ANALYSIS OF MAPPING TECHNIQUES ON A SPATIAL SCAN STATISTIC

Drew McClelland

A Senior Thesis Submitted to the Faculty of
The University of Utah

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Computer Science

School of Computing
The University of Utah
May 4, 2017

_____/_____	_____/_____
Jeff M. Phillips	H. James de St. Germain
Supervisor	Directory of Undergraduate Studies
	School of Computing

_____/_____
Ross Whitaker
Director
School of Computing

Abstract

The discovery of anomalous regions within spatially oriented data can provide valuable insights to the study of disease prevention, crime, and socioeconomic inequality. Spatial scan statistic methods are a common tool used within these fields to discover areas of a data set that have significant variation when compared to the background distribution. However, the effects of data representation on these techniques has not been rigorously studied. In particular, the effects of point-to-region and region-to-point mappings on scan statistics have not been explored in detail.

This thesis will attempt to shed light on the effects of mapping techniques on county and zip code region sets. Data values associated with regions will include cancer rates, educational attainment, and poverty. Additionally, a synthetic data set will be used in an attempt to evaluate the performance of these mapping techniques.

1 Introduction

Anomalous region discovery is a common problem in many real world data sets. The discovery of outlying regions can help to guide socio-economic spending and disease prevention. Many techniques have been explored in an attempt to aid these tasks. The field of scan statistics is commonly applied to such problems and has been heavily focused upon in recent years [1].

A common solution for detecting anomalous regions involves the application of scan statistics. These methods are typically applied to a set of two-dimensional points aligned on a plane. The goal of the scan statistic is to find an anomalous area over such a data set. The focus of this thesis is the analysis of a spatial scan statistic over *regions* rather than points.

It is common for real world data to measure values over regions and this can be caused by a variety of reasons. One example of this is privacy laws in the health care domain; such restrictions may preclude data from being overly-specific in an attempt to prevent identification of any particular individual. Additionally, surveys over populations may only include the region from which the response was taken rather than a single point. Because of this, the analysis of scan statistics over region-based data may produce beneficial results.

This project maps regions to points, allowing compatibility between scan statistics and region-based data sets. Many techniques were explored but the focus was directed towards point-to-region mappings. Initial analysis showed that such an application is not only possible, but can have a significant effect on the accuracy of scan statistics.

A limited amount of research has been done to analyze the effects of region based data sets on scan statistics [2]. The majority of these methods attempt to apply scan statistics to the regions themselves, rather than an approximation or simplification. Such approaches may be limited to the use of regions from which the data was sampled.

Contrary to these approaches, this thesis focused on the approximation of regions by randomly sampling points within each region. By doing this, the point-based scan statistics can be applied directly to a region based data set. This gives the user a simple interface to control a trade-off between precision and speed.

A large portion of the mapping techniques were applied using a two step approach: mapping the region to a point, and mapping the point back to a region. This allows point-based data sets to be directly transformed into regions through application of the second step. This potentially adds to the flexibility of such an approach.

2 Background

This section provides an elementary explanation of a spatial scan statistic used in the experiments. The scan statistic model was first proposed by Joseph Naus [3] in 1965 and an extension of this idea focused on spatial regions was developed by Martin Kulldorff in 1997 [4]. This method provided a strong framework for discovering anomalous regions in geographically located data. An extension of the two-dimensional version of this statistic provided the foundation for techniques used in this project. In particular, the application of this statistic over rectangular areas was considered.

The spatial scan statistic method can be summarized by the following steps:

1. Create a data model
2. Choose a measure function ϕ to evaluate the likelihood that a chosen area is anomalous
3. Scan the data set for an area which approximately maximizes ϕ
4. Determine whether the discovered area is significant through additional analysis

The approximation function ϕ used in this project was the Kulldorff Scan Statistic [4]. This statistic works over a point-based data set by assigning measured and baseline values to each data point. The function can be used to determine the degree of anomalousness found in a particular area. Each rectangular area considered by the scan statistic is drawn from a set of four data points in the original set, each defining a side of the rectangle.

The data model $\mathbb{X}$ in this method is a set of two dimensional points and their associated measured and baseline values. Each data point $x \in \mathbb{X}$ can be summarized as follows:

- A two-dimensional point $(x_1, x_2) \in \mathbb{R}^2$
- A measured value $m(x)$
- A baseline value $b(x)$

A sampling approach over Kulldorff's statistic was developed by Matheny et. al. [5] where only a subset of rectangles based on sampled points are considered. By not considering every rectangular region, this method greatly reduced the required runtime. The number of points considered for the rectangular areas is determined by a function of the desired error limit ϵ as $n = O\left(\frac{1}{\epsilon}\ log(\frac{1}{\epsilon})\right)$. A larger subset of measured and baseline points are then considered. To estimate the data values accumulated with each rectangle, these sets are given an estimated size $s = O\left(\frac{1}{\epsilon^2}\right)$. A point x_d is sampled into the measured set $\mathcal{M}$ and baseline set $\mathcal{B}$ with the probabilities shown in Equations 1 and 2.

$$P(x_d \in \mathcal{M}) = \frac{m(x_d)}{\sum_{i=1}^{n} m(x_i) \cdot s} \tag{1}$$

$$P(x_d \in \mathcal{B}) = \frac{1}{n} \cdot s \tag{2}$$

Each considered rectangle $\mathcal{C}$ has two values associated with it: $m(\mathcal{C})$ which is the faction of measured points within the rectangle, and $b(\mathcal{C})$, the fraction of baseline points within the rectangle. The Kulldorff Scan Statistic over these parameters is shown in Equation 3.

$$\phi_{\mathbb{X}}(\mathcal{C}) = m(\mathcal{C}) \ln \frac{m(\mathcal{C})}{b(\mathcal{C})} + (1 - m(\mathcal{C})) \ \ln \frac{1 - m(\mathcal{C})}{1 - b(\mathcal{C})} \tag{3}$$

This statistic provides a measure of how anomalous an area is by considering the number of measured and baseline points found within. If an area contains a high number of measured points, but few baseline points, the left side of the equation will increase. If there are more baseline points in a considered area, the right side will increase. If the number of measured and baseline points are equal, the log factors on both sides will be zero, giving the entire region a score of zero. Figure 1 provides a graphical representation of this idea. A simple point set example demonstrating this idea is shown in Figure 2, where the red points belong to the measured set and blue to the baseline set.

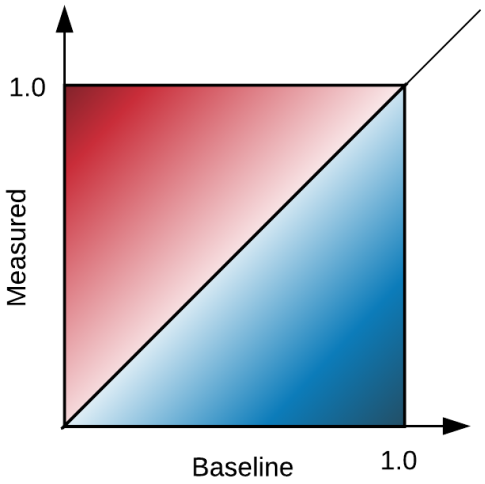

Figure 1: Graphical representation of values over the Kulldorff Scan Statistic

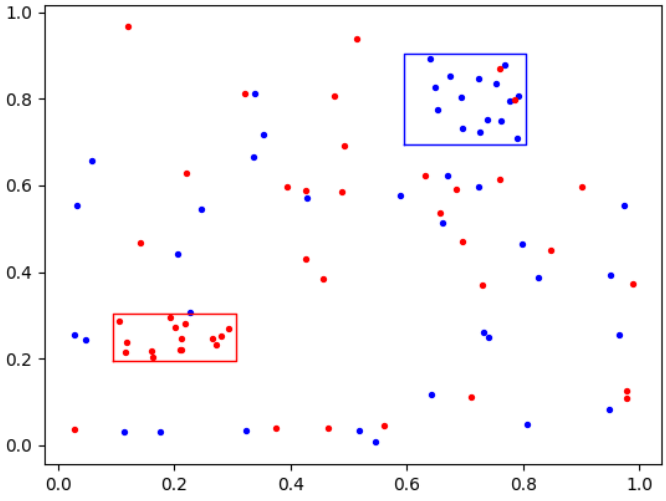

Figure 2: Collection of measured (red), and baseline (blue) points, with a sample of anomalous rectangles

Due to the computational complexity of generating rectangles over the full point set, a sampling approach was necessary for this project. Thus, the Matheny approach was used as a black-box algorithm in each experiment. Some error was introduced as a result of this, but trends in the results remained well-defined.

3 Data Sets

The creation of multiple region-based data sets was a major focus of this project. The type of data focused mainly on U.S. counties but also included zip-code and weather regions. Many of these sets contained extraneous data for the spatial scan statistic and needed to be curated appropriately. All data sets were associated with their specific region and placed in a MySQL database to facilitate mappings between regions and data points. A summary of these data sets and their region counts is shown in Table 1.

Table 1: Considered Data Sets

Data Set	**Region Count**	**Provider**
New York and New Mexico cancer incidence by zip code	1,693	SaTScan [6]
Nationwide cancer incidence by county	3,140	CDC [7]
Diabetes, obesity and inactivity by county	3,140	CDC [7]
Educational attainment and poverty by county	3,140	USDA [8]
Adverse weather phenomena	N/A	NOAA [9]

A choropleth map was generated for each data set created in this project. These maps represent the deviation from the mean of each region for a particular feature. An example of this map is shown for poverty rates over counties in Figure 3. These maps can convey interesting trends on their own and a few will be discussed in the remainder of this section.

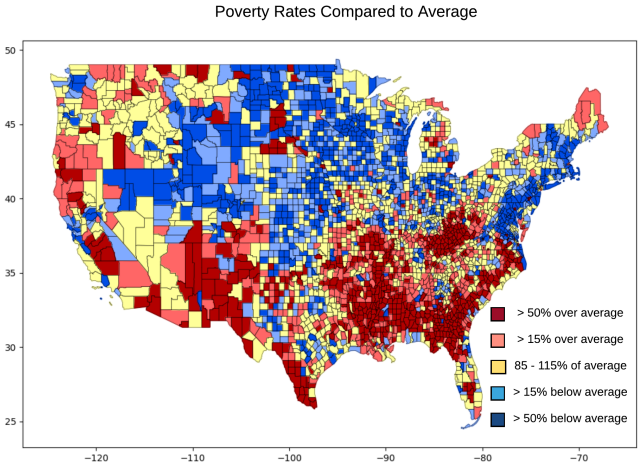

Figure 3: Choropleth map over U.S. poverty rates

The first data set includes the cancer rates of New York and New Mexico zip codes [6]. This data set was used to evaluate the SaTScan approach proposed in the Kulldorff paper [4]. Health

Care data is often protected by privacy laws and this data set provides valuable information over smaller regions.

A cancer data set provided by the Center for Disease Control (CDC) [7] contained total and individual cancer rates for each U.S. county. The choropleth map for lung cancer rates is shown in Figure 4. The gray portions of the map correspond to regions where data was not available; Nevada, Minnesota and Kansas have laws preventing the disclosure of certain heath care statistics so data is not available for these states. One interesting feature of the lung cancer map is the well-defined dark red region. This region has an almost exact overlap with Kentucky, a state known for its heavy tobacco use.

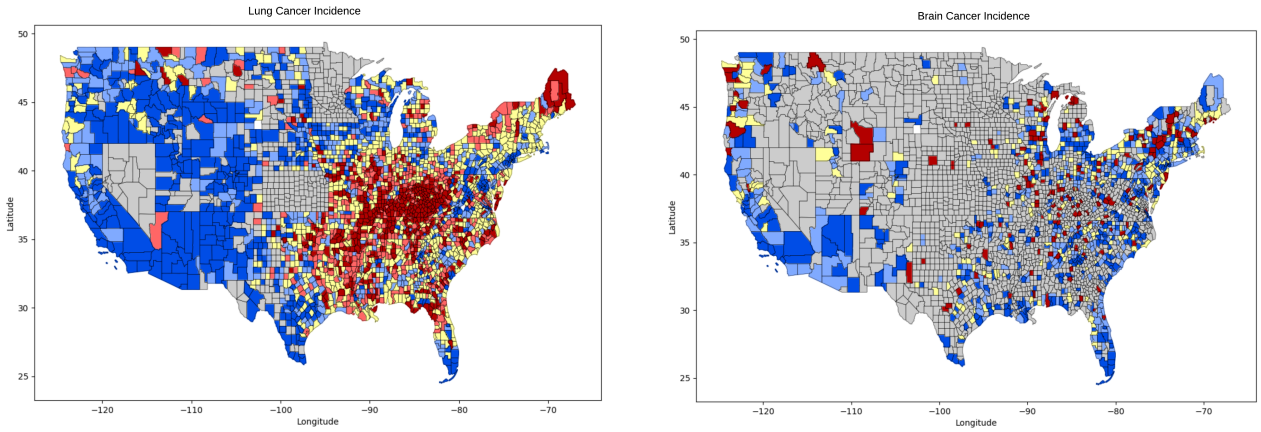

Figure 4: U.S. lung and brain cancer rates by county

Additionally, if a region did not have a statistically significant number of cases, it was not included the data set. This is prevalent for a cancer type such as brain, shown in Figure 4. The rates of this cancer are low enough to cause very few regions to have reliable data.

The CDC also provided rates of diabetes, obesity and inactivity over county regions. This data set was interesting because of the high degree of parallelism between rate categories over regions. The similarity of diabetes and inactivity rates over counties can be inferred from Figure 5.

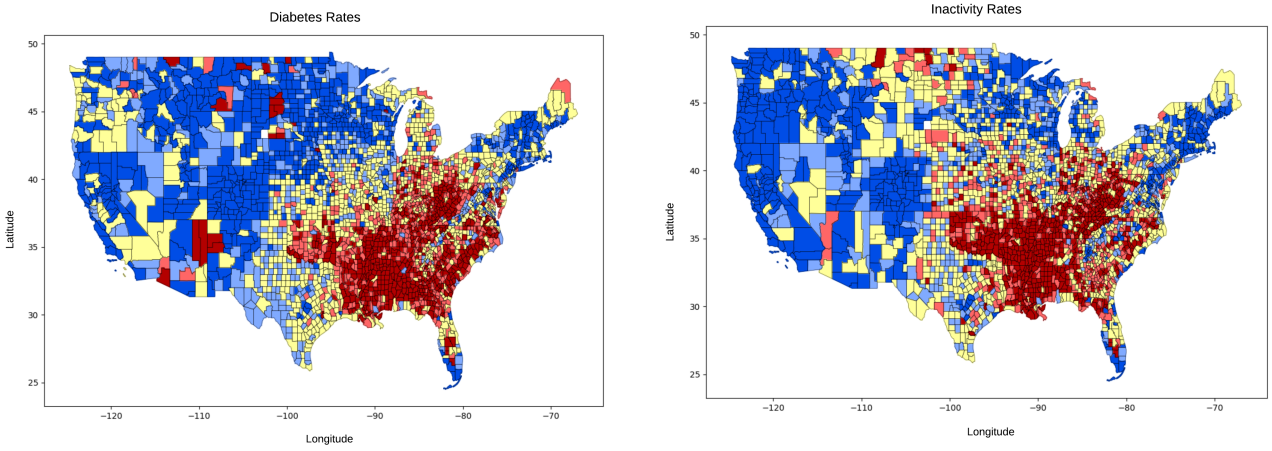

Figure 5: Diabetes and inactivity rates for U.S. counties

The final county-based data set was provided by the United States Department of Agriculture (USDA) [8]. This data set included degrees of educational attainment and percentage of population

living in poverty. The poverty data set was separated into two types: percentage of total population in poverty and minors in poverty. The educational attainment data set was broken into four subsets: completion of high school, obtaining some college education, obtaining an undergraduate degree and obtaining a graduate degree. The educational attainment graphs had noticeable overlap with the poverty and diabetes maps. This may be a result of the effects of poverty; the indigent are more likely to have unhealthy diets and may lack the opportunity pursue higher education.

This project produced an additional data set containing adverse weather phenomena provided by the Northern Oceanic and Atmospheric Administration (NOAA) [9]. This set contained data pertaining to points *and* regions but did not have a clear measured value. As a result, it was not evaluated using the spatial scan statistic.

4 Mapping Techniques

The main focus of this project was the analysis of mapping techniques on a spatial scan statistic. More specifically, a number of region-to-point and point-to-region methods were considered. Each region in a data set represented a U.S. county or zip code polygon provided by the Census Bureau [10]. Additionally, each region was associated with a measured and baseline value as described in the previous section.

4.1 Point Mappings

The region-to-point techniques were quite simple. The first approach computed the **centroid** of each region and assigned the measured value to that point. The centroid is defined over a polygon as the mean in each dimension of every point on the polygon. The **centroid+** approach was also considered, which had the restriction that a computed point must fall within the region. This effect was noticeable on a small number of regions but was not profound enough to have a significant effect on the scan statistic. The centroid mapping approach is shown for Arkansas counties in Figure 6.

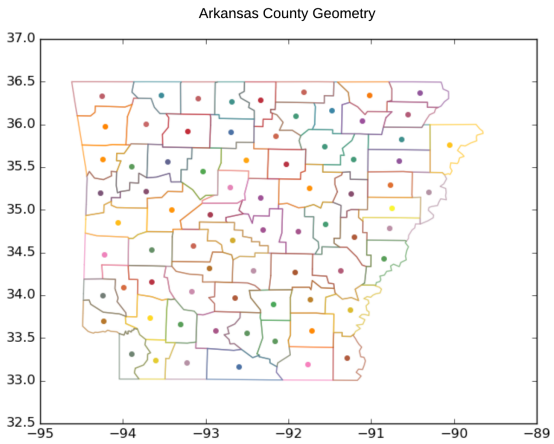

Figure 6: A centroid mapping of Arkansas counties

4.2 Region Mappings

The point-to-region mapping techniques were significantly more involved. These generations relied on an underlying point set, so each county or zip region in the original data was mapped to its centroid point.

The first mapping technique applied was the **Voronoi diagram**. This approach takes a point set and maps each point to a region (called a Voronoi cell) by maximizing the margin between all adjacent points. By doing this, every point in a cell is closer to the generating point than any other point in the original set. An example of this technique is shown for Utah counties in Figure 7.

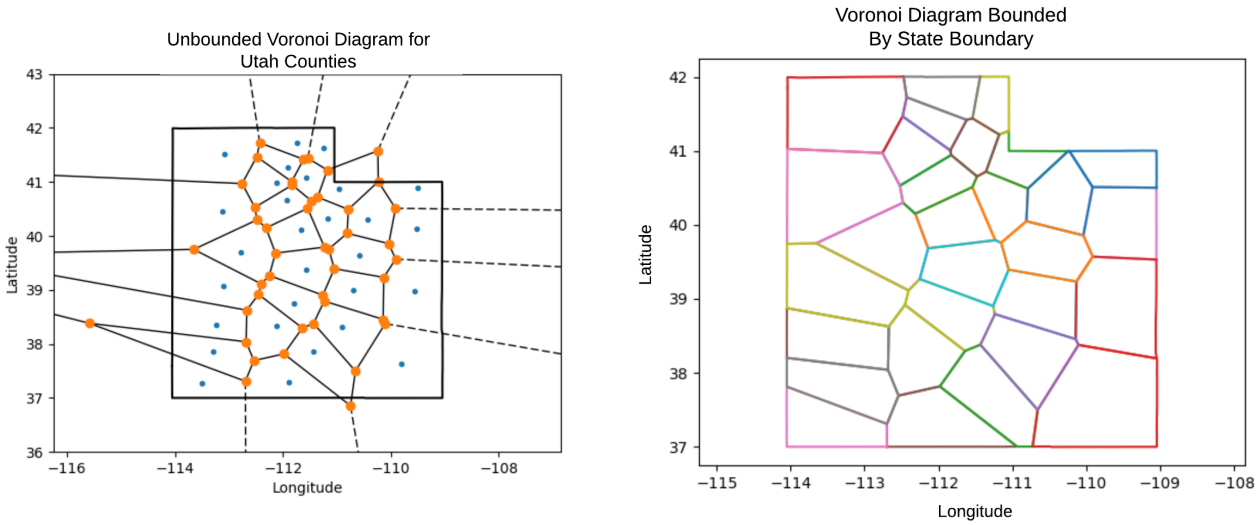

Figure 7: Voronoi diagrams generated from Utah county centroids

It is worth noticing that the boundary points map to cells that extend indefinitely, as indicated by the dotted lines in the left frame of Figure 7. Infinite regions are not desirable and have to be bounded in some way. The first technique involved bounding each cell by the **state boundary** containing the county or zip region. The final result on the Utah data set is shown in the right frame of Figure 7.

This technique produced mostly uniform results across states but still generated complex polygons in states that had rough boundaries. This effect is frequently noticeable for states bordering a river or body of water, as shown with Arkansas in the left frame of Figure 8. Additional regions were generated by bounding the Voronoi cells with the **convex hull** generated by the state polygon. The convex hull is defined as the smallest set of points containing the polygon. This method is again demonstrated using Arkansas counties in the right frame of Figure 8.

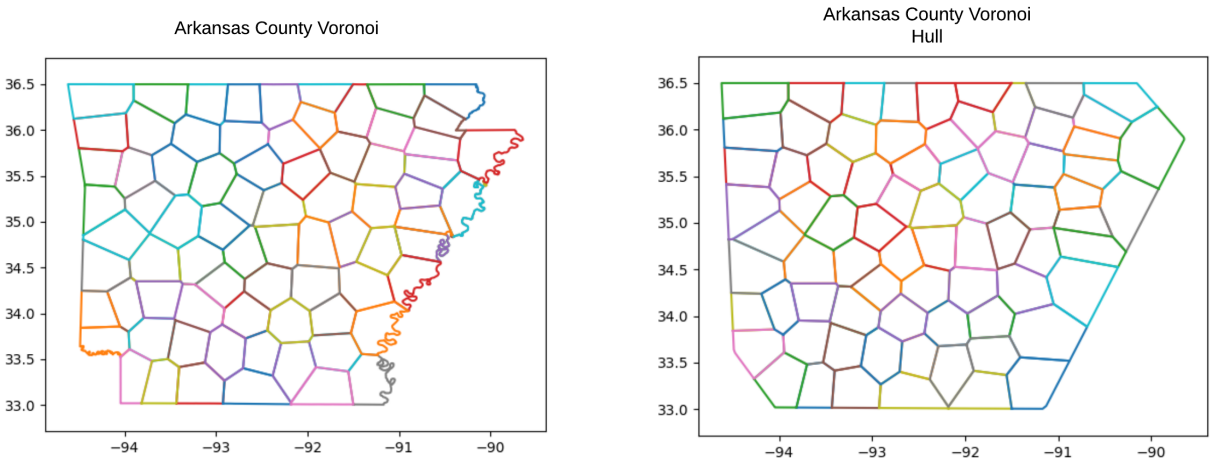

Figure 8: Arkansas Voronoi diagram bounded by state boundary and convex hull

This method was not without its own issues. Some states produced convex hulls with a much larger area than the original polygon. This was noticeable in states with non-rectangular boundaries and is shown for Texas in Figure 9. The increased area produced by this technique had some interesting effects on the experimental results discussed later in this thesis.

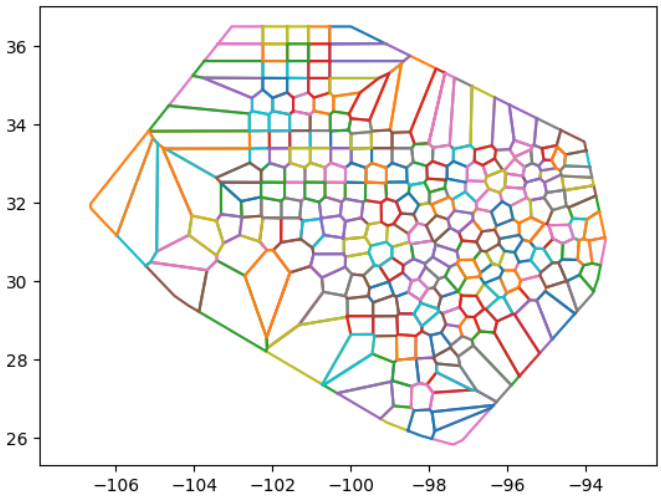

Figure 9: Voronoi diagram over Texas counties bounded by the state convex hull

The last point-to-region mapping technique involved taking the **Delaunay triangulation** over the set of points. This method produces a triangulation in which no ball circumscribed over a triangle contains a point in the original set. A graphical representation of this is shown in the left frame of Figure 10. This triangulation was performed over the set of all U.S. counties and produced the graphic in the right frame of Figure 10.

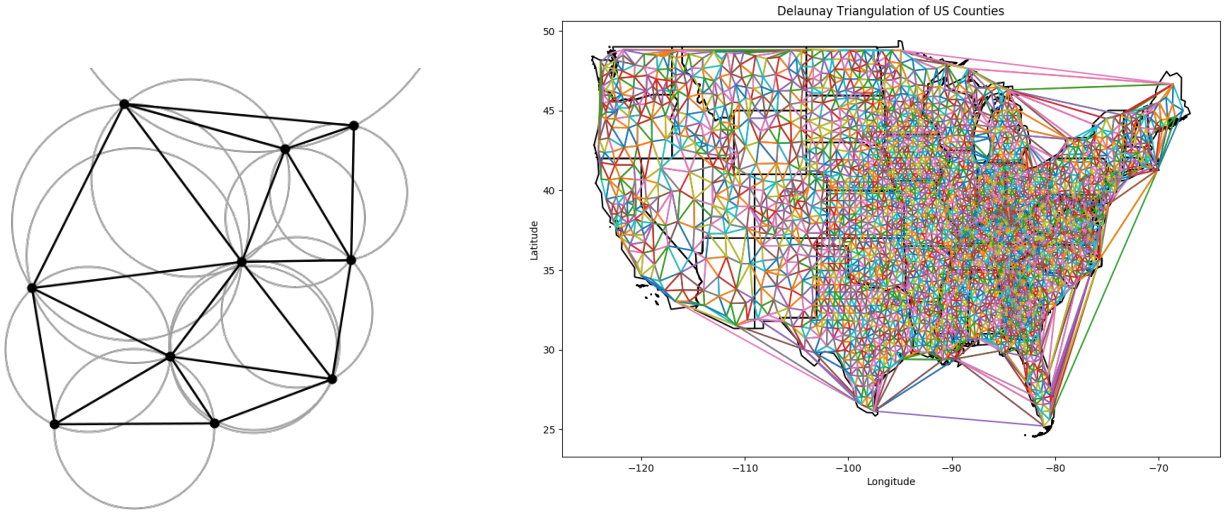

Figure 10: Delaunay triangulation with circumscribed circles and over U.S. Counties

Contrary to the previous techniques, the Delaunay triangulation did not have a one-to-one mapping between county centroids and generated regions. Because of this, the average measured value of each corresponding centroid was assigned to a generated triangle while the baseline values remained constant. In fact, applying this method roughly doubled the number of regions from 3,140 to 6,193, which produced some experimental effects discussed later.

One final region-to-region mapping technique was implemented using the **Douglas-Peucher simplification** algorithm [11]. Each county polygon was simplified using an error parameter between zero and one. The polygon regions were simplified by reducing the number of vertices and edges while attempting to preserve the amount of area covered. The simplified polygons could only deviate from the original by a certain percentage of area, provided as a parameter to the algorithm. An example of this simplification on Arkansas counties using a deviation of 10% is shown in Figure 11. A set of regions were generated and graphed for each state using various error values but were not used in experiments.

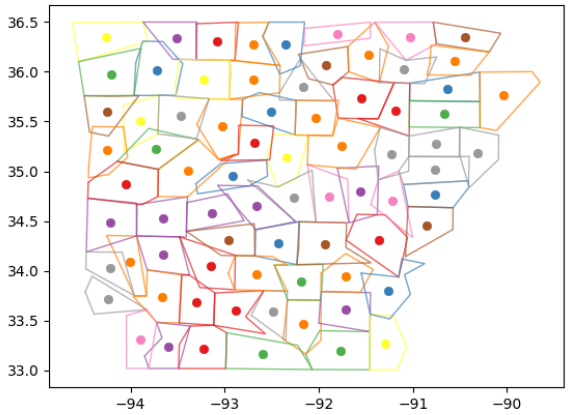

Figure 11: Arkansas county simplification using a deviation of 10%

4.3 Region to Point Mappings

The sample-then-scan spatial scan statistic approach used in this project requires the data set $\mathbb{X}$ to be defined over a set of points rather than regions. This requires one final mapping technique to convert the newly created regions back to point sets. The most common approach in the literature, such as that used by SaTScan [4], is to only consider the centroid. Our proposal is to randomly generate n points within each polygon. An example of this approach using $n = 20$ and 100 is shown for Arkansas in Figure 12. This scheme allows the precision of the region representation to be increased arbitrarily as n increases; the more points that are generated for a region, the more well defined it becomes. A range of values for n were considered in this project and are discussed in more detail in the *Experimental Results* section.

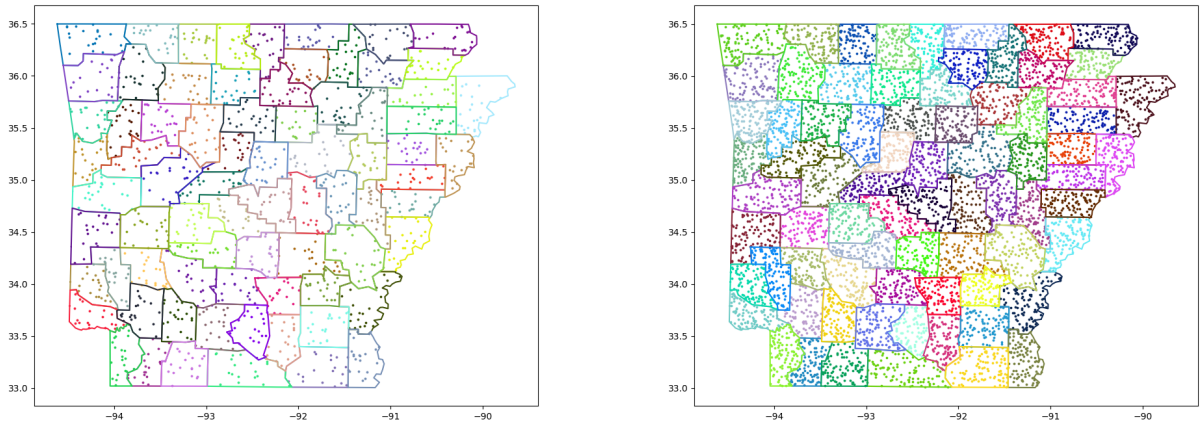

Figure 12: Randomly generated pointsets of size 20 and 100 for Arkansas county boundaries

A special case was necessary for counties that contain multiple polygons, as is common over islands or coastal regions. The approach taken in this project would weight the point count based on the area of each polygon; a polygon containing 10% of the area of the county would receive 10% of the random points. This is shown graphically for the state of Hawaii in Figure 13

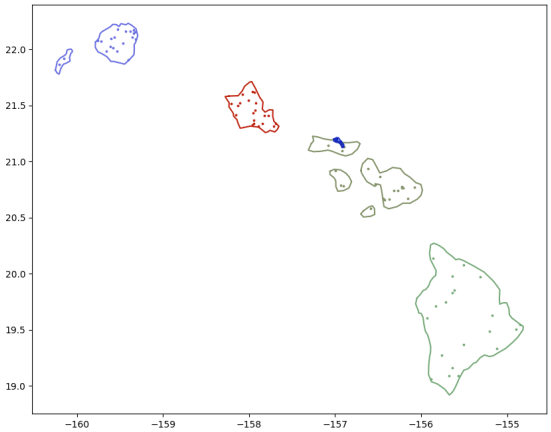

Figure 13: A pointset generated over the non-contiguous Hawaiian counties

5 Statistic Applications

The spatial scan statistic was applied to the data sets created by this project to provide examples of real-world applications. These applications did not have a gold standard, so no quantifiable evaluation was performed. However, some interesting effects of the scan statistic can be observed from the figures. The statistic was applied in multiple iterations over each data set and a select few are discussed here. Each red rectangle corresponds to an anomalous region predicted by the scan statistic over county centroids.

Because the scan statistic used in this project involved sampling there was some non-determinism in the generation of rectangles. Even with this variation it was common to see discovered regions be focused over a particular area. This is shown over the educational attainment data set in the left frame of Figure 14.

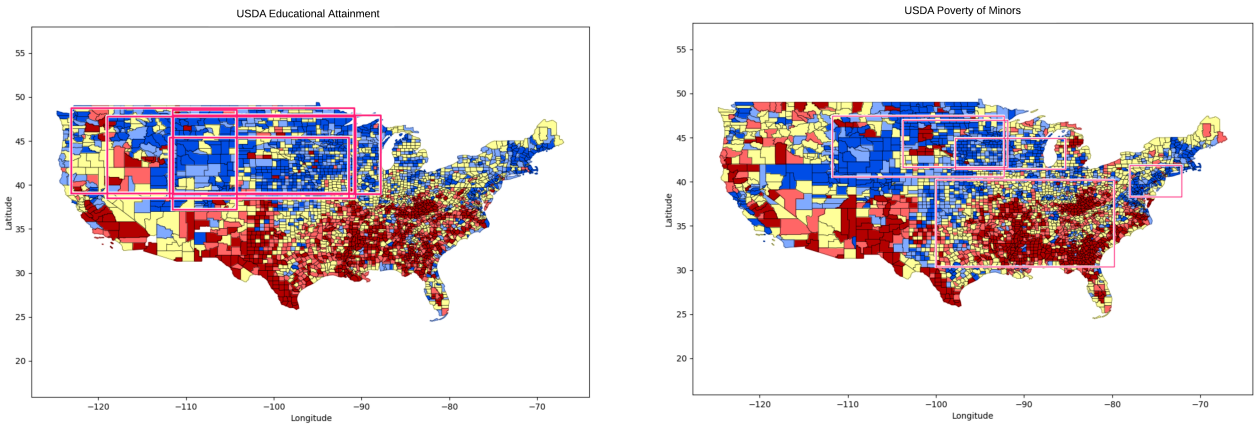

Figure 14: Anomalous regions produced by the spatial scan statistic on educaitonal attainment and poverty of minors

This rectangle grouping effect was not always prevalent. The *Poverty: Minors* data set produced rectangles in regions with higher and lower anomalies, as shown in the right frame of Figure 14. Additionally, many of the New York cancer incidence sets did not have any obvious outlying regions. This set produced anomalous regions with unremarkable significance shown in Figure 15.

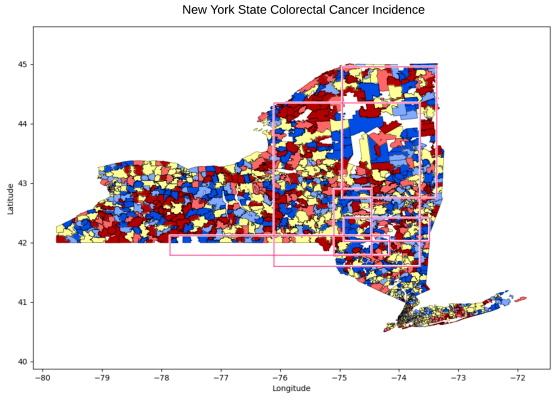

Figure 15: Scan statistic results on colorectal cancer rates in New York State zip codes

The scan statistic was applied to geometries as well as centroids. The statistic produced more focused results as the number of generated points for each region increased. One map demonstrating this is shown in Figure 16, where the red rectangles were generated from county centroids and the green from county geometries with 20 randomly generated points. This effect was also shown in the experimental results described the next section.

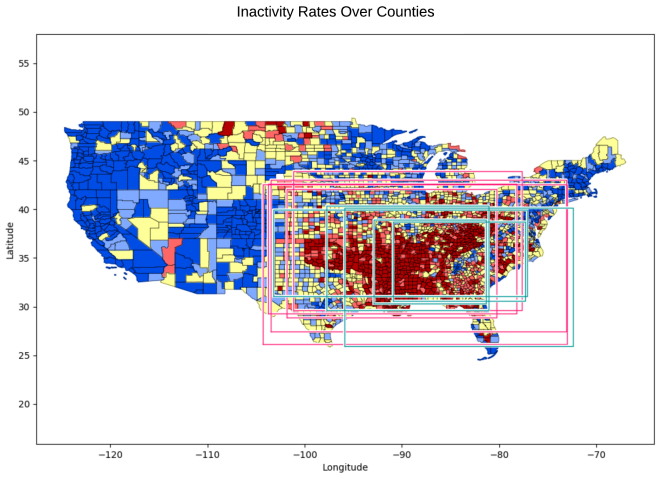

Figure 16: Anomalous regions produced on diabtetes rates using centroid points (red rectangles) and 20 points per county (green rectangles)

6 Experimental Results

The main focus of this thesis was the effects of methods converting data from regions to points, and vice-versa, on spatial scan statistics. The centroid data sets were evaluated using multiple error values. Additionally, each mapping technique was evaluated for multiple point counts generated per region using a fixed error parameter.

6.1 Setup

A synthetic data set was created in order to facilitate the evaluation of the scan statistic. A rectangle was chosen over the continental United States as a target area for the scan statistic. The points falling within the rectangle were chosen into the measured set with one probability and into the baseline set with another. This allowed the degree of anomaly in the target rectangle to be chosen in a deterministic way.

Each point x_i inside the target rectangle was sampled into the measured set M with probability p. Each point x_o outside the rectangle was sampled into the measured set with probability q. All points were sampled into the baseline set with an equal probability. This project used a standard probability of 0.2 from which the values of p and q diverged. The probability of 0.2 was also used to sample points into the baseline set:

- $P(x_i \in M) = p$
- $P(x_o \in M) = q$
- $P(x_i, x_o \in B) = 0.2$

The scan statistic's performance depends on the difference of p and q. Essentially, a higher p and lower q value causes the target rectangle to become more anomalous. The accuracy of the scan statistic increases as these two values diverge. If the values of p and q are equal, the points inside and outside the target rectangle are sampled into the measured set with equal probability. In this case, there is no anomalous region and any overlap between the target rectangle and one returned by the scan statistic is coincidental. This instance is shown in Figure 17 with the target rectangle in blue and five rectangles returned from the scan statistic in red.

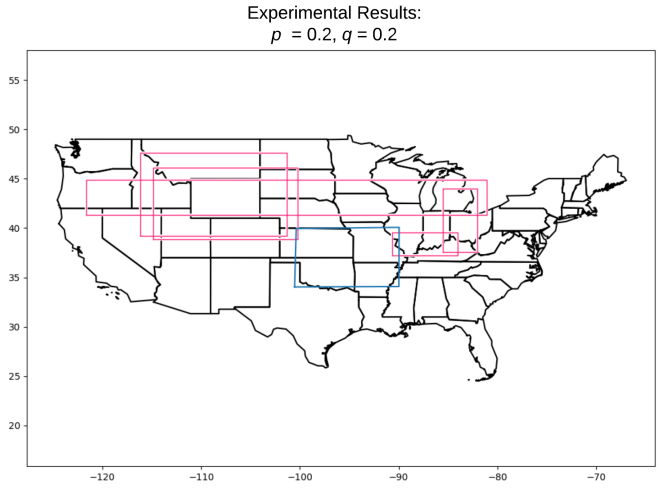

Figure 17: Scan statistic results (red) with a non-anomalous target region (blue)

As the values of p and q diverge, rectangles returned by the scan statistic become more accurate and have more overlap with the target. This is shown in Figure 18 for $p = 0.25, q = 0.15$ and $p = 0.3, q = 0.1$

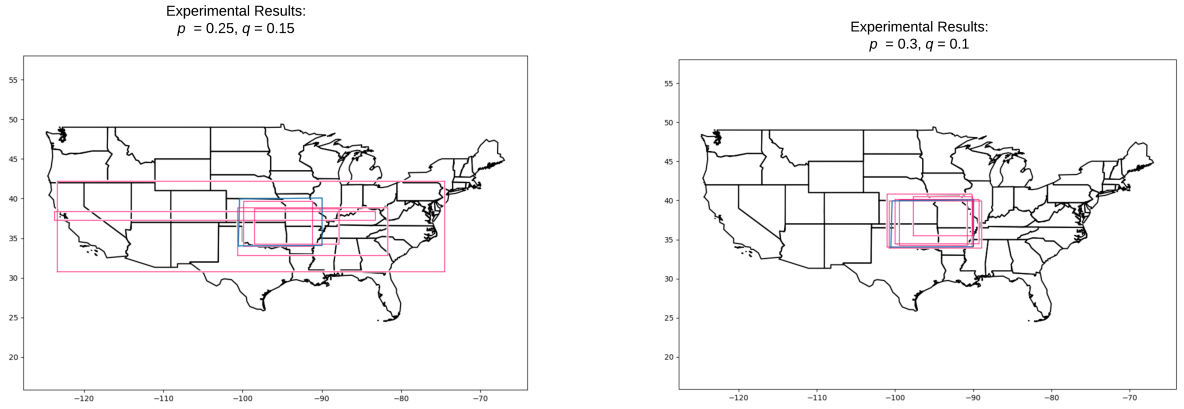

Figure 18: Predicted rectangles as the target region becomes more anomalous

As demonstrated in the previous examples, rectangles containing the same points as the target should be considered more desirable. With this assumption we can choose an accuracy metric to evaluate the scan statistic. This project chose the Jaccard distance, a common metric for determining the difference between two sets. This distance is defined over two sets A and B in Equation 4.

$$d_J(A, B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \tag{4}$$

6.2 Results

The following section contains experimental results on the spatial scan statistic using the previously described framework. These evaluations were made using multiple error values, region mapping techniques, and point counts. Each experimental result was produced by taking the average Jaccard distance over 50 trials.

As a proof of concept, the statistic was evaluated against the target region on multiple values of p and q. This experiment used the county centroids as the underlying point set. A centroid was sampled into the measured set with probability p if it fell within the rectangle and q if it did not. Figure 19 shows the Jaccard distance decreasing as the two values diverge.

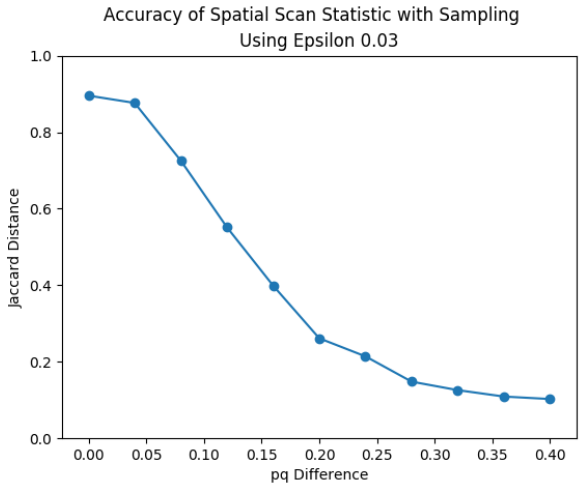

Figure 19: Scan statistic performance on centroids with an increasing anomaly

The previous graph has a well-defined downward trend but never achieves a Jaccard distance of zero. This is due to the sampling technique used in the statistic. Additionally, the graph shows that values of $p = 0.1$ and $q = 0.3$ produce a significant anomaly within the target region. These values were used in the following experiments to evaluate the error parameter and mapping effects.

The error parameter ϵ of the scan statistic was evaluated against the target rectangle. As the value of ϵ decreases, the statistic considers more rectangles and thus has a lower Jaccard distance. These results are shown in Figure 20. Error values below 0.04 were not considered because computational complexity grows polynomially as this value decreases.

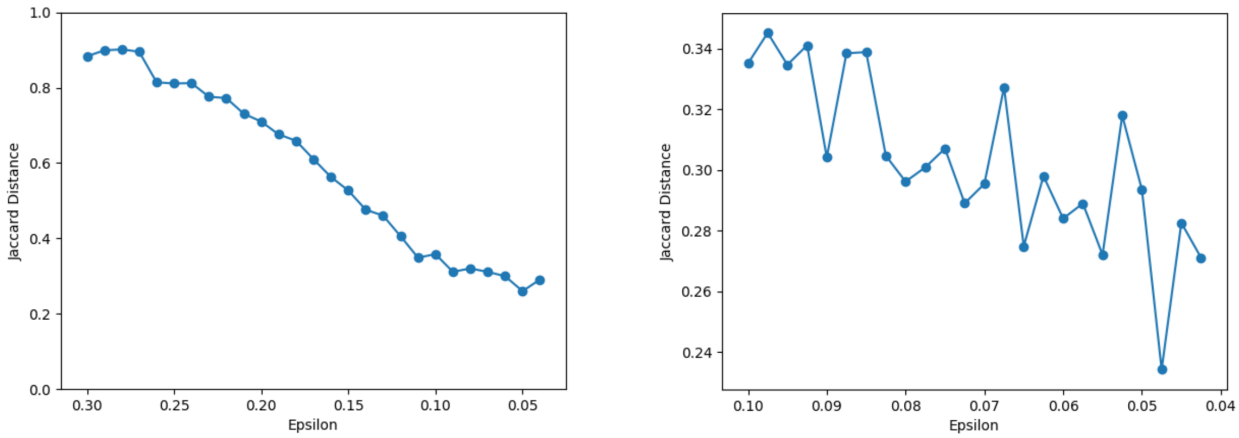

Figure 20: Effects of the error parameter ϵ on the scan statistic

An evaluation was also performed on a higher resolution of error values. This experiment showed that as the error bound decreases the noisy effects of random sampling become more prominent. This result is shown in the right frame of Figure 20.

The remaining experimental results of this project evaluated effects of the mapping techniques described in the *Data Sets* section. The experiments were run over 50 trials, with $p = 0.3$, $q = 0.1$, and an error bound $\epsilon = 0.04$. These techniques involved generating multiple points per region, and special consideration must be made for regions on the boundary of the target rectangle. The percentage C_R of overlapping area between each region R and the target rectangle was computed. Each point generated within a region was given a p and q value proportional to this percentage: $P(x_R \in M) = C_R \cdot p + (1 - C_R) \cdot q$. This allowed regions found on the border of the target rectangle to be given appropriate weight in the measured set. Similarly, the Delaunay triangulation regions were given a weight corresponding to the fraction of vertices falling within the target rectangle.

Each point-to-shape mapping produced in this project was evaluated over a range of point counts. For each mapping, a certain number of points were randomly generated within the region. As the number of points increases, the region becomes more well-defined and the scan statistic has a higher probability of finding the target rectangle. This effect is shown in the left frame of Figure 21 for point counts 5, 20, 100, and 500. The tested regions included the original county geometry, the corresponding Voronoi cell bounded by the state boundary, and the cell bounded by the state convex hull.

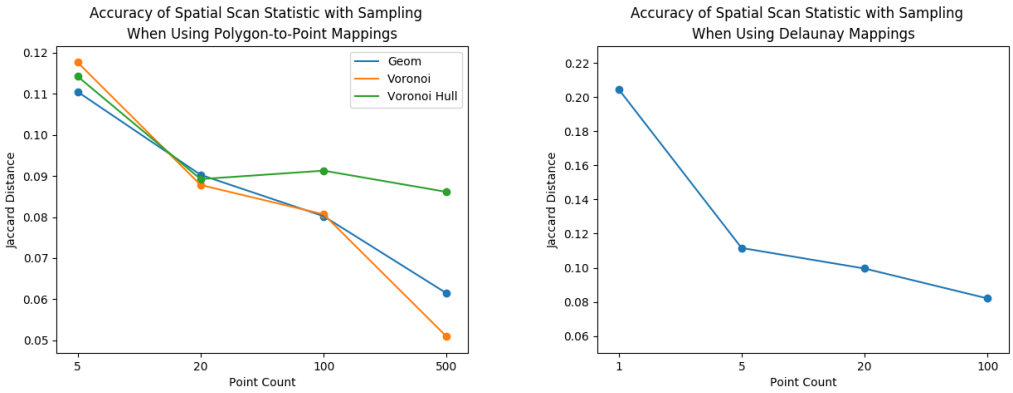

Figure 21: Accuracy of point counts per region type

The Delaunay complex generation approximately doubled the amount of regions being considered. Because of this, these regions were considered separately with point counts 1, 5, 20, and 100. The right frame of Figure 21 shows this result.

These experiments confirm the expectation that increased point counts for each region will increase the resolution of areas produced by the scan statistic. Because of this, the Jaccard distance to the target rectangle decreased as a larger number of points were considered. It is worth noting that the Voronoi regions bounded by the convex hull do not follow a similar trend as the other two mappings. This effect was likely caused by the increased area of the bordering regions, as described in the Texas example of the *Mapping Techniques* section.

The final experiment involved the performance of each mapping technique on diverging values of p and q. Through this method it was possible to evaluate each technique as the target region becomes more anomalous. Each mapping technique can be compared to the original centroid result produced at the beginning of this section.

A comparison of county geometry, state bounded Voronoi cells, and convex hull bounded Voronoi cells was done using point counts of 5 and 20. The experiments showed that each mapping technique performed similarly, with a larger effect being attributed to point counts. A graph of the results is shown in Figure 22. Interestingly, by increasing the number of points per region, the resulting scan statistic is much more accurate. Thus, we recommend this approach over only considering the centroid points.

The Delaunay triangulation method was also evaluated in the same fashion. Again, these experiments showed an increased performance with higher point counts. The right frame of Figure 22 shows this method compared to the centroid baseline.

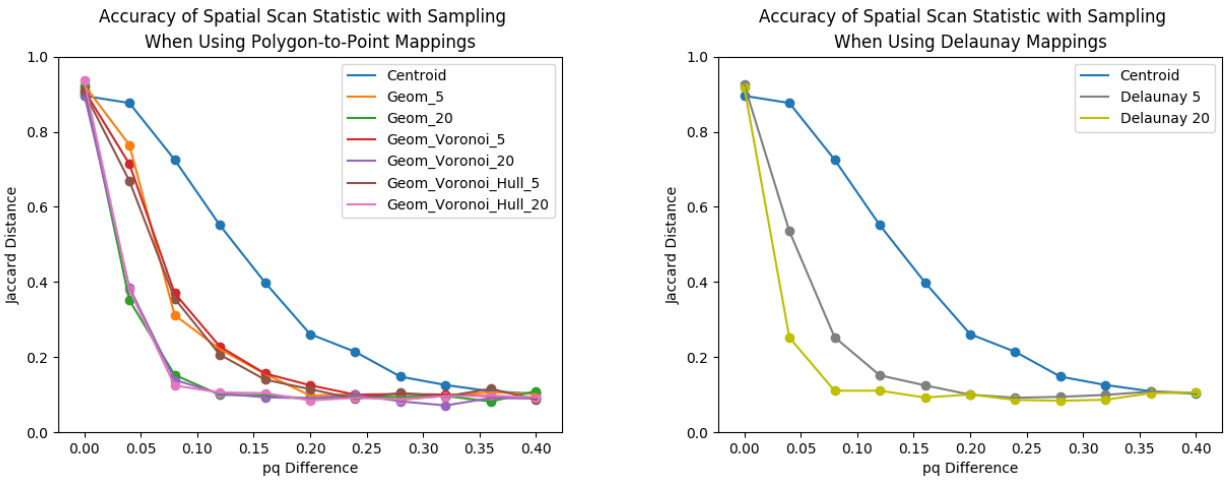

Figure 22: Accuracy of county mapping and delaunay triangulation techniques as the anomalous region diverges

As a final comparison, each mapping technique is plotted in Figure 23. This graph shows that the Delaunay triangulation (in gray and yellow) actually performs *better* than the other methods in their respective point count categories. This is most likely caused by the Delaunay triangulation doubling the amount of regions considered, again showing that increased resolution improves scan statistic performance.

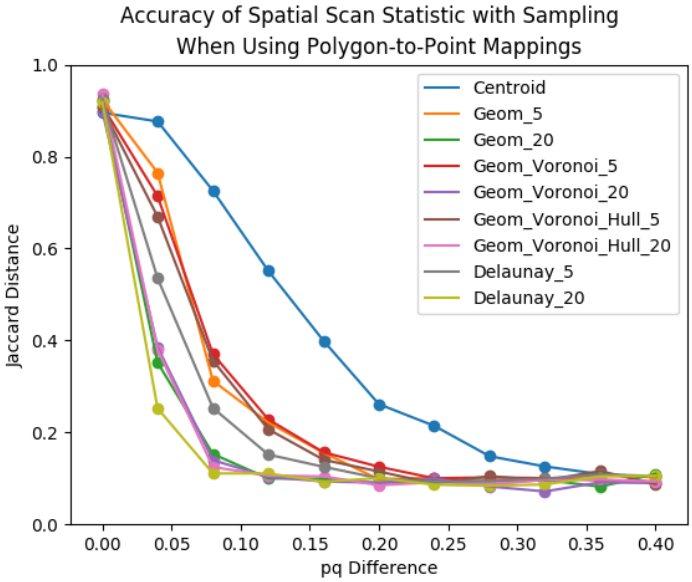

Figure 23: Comparison of all mapping techniques

7 Future Work

The methods and experiments developed in this thesis provide a new analysis on a sampling-based scan statistic. As a result of this, many concepts explored in this project could be extended or modified to produce novel insights. A few possibilities are mentioned in the following section.

Conclusions produced in the experimental section show the importance of resolution when dealing with spatially organized data. Continued evaluation of mapping techniques on additional data sets could provide interesting results, especially as considered regions shrink in size.

Many of the experiments in this project were limited by the time complexity of current scan statistic algorithms. New research aimed at reducing this computational cost could support applications on higher point counts generated per region. This would also allow larger point-based data sets, such as geo-located tweets, to be transformed into regions and experimented with.

Another extension of this work could consider new mapping techniques including largest enclosed and minimum circumscribed balls over polygons. Additionally, methods that shatter a region into multiple subsets based on population distributions may have significant effects on health and socio-economic data sets.

The scan statistic algorithm used in this project only generated rectangular areas, but this assumption is not necessary. Kulldorff's SaTScan[6] and the approach developed by Matheny [5] also consider scans involving disks and ellipsoids. An interesting comparison between rectangular and disk-type areas over region data sets could easily be produced.

This project also limited its scope to algorithms based on point-set data. Some scan statistic techniques can be applied over regions [1], and would not require the mapping techniques described here. Additional experiments could provide insight into the relationship between these two approaches.

8 Conclusion

This project produced a broad spectrum of results related to the problem of anomaly detection in spatially oriented data. Interesting conclusions can be drawn from the simplest analysis on the gathered data sets and observations only deepen as more complex approaches are considered. This project showed that point-to-region mapping techniques have a significant effect on scan statistics and further exploration may provide unseen analytical depth in this field.

Applications in this thesis also provided interesting insight into the proliferation of common diseases facing communities today. As the complexity of relevant algorithms are reduced, the precision and effectiveness of scan statistics will continue to improve. Such methods provide a valuable comprehension of socio-economic data sets, which continue to grow in complexity and size. This knowledge has the potential to positively influence health-care fields and improve societal wellness across nations.

References

[1] T. H. Grubesic, R. Wei, and A. T. Murray, "Spatial clustering overview and comparison: Accuracy, sensitivity, and computational expense," *Annals of the Association of American Geographers*, vol. 104, no. 6, pp. 1134–1156, 2014.

[2] D. Agarwal, A.McGregor, J. M. Phillips, S. Venkatasubramanian, and Z. Zhu., "Spatial scan statistics: Approximations and performance study," in *KDD*, 2006.

[3] J. I. Naus, "Clustering of random points in two dimensions," *Biometrika*, vol. 52, no. 1/2, pp. 263–267, 1965.

[4] M. Kulldorff, "A spatial scan statistic," *Communications in Statistics: Theory and Methods*, vol. 26, 1481–1496, 1997.

[5] M. Matheny, R. Singh, L. Zhang, K. Wang, and J. M. Phillips, "Scalable spatial scan statistics through sampling," in *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ACM, 2016, p. 20.

[6] *New york state cancer data*, 2005-2009. [Online]. Available: `https://www.satscan.org/datasets/nyscancer/index.html`.

[7] *Diabetes county data indicators*, 2015. [Online]. Available: `https://www.cdc.gov/diabetes/data/countydata/countydataindicators.html`.

[8] *County-level data sets*, 2015. [Online]. Available: `https://www.ers.usda.gov/data-products/county-level-data-sets/`.

[9] *Storm events database*, 2015. [Online]. Available: `ftp.ncdc.noaa.gov/pub/data/swdi/stormevents`.

[10] *Cartographic boundary shapefiles*, 2016. [Online]. Available: `https://www.census.gov/geo/maps-data/data/cbf/cbf_counties.html`.

[11] D Douglas and T Peucker, "Algorithm for the reduction of the number of points required to represent a line or its character," *Am Cartogr*, vol. 10, no. 42, pp. 112–123, 1973.