PUBLISHED BY THE INTERNATIONAL SOCIETY FOR DISEASE SURVEILLANCE

Advances in Disease Surveillance

VOLUME 3
NUMBER 2
DATE 2007

ARTICLES

Using Quad Trees to Generate Grid Points for Applications in Geographic Disease Surveillance

Nikolaos Yiannakoulias[1,2], Anthony Karosas[3], Donald P Schopflocher[2,4,5], Lawrence W Svenson[2,4,5], and M John Hodgson[6]

[1] School of Geography and Earth Sciences, McMaster University, Hamilton, Ontario, Canada.
[2] Public Health Surveillance and Environmental Health, Alberta Health and Wellness, Alberta, Canada.
[3] Division of Population Health and Information, Cross Cancer Institute, Edmonton, Alberta, Canada.
[4] Department of Public Health Sciences, University of Alberta, Edmonton, Alberta, Canada.
[5] Department of Community Health Sciences, University of Calgary, Calgary, Alberta, Canada.
[6] Department of Earth and Atmospheric Sciences, University of Alberta, Edmonton, Alberta, Canada.

Received for publication October 1, 2006; accepted for publication March 2, 2007.

In this study, we compare two methods of generating grid points to enable efficient geographic cluster detection when the original geographical data are prohibitively numerous. One method generates uniform grid points, and the other employs quad trees to generate nonuniform grid points. We observe differences in the results of the spatial scan approach to cluster detection for these two grid generation schemes. In both our simulated experiments and our analysis of real data, the grid generation schemes produced different results. Generally speaking, the quad tree scheme is more sensitive to high-resolution spatial clusters than the uniform scheme. The quad tree grid point scheme may be a useful and flexible alternative to the uniform (and other) grid point generation schemes when it is important to set up a timely surveillance system sensitive to finding clusters at unspecified spatial resolutions. The quad tree grid scheme may also be useful in a number of other geographic surveillance applications.

surveillance; medical geography; spatial representation; methodological study

INTRODUCTION

For some applied problems in geographic disease surveillance, high-resolution data can be a mixed blessing. On the one hand, high-resolution data can reveal spatial detail that is undetectable at lower resolutions; on the other hand, these data can create computational and analytical burdens that make timely analysis difficult. When these burdens are a concern, such data must be simplified, which usually involves reducing the number of spatial objects, and averaging, summing or otherwise aggregating attributes associated with these objects. This process almost always results in a loss of precision in location and other attributes of the raw data.

Little research has formally evaluated the effect of aggregation or other methods of data simplification on methods used to detect geographic clusters of disease, though there are some exceptions. For example, Waller and Turnbull analyzed the effects of resolution (or scale) on different focused cluster detection methods (1). Sheehan et al. found similar clusters of late-stage cancer diagnoses for three different geographical aggregation units (2). Gregorio

Correspondence to Nikolaos Yiannakoulias, Public Health Surveillance and Environmental Health, 24th Floor, 10025 Jasper Avenue, Edmonton, Alberta, Canada, T5J 1S6 (e-mail: nwy@ualberta.ca).

TABLE 1. Estimates of time to solve large cluster detection problems

	30,000 original data points		100,000 original data points		1,000,000 original data points	
Search points	1,000	30,000	1,000	100,000	1,000	1,000,000
Total unit operations required	3.00×10^{10}	8.99×10^{11}	9.99×10^{10}	9.99×10^{12}	9.99×10^{11}	9.99×10^{14}
Time*	14 min	7 h	47 min	78 h	8 h	324 d
Relative time	1	30	1	100	1	1,000

* Single 3.0 GHz processor, 500 Mb RAM, estimate of 40 million operations per second.

et al. observed that there are few benefits to the analysis of data at resolutions finer than the U.S. census tract (3). However, more recently, Olson et al. used simulated data to illustrate the benefits (in terms of sensitivity, precision, and likelihood of detection) of high-resolution data when trying to detect geographic clusters (4). The general dearth of research in this area may be limited by data availability—few researchers have had access to data that are so computationally unmanageable as to require simplification for most cluster detection tasks. In this sense, the simplification process has already taken place, and researchers are resigned to working with the spatial representations from which they acquire data.

With the growing availability of high-resolution population and disease data, organizations and researchers with a mandate to perform routine surveillance are increasingly forced to balance complexity with detail. Ideally, data are simplified in a manner that loses as few important details as possible. In this study, we compare two methods of generating grid points to facilitate efficient geographic cluster detection when the original geographic data are prohibitively numerous. One method generates uniform grid points, and the other employs quad trees to generate nonuniform grid points. We use the spatial scan method for detecting geographic clusters as the analytical tool for testing these two schemes, though our observations apply to several different methods of cluster detection and spatial analysis. We use simulated data to test the general performance of these two grid generation options, and then conclude our study with a search for clusters of Parkinson's disease using these two schemes.

METHODS

Spatial scan

The spatial scan approach to cluster detection (5) has gained considerable applied and methodological research attention over the past decade. In its typical form, the spatial scan uses circular windows to scan a region for clusters of high (or low) disease. This involves progressively increasing the radius of these circles and accumulating data points into the windows until a threshold of cluster size is met. The window with the largest likelihood ratio test statistic is treated as a most-likely cluster, and Monte Carlo methods are used to test its significance. By searching a large number of these windows but testing the significance of only the most-likely cluster (and perhaps a few secondary clusters), the method has high sensitivity to detect clusters without the burden of multiple testing or preselection bias common to some other methods of cluster detection (6). Furthermore, if a null hypothesis of constant risk is rejected, the method reports the window that caused the rejection—thus serving as a test for the presence of localized clustering as well as a tool to locate where noteworthy clusters occur.

The circular windows used in the spatial scan are centered at search points, which can be original data points (such as case/control locations or the centroids of polygon areas), or the points in an overlying grid (which we refer to as "grid points"). Among the chief reasons for using grid points rather than original points to search for clusters is to ensure that search processes can be completed in a reasonable time when working with large data sets. Table 1 provides estimates of the time required to find spatial clusters using SaTScan on the basis of the estimation formula offered in the software's documentation (7). Estimates are based on the Poisson model without adjustment for covariates. The traditional spatial scan approach can confront a noteworthy computation burden when applied to large data sets; for example, the analysis of 30,000 original data points can take approximately 7 hours to complete when the original data are used, but a mere 14 minutes if the search process uses 1,000 grid points. For larger problems, or when the scan is over space-time or adjusts for covariates, the solution time might be unreasonably long on a desktop computer without the use of an overlying grid system.

We now discuss two different methods of generating overlying sets of grid points—the uniform grid point method and the quad tree grid point method.

Uniform grid points

Uniform grids are widely used in the spatial analysis of disease. The idea is simple: overlay a uniformly sized tessellation of polygons (usually rectangles, squares, or hexagons) on a study area, and use the centroids or intersecting points of these polygons as the grid points.

Although the exact placement, spacing, and alignment of the uniform grid can vary, the method suggests analytical neutrality; as the grid is uniform, an analyst cannot be accused of analytical gerrymandering. This scheme has some shortcomings, however, the most serious of which relates to the density of points themselves. In study regions where population density or data resolution varies considerably, the choice of a uniform grid that is too blunt in some places (i.e., where the spacing between grid points is too far apart) may preclude the identification of clusters in some small areas. Some scanning windows may find a cluster area, but not without including a proportion of the area

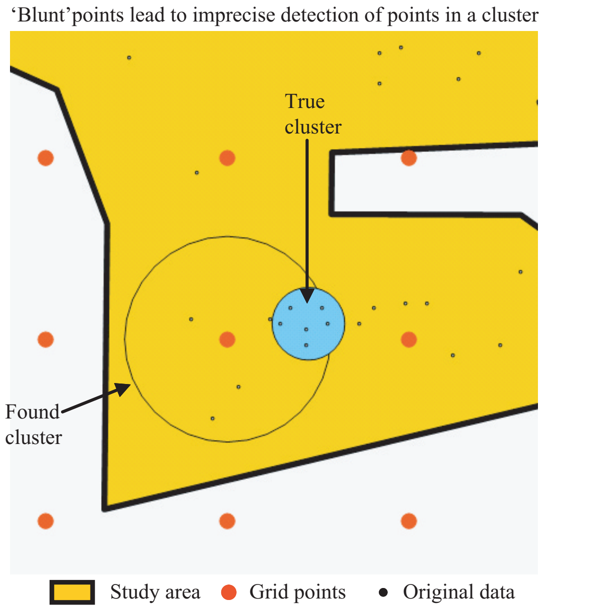
'Blunt'points lead to imprecise detection of points in a cluster
True cluster
Found cluster
Study area
Grid points
Original data

FIGURE 1. Examples of the drawbacks of using a uniform grid.

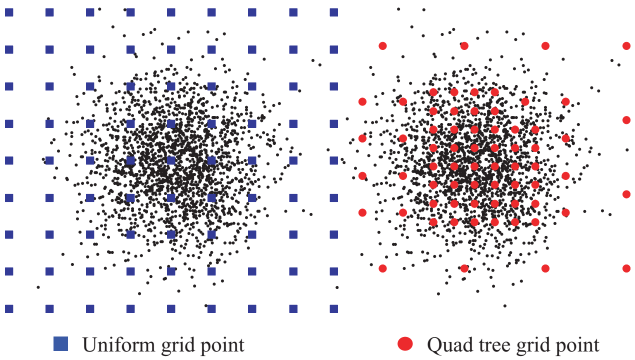

FIGURE 2. Uniform and quad tree grid generation schemes.

that is not part of the true cluster. This decreases both the efficiency and precision with which the method can identify the approximate location of the cluster in the first place (figure 1).

Quad tree grid points

Quad trees (8) have been used for indexing and storage in spatial databases for some time, and recently have been used in multiscale disease modeling (9). In data-indexing applications, the use of quad trees can save time in searching or merging spatial data. Our application uses quad trees to create a tessellation of nonuniform rectangles. As in the uniform grid generation method, this tessellation can then be simplified into a system of grid points (using the center of each rectangle) that can then serve as search seeds in the scan for clusters. Although there are considerably fewer grid points than original points in the system, quad tree-generated grid points will generally reflect the spatial distribution of the original data. We propose the quad tree system as a simple alternative to a uniform grid.

In our application, we create quad trees as follows. First, we define a study space consisting of original point data. Next, we choose a point threshold—for example, a value of 50. This threshold is a matter of convenience that can be determined before the analysis of disease data; lower values create higher-resolution tessellations (with more grid points) and higher values create lower-resolution tessellations (with fewer grid points). If the number of original data points in the study space is larger than the size of the threshold, we split the study space into four separate, equal-area rectangles. We count the number of points within each rectangle, and if the value exceeds the threshold, that rectangle is further subdivided. This is repeated for all rectangles, at all levels of subdivision, until no rectangle has a number of original data points in excess of the threshold. For both methods, the geometric centers of the rectangles can be used as the grid points in the spatial scan. We provide figures of a completed uniform and quad tree grid system for illustration (figure 2).

We test these schemes in two ways: first with an experiment based on simulated data, and then on high-resolution data on Parkinson's disease. We use the term "atomic data" to refer to highest-resolution data available. In our experiments and our application, this refers to centroids of small geographic areas, but can easily apply to case/control data in other settings. Despite the differences in what "atomic data" may represent in different applications, the general term is useful for referring to data that are irreducible because of data availability, privacy, measurement constraints, or metaphysics.

Experiment

We use independent standard normal distributions to generate x and y coordinates of 3,000 centrally clustered atomic data points on a Cartesian plane. Each atomic data point is treated as though it is the centroid of a high-resolution small-area polygon (such as a postal code or census block). For each simulation, we specify a baseline rate and a cluster rate that correspond to a baseline region and a cluster region, respectively. For each synthetic data set, the cluster region is defined by a circle with a center and radius that are systematically varied to observe the effectiveness of the two grid generation schemes across different cluster sizes and locations. The baseline region is always the complement of a synthesized cluster region. Each atomic data point is assigned a population of 1,000 people, and each person in this population is assigned a probability of being a case or noncase based on whether he/she is in a baseline or cluster region. People in the baseline region have a probability of being a case equal to the baseline rate, which is the same for all simulations (0.0005). People in the cluster region have a probability of being a case equal to the cluster rate, which for the first set of simulations is 0.001 and the second set of simulations, 0.00075. In total, 50 different types of synthetic disease cluster scenarios are generated: 2 different cluster rates × 5 different cluster region radii (0.2, 0.4, 0.6, 0.8, and 1.0) × 5 different locations of the cluster region center (where $x = 0, 0.5, 1.0, 1.5,$ and 2.0 and $y = 0$ for all values of x). Each synthetic scenario is repeated 100 times.

We apply the quad tree and uniform schemes to generate grid points and search for clusters in each synthetic disease cluster scenario. The number of quad tree cells is determined dynamically and is dependent on the population threshold and the distribution of the atomic data. To make the grid and quad tree methods comparable, the grid tessellation is generated on the basis of the number of quad tree cells. For example, if there are 100 quad tree cells, we determine the number of grid cells by taking the square root of 100 and create a 10 × 10 uniform grid. When the square root of the number of quad tree cells is not an integer, the counts are rounded in favor of the uniform grid. For example, if there are 1,000 quad tree cells, the grid axis is based on a 32 × 32 cell grid (for a total of 1,024 cells). The outer dimensions of both tessellations are identical, and are based on a bounding box that includes all atomic data. For each tessellation, the grid centroids are defined as the center of the cells.

For each simulated data set, we solve two spatial scan cluster detection problems—once with the quad tree centroids and once with the grid centroids as the centers of the search window. As we synthesize the clusters, we are able to measure the agreement between the two methods with the "true" cluster area. We do this by calculating the proportion of the detected circular cluster that overlaps (and does not overlap) the cluster region. This forms the basis for comparison between the two methods (figure 3). The proportion of true area detected (AB/A) describes the proportion of the found cluster region that correctly intersects the true cluster area. Values close to 1 indicate that over the 100 repetitions of a given scenario, most clusters completely intersect the cluster region; values close to 0 indicate that there is little intersection. The proportion of false area detected ((B − AB)/B) describes the degree to which the found clusters provide false information. Values close to 1 indicate that over the 100 repetitions of a given scenario, most found clusters were located entirely in the baseline region; values near 0 indicate that most of the detected area intersects with the cluster region. For the proportion of true area detected, large values are desirable; for the proportion of the false area detected, small values are desirable.

We generated the grid and quad tree grid points in SAS (10), and used an SAS program to perform the spatial scan and calculate the overlap of the "true" cluster regions and the most-likely clusters found by the two methods. The maximum radius of the most-likely cluster found was set by a 50% population threshold, and 999 Monte Carlo simulations were used to obtain significance estimates associated with the most-likely cluster. The SAS code for generating quad trees code is freely available for download as additional file 1.

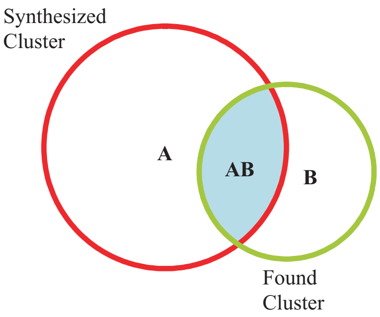

FIGURE 3. Proportion of overlap between synthesized and found clusters.

Application to Parkinson's disease

We use fee-for-service administrative health data to identify people who had received a diagnosis of Parkinson's disease (ICD-9-CM 332.x) (11). We use a public health insurance registration system to identify the number of people in each postal code. Each postal code is assigned a number of cases and population. We use all postal code locations in the province of Alberta with populations greater than 0 as the atomic data set ($N = 40{,}610$).

We define cases of Parkinson's disease in two ways. First, we define people as incident Parkinson's disease cases if they have had two or more diagnoses of Parkinson's disease in 2004, and no previous diagnosis of Parkinson's disease. Second, we define people as prevalent Parkinson's disease cases if they had been an incident case (two or more diagnoses for Parkinson's disease) at any time up to and including 2004. As the smallest resolution unit is the postal code—rather than individual residences—we use the Poisson model to identify most-likely clusters of Parkinson's disease under both of these definitions. We incorporate age and sex as covariates—thereby adjusting found clusters for geographic variations in these variables across the province. We set up the search system to find clusters using SaTScan version 7.0 (12) for the uniform and quad tree grid point generation

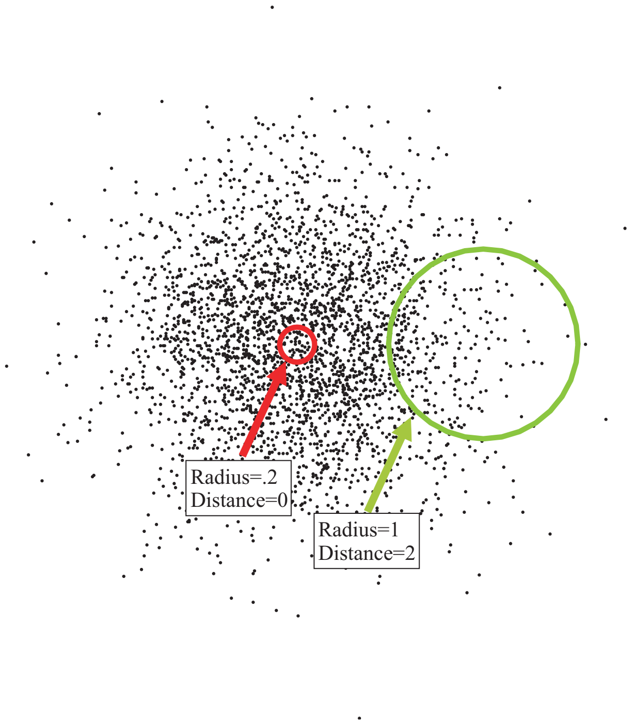

FIGURE 4. Examples of synthesized clusters.

TABLE 2. Experimental results 1 (cluster rate 2 × the baseline rate; cluster rate = 0.001; baseline rate = 0.0005)

Horizontal distance from origin ($x = 0$, $y = 0$)	Radius	Average* proportion of true area found (AB/A) (higher is better)		Average* proportion of found cluster in baseline region (B − AB)/B (lower is better)		Proportion significant at $p \leq 0.001$	
		Quad tree	Uniform	Quad tree	Uniform	Quad tree	Uniform
0	0.2	0.668	0.578	0.455	0.739	0.46	0.27
0	0.4	0.869	0.837	0.178	0.366	1.00	1.00
0	0.6	0.930	0.891	0.111	0.259	1.00	1.00
0	0.8	0.941	0.913	0.088	0.197	1.00	1.00
0	1	0.951	0.933	0.067	0.161	1.00	1.00
0.5	0.2	0.641	0.558	0.554	0.764	0.31	0.14
0.5	0.4	0.866	0.841	0.190	0.371	1.00	0.99
0.5	0.6	0.914	0.884	0.125	0.256	1.00	1.00
0.5	0.8	0.942	0.915	0.089	0.205	1.00	1.00
0.5	1	0.947	0.925	0.067	0.182	1.00	1.00
1	0.2	0.582	0.519	0.613	0.826	0.21	0.15
1	0.4	0.838	0.858	0.226	0.433	0.98	0.94
1	0.6	0.902	0.870	0.114	0.288	1.00	1.00
1	0.8	0.930	0.888	0.101	0.217	1.00	1.00
1	1	0.953	0.933	0.086	0.186	1.00	1.00
1.5	0.2	0.277	0.248	0.849	0.926	0.04	0.03
1.5	0.4	0.776	0.815	0.288	0.400	0.68	0.57
1.5	0.6	0.851	0.844	0.197	0.309	0.99	0.99
1.5	0.8	0.884	0.913	0.126	0.268	1.00	1.00
1.5	1	0.927	0.907	0.125	0.189	1.00	1.00
2	0.2	0.130	0.204	0.977	0.971	0.00	0.00
2	0.4	0.430	0.519	0.705	0.659	0.14	0.18
2	0.6	0.767	0.842	0.388	0.425	0.63	0.75
2	0.8	0.830	0.860	0.287	0.280	0.99	1.00
2	1	0.865	0.907	0.200	0.263	1.00	1.00

* Average calculated over 100 synthetic data sets.

schemes, as well as grid points based on the centroids of Canadian census divisions and census subdivisions (13). Census divisions and subdivisions are large administrative census units that have been used in the provincial-scale analysis of geographic patterns of disease in Canada (e.g., 14–17). For all these analyses, we use the same scan settings: a 50% population threshold, 999 Monte Carlo simulations, and a no-overlapping rule for the identification of secondary clusters.

RESULTS

Experimental results

We present examples of the synthesized clusters for visualization purposes (figure 4). We present summary indicators of the performance of the grid and quad tree methods in tables 2 and 3. Data are sorted by column one, the horizontal distance between the origin ($x = 0$, $y = 0$) and the synthesized cluster region center; larger values indicate that the cluster region is located farther from the densest area of atoms. The "Radius" column indicates the radius of the cluster region. Generally speaking, the quad tree grid method appears to perform as well as or better than the grid method for the majority of scenarios. The most notable exception is for scenarios where the cluster regions are far from the origin (and center of population) and the synthesized cluster radius is large. In these cases, the uniform grid appears superior.

On most occasions, the two methods are fairly similar in their ability to locate a cluster region precisely and differ more with respect to the remaining area discovered (columns 5 and 6). When the radius of the cluster region is small and centrally located, the quad tree method is more efficient and identifies clusters that intersect less of the baseline region than clusters found with the uniform grid scheme. For cluster regions farther from the origin, the pattern is reversed, though less noticeable. This different ability of the two schemes to correctly identify areas as inside and not inside a cluster region is illustrated, albeit indirectly, in the final two columns of tables 2 and 3. These columns indicate the

TABLE 3. Experimental results 1 (cluster rate 1.5 × the baseline rate; cluster rate = 0.00075; baseline rate = 0.0005)

Horizontal distance from origin ($x = 0$, $y = 0$)	Radius	Average* proportion of true area found (AB/A) (higher is better)		Average* proportion of found cluster in baseline region (B − AB)/B (lower is better)		Proportion significant at $p \leq 0.001$	
		Quad tree	Uniform	Quad tree	Uniform	Quad tree	Uniform
0	0.2	0.298	0.231	0.774	0.907	0.05	0.04
0	0.4	0.735	0.677	0.267	0.429	0.59	0.40
0	0.6	0.908	0.825	0.164	0.277	0.97	0.92
0	0.8	0.883	0.856	0.095	0.180	1.00	1.00
0	1	0.921	0.877	0.083	0.151	1.00	1.00
0.5	0.2	0.289	0.303	0.815	0.929	0.05	0.01
0.5	0.4	0.760	0.678	0.318	0.504	0.49	0.39
0.5	0.6	0.877	0.822	0.176	0.273	0.95	0.91
0.5	0.8	0.882	0.852	0.083	0.187	1.00	1.00
0.5	1	0.908	0.894	0.074	0.152	1.00	1.00
1	0.2	0.170	0.166	0.869	0.943	0.03	0.01
1	0.4	0.637	0.655	0.419	0.584	0.31	0.20
1	0.6	0.806	0.787	0.213	0.379	0.80	0.75
1	0.8	0.870	0.856	0.139	0.229	0.98	0.97
1	1	0.915	0.893	0.133	0.239	1.00	1.00
1.5	0.2	0.080	0.118	0.976	0.982	0.01	0.01
1.5	0.4	0.373	0.384	0.697	0.808	0.08	0.05
1.5	0.6	0.666	0.694	0.323	0.450	0.44	0.44
1.5	0.8	0.830	0.852	0.260	0.353	0.86	0.86
1.5	1	0.844	0.885	0.157	0.289	0.99	0.98
2	0.2	0.073	0.136	0.995	0.977	0.01	0.00
2	0.4	0.166	0.278	0.916	0.885	0.00	0.01
2	0.6	0.374	0.478	0.644	0.581	0.05	0.15
2	0.8	0.633	0.722	0.411	0.405	0.41	0.48
2	1	0.750	0.814	0.308	0.339	0.75	0.71

* Average calculated over 100 synthetic data sets.

proportion of times a method finds a cluster that meets a threshold of significance ($p \leq 0.01$). When a large proportion of the baseline region is falsely identified as part of a cluster region, this should decrease the likelihood of detection (as observations with lower disease rates will decrease the likelihood ratio associated with the most-likely cluster). For most sets of synthetic data, the methods are equivalent in their ability to find statistically significant clusters; however, the quad tree grid-generating scheme appears superior when the radius of the cluster region is small and centrally located.

A visual comparison of the results in tables 2 and 3 suggests that the general patterns are comparable.

Parkinson's disease clusters in Alberta

For the quad tree method, the density of grid points is much higher in highly populated areas; more than 60% of the quad tree grid points are in urban areas (figure 5). For example, in the cities of Edmonton and Calgary, there are more than 100 quad tree grid points, but only one or two uniform grid points. The census divisions ($N = 19$) and census subdivisions ($N = 467$) are considerably outnumbered by the uniform and quad tree grid points. To ensure a fair comparison between the uniform and quad tree schemes, the uniform scheme was oversampled (through trial and error) to ensure that there would be a comparable number of uniform grid points after the removal of grid points that fell outside the provincial boundary. After the removal of these points, there were 1,244 uniform grid points and 1,236 quad tree-generated grid points.

The clusters of Parkinson's disease are presented in table 4. A large number of prevalent Parkinson's disease clusters were significant, and we report all found clusters with a Monte Carlo-estimated significance less than or equal to 0.01. No incident clusters met this threshold of significance, but we report primary most-likely clusters for each of the grid generation schemes for illustrative purposes. For prevalent Parkinson's disease, searches based on uniformly distributed grid points and census division grid points found clusters with considerably larger radii than searches based

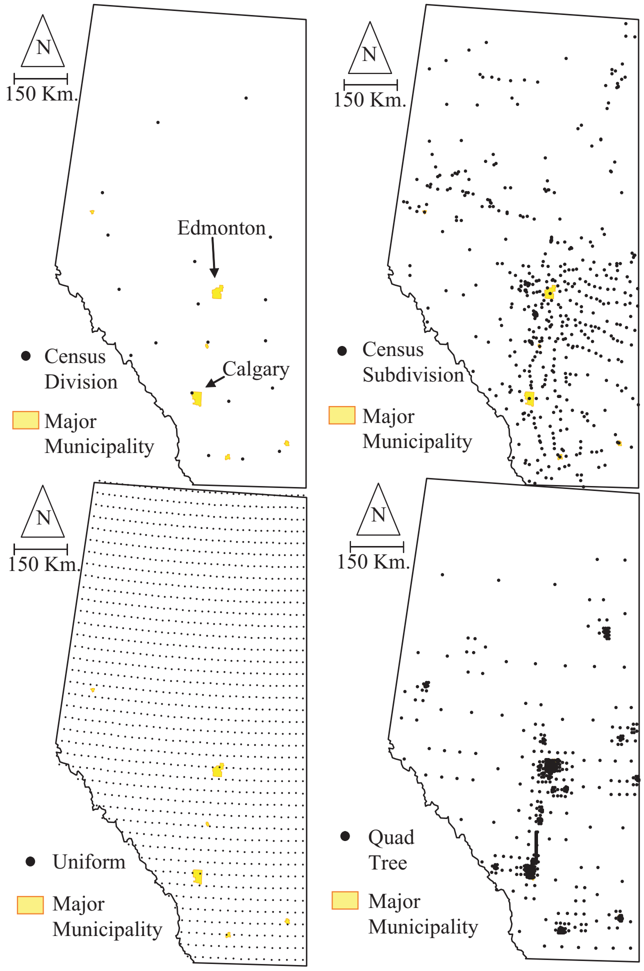

FIGURE 5. Grid points for the four schemes.

TABLE 4. Clusters of Parkinson's disease

Type	Order	Radius	Obs/Exp	*p*
Prevalent quad	Primary	0.600	4.238	0.001
Prevalent quad	Secondary	0.390	3.892	0.001
Prevalent quad	Secondary	0.320	5.487	0.001
Prevalent quad	Secondary	1.140	2.604	0.001
Prevalent quad	Secondary	16.160	3.035	0.001
Prevalent quad	Secondary	0.380	2.920	0.001
Prevalent quad	Secondary	0.750	2.649	0.001
Prevalent quad	Secondary	1.000	2.768	0.001
Prevalent quad	Secondary	1.120	6.254	0.001
Prevalent quad	Secondary	0.480	6.770	0.001
Prevalent uniform	Primary	159.300	1.164	0.001
Prevalent uniform	Secondary	73.160	1.371	0.001
Prevalent uniform	Secondary	41.890	1.460	0.001
Prevalent CD	Primary	181.400	1.142	0.001
Prevalent CSD	Primary	3.500	1.472	0.001
Prevalent CSD	Secondary	0.790	3.839	0.001
Prevalent CSD	Secondary	2.790	1.829	0.001
Prevalent CSD	Secondary	0.800	3.106	0.001
Prevalent CSD	Secondary	0.700	2.504	0.003
Incident quad	Primary	6.940	1.488	0.177
Incident uniform	Primary	44.860	1.261	0.517
Incident CD	Primary	24.990	1.268	0.024
Incident CSD	Primary	8.850	1.347	0.186

on quad tree grid points (figure 6). Clusters based on quad tree and census subdivision grid points were small and mostly located in and around urban areas, whereas the uniform and census division grid points were located in rural areas and covered larger portions of the province. For incident Parkinson's disease, all most-likely clusters were small and located in either Edmonton or Calgary (figure 7).

DISCUSSION

For most scenarios examined, the quad tree grid performs as well as or better than the uniform grid in identifying the location of clusters of events, though the differences are not large. Our experiment suggests that the uniform grids might be insensitive to high-resolution clusters—bluntly reporting clusters of a larger size than necessary. This not only misclassifies regions with normal rates of disease as part of cluster regions but also appears to affect the power of detection under some circumstances. The findings are relatively consistent across different scenarios, with some notable exceptions. As the simulated cluster regions are generated farther from the mass of the population, a search for clusters using the uniform grid points reports a higher success rate than a search for clusters using the quad tree grid points. In areas where the atomic data are sparse, the quad tree points are also sparse and less likely to find true clusters in these locations. The uniform grid retains more points in these areas and appears slightly more sensitive to finding clusters. Were the clusters even farther from the population centers (where $x = 3$ or more, for example), the superiority of the uniform grid generation scheme would be even greater.

The incident clusters of Parkinson's disease were below the threshold of statistical significance, although the cluster found using the quad tree ($p = 0.177$) and census division ($p = 0.023$) grid points may still warrant a field investigation. The prevalent clusters, though significant, may simply reflect the mobility effect; people with serious chronic diseases prefer to live in certain areas (18), and, in particular, urban areas with more services. For both reasons, the clinical implications of our cluster investigation are probably minor. Nevertheless, it is worth noting that the locations and sizes of Parkinson's disease clusters differ considerably depending on the type of overlying grid used. Although the incident clusters overlap, they are of considerably different sizes. Most of the prevalent clusters do not overlap, and they occur at different locations and in different sizes throughout the province. This suggests not only that a hierarchy of variations in Parkinson's disease prevalence may exist (some local and some regional) but also that the choice of overlaying grid is not trivial.

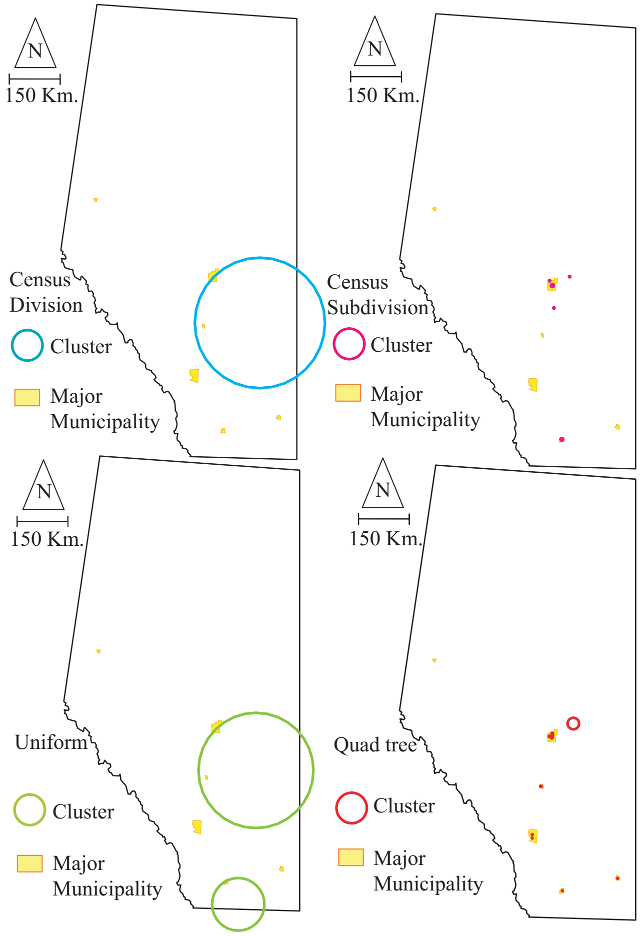

FIGURE 6. Clusters of Parkinson's disease prevalence.

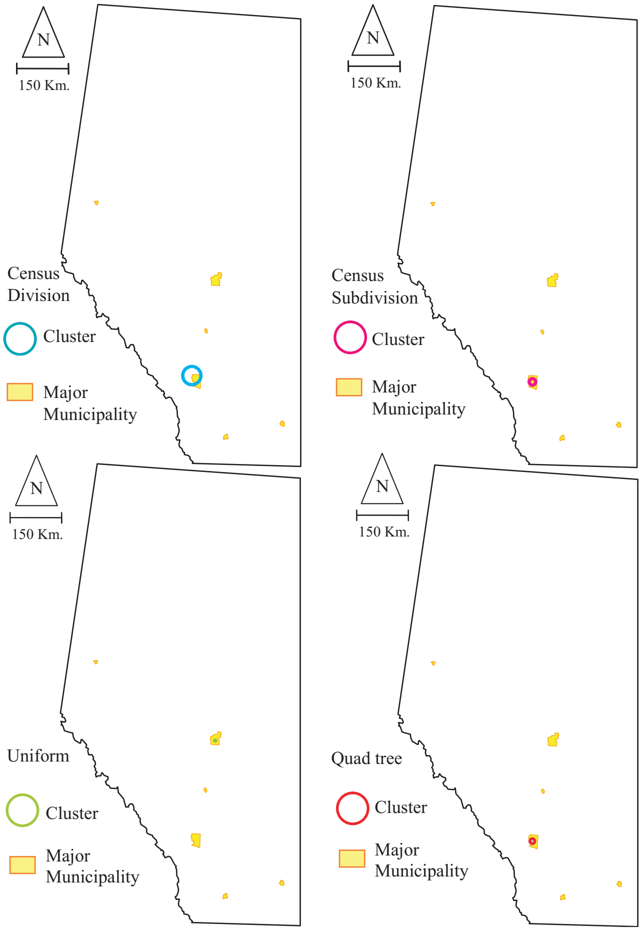

FIGURE 7. Clusters of Parkinson's disease incidence.

Other grid options were available—such as using the centroids of various other administrative census units or a random sample of atomic data points. These other schemes may very well have produced other series of cluster locations or clusters at different resolutions. Unfortunately, it is not easy to determine, at least in the general case, which method is superior. The quad tree method has the advantage of being deterministic (unlike a random sampling approach) and in the control of the user (since the user chooses the criteria for creating the quad tree grid, rather than relying on pre-constructed census-area centroids). In practice, whether these advantages are worth the extra work of generating the quad tree grid points will depend on the application, but the method seems particularly well suited to settings in which population density is highly variable.

Clayton and Gangnon observe that the spatial scan has a tendency preferentially to detect clusters in areas where the search seeds are denser (19, 20). They recommend a penalty be applied to the likelihood ratio to take this tendency into account. Our results provide some indirect evidence that this effect may be of more concern when the search points are nonuniform. Uniform grid points are more evenly distributed, and the circular window searches originate at uniform locations throughout the synthesized study area. When the cluster regions are synthesized farther from the center (where atomic data points are sparsely located) the difference between the grid generation schemes is relatively small, but on some occasions a search for clusters using the uniform grid scheme appears better at finding the synthesized cluster regions.

Fortunately, the quad tree approach provides a framework for directly addressing these issues. The method for generating the quad tree tessellation is not specific about what criteria necessitate the "split" of a cell; in our application, we chose the population of the atomic data to decide whether a cell is split any further. In particular, a cell is split into four new cells if the sum of the atomic population exceeds a certain threshold. Different criteria could be used to influence this splitting process. For example, concerns about edge effects could be managed by imposing a larger weight on atoms located in the margins of a study area, and a smaller weight to atoms located in the middle of a study area. This would create more grid points proportionally in edge regions. Concerns about the tendency of a search to overlook rural areas could be offset in a similar manner; simply apply a larger weight to atoms in rural areas than in urban areas. The manner in which quad trees are derived can be based on any weighting system or even a departure from this paradigm—for example, by basing the decision to split a cell on how homogeneous/heterogeneous the atoms within it are with respect to some attribute(s). As this can be done before running the cluster search algorithm, it does not necessarily undermine the inferential merit of the cluster search results.

In situations when a study region is elongated (e.g., Chile) or irregular in shape (e.g., Mexico), it may be advisable to subdivide the quad tree generation scheme into separate independent components. For example, long study regions could be first divided into a chain of several square (and adjacent) quadrants, within which independent quad tree point sets can be generated. This can take into account regional or local variations in geography—such as islands, peninsulas, or areas on opposite sides of a mountain—that may require an independent grid generation scheme for a meaningful representation. Similar to the use of weights mentioned above, such decisions do not necessarily undermine the inferences of the cluster detection search so long as they are performed *a priori*.

A number of methods have employed uniform grids in geographic disease surveillance (21–23). Many methods of spatial cluster detection can easily employ the quad tree grid approach when such a grid system is required. We suggest that the quad tree scheme is a simple and elegant alternative to the uniform scheme for generating grid points and has the advantage of being in the analyst's control (unlike census or other administrative units) and deterministic (unlike a random sampling strategy). It can also be used as a simple scheme for aggregating atomic data that preserves the spatial distribution of atomic data points.

ACKNOWLEDGMENTS

M.J.H. is supported by a Discovery Grant from the National Sciences and Engineering Research Council. The authors would like to thank the editors and anonymous reviewers for their helpful comments on the manuscript.

REFERENCES

1. Waller LA, Turnbull BW. The effects of scale on tests for disease clustering. Stat Med 1993;12:1869–84.
2. Sheehan TJ, Gershman ST, MacDougall LA. et al. Geographic assessment of breast cancer screening by towns, zip codes and census tracts. J Pub Health Manag Pract 2000;6:48–57.
3. Gregorio DI, DeChello LM, Samociuk H, Kulldorff M. Lumping or splitting: seeking the preferred areal unit for health geography studies. Int J Health Geogr 2005;4:6.
4. Olson KL, Grannis SJ, Mandl KD. Privacy protection versus cluster detection in spatial epidemiology. Am J Pub Health 2006;96:2002–8.
5. Kulldorff M. A spatial scan statistic. Commun Stat Theory Meth 1997;26:1481–96.
6. Kulldorff M. Tango T, Park P. Power comparisons for disease clustering tests. Comput Stat Data Anal 2004;42:665–84.
7. Kulldorff M. SaTScan™ for version 6.1 User Guide, 2006. (Available from: http://www.satscan.org/) (accessed July, 2006).
8. Finkel RA, Bentley JL. Quad trees: a data structure for retrieval on composite keys. Acta Inform 1974;4:1–9.
9. Louie MM, Kolaczyk ED. Multiscale detection of localized anomalous structure in aggregate disease incidence data. Stat Med 2006;25:787–810.
10. SAS Institute: SAS 9.0. SAS Institute Incorporated. Cary, North Carolina, 2002.
11. International Classification of Diseases, 9th Revision. 2006 Los Angeles: Practice Management Information Cooperation.
12. Kulldorff M. Information Management Services, Inc. SaTScan™ v.7.0: Software for the spatial and space-time scan statistics, 2006. (Available from: http:/www.satscan.org/).
13. Statistics Canada. Geography products and services. (Available from http://geodepot.statcan.ca/Diss/Products/Products_e.cfm) (accessed January 31, 2007).
14. Feasby TE, Quan H, Ghali WA. Geographic variation in the rate of carotid endarterectomy in Canada. Stroke 2001;32: 2417–22.
15. Svenson LW, Woodhead SE, Platt GH. Regional variations in the prevalence rates of multiple-sclerosis in the province of Alberta, Canada. Neuroepidemiology 1994;13:8–13.
16. Mao Y, Desmeules M, Semenciw RM, Hill G, Gaudette L, Wigle DT. Increasing brain cancer rates in Canada. Canadian Med Assoc J 1991;145:1583–91.
17. Pearl DL, Louie M, Chui L, Dore K, Grimsrud KM, Leedell D, Martin SW, Michel P, Svenson LW, McEwen SA. The use of outbreak information in the interpretation of clustering of reported cases of *Escherichia coli* O157 in space and time in Alberta, Canada, 2000–2002. Epidemiol Infect 2006;134: 699–711.
18. Bentham G. Migration and morbidity: implications for geographical studies of disease. Soc Science Med 1988;26:49–54.
19. Gangnon RE, Clayton MK. A weighted average likelihood ratio test for spatial clustering of disease. Stat Med 2001; 2:2977–87.
20. Gangnon RE, Clayton MK. Likelihood-based tests for localized spatial clustering of disease. Environmetrics 2004; 15:797–810.
21. Openshaw S, Craft A, Charlton M, et al. Investigation of leukemia clusters by use of a geographical analysis machine. Lancet 1988;331:272–73.
22. Rushton G, Lolonis P. Exploratory spatial analysis of birth defect rates in an urban population. Stat Med 1996;5:717–26.
23. Talbot O, Kulldorff M, Forand SP, Haley VB. Evaluation of spatial filters to create smoothed maps of health data. Stat Med 2000;19:2399–408.