR), (b) using $d = 3$ and FDR, and (c) using $d = 8$ and FDR.

eases. Not being able to identify as many true clusters as possible may result in ineffective control, and identifying too many false positives may increase the human and financial resources required to avoid an epidemic. Second, local statistics of spatial association can be used as an initial exploratory analytic tool, so that critical areas can be identified for further investigation in order to shed some light on what might be the major determinants of either low or high risk of disease transmission. If only a restricted number of locations are declared significant for a clustering pattern because of an overly conservative multiple comparison procedure, important relationships may not come to the attention of scientists and policy makers. Third, if the evaluation is being done in a newly opened settlement area in order to guide the occupation process (so that malarious areas could be avoided until major risk factors are mitigated), then failing to identify critical areas would potentially contribute to malaria outbreaks. These issues are not restricted to health

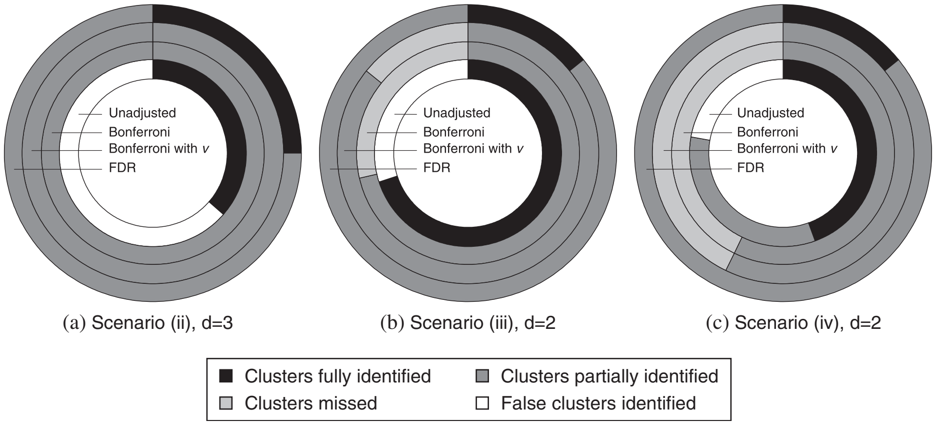

Figure 7. Characterization of clusters identified at distance d by different multiple comparison approaches in the modified scenarios (ii), (iii), and (iv).

applications. The same rationale is valid for any application of local statistics of spatial association. Overall, the goal is to find areas that show significant spatial patterns, and then investigate how these patterns affect the phenomena under study.

The assessment of significance of local statistics of spatial association must consider the problems of multiple comparison and spatial dependence. Currently, this has been addressed using a conservative approach such as the methods of Bonferroni and Sidak. In this article we propose the use of a more powerful procedure, which controls for the FDR (Benjamini and Hochberg 1995). Using empirical and simulated data, we tested different procedures that account for multiplicity in order to enhance the performance of statistics of spatial association. Most conservative MCP methods fail to identify clusters of both high and low values, while the FDR procedure provides a considerable improvement in the analysis. Additionally, we tested the conservative MCP approaches with a correction for spatial dependence (overlap), as proposed by Getis and Ord (2000). Although the results improved, they were still conservative and less powerful than those obtained by controlling the FDR.

The above-mentioned empirical and simulated studies are obviously but two instances of many alternative data sets that we might investigate. Thus, the question arises as to whether or not there is a rigorous general mathematical underpinning to the FDR-based methods that supports the kinds of spatial applications frequently arising in geography. At the present time, the answer to this question is negative. The principal obstacle to a full mathematical understanding of the FDR derives from the complex patterns of spatial dependence that are omnipresent in geographical analyses. A rigorous treatment of the FDR for specific forms of dependence among tests has been previously assessed (Yekutieli and Benjamini 1999; Benjamini and Yekutieli 2001). However, this is not sufficiently general to handle many problems in geography. For example, how would the test perform in a hypothetical case of extremely high spatial dependence among points distributed in an irregular grid?

A broad-based comparative study of new estimation methods in a one-dimensional problem that is far simpler than what we are contemplating is exemplified by the Princeton Monte Carlo study of robust estimators of location (Andrews et al. 1972). Polymer chemistry is another area where two- and three-dimensional spatial arrangements have been simulated and tested (Fixman 1978; Nelson, Rutledge, and Hatton 1997; Doi 2003). However, this is a particular topic supported by a rich physical theory underlying the construction of spatial arrangements. Analogous theories do not exist in the epidemiology of infectious diseases, linked to local ecologies. But, this is the context of our investigation. Our included simulation study lends support for the $G_i^*(d)$—FDR methodology as far as we can presently carry it. We hope that others will pursue this line of inquiry in much greater generality and with accompanying rigorous mathematical theory.

It is useful to point out that there is frequently a substantial time lag between the introduction and simulation-based testing of new statistical methods and the development of a full mathematical theory that underlies the methodology. A good case in point is the 31-year lag between the introduction of the Tukey–Kramer's procedure (1953–1956) for simultaneous confidence intervals in multiple comparisons with unequal sample sizes and the provision of a rigorous mathematical basis for it by Hayter (1984). See Benjamini and Braun (2002) for a fascinating account of this history.

Finally, we recommend that currently available software incorporate an option to address multiple comparisons. Until recently, calculation of LISA was restricted to specific routines, ArcView scripts, and programs such as PPA and GeoDA (the latter two described earlier). The recent incorporation of Spatial Statistical Tools in ESRI® ArcMap™ version 9.0, however, will lead to a dramatic increase in the number of researchers that use local statistics. Considering that some of these users may not be fully aware of the multiplicity problem, the benefits of making these tools available to a large number of users may be overcome by the proliferation of misleading conclusions regarding spatial patterns in analyzed data.

Acknowledgements

We wish to thank Art Getis and Keith Ord for insightful comments and suggestions on earlier drafts. We also thank three anonymous referees for their helpful comments. Thanks for financial support to M. C. C. are due to the Office of Population Research and the Center for Health and Wellbeing at Princeton University, and the Andrew W. Mellon Foundation for support via the Program on Migration and Urbanization at Princeton University. B. H. S. thanks the Rockefeller Foundation and the International Development Research Centre (IDRC) for financial support via the project ''Land Use and Health'' administered thru CEDEPLAR, Federal University of Minas Gerais, Belo Horizonte MG, Brazil.

References

Aldstadt, J., D. Chen, and A. Getis. (1998). Point pattern analysis. http://www.nku.edu/~longa/cgi-bin/cgi-tcl-examples/generic/ppa/ppa.cgi.

Andrews, D. F., P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Turkey (1972). *Robust Estimates of Location: Survey and Advances*. Princeton, NJ: Princeton University Press.

Anselin, L. (1995). ''Local Indicators of Spatial Association—LISA.'' *Geographical Analysis* 27(2), 93–115.

Anselin, L. (1996). ''The Moran Scatter Plot as an ESDA Tool to Assess Local Instability in Spatial Association.'' In *Spatial Analytical Perspectives on GIS*, 111–25, edited by M. M. Fischer, H. J. Scholten, and D. Unwin. London: Taylor & Francis.

Bailey, T. C., and A. C. Gatrell. (1995). *Interactive Spatial Data Analysis*. Harlow Essex, UK: Longman Scientific & Technical; Wiley.

Basford, K. E., and J. W. Tukey. (1999). *Graphical Analysis of Multiresponse Data: Illustrated with a Plant Breeding Trial*. Boca Raton, FL: Chapman & Hall/CRC.

Baumont, C., C. Ertur, and J. Le Gallo. (2004). ''Spatial Analysis of Employment and Population Density: The Case of the Agglomeration of Dijon 1999.'' *Geographical Analysis* 36(2), 146–76.

Benjamini, Y., and H. Braun. (2002). ''John W. Turkey's Contributions to Multiple Comparisons.'' *The Annals of Statistics* 30(6), 1576–94.

Benjamini, Y., and Y. Hochberg. (1995). ''Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.'' *Journal of the Royal Statistical Society B* 57(1), 289–300.

Benjamini, Y., and Y. Hochberg. (2000). ''On the Adaptive Control of the False Discovery Rate in Multiple Testing with Independent Statistics.'' *Journal of Educational and Behavioral Sciences* 25(1), 60–83.

Benjamini, Y., and D. Yekutieli. (2001). ''The Control of the False Discovery Rate in Multiple Testing Under Dependency.'' *The Annals of Statistics* 29(4), 1165–88.

Brown, B. W., and K. Russell. (1997). ''Methods Correcting for Multiple Testing: Operating Characteristics.'' *Statistics in Medicine* 16, 2511–28.

Castro, M. C. (2002). ''Spatial Configuration of Malaria Risk on the Amazon Frontier: The Hidden Reality Behind Global Analysis.'' PhD Thesis. Princeton University, Princeton, NJ.

Cova-Garcia, P. (1961). *Notas sobre los anofelinos de Venezuela y su identificación*. Caracas, Venezuela: Editorial Grafos.

Deane, L. M. (1947). ''Observações sobre a malária na Amazônia Brasileira.'' *Revista do Servico Especial de Saúde Pública* 1, 3–60.

Doi, M. (2003). ''Challenge in Polymer Physics.'' *Pure and Applied Chemistry* 75(10), 1395–402.

Dunnett, C. W., and A. C. Tamhane. (1992). ''A Step-up Multiple Test Procedure.'' *Journal of the American Statistical Association* 87(417), 162–70.

Efron, B., and R. Tibshirani. (2002). ''Empirical Bayes Methods and False Discovery Rates for Microarrays.'' *Genetic Epidemiology* 23, 70–86.

Fixman, M. (1978). ''Simulation of Polymer Dynamics. I. General Theory.'' *Journal of Chemical Physics* 69(4), 1527–37.

Genovese, C. R., N. A. Lazar, and T. Nichols. (2002). ''Thresholding of Statistical Maps in Functional Neuroimaging Using the False Discovery Rate.'' *NeuroImage* 15, 870–78.

Getis, A. (1995). ''Spatial Filtering in a Regression Framework: Examples Using Data on Urban Crime, Regional Inequality, and Government Expenditures.'' In *New Directions in Spatial Econometrics*, 191–203, edited by L. Anselin and R. J. G. M. Florax. Berlin, Germany: Springer.

Getis, A., and J. K. Ord. (1992). ''The Analysis of Spatial Association by Use of Distance Statistics.'' *Geographical Analysis* 24(3), 189–206.

Getis, A., and K. J. Ord. (1996). ''Local Spatial Statistics: An Overview.'' In *Spatial Analysis: Modelling in a GIS Environment*, 261–77, edited by P. Longley and M. Batty. Cambridge, UK: GeoInformation International; New York: distributed in the Americas by J. Wiley, 2617–277.

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

Hayter, A. J. (1984). ''A Proof of the Conjecture that the Tukey–Kramer Multiple Comparison Procedure is Conservative.'' *Annals of Statistics* 12, 61–75.

Hochberg, Y. (1988). ''A Sharper Bonferroni Procedure for Multiple Tests of Significance.'' *Biometrika* 75(4), 800–02.

Holm, S. (1979). ''A Simple Sequentially Rejective Multiple Test Procedure.'' *Scandinavian Journal of Statistics* 6, 65–70.

Hommel, G. (1988). ''A Stagewise Rejective Multiple Test Procedure based on a Modified Bonferroni Test.'' *Biometrika* 75(2), 383–86.

Hommel, G. (1989). ''A Comparison of Two Modified Bonferroni Procedures.'' *Biometrika* 76(3), 624–25.

Jones, L. V., C. Lewis, and J. W. Tukey. (2001). ''Hypothesis Tests, Multiplicity of.'' In *International Encyclopedia of the Social & Behavioral Sciences*, 7127–33, edited by N. J. Smelser and P. B. Baltes. Amsterdam: Elsevier.

Kurtz, T. E., R. F. Link, J. W. Tukey, and D. L. Wallace. (1965). ''Short-Cut Multiple Comparisons for Balanced Single and Double Classifications: Part 1, Results.'' *Technometrics* 7(2), 95–161.

Liu, W. (1996). ''Multiple Tests of a Non-Hierarchical Finite Family of Hypothesis.'' *Journal of the Royal Statistical Society B* 58(2), 455–61.

Miller, R. G. (1981). *Simultaneous Statistical Inference*. New York: Springer-Verlag.

Nelson, P. H., G. C. Rutledge, and T. A. Hatton. (1997). ''On the Size and Shape of Self-Assembled Micelles.'' *Journal of Chemical Physics* 107(24), 10777–81.

Ord, J. K., and A. Getis. (1995). ''Local Spatial Autocorrelation Statistics: Distributional Issues and an Application.'' *Geographical Analysis* 27(4), 286–306.

Packer, C., R. Hilborn, A. Mosser, B. Kissui, M. Borner, G. Hopcraft, J. Wilmshurst, S. Mduma, and A. R. E. Sinclair (2005). ''Ecological Change, Group Territoriality, and Population Dynamics in Seringeti Lions.'' *Science* 307, 390–93.

Paez, A., T. Uchida, and K. Miyamoto. (2001). ''Spatial Association and Heterogeneity Issues in Land Price Models.'' *Urban Studies* 38(9), 1493–508.

Paez, A., T. Uchida, and K. Miyamoto. (2002). ''A General Framework for Estimation and Inference of Geographically Weighted Regression Models: 1. Location-Specific Kernel Bandwidths and a Test for Locational Heterogeneity.'' *Environment and Planning* A34, 733–54.

Rogerson, P. A. (2001). ''A Statistical Method for the Detection of Geographic Clustering.'' *Geographical Analysis* 33(3), 215–27.

Sankoh, A. J., M. F. Huque, and S. D. Dubey. (1997). ''Some Comments on Frequently used Multiple Endpoint Adjustment Methods in Clinical Trials.'' *Statistics in Medicine* 16, 2529–42.

Sawyer, D. R. (1985). *Research Design and Feasibility in the Machadinho Settlement Project*. Belo Horizonte, Brazil: CEDEPLAR.

Sawyer, D. R. (1988). *Frontier Malaria in the Amazon Region of Brazil: Types of Malaria Situations and Some Implications for Control*. Brasília, Brazil: PAHO/WHO/TDR.

Sawyer, D. R., and D. O. Sawyer. (1987). *Malaria on the Amazon Frontier: Economic and Social Aspects of Transmission and Control*. Belo Horizonte, Brazil: CEDEPLAR.

Sidak, Z. (1967). ''Rectangular Confidence Regions for the Means of Multivariate Normal Distributions.'' *Journal of the American Statistical Association* 62(318), 626–33.

Sidak, Z. (1968). ''On Multivariate Normal Probabilities of Rectangles: Their Dependence on Correlations.'' *The Annals of Mathematical Statistics* 39(5), 1425–34.

Sidak, Z. (1971). ''On the Probabilities of Rectangles in Multivariate Student Distributions: Their Dependence on Correlations.'' *The Annals of Mathematical Statistics* 42(1), 169–75.

Simes, R. J. (1986). ''An Improved Bonferroni Procedure for Multiple Tests of Significance.'' *Biometrika* 73(3), 751–54.

Singer, B. H., and M. C. Castro. (2001). ''Agricultural Colonization and Malaria on the Amazon Frontier.'' In *Population Health and Aging: Strengthening the Dialogue Between Epidemiology and Demography*, Vol. 954: 184–222. New York: Annals of the New York Academy of Sciences.

Storey, J. D. (2002). ''A Direct Approach to False Discovery Rates.'' *Journal of the Royal Statistical Society B* 64(3), 479–98.

Storey, J. D. (2003). ''The Positive False Discovery Rate: A Bayesian Interpretation and the *q*-Value.'' *The Annals of Statistics* 31(6), 2013–35.

Storey, J. D., J. E. Taylor, and D. Siegmund. (2004). ''Strong Control, Conservative Point Estimation and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach.'' *Journal of the Royal Statistical Society B* 66(1), 187–205.

Storey, J. D., and R. Tibshirani. (2001). ''Estimating False Discovery Rates under Dependence, with Applications to DNA Microarrays.'' Technical Report 2001–28, Department of Statistics, Stanford University, Stanford, CA.

Storey, J. D., and R. Tibshirani. (2003). ''Statistical Significance for Genomewide Studies.'' *Proceedings of the National Academy of Sciences* 100(16), 9440–45.

Sydenstricker, J. M. (1992). ''Parceleiros de Machadinho: história migratória e as interações entre a dinâmica demográfica e o ciclo agrícola em Rondônia.'' Master degree dissertation, Universidade de Campinas, Campinas SP.

Tobler, W. R. (1979). ''Cellular Geography.'' In *Philosophy in Geography*, 379–86, edited by S. Gale and G. Olsson. Dordrecht, Germany: D. Reidel Publishing Company.

Tukey, J. W. (1991). ''The Philosophy of Multiple Comparisons.'' *Statistical Science* 6(1), 100–16.

Tukey, J. W., J. L. Ciminera, and J. F. Heyse. (1985). ''Testing the Statistical Certainty of a Response to Increasing Doses of a Drug.'' *Biometrics* 41, 295–301.

Van Thiel, P. H. (1962). ''Malaria Problems Arising from the Construction of a Reservoir in the Interior of Surinam.'' *Tropical and Geographical Medicine* 14, 259–78.

Williams, V. S. L., L. V. Jones, and J. W. Tukey. (1999). ''Controlling Error in Multiple Comparisons, with Examples from State-to-State Differences in Educational Achievement.'' *Journal of Educational and Behavioral Sciences* 24(1), 42–69.

Yekutieli, D., and Y. Benjamini. (1999). ''Resampling-Based False Discovery Rate Controlling Multiple Test Procedures for Correlated Test Statistics.'' *Journal of Statistical Planning and Inference* 82, 171–96.