Effects of randomization methods on statistical inference in disease cluster detection

Colleen C. McLaughlin, Francis P. Boscoe*

New York State Cancer Registry, New York State Department of Health, Corning Tower Room 536, Empire State Plaza, Albany, NY 12237, USA

Received 24 March 2005; received in revised form 14 November 2005; accepted 14 November 2005

Abstract

Monte Carlo methods are commonly used to assess the statistical significance of disease clusters. This usually involves permuting the observed outcome measure, such as the rate of disease, across the geographic units within the study area. When the variance of the disease rates is heterogeneous, however, randomizing the disease rate across the geographic units results in over-estimating the p-values in areas of low variance and under-estimating the p-values in areas of high variance. This bias results in under-ascertainment of clusters in urban areas and over-ascertainment of clusters in rural areas. As an alternative, randomizing the number of cases of disease or deaths proportional to the population at risk preserves the variance structure of the study area, therefore resulting in unbiased statistical inference. We compare results from randomizing rates with those from randomizing case counts, using county-level prostate cancer mortality data for the United States and ZIP-Code level prostate cancer incidence data for New York State, using the local Moran’s I statistic.

Keywords: Monte Carlo method; Disease clustering; Local moran’s I; Prostate cancer; United States; New York State

Introduction

Monte Carlo simulation has long been a routinely used tool for spatial point pattern analysis (Besag and Diggle, 1977; Openshaw et al., 1987; Cuzick and Edwards, 1990; Bithell, 1995; Kulldorff and Nagarwalla, 1995; Rushton and Lolonis, 1996) that is increasingly found in spatial analysis software packages. The technique involves repeatedly randomizing (or permuting) the observed values of the outcome measures across the geographic units in a study area, thereby simulating the null distribution and allowing the assignment of a p-value for a test statistic without the need for asymptotic assumptions. It is a powerful tool for spatial analysis because most spatial techniques are non-parametric. Monte Carlo simulation can be used to help describe global and local autocorrelation in the data, identify groups of unusual values (clusters), and identify isolated unusual values (outliers).

In many studies, the outcome measures that are randomized in the Monte Carlo simulation step are rates or proportions. Most commonly, the set of observed rates or proportions in the data set are randomly reallocated to the geographic units in the data set a large number of times. The observed rate is then compared with the set of reallocated rates to determine whether the observed rate exhibits
unusual characteristics. This approach has been applied to identify clusters and outliers in numerous data sets, including county-level low birth weight (Morenoff, 2003), alcohol mortality (Hanson and Wieczorek, 2002), homicide (Baller et al., 2001), suicide (Baller and Richardson, 2002), and child and adolescent amphetamine prescription usage (Lin et al., 2005), and sub-county-level cancer incidence (Jacquez and Greiling, 2003), respiratory disease (Buckeridge et al., 2002), and playground accessibility (Smoyer-Tomic et al., 2004). A problem with this approach, and the focus of our paper, is that randomizing rate data in this manner results in bias toward identifying hot spots in sparsely populated areas, while missing elevated rates or outliers in areas with larger populations. This bias is a consequence of the violation of the assumption of spatially uncorrelated variances in rates, a problem also known as variance instability (Anselin et al., forthcoming). We demonstrate the existence of this bias using prostate cancer data from two data sets at different geographic resolutions. For each data set, we calculate the Local Moran’s $I$ statistic\(^1\) and compare the results obtained by randomizing cases and randomizing rates under two different randomization assumptions, and by randomizing empirical Bayes smoothed rates (Anselin, 2003). We demonstrate that randomly generating the numerator of the rate or proportion, rather than randomizing the rates, preserves the spatial structure of the variance in the simulated null distribution and provides more sensible results.

Although the Local Moran’s $I$ is just one of a large number of techniques for detecting clusters (Kulldorff et al., 2003), we have chosen to use it in this demonstration because it is a popular statistic for which rate randomization is the norm in both published studies using the statistic (Baller et al., 2001; Baller and Richardson, 2002; Buckeridge et al., 2002; Hanson and Wieczorek, 2002; Jacquez and Greiling, 2003; Morenoff, 2003; Smoyer-Tomic et al., 2004; Lin et al., 2005) and in exploratory geographic analysis software packages including ClusterSEER, GeoDa, SpaceStat, CrimeStat, and others (Boots, 2002). We emphasize, however, that the bias we demonstrate is a product of the rate randomization and not the specific test used. While the effect we are illustrating could be demonstrated using simulated data, we chose to use actual prostate cancer incidence and mortality data for this analysis. Prostate cancer has been shown to have wide regional and local geographic variation that has been described in numerous studies (Blair and Fraumeni, 1978; Hanchette and Schwartz, 1992; Kafadar, 1997; Jemal et al., 2002; Grant, 2002; Boscoe et al., 2003, Gregorio et al., 2004, Klassen et al., 2004). In addition, race is a known risk factor for prostate cancer with a high degree of spatial autocorrelation. We had an a priori hypothesis that ZIP-Code level prostate cancer incidence data for New York State should exhibit spatial clustering in the predominantly black urban neighborhoods of New York City. We show that this hypothesis is confirmed only when the numerator of the rates are randomized, while the clusters in the urban areas are missed with rate randomization.

\section*{A hypothetical example}

The bias we describe can be illustrated through a simple hypothetical example. Assume a study area comprised of both large population (cities) and small population (towns) units. In the cities, the observed rates vary from 5 to 15 cases per 1000 residences. In the towns, the observed rates vary from 0 to 20 cases per 1000 residents. An observed rate of 15 cases per 1000 residents would be considered high for a city if it is compared to only other cities, but not if it is compared to cities and towns combined. Conversely, towns with values at the extremes (0 or 20 cases per 1000 residents) have a good chance of being identified as outliers. But because towns have smaller populations, they have less stable rates. If a town had only 50 people at risk for the disease, then it would have a rate of 0 cases per 1000 if no case is observed and 20 cases per 1000 if one case is observed. If, instead of permuting the observed rates across the entire study area, every person is given an equal probability of being a case under the null hypothesis of equal disease risk, then the town with 50 residents would probably be assigned one case in at least some of the permutations, and therefore the observed value of 20 cases per 1000 might not be considered unusual under the null hypothesis.

While consideration of geographic units with as few as 50 people at risk may seem unrealistic, there are many studies where this is done. The well-

\footnote{All calculations were performed using SAS version 8.2. Our source code is available to the interested reader, although since we chose to emphasize clarity over efficiency, the Monte Carlo algorithm is very computationally intensive.}
known upstate New York leukemia data (Waller et al., 1994) that has been used in dozens of papers, for example, contains several very-low population census tracts, including one in the city of Syracuse with 9 people. More importantly, as we will show, the tendency to assign lower \( p \)-values to sparsely populated geographic units when randomizing rates is not simply a problem of small absolute populations, but of small relative populations.

Such bias can have implications for public health planning and policy, particularly as the results of statistical analyses are increasingly used to delineate areas for further study (Wartenberg, 2001). Spatial analysis techniques ideally should be unbiased toward identifying unusual patterns in any particular kind of area, whether sparsely populated areas, urban areas, edge areas or islands (Gelman et al., 2000). While this point is widely acknowledged in theory (Besag and Newell, 1991; Walter, 1992), it is not always put into practice. Our paper offers a straightforward solution to a particular type of bias that tends to be either ignored or addressed in an abstruse manner.

**Data and methods**

The first data set used for this analysis was county-level prostate cancer mortality for the United States. The number of prostate cancer deaths by county among white men for the time period 1970–1994 was obtained from the National Cancer Institute (National Cancer Institute, 2004). These data are available as part of the Atlas of Cancer Mortality (Devesa et al., 1999) and have been used to explore the spatial distribution of prostate cancer mortality in the US (Jemal et al., 2002; Grant, 2002; Boscoe et al., 2003). The parameter of interest is the standardized mortality ratio, which is the ratio of the observed prostate cancer deaths to the age-adjusted expected number of deaths.\(^2\) The standardized mortality ratio has the distribution of a relative risk measure, with a null value of one, elevated risk indicated by values greater than one and decreased risk indicated by values between zero and one.\(^3\)

Because the expected number of deaths in this data set is age-adjusted, additional data were needed to implement the randomization of disease among the population at risk. This included the age-specific number of prostate cancer deaths among white men for the US as a whole and the age-specific population of white men for each county and county-equivalent area in the US. These data were obtained from the National Cancer Institute upon request.\(^4\)

The second data set used for this analysis was prostate cancer incidence for New York State. The number of prostate cancer cases by ZIP Code for New York State for 1995–1999 was obtained from the New York State Cancer Registry in the New York State Department of Health. These data are similar to the data provided as part of the department’s Cancer Surveillance Improvement Initiative (New York State Department of Health, 2004) which have been used by the department and by other researchers and educators to explore the geographic distribution of cancer in New York. The age-specific number of prostate cancer cases for New York State was also obtained from the New York State Cancer Registry; these data are available upon request. Age-specific ZIP-Code populations were obtained from Claritas, Inc. In this data set, the parameter of interest is the standardized incidence ratio, the ratio of the observed number of prostate cancer cases to the age-adjusted expected number.

**Local indicator of spatial autocorrelation**

The local indicator of spatial correlation used in these analyses was the local Moran’s \( I \) (Anselin, 1995), applied to the standardized mortality ratios for the county-level prostate cancer and standardized

\(^2\)For geographic unit \( i \), the SMR is calculated as

\[
\text{SMR}_i = \frac{\sum d_j w_j}{\sum n_j w_j},
\]

where \( j \) are the age strata, \( d \) is the number of observed cases or deaths in \( i \), \( n \) is the population of \( i \) and \( w \) is the disease rate in the standard area (Breslow and Day, 1987).

\(^3\)Our analysis uses the indirect standardization method, though our hypothetical example is intentionally presented in the language of the direct standardization method. While we did not replicate our analysis using the direct standardization method, the two methods have been consistently found to yield largely equivalent results (Goldman and Brender, 2000).

\(^4\)Alaska, Hawaii and the island counties of San Juan in Washington State and Nantucket in Massachusetts were excluded from the analysis because they are not contiguous with the remainder of the United States. Loving County, Texas, was excluded because it had no reported prostate cancer deaths during the time period of the study. In addition, some counties, such as the five counties comprising New York City, were combined by the National Cancer Institute because of past boundary changes or data reporting practices.
incidence ratios for prostate cancer in New York State (hereafter referred to collectively as standardized morbidity ratios, or SMRs). The local Moran’s I measures whether, for each geographic unit of analysis, the SMR is closer to the values for its neighbors than to the mean value for the study area as a whole. The local Moran’s I thus provides a measure of spatial autocorrelation for each areal unit. If the value is strong, positive and significant, the areal unit can be viewed as the center of a cluster. If the value is strong, negative and significant, the area can be viewed as a spatial outlier.

Monte Carlo simulation with rate-based randomization

The Monte Carlo simulation, or permutation, that is commonly used in spatial analysis software involves randomizing the value of the outcome variable across all geographic units in the study area, a technique we refer to as rate-based randomization. For the prostate cancer mortality analysis, the standardized mortality ratios were randomized across all the counties. For the prostate cancer incidence analysis, the standardized incidence ratios were randomized across the ZIP Codes of New York State. Rate-based randomization can be done without constraining the values of the SMRs, such that any given county or ZIP Code can take on the full range of SMRs that were observed in the actual distribution of the data. In this way, the full distribution of the randomized SMRs is identical to the distribution of the original values. Often, however, the value for the geographic unit of interest will be held constant, an option known as conditional randomization. When calculating the local Moran’s I, we used both pure rate-based randomization, in which the central geographic unit was randomized, as well as conditional randomization, in which the central geographic unit retained its observed value.

The randomization was repeated 9999 times, with a new local Moran’s I calculated for each geographic area for each permutation. The simulated local Moran’s I values were then ranked for each areal unit, and the top and bottom percentiles were used to determine the p-value of the observed local Moran’s I. In order to adjust for multiple comparisons, the Bonferroni adjustment was used, whereby alpha was divided by the average number of neighbors for the geographic units (Getis and Ord, 1992). Both the US counties and NYS ZIP Codes had an average number of neighbors between 5 and 6, so that a p-value of 0.01 was used to determine statistical significance with an alpha of 0.05.

Monte Carlo simulation with case-based randomization

An alternative randomization method is to randomize the outcome of interest across the population at risk. This method, which we refer to as case-based randomization, has the effect of randomizing the numerator in proportion to the population while keeping the denominator constant for each unit within the study area. Unlike with rate-based randomization, any adjustment for covariates must be incorporated into the randomization step. For the age-adjusted SMRs, the cases/deaths were randomized by age group according to the age-specific population of each unit within the study area. This required the age-specific population of each geographic unit as well as the total number of cases/deaths occurring in each age group in the study area as a whole.

With the case-based randomization approach, the expected number of cases or deaths (the denominator of the rate) are excluded from the randomization step. For the age-adjusted SMRs, the cases/deaths were randomized by age group according to the age-specific population of each unit within the study area. This required the age-specific population of each geographic unit as well as the total number of cases/deaths occurring in each age group in the study area as a whole.

Confirmation with randomly generated data

To verify that these results were not particular to the observed patterns of incidence or mortality data, we repeated the experiment using the US counties, but assigning random numbers of prostate cancer
deaths to each county. The random assignment was based on the male age-specific population of each county in the same manner as the case-based randomization above. Using this randomly generated data set, it would be expected that 1% of counties would be identified as being in a county with elevated mortality surrounded by neighbors with elevated mortality (a high–high cluster) when a \( p \)-value of 0.01 is used.

**Spatial empirical Bayes smoothed rates**

Rate transformation has been recommended as an alternate method of managing the problem of variance instability (Anselin et al., forthcoming; Goovaerts and Jacquez, 2004). With this approach, the rates are mathematically transformed or smoothed to create a more uniform variance structure. Rate-based randomization is then applied to the transformed or smoothed data. We used a spatial empirical Bayes transformation (Anselin, 2003) to smooth the SMRs, and applied the same rate-based randomization method described above to determine the statistical significance of the local Moran’s \( I \) values for the smoothed SMRs.

**Results**

**National prostate cancer mortality**

The rate-based randomization methods resulted in an unequal distribution of statistically significant results across population size of the counties, as measured by the number of expected deaths per county (Table 1). Using conditional rate-based randomization, 18% of areas with fewer than 10 expected deaths were identified as statistically significant clusters or outliers. The percentage of counties identified as significant declined with increasing numbers of expected deaths, with only 2% of areas with 50 or more expected deaths identified as statistically significant. None of the 300 counties with more than 300 expected deaths was identified as statistically significant by conditional rate-based randomization. For counties identified as statistically significant by conditional rate-based randomization, the mean (47 expected deaths) and median (28 expected deaths) size were both substantially lower than the mean (171 expected deaths) and median (63 expected deaths) for all counties. Similar patterns were found when only significant elevated clusters (SMR for county greater than one and high average SMR for neighboring counties greater than one) were considered. Rate-based randomization without conditional constraint produced similar results as the conditional rate-based randomization (data not shown).

Conversely, case-based randomization resulted in statistically significant results \( (p < 0.01) \) for 5% of counties with fewer than 10 expected deaths and 16% of counties with 50 or more expected deaths. Counties identified as significant by the case-based method had mean of 292 deaths and median of 85 deaths, with a maximum of 5288 expected deaths. The increasing number of counties found to be statistically significant with increasing number of

<table>
<thead>
<tr>
<th>Number of expected deaths in county</th>
<th>Number of Significant Counties(^a)</th>
<th>Significant by case-based method</th>
<th>Significant by rate-based conditional method</th>
<th>Number of high–high Counties(^b)</th>
<th>Significant by case-based method</th>
<th>Significant by rate-based conditional method</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–9</td>
<td>131</td>
<td>6 (5%)</td>
<td>23 (18%)</td>
<td>27</td>
<td>1 (4%)</td>
<td>8 (30%)</td>
</tr>
<tr>
<td>10–19</td>
<td>277</td>
<td>22 (8%)</td>
<td>30 (11%)</td>
<td>78</td>
<td>12 (15%)</td>
<td>13 (17%)</td>
</tr>
<tr>
<td>20–29</td>
<td>289</td>
<td>26 (9%)</td>
<td>25 (9%)</td>
<td>67</td>
<td>11 (16%)</td>
<td>12 (18%)</td>
</tr>
<tr>
<td>30–49</td>
<td>540</td>
<td>58 (11%)</td>
<td>32 (6%)</td>
<td>116</td>
<td>14 (11%)</td>
<td>9 (6%)</td>
</tr>
<tr>
<td>50–99</td>
<td>846</td>
<td>110 (13%)</td>
<td>21 (2%)</td>
<td>215</td>
<td>43 (13%)</td>
<td>10 (2%)</td>
</tr>
<tr>
<td>100–299</td>
<td>659</td>
<td>109 (13%)</td>
<td>20 (2%)</td>
<td>238</td>
<td>52 (13%)</td>
<td>10 (2%)</td>
</tr>
<tr>
<td>&gt;300</td>
<td>308</td>
<td>71 (23%)</td>
<td>0 (0%)</td>
<td>135</td>
<td>39 (29%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Total</td>
<td>3050</td>
<td>402 (13%)</td>
<td>151 (5%)</td>
<td>876</td>
<td>172 (20%)</td>
<td>62 (7%)</td>
</tr>
</tbody>
</table>

\( ^a\)Counties identified as significant clusters or outliers \( (p < 0.01) \), with either elevated or deficit rates.

\( ^b\)Counties identified as significant elevated clusters \( (p < 0.01) \).
expected deaths more appropriately reflects the higher statistical power available with larger populations. To illustrate this further, Fig. 1 shows the distribution of simulated local Moran’s I values generated using the case-based and conditional rate-based methods for two sample counties with disparate population size. Cherry County, Nebraska, which had an average annual population of 3244 males and had 25 expected prostate cancer deaths over the 25-year study period, was statistically significant based on the rate-based method but not the case-based method. Fig. 1 demonstrates that the local Moran’s I distribution generated from the rate-based method is narrower than the distribution from the case-based method. The observed number of deaths for Cherry County was 29, for a standardized incidence ratio of 1.16. The average of the neighboring counties’ SMRs was 1.20. For Monmouth County, New Jersey, with a population of 221,936 males and 1206 expected prostate cancer deaths, the distribution of local Moran’s I from the rate-based randomization is much wider than seen from the case-based method. This county had an observed 1330 deaths, for a standardized incidence ratio of 1.10. The average neighborhood SMR was 1.04.

Applying rate-based randomization to the empirical Bayesian smoothed SMR also over-ascertained clusters in rural areas and under-ascertained clusters in urban areas. This method resulted in statistically significant results for 33% of counties with fewer than 10 deaths and 4% of counties with 300 or more deaths. As with the rate-based randomization of the raw SMR, the counties identified as statistically significant by this method were smaller in size (mean = 82 expected deaths, median = 63 expected deaths) than the country as a whole. Many well-populated urban centers that were identified as areas of high incidence clusters using case-based randomization to assess statistical significance were missed with the rate-based randomization method using both the raw SMR and the smoothed SMR, including Detroit, Baltimore, Seattle and metropolitan Atlanta (data not shown).

In addition, calculating the local Moran’s I on the empirical Bayesian smoothed SMR using the rate-based method of randomization resulted in identification of only 16 counties as outliers (0.5% of counties). Half of these outliers had fewer than 20 expected deaths and only one county identified as an outlier had more than 100 expected deaths. In contrast, the case-based randomization resulted in 81 outlier counties (3.2% of counties), nine of which had fewer than 20 expected deaths and 39 of which had more than 100 expected deaths.

New York State prostate cancer incidence

Moran’s scatterplot maps are often used to display the results of tests for significance of local Moran’s I, as shown for New York State prostate cancer incidence in Fig. 2 for New York State and Fig. 3 for New York City. In addition to the
Fig. 2. (a) Prostate cancer standardized incidence ratios in New York State, by ZIP Code, 1995–1999. (b) Smoothened prostate cancer standardized incidence ratios using empirical Bayesian smoothing. (c) Statistically significant clusters and outliers using case-based randomization. (d) Statistically significant clusters and outliers using conditional rate-based randomization. (e) Statistically significant clusters and outliers using conditional rate-based randomization on the empirical Bayesian smoothed standardized incidence ratios.
Fig. 3. (a) Prostate cancer standardized incidence ratios in New York City, by ZIP Code, 1995–1999. (b) Smoothed prostate cancer standardized incidence ratios using empirical Bayesian smoothing. (c) Percent of total population identified as black in the 2000 census. (d) Statistically significant clusters and outliers using case-based randomization. (e) Statistically significant clusters using conditional rate-based randomization. (f) Statistically significant clusters and outliers using conditional rate-based randomization on the empirical Bayesian smoothed standardized incidence ratios.
Moran’s scatterplot maps using both case-based randomization and rate-based conditional randomization, we also provide maps of the raw standardized incidence ratios and the smoothed ratios from the spatial empirical Bayesian transformation, as well as the Moran’s scatterplot map of the empirical Bayesian smoothed SMR. In examining these maps, we see that the rate-based method identified more clusters and outliers among ZIP Codes in the rural northern areas of New York State, while the case-based method identified more clusters and outliers in New York City, most of which were missed by the rate-based method. The clusters of elevated incidence in New York City correspond to the neighborhoods in the city that are predominately black (Fig. 3c). Because prostate cancer incidence among black men is over 60% higher than among white men, it would be expected that, in the absence of race-adjustment, clusters of prostate cancer should be observed in black neighborhoods. The clusters, however, were largely not identified when statistical significance was assessed using rate-based randomization. Although these clusters were identified to some extent in the scatterplot of the empirical Bayesian smoothed SMR, this method failed to identify the cluster in Harlem, at the northern end of Manhattan, as well as clusters in metropolitan Rochester and Buffalo. As with the national data, this method over-ascertained clusters in rural areas and under-ascertained clusters in urban areas.

**Randomly generated US mortality data**

When the analysis was repeated for the randomly generated United States mortality data, case-based randomization resulted in 1% of counties (31 out of 3050 counties) being identified as statistically significant clusters of elevated mortality (i.e. high county SMR and high neighborhood SMR) with \( p < 0.01 \) (Table 2). For counties with fewer than 10 expected deaths, 0.8% (1 out of 131) were statistically significant high–high clusters. For counties with more than 300 expected deaths, no county was identified as statistically significant using rate-based randomization. As with the observed US mortality data, results were similar when looking at all

<table>
<thead>
<tr>
<th>Number of expected deaths in county</th>
<th>Number of Significant Counties(^{a})</th>
<th>Significant by case-based method</th>
<th>Number of Significant by rate-based conditional method</th>
<th>Number of high–high Counties(^{b})</th>
<th>Significant by case-based method</th>
<th>Significant by rate-based conditional method</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–9</td>
<td>131</td>
<td>3 (2%)</td>
<td>15 (12%)</td>
<td>34</td>
<td>1 (3%)</td>
<td>6 (18%)</td>
</tr>
<tr>
<td>10–19</td>
<td>277</td>
<td>9 (3%)</td>
<td>22 (8%)</td>
<td>65</td>
<td>5 (8%)</td>
<td>5 (8%)</td>
</tr>
<tr>
<td>20–29</td>
<td>289</td>
<td>5 (2%)</td>
<td>6 (2%)</td>
<td>84</td>
<td>3 (4%)</td>
<td>3 (3%)</td>
</tr>
<tr>
<td>30–49</td>
<td>540</td>
<td>4 (1%)</td>
<td>12 (2%)</td>
<td>157</td>
<td>2 (1%)</td>
<td>4 (3%)</td>
</tr>
<tr>
<td>50–99</td>
<td>846</td>
<td>15 (2%)</td>
<td>14 (2%)</td>
<td>276</td>
<td>6 (2%)</td>
<td>2 (1%)</td>
</tr>
<tr>
<td>100–299</td>
<td>659</td>
<td>17 (3%)</td>
<td>5 (1%)</td>
<td>217</td>
<td>10 (5%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>&gt;300</td>
<td>308</td>
<td>5 (2%)</td>
<td>2 (1%)</td>
<td>106</td>
<td>4 (4%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Total</td>
<td>3050</td>
<td>58 (2%)</td>
<td>76 (3%)</td>
<td>939</td>
<td>31 (3%)</td>
<td>20 (2%)</td>
</tr>
</tbody>
</table>

\(^{a}\)Counties identified as significant clusters or outliers \((p < 0.01)\), with either elevated or deficit rates.

\(^{b}\)Counties identified as significant elevated clusters \((p < 0.01)\).
significant counties rather than just the high–high counties.

**Discussion**

Researchers have long recognized that extreme rates (both high and low) are most common in sparsely populated areas as a consequence of higher variance due to smaller populations at risk in these areas (Anselin, 1990; Clayton and Bernardinelli, 1992; Moulton et al., 1994; Gelman and Price, 1999; Gelman et al., 2000). In the field of spatial epidemiology, this problem is exacerbated by the relative rarity of many health outcomes. For example, almost 10% of all counties in the United States average one or fewer prostate cancer deaths per year and another 27% average fewer than three prostate cancer deaths per year. The low-population areas are heavily concentrated in the south, central and western parts of the country, meaning that variances in rates in these areas are higher. The inhomogeneous distribution of population presents a number of challenges for spatial analysis, including low statistical power, spatial autocorrelation of variance, and difficulties with visualization of geographic patterns. In this analysis, we demonstrate that biased statistical significance levels can result if the spatial autocorrelation of variances is not accounted for in Monte Carlo simulation. In our analysis, we demonstrated the presence of bias in the assessment of local Moran’s *I* that resulted in under-ascertainment of clusters in rural areas and over-ascertainment of clusters in urban areas.

In the example of prostate cancer incidence in New York State, clusters of elevated prostate cancer incidence in the black neighborhoods of New York City were entirely missed by the conditional rate-based randomization that is used in most exploratory spatial data analysis software packages. This bias results from the fact that randomizing the rate of disease among the geographic units is equivalent to assuming a random distribution of both the risk of disease and the variance. Population size, and therefore variance, however, is a fixed quantity for any one geographic unit, and should not be included in the randomization process. The correctly defined null hypothesis is that disease risk is distributed uniformly among the population. The null hypothesis with rate-based randomization is that disease risk AND population are uniformly distributed across the geographic areas under study. Clearly, this misspecification will have negative consequences when the second assumption does not hold. A similar conclusion has also recently been reached by Waller and Gotway (2004).

We have also evaluated the empirical Bayesian smoothing approach, which has been proposed as a method for dealing with variance instability. Smoothing changes the distribution of the SMRs by bringing them generally closer to the null value. Because the smoothed rates are transformed in relation to the rates in the neighborhood, the interpretability of local indicators of spatial autocorrelation (LISA), such as the local Moran’s *I*, is problematic. It is counterintuitive to test for spatial pattern in a data set after first imposing spatial pattern on it in the form of smoothing, and LISA values calculated on smoothed data cannot be compared directly with those obtained on unsmoothed data. The empirical Bayesian smoothing approach also yielded biased results. In both New York State and the nation, randomizing the smoothed rates resulted in under-ascertainment of clusters in urban areas and over-ascertainment of clusters in rural areas when compared with the results from applying case-based randomization on unsmoothed SMRs. In addition, the smoothing process resulted in under-ascertainment of outliers, an effect that would be expected when rates are smoothed based on a neighborhood parameter.

On the other hand, an empirical or fully Bayesian smoothing approach alone, without any subsequent statistical simulation and testing applied to the smoothed data, may actually yield a result that is quite similar to the case-based randomization approach that we advocate. This would mean that Fig. 3(b) and (d) would share more similarity than either 3(e) or 3(f). While we did not test this idea specifically, others have reported intriguing findings along these lines, particularly when the smoothed rates are mapped so as to highlight areas with a greater than 95% probability of being elevated (Thomas and Carlin, 2003; Johnson, 2004).

The use of Monte Carlo simulation for assessing statistical significance of disease clusters has become fairly common. Local indicators of spatial correlation such as local Moran’s *I* have also become widely used in social sciences. Studies that cover even a moderate proportion of the population of the nation will likely include areas with wide ranging population sizes, and therefore subject to the bias illustrated here if the randomization techniques do not take this variation into account. A recent paper by Leung et al. offering an exact solution to local
Moran’s $I$ and other local indicators of spatial association means that a Monte Carlo approach need not be relied upon; however, this solution is both complex and computationally intensive (Leung et al., 2003). Appropriate specification of the null hypothesis in the process of the Monte Carlo simulation is a more straightforward and universally applicable solution to this issue.

The proliferation in spatial analysis software (Boots, 2002) has increased the popularity of the Monte Carlo technique, perhaps without all users being fully informed about how the technique is implemented or the interpretation of the results. Although the assumption of homogeneous variance may be mentioned in the software documentation, few exploratory spatial data analysis programs have appropriate adjustments for it in the Monte Carlo routines. Implementing case-based randomization may be slightly more difficult for users, since more data are required to appropriately randomize the cases among the population at risk, particularly when adjustments for covariates is needed, as with our examples which are adjusted for age. It is feasible, however, with SaTScan and certain tests within ClusterSEER being examples of programs that include case-based randomization with adjustments for covariates (multinomial Poisson randomization), although case-based randomization is not available for the Local Moran’s $I$ in either package. We would place the unbiased assessment of statistical significance at a higher priority than increased burden of including the appropriate data needed for the Monte Carlo simulation. A further advantage of including both the numerator and denominator in the input data used for spatial analysis is that the neighborhood averages can then be weighted for population size in calculating the local indicators of spatial correlation, an adjustment that is not routinely implemented in spatial statistics packages.

A remaining issue is how to appropriately adjust for multiple statistical tests inherent in this type of exploratory spatial data analysis and the non-independence of the neighborhoods. We have used the rule of thumb of setting the $p$-value at alpha divided by the average number of neighbors, but this adjustment does not fully account for the multiple tests or independence issues.

Disease cluster techniques are commonly applied to county and ZIP-Code data in the US. The fact that much public health data in the US is only available for these units of analysis is unlikely to change in the near future. Due to the wide variation in population size of counties and ZIP Codes, variance instability will likely be an issue for many analyses that include even moderate coverage of a state or the nation. There is an emerging consensus that a variety of methods should be applied toward the exploration and evaluation of a data set (Hanson and Wieczorek, 2002; Jacquez and Greiling, 2003). While such a position is inarguable, it should not be construed to mean that all methods are of equal merit, or that all give equally informative results. Our illustration highlights the importance of understanding underlying assumptions and correctly specifying the null hypothesis to be tested.

**Acknowledgments**

This publication was made possible in part through a Cooperative Agreement between the Centers for Disease Control and Prevention (CDC) and the Association of Schools of Public Health (ASPH), award number S1355-20/22. Its contents are the responsibility of the authors and do not necessarily reflect the official views of the CDC or ASPH. We thank Martin Kulldorff for his helpful comments on an earlier version of this manuscript, as well as Mike Zdeb for assistance with SAS.

**References**


Anselin, L., 2003. GeoDa 0.9 User’s Guide. Spatial Analysis Laboratory (SAL). Department of Agricultural and Consumer Economics, University of Illinois, Urbana-Champaign, IL.


