Density-equalizing Euclidean minimum spanning trees for the detection of all disease cluster shapes

Shannon C. Wieland*†‡, John S. Brownstein‡§, Bonnie Berger*†¶, and Kenneth D. Mandl‡§¶

*Department of Mathematics and †Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139-4307; ‡Children's Hospital Informatics Program at the Harvard–Massachusetts Institute of Technology Division of Health Sciences and Technology, Children's Hospital Boston, Boston, MA 02115; and §Department of Pediatrics, Harvard Medical School, Shattuck Street, Boston, MA 02115-6092

Edited by Burton H. Singer, Princeton University, Princeton, NJ, and approved April 15, 2007 (received for review October 25, 2006)

Existing disease cluster detection methods cannot detect clusters of all shapes and sizes or identify highly irregular sets that overestimate the true extent of the cluster. We introduce a graph-theoretical method for detecting arbitrarily shaped clusters based on the Euclidean minimum spanning tree of cartogram-transformed case locations, which overcomes these shortcomings. The method is illustrated by using several clusters, including historical data sets from West Nile virus and inhalational anthrax outbreaks. Sensitivity and accuracy comparisons with the prevailing cluster detection method show that the method performs similarly on approximately circular historical clusters and greatly improves detection for noncircular clusters.

biosurveillance | disease cluster detection | graph theory

Tests for the detection of disease clusters (1) are essential tools for identifying emergent infections and elucidating demographic and environmental factors influencing diseases. The shapes of these clusters are unpredictable (2–6). However, the prevailing cluster detection method, a scan statistic that applies a likelihood ratio test to a large number of overlapping circles in a study region, reports only circular clusters (7, 8). Straightforward extensions of the circular scan statistic, such as an elliptical scan (9) and a rectangular scan (10), are also limited to detecting specific outbreak shapes.

Few methods aim to detect clusters of arbitrary shape. One class of methods based on graph theory has recently emerged to address this problem (11–14). However, these have several limitations: they are restricted to clusters that fit inside a circular region of fixed size (11), they attempt to examine a set of potential clusters too large to exhaustively search (12), they have poor specificity (13), or they have yet to be implemented or evaluated (14).

In addition to the difficulties inherent in any disease cluster detection method, such as accounting for the underlying population density and controlling the level of significance given multiple potential clusters of various sizes and in various locations, arbitrary shape cluster detection presents particular challenges. As more shapes are considered, the statistical power declines, and the computational running time may become unreasonable for typical problem sizes (11). Furthermore, if the exact case locations are available, then considering every conceivable shape is problematic; it is always possible to draw a bizarrely shaped region of infinitesimally small total area that includes every case. This problem surfaces when data are aggregated into small regions. Indeed, one study identified excessively large clusters with highly irregular shapes having greater likelihood ratios than the inserted clusters that were the detection targets (13).

In this study, we address these challenges by removing the notion of shape from consideration and replacing it with a mathematical formalization of potential clusters based on intercase distances. We introduce a method to locate clusters of any shape based on Euclidean minimum spanning trees (EMSTs), which have previously found application in heuristic methods to divide other kinds of data into a predetermined number of subsets (15, 16). Application of the method to synthetic, West Nile virus, and anthrax data sets show that sensitivity and accuracy are substantially improved compared with the circular scan statistic method applied to noncircular clusters, which likely include the majority of real disease clusters.

EMST Cluster Detection

Our cluster detection method consists of three sequential tasks. A density-equalizing cartogram of the study region and disease cases is first constructed from a Voronoi diagram of the controls. Second, the family of potential clusters to evaluate is defined, because it is not computationally feasible to consider all 2^n subsets of n cases. Third, the statistical significance of each potential cluster is evaluated. We address each of these tasks.

Cartogram Construction. We begin with the precise spatial coordinates of a set of disease cases and controls and a map of the study area. We first create a Voronoi diagram of the control locations, which subdivides the study area into the regions closest to each control location (17) [see supporting information (SI) Fig. 5]. The density of controls within each Voronoi region is simply the number of controls in the region, which may be more than one if multiple controls can occur at the same location, divided by the region's area. We use this density function to create a density-equalizing cartogram of the Voronoi diagram. Cartograms have previously been used for aggregate data to test for clustering of several diseases (18–22). To construct one, each point on the original map is essentially magnified or demagnified according to its local density. The result is a distorted map on which the density of controls is constant everywhere. Each case is placed on the cartogram at a random location within the region corresponding to its original Voronoi region, and all subsequent analyses are performed by using these new case locations. Under the null hypothesis of constant relative risk, the new locations of the cases on the Voronoi diagram cartogram are uniformly and independently distributed. We use a diffusion-based cartogram construction algorithm (22), although other contiguous cartogram algorithms may also be suitable.

Potential Clusters. We call a potential cluster a subset of points S satisfying the property that every subset of S is "closer" to at least one other point in S than to any other point outside of S. To

Author contributions: S.C.W., J.S.B., B.B., and K.D.M. designed research; S.C.W. performed research; S.C.W., J.S.B., B.B., and K.D.M. analyzed data; and S.C.W., J.S.B., B.B., and K.D.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Abbreviation: EMST, Euclidean minimum spanning tree.

¶To whom correspondence may be addressed. E-mail: bab@mit.edu or kenneth.mandl@childrens.harvard.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0609457104/DC1.

formalize this definition, we begin by defining the distance $\rho(X, Y)$ between two sets X and Y to be the smallest distance separating the sets:

$$\rho(X, Y) = \begin{cases} \min_{\substack{a \in X \\ b \in Y}} \rho(a, b) & \text{if } X \neq \emptyset \text{ and } Y \neq \emptyset \\ \infty & \text{otherwise} \end{cases}, \quad [1]$$

where $\rho(a, b)$ is the Euclidean distance between two points. We also define the internal distance of a nonempty set S to be the maximum distance between any two nonempty subsets of S whose union is S:

$$\rho(S) = \max_{\substack{\emptyset \subsetneq X \subseteq S \\ \emptyset \subsetneq Y \subseteq S \\ X \cup Y = S}} \rho(X, Y). \quad [2]$$

We formally define a potential cluster as follows:

Definition. Let V be a nonempty set of cases of a disease. A potential cluster is a nonempty set $S \subseteq V$ satisfying $\rho(S) < \rho(S, V - S)$.

Note that the entire set V is a potential cluster, as are the sets $\{v\}$ for every $v \in V$. If v is the nearest neighbor of w and w is the nearest of v, then $\{v, w\}$ is a potential cluster.

We want to consider every potential cluster in V, but it is not straightforward from the definition how to locate potential clusters, nor how many of them are present. Progress was made toward finding potential clusters in a different application in bioinformatics (16) by using the minimum spanning tree of V, a connected graph T spanning a set of points having minimal total weight

$$w(T) = \sum_{e \in E(T)} w(e), \quad [3]$$

where $E(T)$ denotes the set of edges of T, and the weight $w(e)$ of an edge e is in this case the Euclidean distance between the endpoints of e. (For a detailed review of graph theoretical definitions, see ref. 23.) Given a set V of n points, every potential cluster is a connected subgraph of the EMST T of V (16). However, even for small epidemiological data sets, the number of connected subgraphs may be extremely large; EMSTs of 50 and 75 random points have approximately 10^6 and 10^8 connected subgraphs, respectively.

We prove that it is not necessary to consider all connected subgraphs of T to find the potential clusters. Remarkably, there are at most $2n - 1$ potential clusters, of which n are trivial sets consisting of only one vertex. Furthermore, the potential clusters may be quickly found from an EMST by using a greedy edge deletion procedure. After constructing an EMST of the set of cartogram case locations V, we iteratively delete the longest remaining edge of T. At each iteration we consider the two newly emergent connected components, each of which is a potential cluster. In this way, we evaluate all $n - 1$ nontrivial potential clusters for statistical significance by using a test described below (see Fig. 1). A proof that this procedure identifies the set of potential clusters is found in *Appendix*.

Statistical Significance. To assign a P value to any potential cluster, a test statistic is required, along with its distribution under the null hypothesis H_0 of independently, uniformly distributed cases on the cartogram. Let Σ be a potential cluster generated under H_0, and let S be an observed potential cluster. We define

$$P_S = Pr\{w(\Sigma) < w(S) \mid card(\Sigma) = card(S)\}, \quad [4]$$

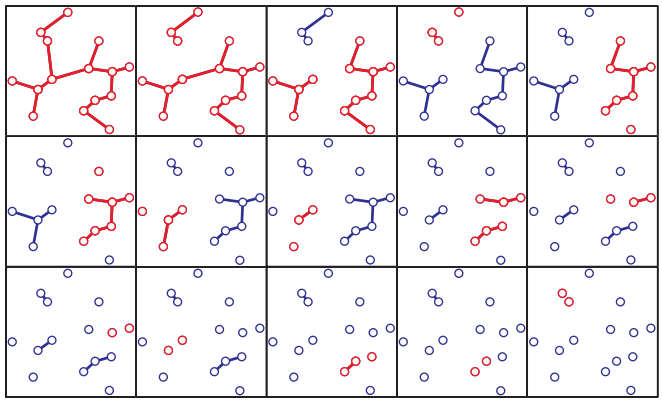

Fig. 1. Procedure to locate potential clusters illustrated for a set of 15 cases. The EMST is first constructed (*Top Left*). This is a tree connecting each case (circle) that minimizes the total summed edge distance. At each step, the longest remaining edge is deleted, forming two new connected components (red). Components that were unchanged from the previous step are shown in blue. The connected components are in one-to-one correspondence with the set of potential clusters.

where w is the weight of the potential cluster subgraph, and *card* denotes the number of cases. P_S is the P value corresponding to the observed candidate cluster weight, conditioned on the number of cases in S. Because cases in a true cluster are closer together than expected, the weight $w(S)$ of a potential cluster S corresponding to a hot spot is likely to be smaller than a random EMST potential cluster subgraph containing the same number of cases. Consequently, a hot spot should have a low value of P_S. We define the test statistic P to be the minimum value of P_S over the set of nontrivial potential clusters containing at most half of the cases. Monte Carlo techniques are used to fit P_S as a function of $w(S)$ to a Gaussian distribution for each possible value of $card(S)$. The null distribution of P is subsequently estimated, again by Monte Carlo, and a cutoff value corresponding to the desired level of significance α is obtained.

The most significant cluster is reported, but the method could easily be modified to report all significant clusters without affecting the asymptotic running time.

Results

We applied the SaTScan circular scan statistic (8) and EMST method to several types of data sets, finding that the EMST method was substantially better able to detect noncircular clusters. The SaTScan Bernoulli model was used with a maximum geographic window size containing 50% of the cases for each data set. For each method and data set, the most significant cluster with a P value of at most 0.05 computed by using 9,999 Monte Carlo replications was reported; thus the specificity, defined as the probability of reporting no significant cluster in data generated under the null hypothesis, was 0.95 for both methods and all data sets. The sensitivity, equal to the fraction of clusters that were detected, was calculated for each data set and method. To quantify the extent of overlap between the most likely cluster and the actual cluster, we defined two other measures. We defined F_{TC} to be the fraction of true cluster cases that were correctly found in the most likely cluster, and F_{MLC} to be the fraction of cases in the most likely cluster that coincided with the true cluster.

West Nile Virus, New York City, 1999. The EMST method and SaTScan had similar performance detecting a 1999 outbreak of West Nile virus in New York City (24). This finding was encouraging because the 56 cases appear to have an approximately circular distribution (see Fig. 2), suggesting an advantage

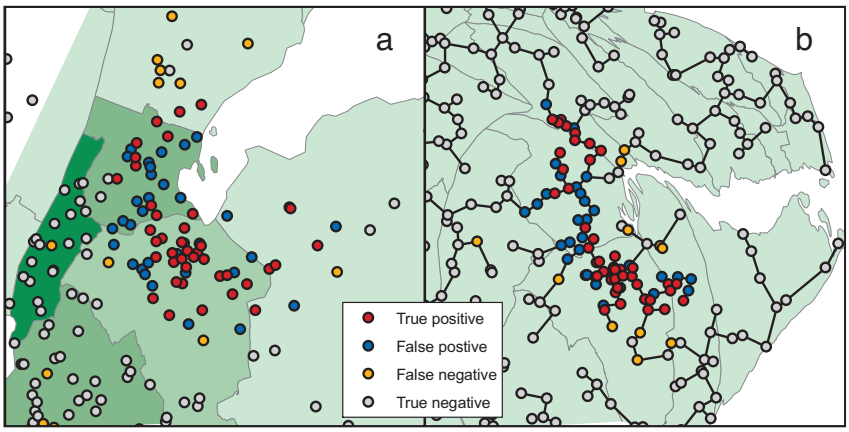

Fig. 2. Detection of 1999 New York West Nile virus cases by SaTScan and the EMST method. (*a*) A typical data set consisting of the 56 West Nile virus cases (red and orange) and 400 background cases (blue and gray) are shown on a map of Connecticut, New Jersey, and New York. Only part of the map is shown for clarity. The West Nile virus case locations have been randomly skewed for privacy (34). The most likely cluster identified by SaTScan is shown (red and blue). The green shading represents the density of controls in each county. (*b*) The Voronoi diagram cartogram of part of the study area is shown along with the transformed case locations. Although the Voronoi diagram cartogram regions are not shown, the distortion of county boundaries induced by the cartogram transformation is apparent. The minimum spanning tree (black edges) connects the most likely cluster identified by the EMST method (red and blue). The control density varies by <2.0% over the entire map.

for the circular scan statistic. We defined a study area consisting of Connecticut, New Jersey, and New York and generated 10,000 controls within the map distributed in proportion to 2000 U.S. census county population data. To evaluate the methods, we required data sets with both outbreak and nonoutbreak cases. In addition to the West Nile virus cases, we generated 400, 600, 800, 1,000, or 1,200 additional nonoutbreak background cases distributed according to the underlying population distribution. As the number of background cases increased, the West Nile virus cluster became harder to detect. We created 1,000 data sets for each background case number. The data sets could represent, for example, emergency visits for neurological symptoms in a multistate surveillance area, with controls drawn from all emergency visits. Fig. 2 shows a typical data set along with its Voronoi diagram cartogram transformation and the most likely cluster obtained by both methods. The results of applying SaTScan and the EMST method to the data sets are summarized in Table 1. Both methods displayed similar comparative performance for all numbers of background cases. The sensitivity of both methods declined from 1.0 for 400 background cases to 0.96 and 0.89 for 1,200 background cases for the EMST method and SaTScan, respectively. The percent change in F_{TC} of the EMST method compared with SaTScan varied from -0.4% to 16%, and the percent change in F_{TC} varied from -14% to -6.8%.

Inhalational Anthrax, Sverdlovsk, Russia, 1979. The EMST method had greater accuracy than SaTScan when applied to a highly noncircular outbreak of 62 cases of inhalational anthrax occurring in Sverdlovsk, Russia in 1979 (2). Because we lacked spatial references for the data necessary to geocode the case locations, we used a uniform distribution within a square study region to generate 10,000 controls. The set of cases consisted of 400, 600, 800, 1,000, or 1,200 uniformly distributed background cases, in addition to the anthrax case locations. These could represent, for example, visits for respiratory complaints to an emergency department, with controls drawn from all visits. For each number of background cases, 1,000 data sets were generated. A typical data set is shown in Fig. 3, along with the most likely cluster detected by SaTScan and the EMST method. The mean sensitivity, F_{TC}, and F_{MLC} are summarized in Table 2. The EMST method had comparable or greater sensitivity than SaTScan for all background population sizes, and it correctly identified a greater fraction of the anthrax cases (F_{TC}) for all background population sizes. Both methods' sensitivity declined as more background cases were added: from 0.98 to 0.52 for the EMST method and from 0.98 to 0.35 for SaTScan. The EMST method had a lower value of F_{MLC} than SaTScan, indicating that it overestimated the cluster to a greater extent than SaTScan. However, the percent decline in F_{MLC} incurred by using the EMST method instead of SaTScan was about half of the gain in F_{TC}.

Circular Clusters, Boston, MA. We also compared the ability of the EMST method and SaTScan to detect circular clusters. Because the circular scan statistic is optimized to detect circular clusters, we were surprised to find that the EMST method was as sensitive as SaTScan. The study area consisted of the 59 zip codes within 10 km of Boston, MA. Ten thousand controls were distributed

Table 1. SaTScan and EMST method applied to West Nile virus

	SaTScan			EMST			Comparisons		
n	SN	F_{TC}	F_{MLC}	SN	F_{TC}	F_{MLC}	Δ SN, %	Δ F_{TC}, %	Δ F_{MLC}, %
400	1.00	0.69	0.61	1.00	0.80	0.53	+0.5	+16	−14
600	1.00	0.63	0.54	1.00	0.69	0.48	+0.2	+9.1	−11
800	0.99	0.58	0.48	1.00	0.61	0.44	+0.7	+5.1	−8.5
1,000	0.99	0.55	0.44	0.99	0.55	0.41	−0.4	−0.1	−6.8
1,200	0.89	0.49	0.40	0.96	0.50	0.38	+8.0	+3.4	−4.6

n, no. of background cases added to cluster cases; SN, average sensitivity; F_{TC}, average fraction of true cluster detected; F_{MLC}, average fraction of most likely cluster coinciding with the true cluster (averaged over data sets for which a significant cluster was found); Δ, percent difference.

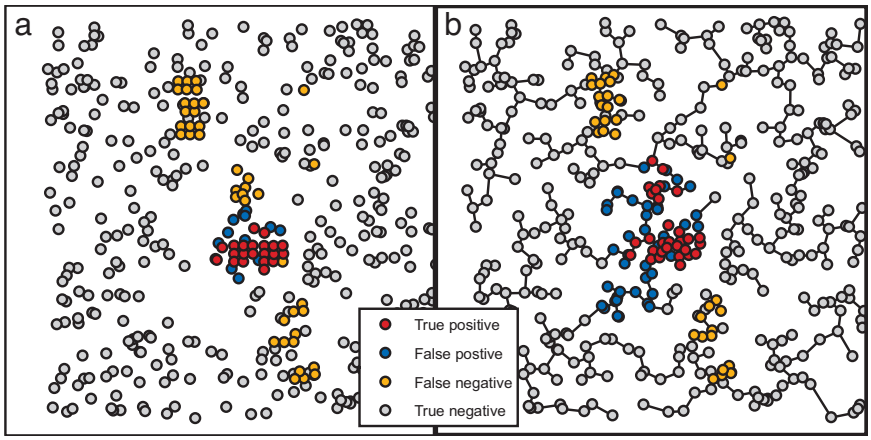

Fig. 3. SaTScan and EMST detection of 1979 Sverdlovsk anthrax outbreak. (*a*) A representative data set of 63 anthrax cases (red and orange) and 400 uniformly distributed background cases (blue and gray) is shown, along with the most likely cluster determined by SaTScan (red and blue). (*b*) The EMST method most likely cluster (red and blue) is shown for the same data set, connected by the minimum spanning tree of the cartogram-transformed cases (black edges).

on the map in proportion to zip code population data from the 2000 U.S. census. Data sets of 500 total cases were created, each containing a synthetic circular cluster in a random location with a radius of 1, 2, or 3 km placed within the study region. We defined the relative cluster density to be the case density within the cluster divided by the case density outside the cluster. This ratio varied from two to five in the data sets. For each combination of outbreak radius and relative cluster density, 1,000 data sets were created. For small clusters containing on average <35 cases, the EMST method had greater sensitivity. However, it is likely that stochastic effects caused such clusters to have noncircular shapes in general. Indeed, the smaller the cluster, the more pronounced the EMST method's relative improvement in sensitivity. For larger clusters, the EMST method had similar sensitivity to SaTScan (0.1% less to 4.1% more) and similar values of F_{TC} (3.4% less to 0.4% more). However, SaTScan always had a larger value of F_{MLC}, indicating that it located large circular clusters with more overall accuracy than the EMST method. See SI Table 4 for detailed results.

Rectangular Clusters, Boston, MA. In a study of rectangular clusters, we found that the EMST method had greater sensitivity than SaTScan. Sets of 500 cases containing artificial rectangular clusters having a height-to-width ratio of 1, 4, or 16 and relative cluster density between two and five were generated within the same study region as above, and 10,000 controls were distributed in proportion to the background population as above. The cluster area was fixed at 20 km^2, and 1,000 data sets were generated for each combination of parameters by randomly placing a rectangular cluster within the study region map. The results are summarized in Table 3. In general, the EMST method had greater sensitivity than SaTScan (0.2% less to 166% more), with the greatest percent increase in sensitivity when the cluster signal strength was weak or the height-to-width ratio was large. The EMST method captured a greater extent of the true cluster (F_{TC}) than SaTScan for all cluster types (2.6% to 419% more). For most cluster types, there was a parallel decline in the fraction F_{MLC} of the most likely cluster coinciding with the true cluster (20% less to +3.2% more).

Arbitrary Shapes. It is possible to gain insight into the EMST method's performance on other cluster shapes without additional intensive computer simulations. The EMST test statistic depends only on the cartogram, the total number of cases, and the cardinality and weight of a potential cluster. Hence, we can extrapolate the *P* value obtained for one potential cluster to others having different shapes, but the same number of cases and weight. To illustrate this, we selected one most likely cluster of 35 cases from one of the Boston analysis data sets. The EMST method assigned a *P* value of 0.0001 to this potential cluster. Fig. 4 shows several configurations of potential clusters having the same number of cases and EMST weight, but very different shapes. If embedded as potential clusters within a Boston data set of 500 total cases, they would each achieve the same *P* value of 0.0001. In fact, any potential cluster of 35 cases of any shape can be scaled in size to have the same weight, illustrating that the method can capture an infinite array of regular and irregular shapes.

Discussion

We find that the EMST method is a powerful and accurate alternative to the circular scan statistic for noncircular clusters.

Table 2. SaTScan and EMST method applied to anthrax

	SaTScan			EMST			Comparisons		
n	SN	F_{TC}	F_{MLC}	SN	F_{TC}	F_{MLC}	Δ SN, %	Δ F_{TC}, %	Δ F_{MLC}, %
400	0.98	0.32	0.65	0.98	0.48	0.49	−0.4	+48	−24
600	0.88	0.28	0.53	0.86	0.39	0.40	−2.3	+38	−25
800	0.60	0.19	0.44	0.72	0.32	0.32	+19	+68	−28
1,000	0.53	0.17	0.37	0.60	0.26	0.26	+12	+55	−31
1,200	0.35	0.11	0.32	0.52	0.21	0.22	+46	+100	−31

n, no. of background cases added to cluster cases; SN, average sensitivity; F_{TC}, average fraction of true cluster detected; F_{MLC}, average fraction of most likely cluster coinciding with the true cluster (averaged over data sets for which a significant cluster was found); Δ, percent difference.

Table 3. SaTScan and EMST method applied to rectangular clusters

Parameters		SaTScan			EMST			Comparisons		
r	*d*	*SN*	F_{TC}	F_{MLC}	SN	F_{TC}	F_{MLC}	Δ SN, %	Δ F_{TC}, %	Δ F_{MLC}, %
1	2	0.56	0.47	0.82	0.61	0.50	0.65	+8.2	+6.0	−20
1	3	0.92	0.82	0.90	0.95	0.86	0.78	+3.2	+4.7	−13
1	4	0.99	0.91	0.93	0.99	0.94	0.85	−0.2	+2.6	−8.9
1	5	1.00	0.93	0.95	1.00	0.97	0.88	+0.2	+4.5	−7.3
4	2	0.43	0.26	0.69	0.58	0.42	0.62	+36	+63	−10.0
4	3	0.95	0.64	0.77	0.97	0.86	0.74	+2.2	+34	−4.4
4	4	1.00	0.73	0.79	1.00	0.95	0.80	+0.1	+29	+0.4
4	5	1.00	0.78	0.81	1.00	0.97	0.84	0.0	+25	+3.2
16	2	0.21	0.06	0.66	0.55	0.31	0.52	+166	+419	−21
16	3	0.82	0.25	0.72	0.98	0.74	0.60	+21	+199	−17
16	4	0.99	0.31	0.76	1.00	0.86	0.67	+0.9	+177	−11
16	5	1.00	0.35	0.77	1.00	0.93	0.73	0.0	+166	−6.0

r = ratio of cluster height to width; *d* = relative cluster density; SN, average sensitivity; F_{TC}, average fraction of true cluster detected; F_{MLC}, average fraction of most likely cluseter coinciding with the true cluster; Δ, percent difference.

At a specificity of 95%, the method had comparable sensitivity to SaTScan applied to large synthetic circular clusters and an approximately circular West Nile virus outbreak. When applied to small circular clusters, synthetic rectangular clusters, and a highly irregular anthrax cluster, the EMST method had greater sensitivity. Although SaTScan had better accuracy detecting large circular clusters, the EMST method had comparable or superior accuracy for all other cluster types. The EMST method is also able to detect a large variety of shapes, including highly irregular ones.

In addition to accurately locating clusters of any shape and size, the EMST method has two unique properties. First, its test statistic is based only on the weight of the potential cluster subgraph. To our knowledge, all other tests that provide the location of any detected clusters while allowing the user to set the level of significance for the test use the likelihood ratio test statistic developed by Kulldorff and Nagarwalla (7). This test statistic requires the area of each region considered, which in turn requires a precise definition, including the shape, of the region. Second, we formally define a cluster in mathematical terms that are independent of cluster geometry, and which depend only on intercase distances. Traditionally, clusters are often imprecisely defined; for example, Knox's frequently cited definition is "a geographically bounded group of occurrences of sufficient size and concentration to be unlikely to have occurred by chance" (25).

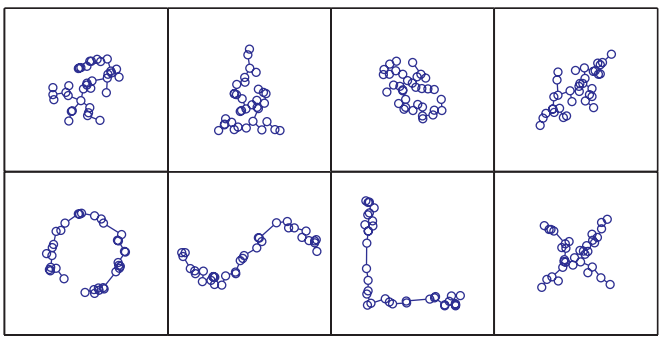

Fig. 4. Equally detectable potential clusters of various shapes. A most likely cluster of 35 points selected from among the Boston circular cluster data sets, along with its minimum spanning tree, is shown in the upper left. Seven other configurations of 35 points, having minimum spanning trees with exactly the same weight, are also shown. Subject to the constraint imposed by the definition of a potential cluster, all eight clusters have equivalent detectability by the EMST method. If embedded as potential clusters in a Boston data set of 500 total cases, all would achieve the same *P* value of 0.0001.

Of other cluster detection methods designed to capture clusters of any shape, the EMST method is most similar mathematically to the upper level set method of Patil and Taillie (14), which examines a well defined family of contiguous administrative regions with high relative rates. Assunção *et al.* (13) used minimum spanning tree of graphs with different vertices, edges, and edge weights to consider contiguous administrative regions having similar disease rates, whether high or low. By contrast, we locate sets of individual cases corresponding to a mathematical formalization of a cluster, using specific subsets of the EMST. General tests of clustering (1) such as Tango's maximized excess events test (26), and disease mapping methods, such as Bayesian partition models (27, 28), kriging (29), and generalized additive models (30, 31), handle arbitrary geometric configurations of cases without difficulty. However, these address separate problems within spatial epidemiology, and comparison of clustering and disease mapping methods to cluster detection methods is not straightforward (32).

The EMST method can easily be extended to analyze regional summary data, consisting of counts of observed and expected disease cases for each region on a map. A cartogram is constructed to equalize the density of expected disease cases, and each observed case is randomly placed on the cartogram within its region of occurrence. After constructing the cartogram, the procedure for case-control data are followed.

One limitation inherent in this and other methods for aggregated data is that exact spatial locations are not used, which decreases cluster detection sensitivity and accuracy (33). This is also a limitation for the procedure detailed above for case-control data, because a loss of spatial information is incurred by randomizing cases within their regions of occurrence on the Voronoi diagram cartogram. Because the expected area of each region on the cartogram tends toward zero as the number of control locations increases, this loss can be minimized by increasing the number of controls. For 10,000 distinct controls on a square map, as used in our study, the loss of spatial information is modest; each case is expected to move ≈1% of the length of one side of the square.

We found that the EMST method gains in F_{TC} for noncircular clusters were partially offset by a decline in F_{MLC}, indicating that the EMST method reports fewer false negatives, but more false positives, than SaTScan. The relative cost to society of false negatives and false positives depends on many factors. The cost of false negative cases includes, for example, an increased risk of

spread of a disease and the possibility that infected individuals who are unaware of the outbreak may not seek early treatment for symptoms, while the cost of false positive cases includes unnecessarily investigating and alarming the community. In retrospective research and prospective surveillance, the shape of true clusters are not known *a priori*. Thus, in most cases, a method that is able to detect clusters of any shape is preferable. Hence the EMST method may represent a practical adjunct to methods currently used in public health practice.

Appendix

We show that potential clusters are in one-to-one correspondence with a small class of subsets of an EMST T. For $w \geq 0$, we define T_w to be the graph derived from T by deleting all edges of T having weight greater than w. We label the $n - 1$ edges of T in order of decreasing weight, so that $w(e_1) \geq w(e_2) \geq \ldots \geq w(e_{n-1}) > 0$. If the edge weights are distinct, then there are n distinct graphs T_w; these are the graphs $T = T_{w(e1)} \supsetneq T_{w(e2)} \supsetneq \ldots \supsetneq T_{w(e_{n-1})} \supsetneq T_0$. $T_{w(e_{k+1})}$ is formed from $T_{w(e_k)}$ by deleting one edge, which splits one connected component of $T_{w(e_k)}$ into two components. Thus $T_{w(e_{k+1})}$ has $k + 1$ connected components, $k - 1$ of which are present in $T_{w(e_k)}$, and two of which are newly created. There are $2n - 1$ total distinct connected components among all of the graphs T_w (see Fig. 1). If the edge weights are not distinct, then a variation of this argument shows that $2n - 1$ is an upper bound on the number of distinct connected components. The following characterizes the connected components:

Lemma 1. *Let V be a nonempty set of points in a plane (representing cases of a disease). Let T be an EMST of V, S a nonempty subset of V, and T_S the subgraph of T induced by S. The set S is a potential cluster if and only if T_S is a connected component of T_0 or of $T_{w(ek)}$ for some k.*

The proof is made easier by two simple lemmas, which we prove in *SI Text*.

Lemma 2. *Let T_S be a connected subgraph of T with vertex set S. Then $\rho(S)$ (Eq. 2) is equal to the maximum weight of an edge in T_S if $|S| > 1$, and 0 otherwise.*

Lemma 3. *If S is a nonempty, proper subset of V, then $\rho(S, V - S)$ is equal to the minimum weight of an edge in T spanning the cut $(S, V - S)$.*

Proof of Lemma 1. We first show that every potential cluster induces a connected component of T_0 or $T_{w(e_k)}$ for some k. Equivalently, we show that if a subgraph H of T is not a connected component of $T_{w(e_k)}$ or T_0, then the vertex set of H is not a potential cluster. Xu *et al.* (16) showed that every potential cluster induces a connected subgraph of T, so that if H is not connected, then its vertex set is not a potential cluster. Suppose H is a connected subgraph of T, which is not a connected component of $T_{w(e_k)}$ for any k, or T_0. H must have at least one edge; let e_j be an edge of H of maximal weight. Let C be the connected component of $T_{w(e_j)}$ containing e_j. Because H is a connected subgraph of $T_{w(e_j)}$ containing e_j, $H \subsetneq C$. We refer interchangeably to a graph and its vertex set to simplify notation. There exists some edge $e \in T$ spanning H and $C - H$, and because $e \in C$, $w(e) \leq w(e_j)$. By *Lemma 2*, $\rho(H) = w(e_j)$, and by *Lemma 3*, $\rho(H, V - H) \leq \rho(H, C - H) \leq w(e) \leq w(e_j)$. Hence $\rho(H, V - H) \leq \rho(H)$ and H is not a potential cluster.

To finish the proof, we must show that every connected component of $T_{w(e_k)}$ for any k or T_0 is a potential cluster. This is trivial for $T_{w(e_1)} = T$ or T_0, whose components are the individual vertices. Let T_S be a connected component of $T_{w(e_k)} \neq T$ with vertex set S. Then $\rho(S) \leq w(e_k)$ by *Lemma 2*. Because $V - S \neq \emptyset$, there must be some edge $e \in T$ spanning S and $V - S$. Because the edge is not in $T_{w(e_k)}$, $w(e) > w(e_k)$. This is true for every spanning edge, so by *Lemma 3*, $\rho(S, V - S) > w(e_k)$. Hence $\rho(S) < \rho(S, V - S)$, and so S is a potential cluster.

Note that the proof does not rely on the uniqueness of T, so degenerate EMSTs do not affect the ability of the method to capture all potential clusters. If the set of cases V are continuously distributed on the cartogram, as in the present study, then in theory the EMST is unique with probability 1. However, degenerate EMSTs may occur with extremely low probability because of the inability of computers to support arbitrary precision.

We thank Lisa Sweeney and Daniel Sheehan of the Massachusetts Institute of Technology Geographic Information Systems Laboratory for their help with Geographic Information Systems software and data and Karen Olson, Chris Cassa, Brad Friedman, and Lenore Cowen for helpful discussions. This work was supported by National Library of Medicine Grant LM007677-03S1.

1. Besag J, Newell J (1991) *J R Stat Soc A* 154:143–155.
2. Meselson M, Guillemin J, Hugh-Jones M, Langmuir A, Popova I, Shelokov A, Yampolskaya O (1994) *Science* 266:1202–1208.
3. Ruiz MO, Tedesco C, McTighe TJ, Austin C, Kitron U (2004) *Int J Health Geogr* 3:8.
4. Diggle P (1990) *J R Stat Soc A* 153:349–362.
5. Keeling MJ, Woolhouse MEJ, Shaw DJ, Matthews L, Chase-Topping M, Haydon DT, Cornell SJ, Kappey J, Wilesmith J, Grenfell BT (2001) *Science* 294:813–817.
6. Elliott P, Wakefield J, Best N, Briggs D (2000) *Spatial Epidemiology: Methods and Applications* (Oxford Univ Press, Oxford).
7. Kulldorff M, Nagarwalla N (1995) *Stat Med* 14:799–810.
8. Kulldorff M (1997) *Commun Stat Theor Methods* 26:1481–1496.
9. Kulldorff M, Huang L, Pickle L, Duczmal L (2006) *Stat Med* 25:3929–3943.
10. Neill DB (2006) PhD thesis (Carnegie Mellon University, Pittsburgh, PA).
11. Tango T, Takahashi K (2005) *Int J Health Geogr* 4:11.
12. Duczmal L, Assunção R (2004) *Comput Stat Data Anal* 45:269–286.
13. Assunção R, Costa M, Tavares A, Ferreira S (2006) *Stat Med* 25:723–742.
14. Patil GP, Taillie C (2004) *Environ Ecol Stat* 11:183–197.
15. Zahn CT (1971) *IEEE Trans Comput* C20:68–86.
16. Xu Y, Olman V, Xu D (2002) *Bioinformatics* 18:536–545.
17. de Berg M, van Kreveld M, Overmars M, Schwarzkopf O (2000) *Computational Geometry: Algorithms and Applications* (Springer, Berlin).
18. Merrill DW, Selvin S, Close ER, Holmes HH (1996) *Stat Med* 15:1837–1848.
19. Merrill D (2001) *Stat Med* 20:1499–1513.
20. Selvin S, Merrill D (2002) *Epidemiology* 13:151–156.
21. Khalakdina A, Selvin S, Merrill DW (2003) *Int J Hyg Environ Health* 206:553–561.
22. Gastner M, Newman M (2004) *Proc Natl Acad Sci USA* 101:7499–7504.
23. Bollobas B (1998) *Modern Graph Theory* (Springer, New York).
24. Brownstein JS, Rosen H, Purdy D, Miller JR, Merlino M, Mostashari F, Fish D (2002) *Vector Borne Zoonotic Dis* 2:157–164.
25. Knox EG (1989) in *Methodology of Enquiries into Disease Clustering*, ed Elliott P (Small Area Health Statistics Unit, London), pp 17–20.
26. Tango T (2000) *Stat Med* 19:191–204.
27. Denison DGT, Holmes CC (2001) *Biometrics* 57:143–149.
28. Ferreira JTAS, Denison DGT, Holmes CC (2002) in *Spatial Cluster Modeling*, eds Lawson AB, Denison DGT (Chapman & Hall, London), pp 125–146.
29. Berke O (2004) *Int J Health Geogr* 3:18.
30. Webster T, Vieira V, Weinberg J, Aschengrau A (2006) *Int J Health Geogr* 5:26.
31. Kelsall JE, Diggle PJ (1998) *J R Stat Soc C* 47:559–573.
32. Diggle PJ (2000) in *Spatial Epidemiology: Methods and Applications*, eds Elliott P, Wakefield J, Best N, Briggs D (Oxford Univ Press, Oxford), pp 87–103.
33. Olson KL, Grannis SJ, Mandl KD (2006) *Am J Public Health* 96:2002–2008.
34. Cassa CA, Grannis SJ, Overhage M, Mandl KD (2006) *J Am Med Inform Assoc* 13:160–165.