Research Paper ■

Power to Detect Spatial Disturbances under Different Levels of Geographic Aggregation

CAROLINE JEFFERY, MA, AL OZONOFF, PHD, LAURA F. WHITE, PHD, MIRIAM NUÑO, PHD, MARCELLO PAGANO, PHD

Abstract **Objective:** Spatio and/or temporal surveillance systems are designed to monitor the ongoing appearance of disease cases in space and time, and to detect potential disturbances in either dimension. Patient addresses are sometimes reported at some level of geographic aggregation, for example by ZIP code or census tract. While this aggregation has the advantage of protecting patient privacy, it also risks compromising statistical efficiency. This paper investigated the variation in power to detect a change in the spatial distribution in the presence of spatial aggregation.

Methods: The authors generated 400,000 spatial datasets with varying location and spread of simulated spatial disturbances, both on a purely synthetic uniform population, and on a heterogeneous population, representing hospital admissions to three community hospitals in Cape Cod, Massachusetts. The authors evaluated the power of the *M*-statistic to detect spatial disturbances, comparing the use of exact spatial locations versus twelve different levels of aggregation, where the *M*-statistic is a comparison of two distributions of interpoint distances between locations.

Results: When the spread of simulated spatial disturbances was contained to a small portion of the study region or affects a large proportion of the population at risk, power was highest when exact locations were reported. If the spatial disturbance was a more modest signal, the best power was attained at an aggregated level.

Conclusions: The precision at which patients' locations are reported has the potential to affect the power of detection significantly.

■ **J Am Med Inform Assoc.** 2009;16:847–854. DOI 10.1197/jamia.M2788.

Introduction

In classical (temporal) disease surveillance, one looks for disturbance in the number of cases, while in a spatiotemporal system, not only the number of cases is observed but their location is also recorded. Considering the spatial component of the data can enhance the detection of an outbreak dramatically,[1,2] and thus it is important to determine how best to collect and analyze these data.

Because data such as residential address or place of work can help identify an individual, privacy concerns dictate that some filtering should occur. One approach to such filtering might be an intentional coarsening of the spatial resolution at which the data are reported. Geographic information might be reported in an aggregate form, for example the number of cases per some administrative unit such as county, ZIP code, or census tract. In this article, we investigate how this coarsening of the data affects the performance of a global spatial test, the *M*-statistic, to detect geographically localized outbreaks, using both a uniform baseline population as well as a heterogeneous population based on Emergency Department (ED) visits to three community hospitals in Cape Cod, Massachusetts.

Background

After reviewing the current literature, we explore some important considerations involved in the study of the effect of spatial resolution on detection of spatial disturbances.

The Effect of Spatial Data Resolution on Detection

One of the most influential maps in the history of medicine is John Snow's depiction of the mortality due to cholera in Soho in the autumn of 1854. Snow created this map to bolster his case against the miasmatists and to support his theory that the cholera is a waterborne disease.[3] To obtain the data for the map, Snow wrote: "I requested permission on the 5th of September, to take a list, at the General Register Office, of the deaths from cholera registered during the week ending the 2nd of September in the subdistricts of Golden Square and Berwick Street, St. James's, and St. Anne's, Soho, which was kindly granted"[4]. Would these data be kindly granted today or would a single number be provided for every Postal Code, or other aggregating device? What information would the aggregated data conceal?

Affiliations of the authors: Department of Biostatistics (CJ, AO, LFW, MN, MP), Harvard School of Public Health, Boston, MA; Department of Biostatistics (AO, LFW), Boston University School of Public Health, Boston, MA.

The research in this paper was funded by a grant from the National Institutes of Health, R01 EB0006195.

The authors thank the Institute of Health Metrics for providing us with their data (http://www.healthmetrics.org).

Correspondence: Caroline Jeffery, MA, Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115.

Received for review: 03/12/08; accepted for publication: 08/01/09.

Several authors have already written on the trade-off between accuracy of data and protection of privacy. Cox[5] described several existing methods, including random perturbation and aggregation, which prevent tabular data from revealing unintentional information. Minot et al.[6] showed that estimates of poverty rates are biased when based on census data that is released at an aggregated level, yet they also found that the bias can be reduced if one can obtain information about the variance of per capita expenditure at the household level. Informatics-based approaches have also appeared as a new direction. For example, Boulos et al.[7] suggested new software agents that would analyze sensitive data without requiring a human intervention, and report only the results to the user while concealing the data from which they were drawn. Armstrong et al.[8] proposed masking techniques applied to matrices of individual attributes. They considered the effect of random perturbation on the power of the Cuzick-Edwards test of spatial clustering[9] in a heterogeneous population and reported that higher levels of perturbation make detection less probable.

Similar results appeared in Cassa et al.[10] where the authors considered the power of Kuldorff's spatial scan statistic to detect spatial clusters injected into a data stream of ED visits from a Boston-area hospital. They first randomly displaced exact locations, and then aggregated both exact and perturbed locations by census tract before performing the analysis under each scenario. Results showed a gradual decrease in power of detection as the average distance between the original and modified locations increases, suggesting that the performance is not dramatically affected. Waller[11] studied the power of three different focused tests of clustering on the uniform distribution in the unit square at four levels of aggregation. Using two different types of clusters ("hot spot" and "clinal" clusters) showed that power decreased as the aggregation became coarser. Waller also reported power results for simulated clinal clusters added to the upstate New York leukemia data, suggesting that an appropriate aggregation level for optimal power might depend on whether the cluster is centered in an urban or a rural county.

Kulldorff et al.[12] introduced the space-time scan statistic (SaTScan) with the New York City emergency department syndromic surveillance system, where they showed that power of detection is reduced when hospital locations are used rather patients' residential ZIP code. Similarly, Olson et al.[13] used the spatial scan statistic on simulated clusters superimposed upon Boston hospital emergency data. They found that the performance in detection improves if spatial data are kept as exact point locations rather than aggregated by ZIP codes or census tracts, except possibly when a cluster falls entirely into one administrative unit. We have also studied the effect of aggregation on the performance of this statistic.[14] Our work considered a uniform baseline distribution in the unit disk and twelve subsequent levels of aggregation. We also found a steady loss of power to detect spatial disturbances as the spatial resolution coarsens.

Methods for Detecting Spatial Disturbances

Spatial methods in prospective surveillance aim to monitor the spatial distribution of incoming cases, and to detect any change in that distribution that might occur. There are several broad reviews of spatial methods in surveillance.[15–17] Rather than duplicate the literature, we focus on aspects that are especially relevant to the present study. One can classify methods to detect spatial disturbances into two categories: those that test whether a change occurs globally, i.e., over the entire study region; or those that test whether a change occurs locally, i.e., in a geographically or otherwise limited area of the region.

Scanning methods such as Kulldorff's spatial or space-time scan statistic[12,18] perform local tests, typically by scanning the study region with a locally defined window. These methods provide a location of any detected change, which is an important advantage for timely response to a potential outbreak. However their implementation usually requires the analyst to define a shape for the scanning tool, which makes some spatial disturbances easier to detect than others; much work has addressed this issue.[19–23] Scanning methods may also have limited capacity to detect multiple clusters, since the likelihood models that such scan statistics typically use ignore the spatial arrangement of disease outside the scanning window.[18] Thus, the very strengths of the local testing approach may in certain situations prove to be a disadvantage.

Conversely, global testing methods typically do not identify local areas for further investigation, but they also do not make any assumption on the specifics of disturbances to detect. Any global disturbance, evaluated against the typical variations expected by chance alone, is a potential sign of unusual disease activity. Such systems might serve as a preliminary warning tool, to raise the attention and sensitivity of other available surveillance tools to a more heightened level than before a disturbance was detected. The *M*-statistic can serve as such a global testing method.[24] Furthermore, it allows for protection of part of the spatial information since, rather than using the recorded locations of patients, this statistic requires as data the interpoint distribution of distances.

Framework for Evaluation of Detection Performance

Before a spatial statistical method can be integrated into a surveillance system, one must consider its performance under different scenarios. In particular it should accurately discern whether any changes in the spatial distribution have occurred or not. Assessing these qualities requires the use of synthetic datasets that reflect real case scenarios as much as possible. In the context of spatial data, this involves generating a baseline population representing normal behavior, and superimposing spatial disturbances atop the baseline. If necessary, complex baseline populations can be achieved either with mixtures of standard probability distributions (e.g., uniform, normal, Poisson), or drawn from existing real datasets. Adding spatial disturbances allows us to define a null and an alternative hypothesis to measure the power of detection of the considered method.

We now describe general frameworks for simulating spatial disturbances and aggregating data. Since the two phenomena occur independently of each other, we treat them in separate sections.

How Can We Generate Spatial Disturbances?

Many of the referenced work in the first part of the background section uses simulated spatial disturbances that can be described within the framework we now introduce.[7,9–11,13,14] We focus on spatial disturbances within a fixed time period, leaving spatiotemporal patterns for future work.

Our goal is to provide sufficient mathematical structure to the description of spatial disturbances so that we can describe a broad range of disease outbreak types with a relatively small number of parameters. This allows us to systematically investigate and compare spatial methods on different outbreak types, using common language and notation. It also helps us to make a connection between simulated data and previously reported historical outbreaks. For example, when modeling possible effects of an outbreak on the population, we might design simulation studies with parameter ranges that more accurately reflect the outbreak characteristics of a particular historical outbreak of concern. Thus, we have aimed for a common framework that accommodates several approaches to describing and simulating outbreaks.

Our framework consists of four main attributes to describe spatial disturbances, chosen in the following order. In the description below, we will denote the study region by R.

Reference Population. This is the spatial distribution of the population at risk, written with population density function g_0. We sometimes refer to this distribution informally as the "null population". When monitoring the current spatial distribution of cases within a prospective surveillance system, we refer back to the null population for our comparisons. When simulating data, we can generate the null population from a combination of known distributions (e.g., uniform, normal) or real spatial data.

Number. The number s of spatial disturbances throughout the region R. Typically we consider a "single hot spot" disturbance with a small geographic region with localized excess, corresponding to $s = 1$. However spatial patterns in disease data often demonstrate more complex patterns with multiple hot and/or cold spots, so we consider the number of disturbances as a separate parameter.

Locations. The locations are the subregions A_l of $R, l \in \{1, \ldots, s\}$ where a spatial disturbance occurs. When $s = 1$ and there is a single disturbance, the subregion A defines the geographic boundaries of that disturbance. Locations can be further described by a *focus point*, a *shape*, and an *extent*. The focus point x_1 represents the 'center' of The subregion. The shape refers to the geometry of the subregion (e.g., square, circle, line along a river or a highway). The extent is the geographic area covered by the subregion. This value might be determined spatially (fixed geographic extent) or cover a fixed proportion of the null population (fixed population extent).

Intensity. The intensity is a function depicting the increase in risk of disease in the particular region A_l. This function g_{A_l} has support A_l (it is zero on any point in the region R outside of the geographic boundaries A_l). For example, suppose there is a single disturbance described by subregion A. We might define the function g_A as constant throughout A_l ("hot spot cluster"), declining according to the distance from the focus point ("clinal cluster"), or varying in some possibly more complex way.

All these attributes are unknown to the analyst if they were to handle real data, except perhaps the reference population. When simulating data, all of these characteristics need to be specified. The intended application will often dictate some of the attributes in a particular way. For example, the extent of a simulated outbreak might be related to the mechanism of disease transmission within the population: does the disease remain contained within a small geographic region, and/or does it affect a certain proportion of the reference population? If the cases are arising due to a local environmental exposure, the geographic extent of the exposure remains fixed regardless of its surrounding population density. If the disease is infectious, the focus x_1 might represent a first individual spreading the disease to others. His/her level of infectiousness and number of contacts determine the proportion of the surrounding population getting infected. If this proportion is fixed regardless of the focus point, it can be represented by a fixed population extent.

The various attributes we have described combine to define a "spatially disturbed" sample, i.e., a new spatial distribution which we aim to compare with the reference population. We consider this spatially disturbed sample as a mixture of distributions, defined for $x \in R$ as:

$$g(x) = q_0\, g_0(x) + \sum_{l=1}^{s} (1 - q_1) g_{A_l}(x)$$

where $q_0, \ldots, q_s \geq 0$ and $q_0 + q_1 + \ldots + q_s = 1$ to guarantee that g is a probability density function. The parameters $q_0, \ldots, q_s$ control the strength of the disturbance to detect. For example, given a fixed extent, large q_0 and small $q_l, l \in \{1, \ldots, s\}$ confer weak signals.

Real examples such as the Sverdlovsk,[25] Woburn[26] and Milwaukee[27,28] outbreaks can be described according to this framework. In the Sverdlovsk anthrax outbreak, the single focal point was a nearby military facility and the geographic spread was shaped as a plume defined by the prevailing northerly wind. This event can be characterized with $s = 1$ and an intensity function g_A defined nonconstantly over the plumed-shaped region, i.e., according to the wind pattern.

The investigation of the Woburn outbreak of childhood leukemia showed that the high rates of cases were significantly associated with exposure to the water supply serviced by two of the municipal wells. Both focus points were located in contiguous regions of the eastern part of Woburn and the exposure of households to both contaminated supplies was expressed as a single variable. The exposure variable varied depending on the location of households, hence this scenario can be represented by $s = 1$ and a nonconstant intensity g_A.

Finally, Milwaukee experienced a massive outbreak of *Cryptosporidiosis* where the primary mechanism of exposure was delivery of contaminated drinking water via the public water supply. In particular the population receiving residential drinking water from the southern treatment plant rather than the northern plant was the most affected. This incident can be represented as two focus points ($s = 2$), each covering half of the study region (A_{south}, A_{north}). Assuming the contaminated water is reaching all households, the corresponding intensities $g_{A_{south}}$, $g_{A_{north}}$ are both constant. The higher risk from the southern plant is expressed with a higher value for the corresponding q_{south} compared with q_{north}.

In all three examples, an appropriate reference population g_0 also needs to be specified. This parameter should be representative of the population at risk, like a recent census before the outbreak in Milwaukee for this last case.

What Aggregation Schemes Can We Consider?

Spatial information of cases is in aggregated form when cases within a particular region are assigned to the same location. This can happen in more than one way, as we now describe. In surveillance settings, spatial data are typically available in the form of household address, work/school address, billing address, or simply the location of the health facility where a medical visit was recorded. Regardless of the accuracy of the spatial information, multiple cases might have the same location, and thus appear aggregated, simply because they live/work in the same building or visit the same health facility. In such instances, we would need nonspatial covariates to distinguish individuals further. In this paper we assume that this does not occur, and thus the available spatial data allows us to differentiate any two individuals. Although a strong assumption, we prefer to focus on "man-made" aggregation, i.e., the kind that results from collecting or analyzing the data in an intentionally aggregated form despite the fact that there exists more accurate information.

When spatial data are intentionally aggregated, it is typically done at an administrative level, for example in the United States at the level of zip codes or census tracts. Aggregating data along administrative boundaries is convenient and easy to execute, but these boundaries are unrelated to the spatial spread of any disease, and might also not match the geographic organization of health facilities. As an alternative to aggregation along real administrative units, we consider superimposing the study region with several regular rectangular grids of varying spacing. For a given grid spacing, we reassign locations falling in a particular grid square to its center. The average side length of a grid square defines an aggregation level and thus the spatial resolution of the data after alteration.

This aggregation scheme, first described in references 14 and 29,[14,29] offers several advantages for our intended study of the effects of aggregation. As just noted, the level of aggregation and spatial resolution is easily indexed with a single value, the side length of a grid square. This facilitates systematic study and quantitative analysis of the effect of aggregation. We can calculate power and false detection over a range of indexed levels to understand better the features of the considered detection method. Finally, the level of aggregation is a continuous parameter which can range widely. Familiar geographic scales such as ZIP codes and census tracts are represented at one end of this continuum, as well as intermediate levels that might result in satisfactory detection power.

Research Question

Most of the published work mentioned in the first part of the background section focuses on local methods. Global clustering methods are a distinct family of spatial methods with their own strengths and weaknesses (see second part of background section). We propose to investigate the effect of aggregation on the power of detection of a global spatial test, the *M*-statistic. We extend our previous work[14,29] in two ways. First we have developed a general framework described in the background to simulate spatial disturbances, which includes a much larger variety of scenarios and allows a more systematic study of our methods. Second we simulate data using both a homogeneous and heterogeneous baseline. While the former allows us to study the effect of aggregation on power without interference from the complex features of real data, the latter is more representative of authentic surveillance data.

Methods

We now give a detailed description of the *M*-statistic and our simulations.

The *M*-Statistic

The *M*-statistic is a global spatial test introduced by Bonetti and Pagano.[30] It compares the interpoint distribution of distances between cases to an expected distribution. The *M*-statistic can accommodate spatial data either with exact locations or aggregated at a collection of discrete locations. Also, it does not use the recorded location of cases, rather some transformation of the spatial information from which the original data cannot be recovered. Studies have shown the *M*-statistic is a flexible and extensible spatial statistic; for example other work has described its use to compare two spatial distributions;[31] with genetic data to identify patterns of mutations associated with HIV-ARV resistance;[32] extensions to multiple addresses;[33,34] and in combination with temporal methods to perform prospective spatiotemporal surveillance.[2]

Given a set of n independent locations $X_1,\ldots, X_n$ in the plane, there are $n(n\text{-}1)/2$ pairwise or interpoint distances d_{ij}, with cumulative distribution function (cdf) $F(d)$. The empiric cumulative distribution function (ecdf) of their distribution can be written as:

$$F_n(d) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} I(dist(X_i, X_j) \le d),$$

where *dist* is the Euclidean distance in the plane. Using the theory of *U*-statistics, they[30] prove that the distribution of $F_{\mathrm{n}}(d)$ computed at a finite set of values, $d_1,\ldots, d_k$, converges to a multivariate normal distribution as $n \to \infty$.

Now suppose exact locations are actually reported as in an aggregated fashion. More specifically suppose the study region is broken up into m distinct areas represented by fixed locations $l_1,\ldots,l_m$, and define $X_i^* = l_j$ if X_{i} falls into area j. Then, $X_i^*,\ldots, X_n^*$ independently arise from any of $l_1,\ldots, l_m$ with probabilities $p_1,\ldots, p_m$, where p_{j} is the probability that a sampled case's location is in area j. We can now write the cdf as:

$$F_n(d) = F(d|p) = \frac{1}{n^2}\sum_{i=1}^{m}\sum_{j=1}^{m} p_i p_j I(dist(X_i, X_j) \le d).$$

Conditioning on n, the total number of observed cases in the study region, let N_i, $i = 1,\ldots, m$ be the random variable representing the number of individuals arising at locations $l_1,\ldots, l_m$, with values n_{i} observed at each location. Then, these N_{i} follow a multinomial distribution with probabilities $p = (p_1,\ldots, p_m)$. The numbers of observed cases n_i, at each location $i = 1,\ldots, m$ provide consistent estimators for probabilities $\hat{p}_{\mathrm{i}} = n_i/n$, and define the ecdf $F_n(d) = F_n(d|\hat{p})$.

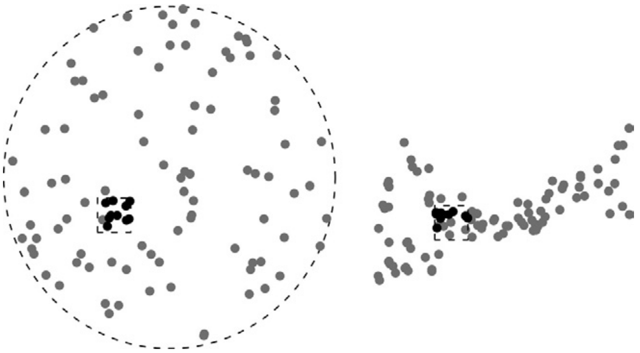

Figure 1. Ninety points distributed according to reference population (LEFT: uniform in unit disk, RIGHT: Cape Cod), and ten additional "outbreak" points from the square left of center.

The asymptotic result for the continuous case described above also holds in this discrete setting.

The M-statistic is a goodness-of-fit statistic calculating deviations between the observed ecdf F_n and the a priori specified cdf F, using a Mahalanobis-type distance. In the context of global spatial clustering, F represents the spatial distribution of incoming cases under conditions of "normalcy" (the null distribution). The approach is to discretize the distribution of interpoint distances such that F can be represented by a $k \times 1$ vector of cumulative probabilities where the k^{th} entry is 1. This discretization is given by a set of finite values $d = (d_1, \ldots, d_k)$ such that $P_F(dist(X_1, X_2) \leq d_j) = j/k$ for $j = 1, \ldots, k$. The authors[30] formulate their statistic as:

$$M = (F_n(d) - F(d))^T \textstyle\sum_F^- (F_n(d) - F(d)) \tag{4}$$

where $\sum_F^-$ is the generalized inverse of rank k-1 of the variance covariance matrix of $Fn(.)$. One can also use the successive differences of the cdf, $\{F_n(d_1), F_n(d_2) - F_n(d_1), \ldots, 1 - F_n(d_{k-1})\}$, to define a similar statistic. Asymptotically, M follows a χ^2 distribution with degrees of freedom equal to $rank(\sum_F^- \sum_F)$.

To estimate$\sum_F$, we use a resampling-based procedure. We first draw s random samples from F, and then estimate $\sum_F$ with

$$\frac{1}{s}\sum_{l=1}^{s} (o_l - e)(o_l - e)^T \tag{5}$$

where o_1 is the vector of observed cell counts defined by $d = (d_1, \ldots, d_k)$ from the s^{th} sample. Once we estimate the $\sum_F$ matrix, we can then estimate the sampling distribution of the M-statistic under the null hypothesis again via resampling from the reference population F. The statistic is implemented with the simulations described below using $k = 50$ bins. Power of the M-statistic is defined as the proportion of simulations with test statistic greater than the 0.05 threshold established from the null.

Simulations

Data

We consider two different sampling frames: the uniform distribution in the unit disk and a more heterogeneous distribution drawn from three community hospitals serving Cape Cod, Massachusetts. These data consist of spatial locations of patients arriving for emergency care (geocoded billing address where coordinates were sufficiently altered to protect anonymity) between 1994 and 1999. For ease of comparison with the uniform disk, the coordinates were transformed to range between −1 and 1.

To simulate spatial disturbances, we follow the framework proposed in the background section. The reference population is either the uniform or heterogeneous populations just described. We create a single square spatial disturbance ($s = 1$). The focus point x is a case location randomly selected from the reference population, and the geographic extent of the disturbance remains fixed and indexed by the side length of the square where the disturbances occurs, an index which we will call "diameter". We consider ten values for the diameter, ranging from 0.05 to 0.5. The intensity function remains constant throughout A:$g_A(x) = g_0(x)/\int_A g_0(y)dy$ for $x \in A$ and 0 otherwise. Finally, q is fixed to one of four values: 0.80, 0.85, 0.90 and 0.95. Each of the resulting $2 \times 10 \times 4 = 80$ spatial disturbance schemes contains 10,000 simulated datasets, within which the focus is the only parameter that varies.

Figure 1 presents an illustration of one simulated dataset using each reference population for $q = 0.90$. Each simulation contains 100 points: 90 cases are drawn from the reference population and 10 are sampled from a small square in the region to represent an increased risk of limited geographic range.

Aggregation Scheme

As presented in the background, the study region defined by the reference population is superimposed by a rectangular grid (Figure 2). Spatial locations falling in a particular grid square are reassigned to the center of that grid square. The term grid square here refers to a single square from the superimposed grid. A small bivariate jitter is added to the center of each grid square to avoid too regular an aggregation and give a more diverse range of interpoint distances. We consider a sequence of twelve levels of aggregation, varying the coarseness of the grid at each level. We use the side length of a grid square as an index of spatial aggregation; the number of grid square per side ranges from fifteen to four (corresponding side length ranges from 0.133 to 0.5). For each side length value, Table 1 gives the number of grid squares where data drawn from the reference populations

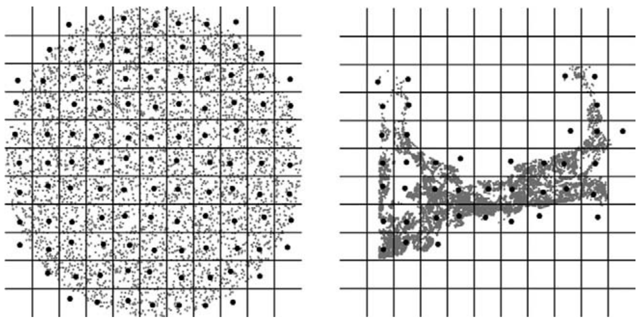

Figure 2. One level of aggregation for the unit disk and the Cape Cod populations, where the exact locations are reassigned to a single point in the corresponding grid square.

Table 1 ▪ Number of Grid Squares per Aggregation Level

Side Length	0.13	0.14	0.15	0.17	0.18	0.20	0.22	0.25	0.29	0.33	0.40	0.50
Uniform	199	172	149	132	109	88	77	60	45	36	25	16
Cape Cod	71	68	59	48	40	39	32	26	25	19	17	12

gets observed. For example, the illustration in Figure 2 corresponds to a side length of 0.18. Also, one can make a connection with administrative units for the Cape Cod population. This region contains 64 ZIP codes and 17 towns, which can be represented by side lengths 0.145 and 0.40 respectively.

Results

Results are presented in the eight panels of Figure 3 (available as on online data supplement at http://www.jamia.org). The power results for the uniform and Cape Cod populations are displayed in the top and bottom rows respectively, where each column corresponds to a value for the mixture proportion q. On each panel, the power of the M-statistic (vertical axis) is reported as a function of the side length of a grid square (horizontal axis). The limits of the vertical axis are varied to accommodate for the different ranges in power. A side length equal to zero means the locations are not aggregated, i.e., exact (point) locations. Each colored curve corresponds to one of the ten diameters.

Figure 3 illustrates two features common to both reference populations. First for fixed q and side length, the power decreases as the diameter increases, except for a few scenarios with $q = 0.95$. Second, for fixed diameter and side length, power of detection decreases as q increases. Both confirm that the signal is harder to detect for large values of q or large diameter.

Uniform Population

The results for the uniform population can be further categorized into four patterns when q is fixed: for the smaller diameters, power decreases as the aggregation level increases (pattern **a**); for slightly higher diameter values, power first decreases, then increases to sometimes a higher value than with exact locations, before decreasing again (pattern **b**); for a medium size diameter, power first increases than decreases (pattern **c**); finally for the larger diameters, power increases with the side length (pattern **d**). The range of diameters for which each pattern is observed varies with q (Table 2).

We see that pattern **a** appears more often when the signal to detect is strong, both in terms of small q and small diameter. As the signal becomes weaker, results change to patterns **b**, **c** and then **d**. In all of these "increasing then decreasing" patterns (**b, c, d**), the side length value at which the maximum power is attained usually increases with the diameter (Figure 4, left panel), and in particular tends to occur when the diameter is somewhat larger than the side length.

Table 2 ▪ Range of Diameter for Each q and Curve Pattern (Uniform Population)

	$q = 0.80$	$q = 0.85$	$q = 0.90$	$q = 0.95$
Pattern **a**	[0.05.0.40]	[0.05.0.30]	[0.05.0.15]	0.05
Pattern **b**	[0.45.0.50]	[0.35.0.45]	0.2	
Pattern **c**		0.5	[0.25.0.40]	[0.10.0.30]
Pattern **d**			[0.45.0.50]	[0.35.0.50]

Regardless of the specifics of the spatial disturbance considered, the patterns just described are smoothed interpretations of the irregularities observed when the side length ranges between 0.13 and 0.33. These irregularities might be explained by the perturbation added to the center of the grid square.

Cape Cod Population

With the Cape Cod population, power tends to be lower than for the homogeneous population. In all four bottom panels of Figure 3 we usually see an increase in power between side length values 0.33 and 0.40, except for three spatial disturbance schemes where there are no noticeable difference between the two aggregation levels ($q = 0.80$ and diameter = 0.35, 0.50 and $q = 0.85$ and diameter = 0.50). These results possibly reflect the heterogeneity of the population. Aside from this common pattern, the results for the Cape Cod population can be categorized into patterns similar to **a, b, c, d** when q is fixed: a decreasing trend between power and side length (pattern **e**); an increasing then decreasing trend (pattern **f**); an increasing then slightly decreasing trend with all power values less than 0.20 (pattern **g**); power values less than 0.10 regardless of the aggregation level (pattern **h**). Values of the diameter for each pattern are summarized in Table 3.

We see that pattern **e** is mostly observed with the smallest diameter while other scenarios have highest power at an aggregated level. As with the homogeneous population, for the "increasing then decreasing" patterns **f** and **g**, the side length value at which the maximum power is attained tends to increase with the diameter, and mostly occurs when the diameter is larger than the side length (Figure 4, right panel).

Discussion

In this work we have investigated the effect of spatial aggregation on the power of a method to detect global clustering in the presence of various spatial disturbances. The level at which aggregation occurs is a measure of the accuracy of the spatial resolution. At one extreme lies unaggregated spatial information about each case, as with geocoded addresses; at the other extreme we have no information about location, just that the patient exists. To measure the effect on power, we have developed our simulations within frameworks that can be connected to real examples of outbreaks with geographic component and real aggregation schemes. In particular we vary widely some of the parameters and use two different baseline populations.

Our results confirm previous studies, in that the degree of aggregation at which location gets reported is important. However they do not confirm the intuitive idea that coarser spatial resolution results in a loss of power. In fact the relationship is more complex and also depends on the

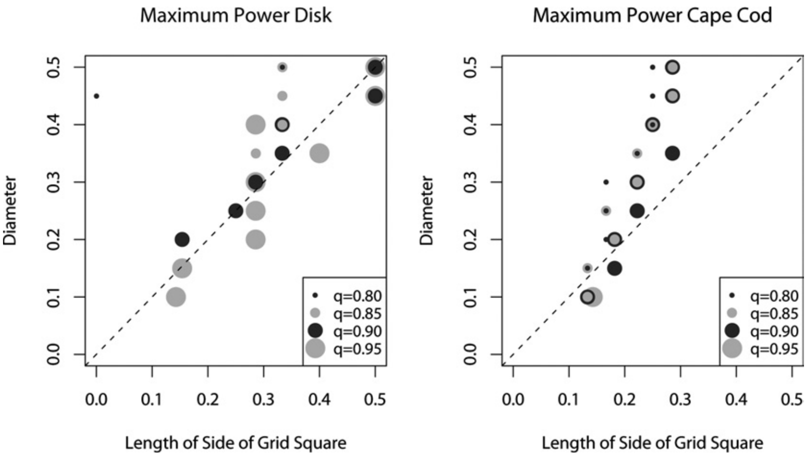

Figure 4. Schemes where maximum power occurs in patterns **b, c, d** for unit-disk population (LEFT) and patterns **f, g** for Cape Cod population (RIGHT). Dashed line means side length of grid square equals diameter of spatial disturbance.

geographic scale of the spatial disturbance. With both baseline populations, we showed that power is highest using exact locations for strong signals, while aggregation at the right level may improve the ability to detect weak signals. The best performing level of aggregation depends on the geographic extent of the spatial disturbance, but typically occurs when the extent is greater than or equal to the dimension of an aggregation grid square.

This phenomenon was already observed in the simulations results from Waller,[11] where the power to detect an increase in relative risk improved as the signal got stronger, but only if the level of aggregation resulted in grid squares of the same geographic size or smaller than that of the cluster. We conclude that this interaction between the two geographic scales might not pertain to particular detection methods. As a side note, our previous simulation results did not show such an interaction.[14] However we had only considered three types of spatial disturbance, and under these three scenarios, the *M*-statistic generally shows a decrease in power ($q = 0.90$ and diameter = 0.05, 0.10, 0.20). This further points to the need for considering a wide range of parameters when simulating spatial disturbances.

A few limitations should be mentioned. Our framework to simulate spatial disturbances allows a diversity of scenarios, but we do not vary all possible parameters. However, our results already shed extensive light on particular aspects on the problem of interest. The situation is even more complex when considering spatiotemporal disturbances, and we have not explored the effect of space-time aggregation which is a natural extension to this work. Furthermore, most studies using spatial information assume that each individual has only a single location, while in reality human beings are mobile. Several extensions of distance-based methods have been proposed,[33,34] but the effects of aggregation in these more complex settings will need to be studied.

Table 3 ■ Range of Diameter for Each *q* and Curve Pattern (Cape Cod Population)

	$q = 0.80$	$q = 0.85$	$q = 0.90$	$q = 0.95$
Pattern **e**	[0.05.0.10]	0.05	0.05	0.05
Pattern **f**	[0.15.0.50]	[0.10.0.50]	[0.10.0.30]	
Pattern **g**			[0.35.0.50]	0.1
Pattern **h**				[0.15.0.50]

Finally, this study limits itself to synthetically generated "outbreaks". The use of synthetic data allows us to define a null and an alternative, and thus to measure the power of detection, which is a well defined and understood metric to evaluate a statistical test. By contrast, artificial data gives a limited representation of reality. While investigating the performance of our method on real data would give further insight on its capability in a complex setting, datasets containing a large number of real spatial disturbances are not always readily available, partly because there is not always consensus to define a spatial disturbance. The use of real data can also be a more convincing way to evaluate other approaches aimed at relating the occurrence of outbreaks to geographic areas. Epidemiological methods, such as case-control studies, allow us to assess whether a particular exposure in specific locations is associated with a high number of cases.[35,36] The advantage of these studies is that they bypass the detection of the spatial disturbance by directly pointing to the actual exposure responsible for the outbreak. Yet they take place after the first alert of an excess in the number of cases, while our method is intended to be applied earlier, for example by accompanying a temporal surveillance system generating such alerts.[2]

Conclusions

We have shown that reporting patients' spatial information at an aggregated level affects the power of detecting clusters differently, depending on the strength of the signal. Regardless of the underlying population, strong signals are better detected when exact locations are reported, while weaker signals are better detected at an aggregated level of similar geographic extent. Because the strength of a putative signal is not known when operating a surveillance system prospectively, one might consider simultaneously several levels of precision, rather than choosing a single one. This would probably still require having exact locations at hand for analysis, and adjusting for multiple testing. Thus, to maximize the chances of detecting an outbreak, we should seek alternatives that ensure patients' privacy, and at the same time provide as much spatial information as possible. Novel uses of informatics may provide some of these alternatives,[7,37] and thus avoid an unnecessary trade-off of statistical efficiency for the sake of individual privacy. Furthermore, the audiences from which the information should be concealed and those to whom it can be disclosed may guide the development of different approaches.[38]

References ■

1. Kulldorff M, Tango T, Park PJ. Power comparisons for disease clustering tests. Comput Stat Data Anal 2003;42(4):665–84.
2. Ozonoff A, Forsberg L, Bonetti M, Pagano M. A bivariate method for spatiotemporal syndromic surveillance. Morb Mortal Wkly Rep 2004;53 (Suppl):61–6.
3. Koch T. The Map as intent: Variations on the theme of John Snow. Cartographica: The International Journal for Geographic Information and Geovisualization. 2004;39(4):1–14.
4. Snow J. In Report on the Cholera Outbreak in the Parish of St. James, Westminster during the Autumn of 1854, by the Cholera Inquiry Committee. London: Churchill; 1855.
5. Cox L. Protecting confidentiality in small population health and environmental statistics. Stat Med 1996;15:1895–905.
6. Minot N, Baulch B. Poverty mapping with aggregate census data: What is the loss in Precision? Review Dev Econ 2005;9(1): 5–24.
7. Boulos M, Cai Q, Padget J, Rushton G. Using software agents to preserve individual health data confidentiality in micro-scale geographic analyses. J Biomed Inform 2006;39:160–70.
8. Armstrong MP, Rushton G, Zimmerman DL. Geographically masking health data to preserve confidentiality. Stat Med 1999; 18:497–525.
9. Cuzick J, Edwards R. Spatial clustering for inhomogeneous populations. J R Statist Soc B 1990;52:73–104.
10. Cassa C, Grannis S, Overhage J, Mandl K. A novel, context-sensitive approach to anonymizing spatial surveillance data: Impact on outbreak detection. Adv Dis Surveill 2006;1:10.
11. Waller L. Statistical power and design of focused clustering studies. Stat Med 1996;15:765–82.
12. Kulldorff M, Heffernan R, Hartman J, Assuncão R, Mostashari F. A space-time permutation scan statistic for disease outbreak detection. PLoS Med 2005;2(3):216.
13. Olson KL, Grannis SJ, Mandl KD. Privacy protection versus cluster detection in spatial epidemiology. Am J Pub Health 2006;96(11):2002.
14. Ozonoff A, Jeffery C, Manjourides J, White LF, Pagano M. Effect of spatial resolution on cluster detection: A simulation study. Internation J Health Geogr 2007;6:52.
15. Lawson A. Spatial and syndromic surveillance for public health. In: Lawson A, Kleinman K, eds., Wiley, 2005.
16. Lawson A. Statistical Methods in Spatial Epidemiology, Wiley, 2006, 2ed.
17. Kulldorff M. Tests of spatial randomness adjusted for an inhomogeneity: A general framework. J Am Stat Assoc 2006;101(475): 1289–305.
18. Kulldorff M. A spatial scan statistic. Commun Statist Theory Methods 1997;26:1481–96.
19. Patil G, Taillie C. Upper level set scan statistic for detecting arbitrarily shaped hotspots. Environ Ecol Stat 2004;11(2):183–97.
20. Duczmal L, Assunção R. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. Comput Stat Data Anal 2004;45(2):269–86.
21. Tango T, Takahashi K. A flexibly shaped spatial scan statistic for detecting clusters. Internation J Health Geogr 2005;4(11).
22. Kulldorff M, Huang L, Pickle L, Duczmal L. An elliptic spatial scan statistic. Stat Med 2006;25(22):3929–43.
23. Dematteï C, Molinari N, Daurès JP. Arbitrarily shaped multiple spatial cluster detection for case event data. Comput Stat Data Anal 2007;51(8):3931–45.
24. Olson KL, Bonetti M, Pagano M, Mandl KD. Real time spatial cluster detection using interpoint distances among precise patient locations. BMC Med Inform Decis Mak 2005;5(1):19.
25. Meselson M, Guillemin J, Hugh-Jones M et al. The Sverdlovsk anthrax outbreak of 1979. Science 1994;266(5188):1202.
26. Lagakos S, Wessen B, Zelen M. An analysis of contaminated well water and health effects in Woburn, Massachusetts. J Am Stat Assoc 1986:583–96.
27. Mac Kenzie WR, Hoxie NJ, Proctor ME, et al. A massive outbreak in Milwaukee of *Cryptosporidium* infection transmitted through the public water supply. N Engl J Med 1994;331(3): 161–7.
28. Naumova EN, Egorov AI, Morris RD, Griffiths JK. The elderly and waterborne *Cryptosporidium* infection: Gastroenteritis hospitalizations before and during the 1993 Milwaukee outbreak. Emerg Infect Dis 2003;9(4):418–25.
29. Jeffery C, Ozonoff A, Forsberg L, Nuño M, Pagano M. The cost of obfuscation when reporting locations of cases in syndromic surveillance systems. Adv Dis Surveill 2006;1:36.
30. Bonetti M, Pagano M. The interpoint distance distribution as a descriptor of point patterns, with an application to spatial disease clustering. Stat Med 2005;24(5):753–73.
31. Manjourides J, Pagano M. A test of the difference between two interpoint distance distributions, submitted.
32. Kowalski J, Pagano M, De Gruttola V. A nonparametric test of gene region heterogeneity associated with phenotype. J Am Stat Assoc 2002;97(458):398–409.
33. Ozonoff A, Bonetti M, Forsberg L, Pagano M. The Use of Multiple Addresses to Enhance Cluster Detection. Proceedings of the American Statistical Association, Biometrics Section, CDROM, 2003.
34. Manjourides J, Pagano M. Improving the power of chronic disease surveillance by incorporating residential history, submitted.
35. Hennessy TW, Hedberg CW, Slutsker L, et al. A national outbreak of *Salmonella enteritidis* infections from ice cream. N Engl J Med 1996;334(20):1281–6.
36. Wheeler C, Vogt TM, Armstrong GL, et al. An outbreak of hepatitis A associated with green onions. N Engl J Med 2005; 353(9):890–7.
37. Lombardo JS, Buckeridge DL. Disease surveillance: A public health informatics approach, Wiley-Interscience; 2007.
38. El Emam K, Dankar FK. Protecting privacy using k-anonymity. J Am Med Inform Assoc 2008;15(5):627–37.