A comparison of spatial scan methods for cluster detection

Joshua P. French, Lauren M. Hall, Minh C. Nguyen, Mohammad Meysami,
Lee Panter, Nicholas Weaver

April 24, 2021

Abstract

The spatial scan method is extremely popular for identifying disease clusters using disease count data. The original spatial scan method is relatively simple, very fast, and has high power for detecting circular clusters. Free, publicly-available software led to its widespread use in a variety of contexts. However, the original spatial scan method can struggle to identify non-circular clusters. Many extensions of the original method have been proposed to better detect irregularly-shaped clusters compared to the circular scan method. We briefly describe many of the popular spatial scan method extensions (e.g., Upper Level Set, Flexibly-shaped, Dynamic Minimum Spanning Tree, Fast Scan, etc.). We then compare the performance of the various methods using power, specificity, and sensitivity by applying these methods to 61 publicly-available benchmark data sets that utilize 41 different cluster shapes. The comparisons go into more depth and have more competing methods than previous studies of this topic, allowing us to draw broader conclusions about the best performing methods.

Key words: spatial scan statistics, disease clusters, statistical power, most likely cluster, likelihood ratio statistics

1 Introduction

Data related to health, crime, and other events of interest are frequently reported as counts within pre-specified enumeration regions such as counties or census tracts. This helps to preserve the privacy of individuals associated with the event of interest while allowing for pattern detection. Frequently, researchers are most interested in identifying hotspot clusters where incidence rates are higher than those of surrounding regions.

Statistical methods for cluster detection are frequently proposed in the context of disease outbreak. Waller and Gotway (2004), Tango (2010), and Bivand et al. (2013) provide helpful overviews of many of the popular methods available for cluster detection. Some well-known historical methods for cluster detection include Moran's I (Moran, 1950), Geary's C (Geary, 1954), and subsequent extensions. Early exploratory methods for cluster detection include the Geographical Analysis Machine proposed by Openshaw et al. (1988) and a method based on overlapping local incidence proportions proposed by Rushton and Lolonis (1996). Turnbull et al. (1990) and Besag and Newell (1991) proposed some of the early statistical tests for cluster identification based, each with a different approach for accounting for the variability in local incidence rates. Tango (2010) and Mclafferty (2015) provide many examples of methodological development in disease cluster identification since that time. In what follows, we will focus our discussion on a specific stream of research in disease cluster detection, which are broadly known as spatial scan methods.

Two main weaknesses prevalent with early cluster detection techniques were that they (i) were global tests that identified a general discrepancy between observed and expected incidence rates, but failed to identify a specific set of regions having an unusually large incidence rate or (ii) did not satisfactorily address the problem of multiple comparisons. Kulldorff and Nagarwalla (1995) proposed the spatial scan method, which addressed both of these issues in a well-defined statistical framework. This utility of this method was quickly recognized, and the method became a popular choice for disease cluster identification. As of early January 2019, the Kulldorff and Nagarwalla (1995) article has over 1,300 citations. Kulldorff (1997) provided additional exposition of the spatial scan method and has over 3,100 citations (Google Scholar, 2019). Despite the fact that the spatial scan method is over 20 years old, it continues to be widely applied even today.

The spatial scan method's popularity stems from it's simplicity, computational efficiency, the availability of a free implementation of the methodology (Kulldorff, 2018), and its power to detect disease clusters. Naturally, the popularity of the spatial scan method encouraged other researchers to propose extensions and variants to address different context and/or improve the accuracy of the method. Overall, the goal of this article is an in-depth comparison of the many scan-based methods. In Section 2, we describe the original spatial scan method in further detail. We then describe most of the subsequent variants related to identifying disease clusters using regional count data. In Section 3, we provide the results of applying the many spatial scan variants to a benchmark data set made available by Kulldorff et al. (2003). While some of the results are available through previous research, we compare more variants than has ever been done, and for a more extensive set of data. Additionally, we have created the `smerc` R package (French, 2019) to provide a free, open-source implementation of the various benchmarked methods. This also allows us to make fairer timing comparisons of the various scan-based methods. We provide a concluding discussion of our findings in Section 4.

2 Methodology

Consider a study area A that is partitioned into N disjoint regions. The at-risk populations of the regions are $n_1, n_2, \ldots, n_N$, respectively, with $n := \sum_{i=1}^{N} n_i$. For each region, we observe the number of observed disease cases during the relevant time period, with $Y_i, i = 1,2,\ldots,N$, denoting the counts in each region and $y := \sum_{i=1}^{N} y_i$.

We consider modeling the case counts with either a Poisson or Binomial distribution. If the cases are modeled with a Poisson distribution, then $Y_i \sim \text{Poisson}(n_i\theta_i), i = 1,2,\ldots,N$, where θ_i is the risk of catching the disease in region i, and the counts are independent for $i = 1,2,\ldots,N$. If the cases are modeled with a Binomial distribution, then $Y_i \sim \text{Binomial}(n_i, \theta_i), i = 1,2,\ldots,N$, where θ_i is the risk of catching the disease in region i, and the counts are independent for $i = 1,2,\ldots,N$

We intend to test whether there is a collection of (contiguous) regions $Z \subset \{1,2,\ldots,N\}$ such that $\theta_i = \theta_Z$ for $i \in Z$ and $\theta_i = \theta_0$ for $i \in Z^c$, with $\theta_Z > \theta_0$. Define $y_Z = \sum_{i \in Z} y_i$ and $n_Z = \sum_{i \in Z} n_i$.

Assuming the counts come from the previously described Poisson distribution, Kulldorff (1997) derived a likelihood ratio test statistic as

$$\sup_{Z \in \mathcal{Z}} \frac{\left(\frac{y_Z}{n_Z}\right)^{y_Z} \left(\frac{y-y_Z}{n-n_Z}\right)^{y-y_Z}}{\left(\frac{y}{n}\right)^{y}} I\left(\frac{y_Z}{n_Z} > \frac{y-y_Z}{n-n_Z}\right), \tag{1}$$

and 1 otherwise, where $\mathcal{Z}$ is the set of all potential zones under consideration. Similarly, under the Binomial modeling assumption, a likelihood ratio test statistic can be derived as

$$\sup_{Z \in \mathcal{Z}} \frac{\left(\frac{y_Z}{n_Z}\right)^{y_Z} \left(\frac{n_Z-y_Z}{n_Z}\right)^{n_Z-y_Z} \left(\frac{y-y_Z}{n-n_Z}\right)^{y-y_Z} \left(\frac{n-n_Z-(y-y_Z)}{n-n_Z}\right)^{n-n_Z-(y-y_Z)}}{\frac{y^y(n-y)^{n-y}}{n^n}} I\left(\frac{y_Z}{n_Z} > \frac{y-y_Z}{n-n_Z}\right), \tag{2}$$

and 1 otherwise (Duczmal and Assuncao, 2004).

Scan-based methods almost exclusively rely on the statistics proposed in Equations (1) and (2). The substantive differences between methods are the approaches for determining $\mathcal{Z}$, the set of zones under consideration. For even moderately large N, determining all possible combinations of (connected) zones is computationally infeasible. Thus, researchers have proposed many different approaches for strategically choosing candidate zones to make computation feasible, but flexible enough to identify clusters of many different shapes.

We now outline the basic details of the scanning methods we will compare.

2.1 The circular scan test

Kulldorff and Nagarwalla (1995) and Kulldorff (1997) proposed the original spatial scan test. It is frequently referred to as the *circular* scan test, as the potential zones in $\mathcal{Z}$ tend to be circular in shape. The candidate zones used by the circular scan test are based on the sequences of k nearest neighbors for each region (including the region itself), typically subject to a population or distance constraint.

We briefly describe the candidate zones for the circular scan method in more detail. Let $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N$ denote a representative location within $Z_1, Z_2, \ldots, Z_N$. Let d_{ij} denote the intercentroid distance between

$\mathbf{s}_i$ and $\mathbf{s}_j$. Additionally, let $d_{i(j)}$ denote the jth largest intercentroid distance associated with region i and $I_{i(j)}$ denote the index of the region associated with the distance $d_{i(j)}$. By definition, $I_{i(1)} = i$, since $d_{i(1)}$ is necessarily 0 for d_{ii}. For each starting centroid $i = 1,2,\ldots,N$, we consider the set of potential zones $\{\{I_i(1)\},\{I_i(1),I_i(2)\},\{I_i(1),I_i(2),I_i(3)\},\ldots\}$. The potential zones continue to sequentially add the regions nearest to the starting centroid until relevant constraints are violated. Frequently, the zones are allowed to increase in size until no more than half the total population resides in the cluster; alternatively, one could restrict the maximum radius of the cluster.

The circular scan test is still an extremely popular method for cluster detection. It is relatively simple to understand, is very fast to implement due to the special structure of the candidate zones, and is available as part of the free, publicly-available SaTScan software (Kulldorff, 2018).

2.2 The elliptic scan test

The circular window in the scanning method is very popular and commonly used in detecting geographical clusters. However, other window shapes are also available, especially when we suspect that the true shape of clusters may not be circular. Kulldorff et al. (2006) presented the elliptic window as an alternative for the circular window in the scanning method. The mathematical principles for the elliptic scan statistics are identical to those of the circular scan statistics. While expanding the window, a likelihood ratio test statistic of each region is calculated as mentioned in the methodology section. However, unlike a directional - free tendency of a circle, regions included in an elliptic window may be largely different if the ellipse is rotated around its center.

The center of an elliptic window is usually chosen at the centroid of a region $Z_i, i \in \{1,2,\ldots,N\}$. Two parameters that help identify an ellipse are the ratio between the lengths of its major-axis a and minor-axis b (which is called the shape $s = \frac{a}{b}$), and the angle θ between the major-axis and the horizontal axis. The smallest value of s is 1 and it represents a circular shape. Increasing the value of s indicates a longer and narrower ellipse. The angle θ shows the rotational status of an ellipse around its center. For a chosen shape s and a chosen rotational angle θ, an ellipse is enlarged by increasing the lengths of its major-axis and minor-axis at the same factor until the combined population of all regions inside the ellipse reaches the population upperbound. Each zone in the ordered set $\{I_i(1),I_i(2),\ldots,I_i(N), i \in \{1,2,\ldots,N\}$ mentioned in section 2.1 needs is included in the calculations if its centroid is physically inside a considered ellipse. Namely, a zone $I_i(j)$ where its centroid has coordinates (x_j, y_j) is in an ellipse centered at (x_i, y_i) if

$$\frac{[cos(\theta)(x_j - x_i) + sin(\theta)(y_j - y_i)]^2}{a^2} + \frac{[cos(\theta)(x_j - x_i) - sin(\theta)(y_j - y_i)]^2}{b^2} \leq 1$$

.

The procedure of making an elliptic window can be summarized as below:

1. Select a center region and calculate the distance from this point to all other points.
2. Select a maximum shape and consider a discrete set of shapes smaller or equal to the selected maximum shape.
3. Choose the number of angles between the horizontal axis and the major axis.
4. For each combination of a shape ratio and an angle, increase the size of the elliptic windows until the population limit is met.
5. Repeat steps 1-4 for all other regions.
6. Remove any identical subsets so that only unique subsets are under consideration.
7. Take the maximum over all subsets.

It is also recommended in Kulldorff et al. (2006) that the number of angles $\frac{180^o}{\theta}$ could be at least three times the shape s so that when rotating, 70% or more of an ellipse will be overlapped. Technically, one can choose the shape s ahead of time, then decide the number of angles and the angle of rotation.

When the shape is large, the ellipse is long and narrow, and regions in a cluster may not be neighboring. To prevent this eccentric tendency, Kulldorff et al. (2006) recommended a penalized version for log-likelihood function:

$$LLR_{adj} = LLR \times \left(\frac{4s}{(s+1)^2}\right)^k$$

where LLR_{adj} is the adjusted log-likelihood ratio, LLR is the original adjusted log-likelihood ration, s is the shape of the ellipse, and $k \geq 0$ is a tuning parameter. In the circular case when $s = 1$, the adjusted log-likelihood is exactly the same as the original log-likelihood since there is no needs to fix the eccentricity. Otherwise, we note that $(s+1)^2 \geq 4s$ for all values of s, thus the penalty is stronger when increasing k with a fixed $s \neq 1$. Moreover, when $k = 0$, no penalty is applied to the log-likelihood, and when $k \to \infty$, the penalty is so strong that only circular clusters are considered Kulldorff et al. (2006) .

2.3 The Upper Level Set (ULS) scan test

The Circular and Elliptical scanning methods described previously impose an artificial restriction on the windows evaluated for cluster content. Each window under consideration must conform to the geometric restrictions of the underlying search methodology. Consequently, so must any discovered hotspots. In the context of the Circular and Elliptical scanning methods, this realization leads to the logical conclusion that the clusters discovered using these methods would be either circular or elliptical, respectively (or conforming to these geometries, i.e. non-circular/non-elliptical “subsets" of the circular/elliptical scanning window).

The Upper Level Set Scan for detecting arbitrarily shaped hotspots proposed by G. P. Patil and C. Tallie Patil and Taillie (2004) uses the underlying spatial connectivity of the region under investigation (A), in combination with a population-scaled response value across this region to determine a reduced parameter space over which an exhaustive search can be completed for a most likely cluster or set of clusters using the Likelihood Ratio Statistic (LRS).

We define the population-scaled response values as:

$$G_i = \left\{ \frac{Y_i}{n_i} \;\middle|\; i = 1,\ldots,N \right\}$$

The values of G_i constitute a function over the nodes of the connectivity graph that is dual to the plane graph of the region A:

$$G_i : i \mapsto \frac{Y_i}{n_i} \quad for\ i = 1,\ldots,N$$

we will also define the set:

$$G = \{G_i \mid i = 1,\ldots,N\}$$

If we define the distinct values of G to be the values: r_j for $j = 1,\ldots,M$ where we order the values of r_j according to:

$$r_1 > \cdots > r_M$$

then we have $M \leq N$. Additionally, for each of the distinct population-scaled response values given by a corresponding value of r_j we define the set of indices $T_j \subseteq A$ that satisfy:

$$T_j = \{i \in A \mid G_i = r_j\} \quad for\ \ j = 1,\ldots,M$$

From each of the values T_j, we may define an Upper Level Set U_j for $j = 1,\ldots,M$ by:

$$U_j = T_1 \cup \cdots \cup T_j \qquad \text{for} \quad \text{j=1,...,M}$$

$$= \{i \in A \mid G_i \geq r_j\}$$

The Upper Level Set U_j corresponding to the distinct population-scaled response value r_j can be interpreted as: the set of regional indices $i \in A$ which have a population-scaled response value as high or higher than r_j. This interpretation means that for any given $j = 1,\ldots,M$ the set U_j contains only those regions with the highest values of population-scaled response values.

Special consideration should be given to each of the connected subsets within each U_j. Let C_{j_k} represent a set of connected regions within U_j. According to this definition, we know that $1 \le k \le |U_j|$ where $|U_j|$ is the number of elements in U_j.

The defining properties of U_j mean that each set of connected regions C_{j_k} have elevated values of population-scaled response in comparison to those regions not included within U_j. This property, along with the contiguity of each C_{j_k} qualifies these regions as candidates for most likely cluster. We therefore define the new reduced parameter space Z_{ULS} as:

$$Z_{ULS} = \bigcup_{j_k} C_{j_k} \quad for\ j = 1,\ldots,M \quad and\ 1 \le k \le |U_j|$$

The parameter space reductions made in the search of Z_{ULS} can be shown to significantly decrease the amount of computational requirements in order to calculate an estimate using the LRS. Whereas an exhaustive search over the original region using the LRS would require calculating 2^N LRS values, the ULS method has the advantage that: the most likely cluster is in Z_{ULS} under assumption. This implies that the ULS-LRS search calculations can be bounded using the fact that:

$$|Z_{ULS}| = \left|\bigcup_{j_k} C_{j_k}\right| = \sum_{j_k} |C_{j_k}| \le \sum_{j=1}^{M} |U_j|$$

But since we defined:

$$U_j = \bigcup_{i=1}^{j} T_i$$

We have:

$$|Z_{ULS}| \le \sum_{j=1}^{M} \left|\bigcup_{i=1}^{j} T_i\right| = \sum_{j=1}^{M} \sum_{i=1}^{j} |T_i|$$

According to the definition of T_i we know that

$$\sum_{i=1}^{j} |T_i| = N$$

Which implies that:

$$|Z_{ULS}| \le \sum_{j=1}^{M} N = M * N$$

2.4 The flexible scan test

The literature shows that the circular spatial scan statistic proposed by Kulldorff (1997) has high power to detect circular clusters in rare diseases. However, it does not perform well in identifying non-circular clusters as circular windows are used to find most likely clusters (MLC) but have difficulty to detect noncircular clusters. Duczmal et al. (2006a) proposed a noncircular spatial scan statistic for detecting irregularly shaped cluster. However, their method tends to detect most likely clusters that were much larger than true clusters by including nonsignificant regions into MLC candidates. The flexible spatial scan statistic proposed by Tango and Takahashi (2005) performs well when detecting noncircular clusters by exhaustively searching all potential clusters in windows with a pre-set maximum length K regions.

Let's denote $\mathcal{Z}_1$ and $\mathcal{Z}_2$ the set of all zones that need to be scanned by the circular and flexible method, respectively. For each region $i \in \{1,...,N\}$ and prespecified K nearest neighbors to region i, $\mathcal{Z}_1$ consists of K concentric circles starting at i. Therefore, NK zones are included in $\mathcal{Z}_1$. For the flexible method, in addition to K concentric circles, all the connected zones inside a circle with center i and radius K should also be included. Therefore $\mathcal{Z}_1 \subset \mathcal{Z}_2$ and $\mathcal{Z}_2$ is substantially larger than $\mathcal{Z}_1$. This size difference increases computational load for the flexible method compared to the circular scan and makes it slower. Nevertheless Tango and Takahashi proposed the following method in order to find $\mathcal{Z}_2$.

1. For each region $i \in \{1,...,N\}$, Define the set $W_i = \{i, i_1, ..., i_k\}$ such that i_k is the k^{th} nearest region to the region i. Let A be a set in the power set of W_i which includes region i.
2. Split the set A to two subsets $A_1^* = \{i\}$ and $A_1 = \{i_1, ..., i_k\}$.
3. Split set A_1 to two subsets A_2 and A_2^* such that A_2^* contains all the regions of A_1 that are connected to set A_1^*, and A_2 contains all the regions that are not connected to A_1^*. The process continues until either A_j^* or A_j becomes a null set for a $j \in \mathbb{N}$. This step is illustrated in Figure 1.
4. A is a set of connected regions if A_j becomes a null set first, otherwise A is disconnected. If A is a connected zone, it will be added to $\mathcal{Z}_2$.

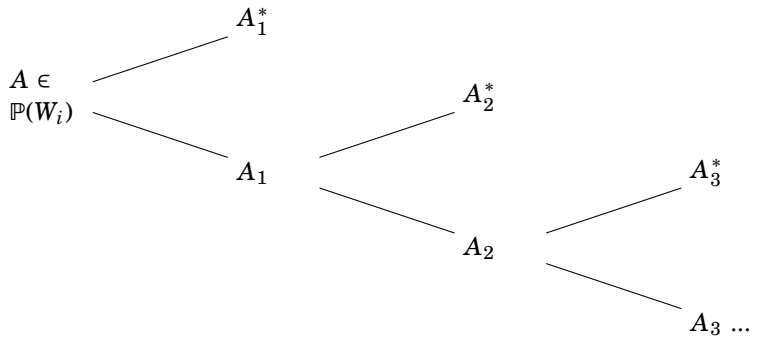

Figure 1: Set splitting process to classify set $A \in \mathbb{P}(W_i)$ as a connected or disconnected set.

In order to find most likely clusters, equation (1) is used as a likelihood ratio test for each element of $\mathcal{Z}_2$ This exhaustive search in $\mathcal{Z}_2$ covers all circular and irregularly shaped regions, allowing the flexible method to detect non-circular clusters. However, the main issue of the proposed method is the computational time that increases exponentially as K gets larger. For instance, for $K = 10$ and $K = 15$ it takes three and 25 days, respectively. Furthermore, it is not feasible for large K, say $K = 30$. Moreover, in cases where the true cluster is circular, the flexible method is tends to detect clusters larger than the true cluster.

2.5 The restricted flexible scan test

In the previous section, the flexible method was used to detect circular and non-circular clusters. Due to the computation inefficiency of the flexible method, Tango and Takahashi proposed a restricted likelihood ratio to decrease the computation time while still to detecting larger clusters (we implemented $K = 90$ regions). Furthermore, the flexible method tends to detect much larger clusters than the true cluster through including some nonelevated risk regions. To eliminate nonelevated risk regions from the cluster candidates, Tango proposed the restricted likelihood ratio that takes into account the risk of each individual regions.
To accomplish this goal, Equation (1) was adjusted to the following restricted likelihood spatial scan statistic:

$$\sup_{Z \in \mathcal{Z}} \frac{\left(\frac{y_Z}{n_Z}\right)^{y_Z} \left(\frac{y-y_Z}{n-n_Z}\right)^{y-y_Z}}{\left(\frac{y}{n}\right)^{y}} I\left(\frac{y_Z}{n_Z} > \frac{y-y_Z}{n-n_Z}\right) \prod_{Z \in \mathcal{Z}} I(p_i < \alpha_1), \tag{3}$$

and p_i is middle p-value given by

$$p_i = P(Y_i \geq y_i + 1) + \frac{1}{2} P(Y_i = y_i); \tag{4}$$

where $Y_i \sim \text{Poisson}(n_i \theta_i)$ and α_1 is the level of significance for each region. In this paper, the restricted method was implemented at $\alpha_1 = 0.2$. Power was computed for two different type one errors $\alpha_0 = 0.01$ and $\alpha_0 = 0.05$. Also, prespecified K for K-nearest neighbors was set to $K = 90$ based on the largest cluster in Duczmal et al. (2006a) that includes 78 regions.
Based on equations (3) and (4), only those collection of regions could be MLC candidates that besides having a large likelihood ratio test, each individual region have a significantly elevated risk. Otherwise, $I(p_i < \alpha_1)$ becomes zero and therefore the entire region considered to be nonsignificant. ...

2.6 Minimum spanning trees

Assunção et al. (2006) propose a different method for detecting irregularly shaped clusters that utilizes minimum spanning trees. The goal is to efficiently scan a subset of potential zones without needing an arbitrarily selected tuning parameter (as seen in other methods). We briefly describe the intuition for the approach as well as the construction of candidate zones.

The key focus in this method is determining the potential zones which should belong within the subset to be scanned. The suggestion is to represent the data as a weighted graph, $\mathcal{G} = (\mathbf{V}, \mathbf{E}, \mathbf{W})$, where $\mathbf{V}$ is the set of region centroids, $\mathbf{E}$ is the set of edges (if regions share a boarder, an edge will exist between them on the graph), and $\mathbf{W}$ is the set of edge weights. Since the sets $\mathbf{V}$ and $\mathbf{E}$ are fixed, the goal is to determine edge weights so that they provide information about likely clusters. WLOG, we specify small weights to indicate regions that should be clustered together (i.e. they have similar disease risk) and large weights to indicate regions that should not be clustered together. A good guess for disease clusters can then be represented by connected zones with a small overall weight. Therefore, constructing a minimum spanning tree of $\mathcal{G}$ should allow us to find a small subset of potential clusters that are a decent representation of the overall most-likely cluster.

Assunção et al. (2006) develop two methods to accomplish the goal above. Then Costa et al. (2012) propose three modifications to the construction of minimum spanning trees in an attempt to resolve the octopus effect present in previous methods. We introduce one method from Assunção et al. (2006) and the three methods from Costa et al. (2012) in the following sections.

2.6.1 The dynamic minimum spanning tree (dMST) scan test

Assunção et al. (2006) developed the dMST method as an initial implementation of minimum spanning trees for disease cluster detection. The method begins by assigning weights to every edge of $\mathcal{G}$:

$$w(i,j) = log\left(\frac{MLE_\alpha}{MLE_0}\right)$$

Where i and j are two regions that share a border, MLE_α is the maximum likelihood estimate if the two regions have different disease rates, and MLE_0 is the maximum likelihood estimate if the two regions have the same disease rate. Then we use the following algorithm to construct a minimum spanning tree:

1. Arbitrarily select an initial centroid (vertex) to begin the tree
2. Add the edge (and corresponding vertex) with minimum weight that is connected to the current tree (such that the corresponding vertex is not currently in the tree)
3. Calculate new edge weights such that $w(i,j) = -log(l_{ij})$ where l_{ij} is the numerator of equation (1) or equation (2) (depends upon choice of a Poisson or a Binomial model)
4. Repeat steps 2 and 3 until a spanning tree is constructed (or a stopping criterion is met)
5. The most likely cluster from this tree is represented by the cluster that had the largest l_{ij} value
6. Repeat steps 1-5 until all possible initial vertices have been considered

The most likely cluster is the zone that provided the largest l_{ij} over all of the spanning trees. Assunção et al. (2006) discovered that the above method tends to favor large clusters that possess an octopus effect. Additionally, the algorithm takes a substantial amount of time to complete and may not be worth the minimal benefits of finding irregularly shaped clusters.

2.6.2 The early stopping dynamic minimum spanning tree (e-dMST) scan test

The e-dMST method makes one modification to the dMST method. In step 2 of the algorithm, a vertex will be added to the tree only if the addition increases the MLE. This effectively stops the algorithm from branching off into long reaching clusters, a common issue of the dMST test. As a consequence of this modification, the cluster of interest is simply the tree that is present when the algorithm stops. Thus, step 5 of the algorithm does not require any additional computation and we can simply move to step 6. This efficiently reduces the computational time and complexity of the dMST test.

2.6.3 The double scan test

Similar to the e-dMST, the double scan test includes the early stopping criterion. To increase the compactness of the tree throughout the algorithm, the double scan test will include an additional requirement. When constructing the tree, new vertices will be considered for addition if they have at least two edges incident to the current tree. The belief is that this additional constraint will force the tree to maintain a more plausible shape throughout construction. Again, the early stopping criterion significantly reduces the computational time required to complete the algorithm.

2.6.4 The m-link scan test

In an attempt to remove the octopus effect without using an early stopping criterion, the m-link method was introduced. Again, suppose we are constructing a minimum tree as described for the dMST. The m-link algorithm will seek to add a vertex not currently in the tree such that this vertex has a maximum connection to the current tree. Note, if two or more vertices share the maximum connection, the vertex that provides the largest MLE is selected. Otherwise the MLE is not used when constructing the tree. A stopping criterion is required (typically a population size or number of regions, but not the early stopping criterion used by the e-dMST and double tests) to determine when the tree is large enough to contain a good estimate for a cluster. We then use the same frame work as explained for the dMST to calculate the most likely cluster.

Notice, this method should be similar in computational complexity to the dMST method because a similar amount of calculations are made in the two methods. As we will see, this is not a trivial observation, as both the dMST and mlink tests take a significant amount of time to run.

2.7 The fast subset scan test

While methods such as the elliptical scan statistic or flexible scan statistic have high power to detect disease clusters by searching a large set of potential subsets of the study area, they tend to suffer from long computation times, with the number of potential clusters increasing rapidly as the size of the study area increases. Neill (Neill, 2012) proposed a fast subset scan, which seeks the maximum scan statistic over a small set of potential clusters, doing so in linear time.

Let $Z \subseteq A$ be a potential cluster, and define $S(Z)$ to be the statistic (1) for Z. In this method, each region $i \in \{1,\ldots,N\}$ in the study area is assigned a priority based on some priority function $G(i)$, and sorts the areas according to priority. Let $R_{(j)}$ be the region with the j^{th} highest priority. Neil defines a statistic S and priority function G to have the linear time subset scanning (LTSS) property as follows: A score function S and priority function G satisfy the LTSS property if and only if

$$\max_{Z \subseteq A}\{S(Z)\} = \max_{j=1,\ldots,N}\left[S(\{R_{(1)},\ldots,R_{(j)}\})\right].$$

In other words, if the maximum over all statistics occurs on a subset of the first j regions ordered by priority, then that statistic S and priority function G have the LTSS property and the global maximum of S can be found in linear time.

Neill proves that Kulldorff's scan statistic as defined by (1) has the LTSS property when accompanied by the priority function $G(i) = y_i/e_i$, the ratio of case count to expected case count in each region, where $e_i = \left(\frac{y}{n}\right) n_i$.

The unrestricted fast subset scan method can be implemented as follows:

1. Calculate the expected counts $e_i = \left(\frac{y}{n}\right) n_i$ for each region.
2. Calculate the priority score $G(i) = y_i/e_i$ for each region.
3. Sort the regions in descending priority order. This can be done in $O(N \log N)$ time.
4. Calculate $S(\{R_{(1)},\ldots,R_{(j)}\})$ for each $j = 1,\ldots,N$, for a total of N statistics.
5. Take the maximum over the N statistics.

As only N statistics need to be calculated, the maximum scan statistic can be computed in $O(N)$ time. When applied to the northeast benchmark data, the unconstrained fast scan computed the most likely cluster in approximately 2-4 seconds per 10,000 alternative data sets and 15 seconds per null model.

Neill's fast subset scan is similar to the ULS method in that both order the regions by priority, and consider subsets based on the ordering. However, the fast subset scan does not impose any connectivity constraints. This means that the unrestricted fast scan often returns a set of disconnected regions as the "most likely cluster," due to those regions having the highest priority scores.

Variations on the fast scan exist which impose proximity constraints, forcing the fast subset scan to return a set of spatially close regions as the most likely cluster, though the regions may still be disconnected. Neill proposes two spatially constrained local subset scan methods: Fixed k neighborhood, and fixed radius r.

For the fixed k approach, a region i and its $k-1$ nearest neighbors form a local neighborhood. The fast subset scan is performed on each of the N neighborhoods defined by the N regions in the study area, and the maximum taken over all the statistics. The fixed r method is similar in execution, but the local neighborhoods are defined by all regions within a fixed distance r of region i.

Both the fixed k and the fixed r local subset scans can be repeated with multiple values of k and r, which Neill names the multiscan k and multiscan r methods. Once each set of scans has been performed, a Pareto set of the scan statistics is created containing all potential clusters Z of size k_Z or r_Z such that no other cluster Z' of size $k_{Z'}$ or Z' has the following properties:

$$S_{Z'} > S_Z \text{ and } k_{Z'} \leq k_Z \text{ or } r_{Z'} \leq r_Z,$$
$$S_{Z'} = S_Z \text{ and } k_{Z'} < k_Z \text{ or } r_{Z'} < r_Z.$$

The most likely cluster is then selected from those in the Pareto set by computing $S_Z - Lk_Z$ or $S_Z - Lr_Z$ for some constant L and taking the maximum. Larger values of L penalize larger clusters and result in smaller clusters with higher scores, while values of L at or near zero penalize size less and result in larger clusters.

The mutiscan k and multiscan r approaches take considerably more time to compute than the unrestricted fast scan, as instead of searching N potential clusters, the multiscan k searches N^2k potential clusters for each value of k, and the multiscan r searches $N^2\bar{k}$ potential clusters for each value of r, where $\bar{k}$ is the average neighborhood size for that r.

When applies to the benchmark data, the multiscan k approach with $k = 5, 10, \ldots, 60$ and $L = 0.5$ averages six minutes per 10,000 alternative data sets and 40 minutes per 99,999 null data sets using multicore processing and 32 cores.

2.8 Other methods

Static Minimum Spanning Tree (Nick) Problem: no power MultiScan, Fixed k, Fixed r, etc. (other versions of the past scan) (Lauren) GraphScan (Lauren) Spatial significant cluster detection (Lee) Problem: proprietary software PULSE (Lee): No software available and appears to be computationally expensive Duczmal and Assuncao (2004): Complicated, computationally expensive, lower power (Mohammad) Duczmal et al. (2007): Also complicated? (Mohammad) Duczmal (2008): - Chau what papers cited? Why didn't we implement? Focused elliptical scan - no software?

2.8.1 The confocal elliptic scan

Another version of the elliptic scan method was introduced by Christiansen et al. (2006),in which, an ellipse is constructed based on focal points. The centroid of each region is treated as the first focal and the second focal is gradually chosen based on its distance to the first focal until some specified proportion is met. Using the constant sum property of an ellipse, other centroids are ranked as the next nearest points based on the sum of the distances to the chosen focal points until some other proportion is met.

The confocal elliptic scan method, in general, is not much different from the elliptic scan described in section 2.2 except an ellipse is formed from the foci instead of from the center. In the confocal version, the combination of sets of second focal points and sets of regions make the number of testing regions large. Moreover, up till the time of this paper, there is no software publicly available for this method. It may not be worth to put further investment on this version.

3 Benchmark evaluations

3.1 Background

We now benchmark the methods described in Section 2 using publicly-available benchmark data sets constructed by Kulldorff et al. (2003) and Duczmal et al. (2006b). The data sets are inspired by breast cancer mortality data from the northeastern United States during the years 1988-1992. The data were previously studied in Kulldorff et al. (1997) and elsewhere. The original breast cancer data set included mortality counts for 245 regions spread throughout Connecticut, Delaware, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont, and the District of Columbia. The population in each region corresponds to the number of women recorded in the region for the 1990 United States Census. A centroid coordinate is associated with each region.

Kulldorff et al. (2003) provided data sets based on 70 different clustering models. Kulldorff et al. (2003) simulated clusters centered around three area types: a rural area (Grand Isle County in Vermont), a "mixed" area (Allegheny County in Pennsylvania), and an urban area (New York City). Clusters with 1, 2, 4, 8, and 16 regions were simulated for each area type. For each of the 15 clusters, the risk within the cluster was greater than the regions outside the cluster. In addition, Kulldorff et al. (2003) created additional scenarios by combining the clusters previously described. Specifically, there were scenarios where rural and urban, rural and mixed, mixed and urban, and rural, mixed, and urban clusters occurred simultaneously, each occurring with the same number of regions in each cluster. For each cluster model, data sets were simulated with both 600 and 6000 cases distributed across the regions in proportion to population, except in the in the clusters, which had a greater risk. simulated in each region according to a multinomial distribution where the risk in each region was n_i/n All the clusters simulated by Kulldorff et al. (2003) were circular in shape. Duczmal et al. (2006b) simulated 11 different irregularly-shaped clusters using similar principles. For each of the 81 different cluster models (35 regions $\times$ 2 case sized + 11 irregularly-shaped clusters), 10,000 data sets were generated.

For each set of observed cases (600 and 6000), 99,999 data sets were generated under the constant risk hypothesis. These data sets are used to determine the null distribution.

The relative risk of each cluster was chosen so that the null hypothesis (of no cluster) would be rejected with probability 0.999 using a standard binomial test, assuming the cluster regions were known ahead of time. The mean associated with each cluster under the null hypothesis, the alternative hypothesis, and the associated relative risk under the alternative hypothesis are provided in Table .

4 Discussion

References

Assunção, R., Costa, M., Tavares, A., and Ferreira, S. (2006). Fast detection of arbitrarily shaped disease clusters. *Statistics in Medicine*, 25(5):723–742.

Besag, J. and Newell, J. (1991). The detection of clusters in rare diseases. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 154(1):143–155.

Bivand, R. S., Pebesma, E., and Gomez-Rubio, V. (2013). *Applied spatial data analysis with R, Second edition*. Springer, NY.

Christiansen, L. E., Andersen, J. S., Wegener, H. C., and Madsen, H. (2006). Spatial scan statistics using elliptic windows. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(4):411.

Costa, M. A., Assunção, R. M., and Kulldorff, M. (2012). Constrained spanning tree algorithms for irregularly-shaped spatial clustering. *Computational Statistics & Data Analysis*, 56(6):1771–1783.

Duczmal, L. and Assuncao, R. (2004). A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics & Data Analysis*, 45(2):269–286.

Duczmal, L., Kulldorff, M., and Huang, L. (2006a). Evaluation of spatial scan statistics for irregularly shaped clusters. *Journal of Computational and Graphical Statistics*, 15(2):428–442.

Duczmal, L., Kulldorff, M., and Huang, L. (2006b). Evaluation of spatial scan statistics for irregularly shaped clusters. *Journal of Computational and Graphical Statistics*, 15(2):428–442.

French, J. P. (2019). *smerc: Statistical Methods for Regional Counts*. R package version 1.0.

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician*, 5(3):115–146.

Google Scholar (2019). Accessed January 11, 2019.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics - Theory and Methods*, 26(6):1481–1496.

Kulldorff, M. (2018). SaTScan, version 9.6.

Kulldorff, M., Feuer, E. J., Miller, B. A., and Freedma, L. S. (1997). Breast cancer clusters in the northeast united states: A geographic analysis. *American Journal of Epidemiology*, 146(2):161–170.

Kulldorff, M., Huang, L., Pickle, L., and Duczmal, L. (2006). An elliptic spatial scan statistic. *Statistics in Medicine*, 25(22):3929–3943.

Kulldorff, M. and Nagarwalla, N. (1995). Spatial disease clusters: Detection and inference. *Statistics in Medicine*, 14(8):799–810.

Kulldorff, M., Tango, T., and Park, P. J. (2003). Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis*, 42(4):665–684.

Mclafferty, S. (2015). Disease cluster detection methods: recent developments and public health implications. *Annals of GIS*, 21(2):127–133.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23.

Neill, D. B. (2012). Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):337–360.

Openshaw, S., Charlton, M., Craft, A. W., and Birch, J. (1988). Investigation of leukaemia clusters by use of a geographical analysis machine. *The Lancet*, 331(8580):272–273.

Patil, G. P. and Taillie, C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological statistics*, 11(2):183–197.

Rushton, G. and Lolonis, P. (1996). Exploratory spatial analysis of birth defect rates in an urban population. *Statistics in Medicine*, 15(7-9):717–726.

Tango, T. (2010). *Statistical methods for disease clustering*. Springer-Verlag, New York.

Tango, T. and Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters. *International journal of health geographics*, 4(1):11.

Turnbull, Bruce W.and Iwano, E. J., Burnett, W. S., Howe, H. L., and Clark, L. C. (1990). Monitoring for clusters of disease: Application to leukemia incidence in upstate new york. *American Journal of Epidemiology*, 132(supp1):136–143.

Waller, L. A. and Gotway, C. A. (2004). *Applied spatial statistics for public health data*. John Wiley & Sons, Hoboken.

	600															
	Circular		**Flex** $K = 10$		**Flex** $K = 15$		**Restricted**		**ULS**		—		—		—	
Counties	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01						
mixed 01	0.93560	0.871	0.929	0.858	—	—	—	—	0.1092	0.0379	—	—	—	—	—	—
mixed 02	0.93910	0.871	0.931	0.858	—	—	—	—	0.1239	0.0418	—	—	—	—	—	—
mixed 04	0.93720	0.873	0.930	0.855	—	—	—	—	0.1416	0.0515	—	—	—	—	—	—
mixed 08	0.94130	0.876	0.934	0.857	—	—	—	—	0.1386	0.0524	—	—	—	—	—	—
mixed 16	0.94850	0.886	0.913	0.818	—	—	—	—	0.1791	0.0670	—	—	—	—	—	—
rural 01	0.99840	0.992	0.998	0.992	—	—	—	—	0.7206	0.5932	—	—	—	—	—	—
rural 02	0.99120	0.986	0.988	0.971	—	—	—	—	0.3856	0.1923	—	—	—	—	—	—
rural 04	0.97250	0.946	0.973	0.941	—	—	—	—	0.1555	0.0462	—	—	—	—	—	—
rural 08	0.97110	0.937	0.973	0.938	—	—	—	—	0.1734	0.0498	—	—	—	—	—	—
rural 16	0.96950	0.936	0.954	0.888	—	—	—	—	0.2115	0.0647	—	—	—	—	—	—
urban 01	0.92220	0.818	0.896	0.775	—	—	—	—	0.1414	0.0431	—	—	—	—	—	—
urban 02	0.90280	0.822	0.881	0.791	—	—	—	—	0.1326	0.0403	—	—	—	—	—	—
urban 04	0.89230	0.794	0.864	0.750	—	—	—	—	0.1197	0.0377	—	—	—	—	—	—
urban 08	0.91300	0.824	0.876	0.767	—	—	—	—	0.1221	0.0391	—	—	—	—	—	—
urban 16	0.92620	0.836	0.709	0.515	—	—	—	—	0.1242	0.0366	—	—	—	—	—	—
urbmix 01	0.98700	0.950	0.983	0.937	—	—	—	—	0.1228	0.0514	—	—	—	—	—	—
urbmix 02	0.98390	0.950	0.979	0.938	—	—	—	—	0.1231	0.0524	—	—	—	—	—	—
urbmix 04	0.96560	0.901	0.957	0.880	—	—	—	—	0.1054	0.0459	—	—	—	—	—	—
urbmix 08	0.95450	0.871	0.942	0.846	—	—	—	—	0.0992	0.0413	—	—	—	—	—	—
urbmix 16	0.93490	0.811	0.850	0.655	—	—	—	—	0.1015	0.0428	—	—	—	—	—	—
rurmix 01	0.99972	0.999	1.000	0.999	—	—	—	—	0.6941	0.5447	—	—	—	—	—	—
rurmix 02	0.99900	0.997	0.998	0.993	—	—	—	—	0.3573	0.1659	—	—	—	—	—	—
rurmix 04	0.99660	0.987	0.997	0.982	—	—	—	—	0.1692	0.0595	—	—	—	—	—	—
rurmix 08	0.99650	0.986	0.997	0.984	—	—	—	—	0.1741	0.0786	—	—	—	—	—	—
rurmix 16	0.99570	0.982	0.991	0.962	—	—	—	—	0.2255	0.1216	—	—	—	—	—	—
	6000															
mixed 01	0.88450	0.783	0.875	0.772	—	—	—	—	0.1434	0.0527	—	—	—	—	—	—
mixed 02	0.89050	0.784	0.879	0.777	—	—	—	—	0.1475	0.0549	—	—	—	—	—	—
mixed 04	0.89140	0.784	0.873	0.766	—	—	—	—	0.1755	0.0715	—	—	—	—	—	—
mixed 08	0.90500	0.811	0.888	0.783	—	—	—	—	0.1795	0.0685	—	—	—	—	—	—
mixed 16	0.92280	0.830	0.859	0.727	—	—	—	—	0.1941	0.0735	—	—	—	—	—	—
rural 01	0.99050	0.974	0.990	0.974	—	—	—	—	0.2026	0.0815	—	—	—	—	—	—
rural 02	0.95520	0.901	0.935	0.891	—	—	—	—	0.0753	0.0182	—	—	—	—	—	—
rural 04	0.92030	0.844	0.919	0.838	—	—	—	—	0.0927	0.0244	—	—	—	—	—	—
rural 08	0.92880	0.846	0.924	0.843	—	—	—	—	0.1176	0.0357	—	—	—	—	—	—
rural 16	0.93560	0.850	0.882	0.749	—	—	—	—	0.1359	0.0411	—	—	—	—	—	—
urban 01	0.84070	0.733	0.800	0.671	—	—	—	—	0.1137	0.0425	—	—	—	—	—	—
urban 02	0.84840	0.728	0.816	0.674	—	—	—	—	0.1192	0.0422	—	—	—	—	—	—
urban 04	0.86220	0.731	0.822	0.683	—	—	—	—	0.0995	0.0383	—	—	—	—	—	—
urban 08	0.89570	0.782	0.843	0.711	—	—	—	—	0.0961	0.0367	—	—	—	—	—	—
urban 16	0.91840	0.817	0.679	0.485	—	—	—	—	0.0946	0.0334	—	—	—	—	—	—
urbmix 01	0.96830	0.907	0.959	0.884	—	—	—	—	0.1705	0.0853	—	—	—	—	—	—
urbmix 02	0.96550	0.897	0.955	0.872	—	—	—	—	0.1801	0.0878	—	—	—	—	—	—
urbmix 04	0.95780	0.873	0.946	0.851	—	—	—	—	0.1560	0.0773	—	—	—	—	—	—
urbmix 08	0.94420	0.842	0.925	0.809	—	—	—	—	0.1442	0.0673	—	—	—	—	—	—
urbmix 16	0.93440	0.816	0.829	0.638	—	—	—	—	0.1471	0.0638	—	—	—	—	—	—
rurmix 01	0.99850	0.994	0.998	0.994	—	—	—	—	0.2719	0.1191	—	—	—	—	—	—
rurmix 02	0.99310	0.975	0.989	0.973	—	—	—	—	0.1712	0.0717	—	—	—	—	—	—
rurmix 04	0.98990	0.961	0.990	0.958	—	—	—	—	0.2156	0.1044	—	—	—	—	—	—
rurmix 08	0.98960	0.958	0.986	0.951	—	—	—	—	0.2627	0.1361	—	—	—	—	—	—
rurmix 16	0.99070	0.961	0.976	0.909	—	—	—	—	0.3271	0.1818	—	—	—	—	—	—
	Irregulary Shaped															
a	0.85270	0.724	0.878	0.758	—	—	—	—	0.2205	0.0778	—	—	—	—	—	—
b	0.78680	0.610	0.776	0.583	—	—	—	—	0.2953	0.1049	—	—	—	—	—	—
c	0.87510	0.755	0.876	0.768	—	—	—	—	0.1926	0.0709	—	—	—	—	—	—
d	0.86050	0.712	0.852	0.688	—	—	—	—	0.3230	0.1169	—	—	—	—	—	—
e	0.80650	0.632	0.739	0.514	—	—	—	—	0.3692	0.1479	—	—	—	—	—	—
f	0.69460	0.489	0.548	0.317	—	—	—	—	0.1982	0.0779	—	—	—	—	—	—
g	0.45850	0.238	0.503	0.274	—	—	—	—	0.3642	0.1622	—	—	—	—	—	—
h	0.65630	0.445	0.638	0.418	—	—	—	—	0.3275	0.1415	—	—	—	—	—	—
i	0.77440	0.592	0.637	0.398	—	—	—	—	0.2602	0.1050	—	—	—	—	—	—
j	0.68630	0.474	0.538	0.292	—	—	—	—	0.4524	0.2211	—	—	—	—	—	—
k	0.79720	0.610	0.438	0.200	—	—	—	—	0.5001	0.2694	—	—	—	—	—	—

Table 1: Basic Power

	600															
	Circular		**Flex** $K = 10$		**Flex** $K = 15$		**Restricted**		**ULS**		—		—		—	
Counties	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01						
mixed 01	—	—	—	—	—	—	—	—	0.1077	0.0379	—	—	—	—	—	—
mixed 02	—	—	—	—	—	—	—	—	0.1226	0.0417	—	—	—	—	—	—
mixed 04	—	—	—	—	—	—	—	—	0.1412	0.0514	—	—	—	—	—	—
mixed 08	—	—	—	—	—	—	—	—	0.1382	0.0523	—	—	—	—	—	—
mixed 16	—	—	—	—	—	—	—	—	0.1789	0.0670	—	—	—	—	—	—
rural 01	—	—	—	—	—	—	—	—	0.7145	0.5921	—	—	—	—	—	—
rural 02	—	—	—	—	—	—	—	—	0.3781	0.1912	—	—	—	—	—	—
rural 04	—	—	—	—	—	—	—	—	0.1510	0.0460	—	—	—	—	—	—
rural 08	—	—	—	—	—	—	—	—	0.1715	0.0497	—	—	—	—	—	—
rural 16	—	—	—	—	—	—	—	—	0.2112	0.0647	—	—	—	—	—	—
urban 01	—	—	—	—	—	—	—	—	0.1404	0.0430	—	—	—	—	—	—
urban 02	—	—	—	—	—	—	—	—	0.1322	0.0403	—	—	—	—	—	—
urban 04	—	—	—	—	—	—	—	—	0.1197	0.0377	—	—	—	—	—	—
urban 08	—	—	—	—	—	—	—	—	0.1221	0.0391	—	—	—	—	—	—
urban 16	—	—	—	—	—	—	—	—	0.1242	0.0366	—	—	—	—	—	—
urbmix 01	—	—	—	—	—	—	—	—	0.1053	0.0493	—	—	—	—	—	—
urbmix 02	—	—	—	—	—	—	—	—	0.1069	0.0500	—	—	—	—	—	—
urbmix 04	—	—	—	—	—	—	—	—	0.0975	0.0451	—	—	—	—	—	—
urbmix 08	—	—	—	—	—	—	—	—	0.0892	0.0395	—	—	—	—	—	—
urbmix 16	—	—	—	—	—	—	—	—	0.0967	0.0424	—	—	—	—	—	—
rurmix 01	—	—	—	—	—	—	—	—	0.6655	0.5365	—	—	—	—	—	—
rurmix 02	—	—	—	—	—	—	—	—	0.3098	0.1524	—	—	—	—	—	—
rurmix 04	—	—	—	—	—	—	—	—	0.1219	0.0506	—	—	—	—	—	—
rurmix 08	—	—	—	—	—	—	—	—	0.1426	0.0733	—	—	—	—	—	—
rurmix 16	—	—	—	—	—	—	—	—	0.2016	0.1185	—	—	—	—	—	—
	6000															
mixed 01	—	—	—	—	—	—	—	—	0.1427	0.0527	—	—	—	—	—	—
mixed 02	—	—	—	—	—	—	—	—	0.1471	0.0549	—	—	—	—	—	—
mixed 04	—	—	—	—	—	—	—	—	0.1753	0.0715	—	—	—	—	—	—
mixed 08	—	—	—	—	—	—	—	—	0.1792	0.0685	—	—	—	—	—	—
mixed 16	—	—	—	—	—	—	—	—	0.1941	0.0735	—	—	—	—	—	—
rural 01	—	—	—	—	—	—	—	—	0.1937	0.0796	—	—	—	—	—	—
rural 02	—	—	—	—	—	—	—	—	0.0681	0.0167	—	—	—	—	—	—
rural 04	—	—	—	—	—	—	—	—	0.0900	0.0242	—	—	—	—	—	—
rural 08	—	—	—	—	—	—	—	—	0.1170	0.0356	—	—	—	—	—	—
rural 16	—	—	—	—	—	—	—	—	0.1359	0.0411	—	—	—	—	—	—
urban 01	—	—	—	—	—	—	—	—	0.1121	0.0423	—	—	—	—	—	—
urban 02	—	—	—	—	—	—	—	—	0.1184	0.0422	—	—	—	—	—	—
urban 04	—	—	—	—	—	—	—	—	0.0995	0.0383	—	—	—	—	—	—
urban 08	—	—	—	—	—	—	—	—	0.0961	0.0367	—	—	—	—	—	—
urban 16	—	—	—	—	—	—	—	—	0.0946	0.0334	—	—	—	—	—	—
urbmix 01	—	—	—	—	—	—	—	—	0.1614	0.0833	—	—	—	—	—	—
urbmix 02	—	—	—	—	—	—	—	—	0.1719	0.0870	—	—	—	—	—	—
urbmix 04	—	—	—	—	—	—	—	—	0.1516	0.0772	—	—	—	—	—	—
urbmix 08	—	—	—	—	—	—	—	—	0.1414	0.0671	—	—	—	—	—	—
urbmix 16	—	—	—	—	—	—	—	—	0.1445	0.0635	—	—	—	—	—	—
rurmix 01	—	—	—	—	—	—	—	—	0.2347	0.1082	—	—	—	—	—	—
rurmix 02	—	—	—	—	—	—	—	—	0.1367	0.0611	—	—	—	—	—	—
rurmix 04	—	—	—	—	—	—	—	—	0.1965	0.0997	—	—	—	—	—	—
rurmix 08	—	—	—	—	—	—	—	—	0.2538	0.1339	—	—	—	—	—	—
rurmix 16	—	—	—	—	—	—	—	—	0.3232	0.1881	—	—	—	—	—	—
	Irregulary Shaped															
a	—	—	—	—	—	—	—	—	0.2205	0.0778	—	—	—	—	—	—
b	—	—	—	—	—	—	—	—	0.2953	0.1049	—	—	—	—	—	—
c	—	—	—	—	—	—	—	—	0.1924	0.0709	—	—	—	—	—	—
d	—	—	—	—	—	—	—	—	0.3230	0.1169	—	—	—	—	—	—
e	—	—	—	—	—	—	—	—	0.3692	0.1479	—	—	—	—	—	—
f	—	—	—	—	—	—	—	—	0.1982	0.0779	—	—	—	—	—	—
g	—	—	—	—	—	—	—	—	0.3642	0.1622	—	—	—	—	—	—
h	—	—	—	—	—	—	—	—	0.3275	0.1415	—	—	—	—	—	—
i	—	—	—	—	—	—	—	—	0.2602	0.1050	—	—	—	—	—	—
j	—	—	—	—	—	—	—	—	0.4524	0.2211	—	—	—	—	—	—
k	—	—	—	—	—	—	—	—	0.5001	0.2694	—	—	—	—	—	—

Table 2: Intersect Power

	600															
	Circular		**Flex $K = 10$**		**Flex $K = 15$**		**Restricted**		**ULS**		—		—		—	
Counties	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01						
mixed 01	—	—	—	—	—	—	—	—	0.1077	0.0379	—	—	—	—	—	—
mixed 02	—	—	—	—	—	—	—	—	0.1168	0.0400	—	—	—	—	—	—
mixed 04	—	—	—	—	—	—	—	—	0.1209	0.0454	—	—	—	—	—	—
mixed 08	—	—	—	—	—	—	—	—	0.0522	0.0198	—	—	—	—	—	—
mixed 16	—	—	—	—	—	—	—	—	0.0037	0.0016	—	—	—	—	—	—
rural 01	—	—	—	—	—	—	—	—	0.7145	0.5921	—	—	—	—	—	—
rural 02	—	—	—	—	—	—	—	—	0.3224	0.1681	—	—	—	—	—	—
rural 04	—	—	—	—	—	—	—	—	0.0490	0.0144	—	—	—	—	—	—
rural 08	—	—	—	—	—	—	—	—	0.0159	0.0045	—	—	—	—	—	—
rural 16	—	—	—	—	—	—	—	—	0.0002	0.0000	—	—	—	—	—	—
urban 01	—	—	—	—	—	—	—	—	0.1404	0.0430	—	—	—	—	—	—
urban 02	—	—	—	—	—	—	—	—	0.1286	0.0397	—	—	—	—	—	—
urban 04	—	—	—	—	—	—	—	—	0.1079	0.0347	—	—	—	—	—	—
urban 08	—	—	—	—	—	—	—	—	0.0738	0.0249	—	—	—	—	—	—
urban 16	—	—	—	—	—	—	—	—	0.0136	0.0032	—	—	—	—	—	—
urbmix 01	—	—	—	—	—	—	—	—	0.1053	0.0493	—	—	—	—	—	—
urbmix 02	—	—	—	—	—	—	—	—	0.1025	0.0478	—	—	—	—	—	—
urbmix 04	—	—	—	—	—	—	—	—	0.0817	0.0373	—	—	—	—	—	—
urbmix 08	—	—	—	—	—	—	—	—	0.0292	0.0123	—	—	—	—	—	—
urbmix 16	—	—	—	—	—	—	—	—	0.0011	0.0005	—	—	—	—	—	—
rurmix 01	—	—	—	—	—	—	—	—	0.6655	0.5365	—	—	—	—	—	—
rurmix 02	—	—	—	—	—	—	—	—	0.2638	0.1331	—	—	—	—	—	—
rurmix 04	—	—	—	—	—	—	—	—	0.0372	0.0148	—	—	—	—	—	—
rurmix 08	—	—	—	—	—	—	—	—	0.0108	0.0065	—	—	—	—	—	—
rurmix 16	—	—	—	—	—	—	—	—	0.0002	0.0001	—	—	—	—	—	—
	6000															
mixed 01	—	—	—	—	—	—	—	—	0.1427	0.0527	—	—	—	—	—	—
mixed 02	—	—	—	—	—	—	—	—	0.1426	0.0532	—	—	—	—	—	—
mixed 04	—	—	—	—	—	—	—	—	0.1508	0.0606	—	—	—	—	—	—
mixed 08	—	—	—	—	—	—	—	—	0.0800	0.0307	—	—	—	—	—	—
mixed 16	—	—	—	—	—	—	—	—	0.0080	0.0030	—	—	—	—	—	—
rural 01	—	—	—	—	—	—	—	—	0.1937	0.0796	—	—	—	—	—	—
rural 02	—	—	—	—	—	—	—	—	0.0641	0.0158	—	—	—	—	—	—
rural 04	—	—	—	—	—	—	—	—	0.0622	0.0168	—	—	—	—	—	—
rural 08	—	—	—	—	—	—	—	—	0.0481	0.0146	—	—	—	—	—	—
rural 16	—	—	—	—	—	—	—	—	0.0093	0.0027	—	—	—	—	—	—
urban 01	—	—	—	—	—	—	—	—	0.1121	0.0423	—	—	—	—	—	—
urban 02	—	—	—	—	—	—	—	—	0.1143	0.0412	—	—	—	—	—	—
urban 04	—	—	—	—	—	—	—	—	0.0867	0.0344	—	—	—	—	—	—
urban 08	—	—	—	—	—	—	—	—	0.0570	0.0208	—	—	—	—	—	—
urban 16	—	—	—	—	—	—	—	—	0.0096	0.0028	—	—	—	—	—	—
urbmix 01	—	—	—	—	—	—	—	—	0.1614	0.0833	—	—	—	—	—	—
urbmix 02	—	—	—	—	—	—	—	—	0.1646	0.0831	—	—	—	—	—	—
urbmix 04	—	—	—	—	—	—	—	—	0.1292	0.0658	—	—	—	—	—	—
urbmix 08	—	—	—	—	—	—	—	—	0.0526	0.0252	—	—	—	—	—	—
urbmix 16	—	—	—	—	—	—	—	—	0.0030	0.0017	—	—	—	—	—	—
rurmix 01	—	—	—	—	—	—	—	—	0.2347	0.1082	—	—	—	—	—	—
rurmix 02	—	—	—	—	—	—	—	—	0.1282	0.0572	—	—	—	—	—	—
rurmix 04	—	—	—	—	—	—	—	—	0.1306	0.0664	—	—	—	—	—	—
rurmix 08	—	—	—	—	—	—	—	—	0.1008	0.0539	—	—	—	—	—	—
rurmix 16	—	—	—	—	—	—	—	—	0.0168	0.0097	—	—	—	—	—	—
	Irregulary Shaped															
a	—	—	—	—	—	—	—	—	0.0011	0.0004	—	—	—	—	—	—
b	—	—	—	—	—	—	—	—	0.0018	0.0010	—	—	—	—	—	—
c	—	—	—	—	—	—	—	—	0.1044	0.0410	—	—	—	—	—	—
d	—	—	—	—	—	—	—	—	0.0084	0.0040	—	—	—	—	—	—
e	—	—	—	—	—	—	—	—	0.0003	0.0001	—	—	—	—	—	—
f	—	—	—	—	—	—	—	—	0.0000	0.0000	—	—	—	—	—	—
g	—	—	—	—	—	—	—	—	0.0008	0.0007	—	—	—	—	—	—
h	—	—	—	—	—	—	—	—	0.0000	0.0000	—	—	—	—	—	—
i	—	—	—	—	—	—	—	—	0.0000	0.0000	—	—	—	—	—	—
j	—	—	—	—	—	—	—	—	0.0000	0.0000	—	—	—	—	—	—
k	—	—	—	—	—	—	—	—	0.0000	0.0000	—	—	—	—	—	—

Table 3: Contain Power

	600															
	Circular		**Flex** $K = 10$		**Flex** $K = 15$		**Restricted**		**ULS**		—		—		—	
Counties	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01						
mixed 01	—	—	—	—	—	—	—	—	0.1077	0.0379	—	—	—	—	—	—
mixed 02	—	—	—	—	—	—	—	—	0.1197	0.0409	—	—	—	—	—	—
mixed 04	—	—	—	—	—	—	—	—	0.1357	0.0498	—	—	—	—	—	—
mixed 08	—	—	—	—	—	—	—	—	0.1223	0.0465	—	—	—	—	—	—
mixed 16	—	—	—	—	—	—	—	—	0.1419	0.0537	—	—	—	—	—	—
rural 01	—	—	—	—	—	—	—	—	0.7145	0.5921	—	—	—	—	—	—
rural 02	—	—	—	—	—	—	—	—	0.3503	0.1797	—	—	—	—	—	—
rural 04	—	—	—	—	—	—	—	—	0.1234	0.0376	—	—	—	—	—	—
rural 08	—	—	—	—	—	—	—	—	0.1360	0.0394	—	—	—	—	—	—
rural 16	—	—	—	—	—	—	—	—	0.1513	0.0470	—	—	—	—	—	—
urban 01	—	—	—	—	—	—	—	—	0.1404	0.0430	—	—	—	—	—	—
urban 02	—	—	—	—	—	—	—	—	0.1304	0.0400	—	—	—	—	—	—
urban 04	—	—	—	—	—	—	—	—	0.1162	0.0369	—	—	—	—	—	—
urban 08	—	—	—	—	—	—	—	—	0.1137	0.0370	—	—	—	—	—	—
urban 16	—	—	—	—	—	—	—	—	0.1065	0.0317	—	—	—	—	—	—
urbmix 01	—	—	—	—	—	—	—	—	0.1053	0.0493	—	—	—	—	—	—
urbmix 02	—	—	—	—	—	—	—	—	0.1047	0.0489	—	—	—	—	—	—
urbmix 04	—	—	—	—	—	—	—	—	0.0934	0.0431	—	—	—	—	—	—
urbmix 08	—	—	—	—	—	—	—	—	0.0775	0.0342	—	—	—	—	—	—
urbmix 16	—	—	—	—	—	—	—	—	0.0741	0.0327	—	—	—	—	—	—
rurmix 01	—	—	—	—	—	—	—	—	0.6655	0.5365	—	—	—	—	—	—
rurmix 02	—	—	—	—	—	—	—	—	0.2868	0.1428	—	—	—	—	—	—
rurmix 04	—	—	—	—	—	—	—	—	0.0993	0.0412	—	—	—	—	—	—
rurmix 08	—	—	—	—	—	—	—	—	0.1126	0.0584	—	—	—	—	—	—
rurmix 16	—	—	—	—	—	—	—	—	0.1397	0.0834	—	—	—	—	—	—
	6000															
mixed 01	—	—	—	—	—	—	—	—	0.1427	0.0527	—	—	—	—	—	—
mixed 02	—	—	—	—	—	—	—	—	0.1449	0.0541	—	—	—	—	—	—
mixed 04	—	—	—	—	—	—	—	—	0.1691	0.0688	—	—	—	—	—	—
mixed 08	—	—	—	—	—	—	—	—	0.1624	0.0622	—	—	—	—	—	—
mixed 16	—	—	—	—	—	—	—	—	0.1592	0.0600	—	—	—	—	—	—
rural 01	—	—	—	—	—	—	—	—	0.1937	0.0796	—	—	—	—	—	—
rural 02	—	—	—	—	—	—	—	—	0.0661	0.0163	—	—	—	—	—	—
rural 04	—	—	—	—	—	—	—	—	0.0828	0.0223	—	—	—	—	—	—
rural 08	—	—	—	—	—	—	—	—	0.1058	0.0322	—	—	—	—	—	—
rural 16	—	—	—	—	—	—	—	—	0.1146	0.0348	—	—	—	—	—	—
urban 01	—	—	—	—	—	—	—	—	0.1121	0.0423	—	—	—	—	—	—
urban 02	—	—	—	—	—	—	—	—	0.1164	0.0417	—	—	—	—	—	—
urban 04	—	—	—	—	—	—	—	—	0.0958	0.0373	—	—	—	—	—	—
urban 08	—	—	—	—	—	—	—	—	0.0893	0.0343	—	—	—	—	—	—
urban 16	—	—	—	—	—	—	—	—	0.0807	0.0286	—	—	—	—	—	—
urbmix 01	—	—	—	—	—	—	—	—	0.1614	0.0833	—	—	—	—	—	—
urbmix 02	—	—	—	—	—	—	—	—	0.1683	0.0851	—	—	—	—	—	—
urbmix 04	—	—	—	—	—	—	—	—	0.1457	0.0743	—	—	—	—	—	—
urbmix 08	—	—	—	—	—	—	—	—	0.1253	0.0595	—	—	—	—	—	—
urbmix 16	—	—	—	—	—	—	—	—	0.1137	0.0500	—	—	—	—	—	—
rurmix 01	—	—	—	—	—	—	—	—	0.2347	0.1082	—	—	—	—	—	—
rurmix 02	—	—	—	—	—	—	—	—	0.1325	0.0592	—	—	—	—	—	—
rurmix 04	—	—	—	—	—	—	—	—	0.1794	0.0911	—	—	—	—	—	—
rurmix 08	—	—	—	—	—	—	—	—	0.2275	0.1202	—	—	—	—	—	—
rurmix 16	—	—	—	—	—	—	—	—	0.2692	0.1515	—	—	—	—	—	—
	Irregulary Shaped															
a	—	—	—	—	—	—	—	—	0.1525	0.0550	—	—	—	—	—	—
b	—	—	—	—	—	—	—	—	0.2276	0.0821	—	—	—	—	—	—
c	—	—	—	—	—	—	—	—	0.1728	0.0648	—	—	—	—	—	—
d	—	—	—	—	—	—	—	—	0.2619	0.0966	—	—	—	—	—	—
e	—	—	—	—	—	—	—	—	0.2742	0.1122	—	—	—	—	—	—
f	—	—	—	—	—	—	—	—	0.1222	0.0492	—	—	—	—	—	—
g	—	—	—	—	—	—	—	—	0.2676	0.1238	—	—	—	—	—	—
h	—	—	—	—	—	—	—	—	0.2058	0.0937	—	—	—	—	—	—
i	—	—	—	—	—	—	—	—	0.1665	0.0687	—	—	—	—	—	—
j	—	—	—	—	—	—	—	—	0.2743	0.1393	—	—	—	—	—	—
k	—	—	—	—	—	—	—	—	0.2820	0.1565	—	—	—	—	—	—

Table 4: Overlap Power