EFFECTS OF NON-HOMOGENEOUS POPULATION DISTRIBUTION ON SMOOTHED MAPS
PRODUCED USING KERNEL DENSITY ESTIMATION METHODS

Jesse Jack Jones

Thesis Prepared for the Degree of
MASTER OF SCIENCE

UNIVERSITY OF NORTH TEXAS

December 2014

APPROVED:

Chetan Tiwari, Major Professor
Pinliang Dong, Committee Member
Joseph Oppong, Committee Member
Paul Hudak, Chair of the Department of Geography
Mark Wardell, Dean of the Toulouse Graduate School

Understanding spatial perspectives on the spread and incidence of a disease is invaluable for public health planning and intervention. Choropleth maps are commonly used to provide an abstraction of disease risk across geographic space. These maps are derived from aggregated population counts that are known to be affected by the small numbers problem. Kernel density estimation methods account for this problem by producing risk estimates that are based on aggregations of approximately equal population sizes. However, the process of aggregation often combines data from areas with non-uniform spatial and population characteristics. This thesis presents a new method to aggregate space in ways that are sensitive to their underlying risk factors. Such maps will enable better public health practice and intervention by enhancing our ability to understand the spatial processes that result in disparate health outcomes.
TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>v</td>
</tr>
<tr>
<td>CHAPTER 1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>CHAPTER 2 LITERATURE REVIEW</td>
<td>5</td>
</tr>
<tr>
<td>CHAPTER 3 DATA AND METHODS</td>
<td>14</td>
</tr>
<tr>
<td>Tools</td>
<td>15</td>
</tr>
<tr>
<td>Synthetic Data</td>
<td>15</td>
</tr>
<tr>
<td>Fishnet</td>
<td>18</td>
</tr>
<tr>
<td>Cluster Analysis</td>
<td>20</td>
</tr>
<tr>
<td>Kernel Density Estimation</td>
<td>23</td>
</tr>
<tr>
<td>CHAPTER 4 RESULTS AND DISCUSSIONS</td>
<td>29</td>
</tr>
<tr>
<td>CHAPTER 5 CONCLUSIONS</td>
<td>42</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>44</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1:</td>
<td>Cluster Intensity Matrix</td>
<td>21</td>
</tr>
<tr>
<td>Table 2:</td>
<td>Cluster Intensity Matrix for Dataset 1</td>
<td>30</td>
</tr>
<tr>
<td>Table 3:</td>
<td>Cluster Intensity Matrix for Dataset 1, Standard KDE vs DAAM</td>
<td>33</td>
</tr>
</tbody>
</table>
**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>Controls file</td>
<td>17</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Fishnet overlaid with cases</td>
<td>20</td>
</tr>
<tr>
<td>Figure 3</td>
<td>ArcGIS model builder for nearest neighbor index</td>
<td>22</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Fishnet divided into 4 partitions with cases overlaid</td>
<td>26</td>
</tr>
<tr>
<td>Figure 5</td>
<td>KDE grids for standard KDE (red) and DAAM (blue)</td>
<td>28</td>
</tr>
<tr>
<td>Figure 6</td>
<td>Map of synthetic dataset 1</td>
<td>29</td>
</tr>
<tr>
<td>Figure 7</td>
<td>Standard kernel density estimation for synthetic dataset 1</td>
<td>31</td>
</tr>
<tr>
<td>Figure 8</td>
<td>Disassociated aggregation and analysis method for synthetic dataset 1</td>
<td>31</td>
</tr>
<tr>
<td>Figure 9</td>
<td>Difference map for synthetic dataset 1</td>
<td>32</td>
</tr>
<tr>
<td>Figure 10</td>
<td>Standard kernel density estimation for synthetic dataset 2</td>
<td>35</td>
</tr>
<tr>
<td>Figure 11</td>
<td>Disassociated aggregation and analysis method for synthetic dataset 2</td>
<td>35</td>
</tr>
<tr>
<td>Figure 12</td>
<td>Difference map for synthetic dataset 2</td>
<td>36</td>
</tr>
<tr>
<td>Figure 13</td>
<td>Standard kernel density estimation for synthetic dataset 3</td>
<td>37</td>
</tr>
<tr>
<td>Figure 14</td>
<td>Disassociated aggregation and analysis method for synthetic dataset 3</td>
<td>37</td>
</tr>
<tr>
<td>Figure 15</td>
<td>Difference map for synthetic dataset 3</td>
<td>38</td>
</tr>
<tr>
<td>Figure 16</td>
<td>Standard kernel density estimation for synthetic dataset 4</td>
<td>39</td>
</tr>
<tr>
<td>Figure 17</td>
<td>Disassociated aggregation and analysis method for synthetic dataset 4</td>
<td>39</td>
</tr>
<tr>
<td>Figure 18</td>
<td>Difference map for synthetic dataset 4</td>
<td>40</td>
</tr>
</tbody>
</table>
CHAPTER 1
INTRODUCTION

Maps—and the creation there-of—are a principle concern of the geographer, from the navigators of the ancient world through John Snow's mapping of the 1854 London cholera outbreak to the modern world of Geographic Information Systems (GIS). Over the centuries, the discipline has produced numerous mapping methods that present a variety of quantitative and qualitative information in increasingly useful fashion, as well as techniques to map the results of complex spatial analyses and statistical models. Commonly used map types include: choropleth, isopleth, dasymetric, isarithmic, proportional symbol/dot, and more (Lawson and Williams, 2001). All of these map types are typically constructed by summarizing disease outcomes as a function of the underlying at-risk population. As population data is typically collected by the Census, the unit of analysis on a map is restricted to small area Census units such as zip codes, Census tracts, or, in some cases, counties. Data aggregation is popular, because it addresses privacy concerns while helping to improve reliability in the face of limited numbers of observations (Bithell, 1990; Rushton and Lolonis, 1994; Talbot et al. 2000; Tiwari and Rushton, 2009; Beyer et al. 2012). However, it also requires decisions about how the underlying data—be it point data or small areas—is to be aggregated.

Unfortunately, maps produced using aggregated data—especially in areas with sparse population densities—are affected by the so-called small numbers problem (Rushton and Lolonis, 1994; Rushton, 2003), where-in any rates calculated using only a small number of cases are subject to the specific variances of those cases. This produces rates with a low degree of confidence that can exhibit local characteristics that would be eliminated by a larger sample
size. For instance, if you took a sample size of only 2 individuals and one happened to be Caucasian while the other was African-American, you would end up with data telling you that 50% of all individuals with the condition were African-American. But if you were to take a sample size of 100 individuals, you might discover that there were only a total of 4 African-Americans with the condition, dropping your rate down to 4%. With the small numbers problem, chance and coincidence play such a large part in the possible results that they cannot be used to draw any reliable, statistically-valid results.

Additionally, the use of areal units on a map introduces the well-known modifiable areal unit problem (MAUP) (Rushton and Lolonis, 1994; Rushton, 2003) that can be introduced when deciding on the scale of the final aggregation units. In this case, rather than being a problem inherent with the dataset being used, the MAUP is a bias that can be inadvertently introduced by a researcher's decision about how to analyze their data. If a researcher is looking at the rates at which a given disease occurs, they may decide to choose the 'county' as their level of analysis and discover that one county has a rate of 20%. However, if that researcher were to take the same data and look at it on the level of 'zip code' instead, they could discover that the eastern half of that same county has zip codes where the disease rate is only 5% while the western half of the county has zip codes where the disease is 30 to 40%, indicating that the problem is both more severe and more localized than the county-level analysis indicated. When dealing with MAUP, it is always a battle between how fine the data's granularity is and the resources you have available to conduct your analysis.

Smoothing techniques are commonly used as a means of addressing these problems, particularly the small numbers problem (Bithell 1990; Talbot et al. 2000). The purpose of
smoothing is to capture spatial patterns present in the underlying data while filtering out statistical noise caused by small population counts. Smoothing typically addresses the problem of small population counts by borrowing information or strength from surrounding areas.

Smoothing can be done in one of two ways: (1) rates computed for local areas are averaged using a local moving window, or (2) populations are aggregated using a local moving window before computing a rate. The first approach is based on the premise that a single, isolated observation of high rates is likely caused by data artifacts and is not truly representative of clusters. Techniques within this approach include geostatistical methods like kriging or statistical inference methods such as empirical Bayes (Goovaerts, 1997; Mungiole et al. 1999). The second approach relies on methods such as kernel density estimation, which are divided into two basic types: fixed and adaptive filters. Fixed filter methods aggregate populations using a fixed-size circle characterized by some predefined density function such as uniform or Gaussian (Rushton and Lolonis, 1996). Adaptive filters aggregate populations based on some minimum population threshold and use circles of varying sizes that grow or shrink depending on the underlying population density (Talbot et al., 2000; Tiwari and Rushton, 2005). While fixed-filters account for problems of rate instability by aggregating populations before any rate is calculated on the map, they can introduce new problems in the form of undersmoothing and oversmoothing the data in question.

Undersmoothing (or overfitting) is a product of noise in the data and can occlude meaningful features in the data distribution, while oversmoothing (or underfitting) is a situation where those meaningful features are concealed by the larger pattern. Adaptive filters take care of these issues. However, as both fixed and adaptive filters are typically circular in shape and
use a greedy approach to aggregate populations until a certain population threshold has been met, they, introduce new problems of their own:

1. Spatial filters may be formed using an aggregation of non-homogeneous population densities. For example, in urban-rural transition zones, spatial filters may include densely populated suburban areas along with clusters of rural populations. The problem with such filters is that they distribute disease risk over large geographic areas, thus resulting in maps that have suboptimal levels of geographic detail.

2. Spatial filters may be created using aggregations of populations with different population characteristics. It is important to note that these filters are ultimately used for calculating disease incidence or mortality rates, which are computed by dividing the cases that fall within each spatial filter by its corresponding population. As spatial filters are not yet designed to respect population characteristics, rates may be computed by aggregating populations that belong to very different demographic groups. This is problematic as previous literature has clearly shown that underlying population characteristics influence the risk of disease. Thus, such mixing of populations may result in inaccurate representations of disease risk across geographic space. Public health efforts that rely on such maps for intervention or policy planning may result in suboptimal or biased solutions.

My research outlines a new method to (1) quantify population clustering introduced by kernel density estimation and (2) produce smoothed maps from spatial filters that are sensitive to underlying population characteristics.
Spatial epidemiology methods at small-area scale can be categorized into three basic approaches: disease mapping, geographic correlation studies, and clustering (Elliott and Wartenburg, 2004). Disease mapping primarily focuses on issues relating to the representation of disease outcomes such as incidence or mortality and associated processes over space and time. Many of these methods commonly represent disease burdens in the form of an image that quantifies “the deviation from the background level of disease expected for the community of interest” (Lawson and Williams, 2001). Focusing less on representation, geographic correlation studies are more interested in the variations across space between population groups and predefined variables such as demographics, lifestyle, and environmental factors (Elliot et al. 2000). The third approach to small-area scale spatial epidemiology relies on statistics rather than mapping or correlation, with cluster analysis detecting disease by taking advantage of the "tendency for disease cases to occur in a nonrandom spatial pattern relative to the pattern of the noncases" (Elliott and Wartenburg, 2004). Methods developed under these three categories mostly operate on the assumption that any data input will likely be available in an aggregated form due to reasons of privacy concerns or data unavailability. The U.S. Census Bureau labors under the privacy strictures of Title 13 of the U.S. Code; the CDC and other medical outlets must answer to HIPAA; other bodies both public and private are bound by their own mixture of ethical and legal codes that mean access to precise, point-level data is almost unheard of amongst geographers (especially medical geographers). However, recent developments in GIS and computer science showcase the use of automated systems to
process and analyze individual-level health data within protected data environments producing summarized results that guarantee individual privacy. Tiwari and Rushton (2009) developed an automated system for mapping disease burdens using spatial filtering methods and then linking those outcomes to exposure estimates obtained from environmental models. The system, called the Environmental Health Surveillance System, is capable using point-level data stored in secure health databases to produce smoothed surfaces of disease risk that guarantee anonymity of the input data points. Anselin (2012) provides a comprehensive review of the changing nature of how spatial analysis methods and tools are being incorporated into fully or semi-automated spatial analysis systems. Increasing interest and research in the development of such systems suggests a need for studies that evaluate the applicability of existing spatial analysis methods for mapping and understanding fine-scaled variations in disease risk. This thesis evaluates the types of spatial biases that are introduced by disease mapping methods such as the ones used in the Environmental Health Surveillance System (2009) when individual or small-area input data are used. Methods to remove those spatial biases are also developed in this research.

Within disease mapping, meanwhile, there are two major factions driving the discipline: rate modeling and rate computing. Modeling the rate—or geostatistics—is used to compensate for the uncertainty of unknown data by creating a model that takes into account densely sampled secondary data and "reproduce[s] aspects of patterns of spatial dependence and other statistics deemed consequential for the problem at hand" (Goovaerts, 1997). Computing the rate is more useful when the necessary data is available and the rigor of the method is focused extrapolating meaning from the data rather than compensating for a lack of
available information. This introduces concerns about data reliability and the scale on which the data is available (Forand et al. 2002), compounded by the assumptions of aggregation that continue to haunt small-area scale spatial epidemiology.

Of the many ways of condensing datasets into usable models, one of the most pervasive is the cluster analysis of point data often used in the detection and analysis of disease patterns (Conley, 2005). When performing these sorts of analysis, the spatial distribution of individual disease incidents is merely a precursor to defining the disease's geographic nature and, by extension, understanding where it thrives, how it spreads, and what its future growth might be. Many methods have arisen whose primary focus is taking this point data and underlying factors such as the age, density, and level of poverty of the potentially afflicted populations and environmental aspects that might affect disease spread and converting that information into a form that can be useful in identifying regions with an elevated risk of the disease.

A number of the available techniques for defining these areas have been put forward over the years, each with its own unique strengths and drawbacks. Perhaps the simplest method is to aggregate data based on some predefined area—usually some sort of administrative district such as a Census tract, district, or country—to create a choropleth map (Wright, 1938). Though it enjoys advantages of simplicity and conforming to boundaries that can be generally recognized by policymakers and the general public, it also faces numerous problems with regards to its method of data aggregation. One of the chief issues is that all spatial information within each aggregation area is completely lost; the entire region is assigned homogenous characteristics (Murray and Shyy, 2000). Other problems include the inability to address multivariate data, arbitrary cut-points (Indrayan and Kumar, 1996), and numerous
other issues related to the choices behind how data is selected, calculated, and displayed (Walter and Birnie, 1991).

Basic kernel density estimations (KDEs) are amongst the most widely-used and time-tested of methods, but often fall victim to oversmoothing and sparsely-populated examination areas can lead to large variances in disease rate (Cai et al. 2012). They compensate for many of the disadvantages in traditional choropleth mapping approaches, since they are not constrained by administrative boundaries and are able to display information as a gradient, rather than as a single, homogenous block. Even they are not without their own issues, however, such as edge effects and—of greatest concern to this thesis—a vulnerability to the adverse effects of spatial clustering (Gatrell et al. 1996). Poisson regression modeling has the advantage of detecting and illustrating relationships versus the kernel estimation's illustration of simple levels, but suffers from the same flaws and drawbacks as the other as well (Nakaya, 2005).

Many currently available techniques are also computationally slow and tend to produce many positive solutions (i.e. circles or ellipses) for each cluster in the dataset. This over-reporting of clusters can lead to a prohibitively large workload for postprocessing. Finally, many methods make assumptions about the distribution and shape of clusters that may not apply to the dataset. Therefore, techniques that can account for a spatially variable background population, scale easily to large datasets, minimize or generalize assumptions, and minimize the number of solutions returned per cluster are needed.

There is an established history of employing statistical methods to detect clustering, such as the nearest neighbor statistics (Clark and Evans, 1954). Not only can this method detect
clusters in both space and time (Ederer et al. 1964) but it can be scaled for any number of nearest neighbors (Thompson, 1956). Another traditional method is the use of quadrat counts in detecting clusters in the distribution of towns (Dacey, 1964), land use patterns (Getis, 1964), or plant distributions (Skellam, 1954). More recently, grid-based cluster detection has served in a more exploratory role (Wang et al., 1997). As the field grows and matures, new methods enter into the geographer's toolbox, including the Gi* statistic, geographically weighted regression, the Moran’s I, and model-based approaches (Conley, 2005). Unfortunately, many of these methods prove difficult when used to attempt cluster detection in a geographic context for a number of reasons put forth by Conley (2005):

1. The need to account for a spatially variable background population
2. The exponential increase in the size of datasets
3. Assumptions made by statistical method
4. The need to reduce over-reporting of clusters

When working in a health geography context, such as the detection of disease clusters, spatial variations in population must be accounted for, which highlights the importance of ratios versus an analysis of pure population (Levison and Hadden, 1965). These population adjustments have a number of solutions, though some researchers choose to ignore the issue, claiming "while geographers see a high concentration of disease as not a cluster when it overlays on a densely populated city, data mining sees it as a cluster explained by the presence of a city” (Estivill-Castro and Lee 2002). Still, it is commonly accepted that cluster detection methods have greater utility if they can distinguish between situations where the cause of clustering is already known—such as high-population areas—versus unexpected causes that
indicate unusual activity.

In modern mapping techniques, certain considerations of reliability and consistency are neglected as a result of common problems with the underlying data. Early data often suffered from a lack of precision—technological hurdles in data collection meant that much of the data collected could not be relied upon to be 100% accurate—or else it fell victim to the increased demands on time that precise spatial analysis incurred before the advent of computers. More recently, privacy concerns by data-harvesting operations such as the U. S. Census have meant that the information researchers have available to them often begins its life at one or more removes from perfect precision; data acquired from specific points or individuals being aggregated to Census blocks or zip codes or counties before any researcher ever gets a chance to look at it. With government regulations such as HIPAA and private concerns over litigation and patient security, health data also suffers from aggregation even where refined data is available as a mechanism of privacy protection.

Because of these limitations on the basic data, the techniques to analyze it almost universally suffer from a lack of precision themselves. It can be difficult to notice with aggregated data, but as the growing reach of technology promises ever-more-accurate data to researchers, techniques specifically designed to account for higher degrees of precision will become necessary. Problems of data precision have been addressed to some degree, with methods such as kernel density estimation, which makes assumptions about a population based off of a finite sample of that population. As discussed above, such methods are capable of producing more reliable surfaces by accounting, to some degree, for the small numbers problem and the modifiable areal unit problem (MAUP) to some degree. However, they are
also subject to the problems of undersmoothing and oversmoothing of data.

Meanwhile, as datasets continue to grow in size and scope (Gahegan, 2003), methods that have been relied on for decades are becoming unusable due to the time it takes for the algorithms to be computed (Conley, 2005). With a growing concern for methods that are both accurate and less hardware-intensive, geographers continue to push the boundaries of data analysis, with the spatial scan statistic (Patil and Taillie, 2004), partitioning methods, and hierarchical methods such as DIANA and AGNES (Kaufman and Rousseeuw, 1990).

Similar approaches using graph theory include using the minimum spanning tree (Zahn, 1971), a k-nearest neighbor graph (Karypis et al. 1999), adjacency graphs of administrative regions (Duczmal and Assunção, 2004), Delaunay triangulations (Eldershaw and Hegland, 1997), and Voronoi/Delaunay diagrams (Estivill-Castro and Lee 2002). These graph-based methods are able, more often, to detect clusters of arbitrary shapes; unfortunately, they remain unable to account for variation in the point density. Some algorithms seek to define a cluster as a circular buffer containing a greater-than-average point density. This approach has the advantage that the background population can be accounted for by establishing a ratio of points in the circle with the background population contained within the same circle. This allows the research to determine whether that ratio is high enough to consider the circle a cluster of high disease incidence.

Given limitations in the availability of fine-scaled geospatial health data, the extrapolation of two-dimensional information from collections of one-dimensional points or smaller two-dimensional area can call into question the validity of data aggregation. But in the past few decades, the amount of geographic information—and the amount of information with
a geographic component—has begun to grow at exponential rates. Be it the ever-increasing number of satellites streaming fantastically detailed information on land cover, topography, and climate back to the planet or the increasing geographic fidelity of traditional information sources such as the U.S. Census, with their use of hand-held GPS computers to accurately pinpoint every residence in the United States; be it the ubiquity of cell phone cameras taking pictures and video with geotagged metadata or the expanding commercial databases tying hundreds or thousands of marketing metrics to customer addresses, the sheer mass of information available threatens to overwhelm the human capacity to process it all into something meaningful.

Furthermore, as technology increases, so does the common researcher's access to more spatially precise data, either because precise data becomes easier for the individual to legally collect or else because traditional sources of data become more capable of providing specific data without violating their privacy mandates. These two convergent factors conspire to make traditional methods of data aggregation and analysis less applicable than they have historically proven to be; assumptions that were once universal—such as the low resolution of data aggregated to the county level—are becoming less prevalent. Therefore, techniques that make those assumptions will, more and more, be found to produce results that are below an acceptable level of reliability. This creates the stage for new methods that are tailored from the outset to deal with data at the most precise levels possible.

The purpose of this proposal is to formalize the types of oversmoothing that can occur when kernel density estimation methods are used to produce disease maps, develop an algorithm for quantifying the degree of oversmoothing, and investigate methods to minimize
oversmoothing. As data is scaled up from the smallest level of available information (i.e., as individual points or small areas such as census blocks or zipcodes) to larger analysis areas, it must be grouped in a way that is both geographically logical—usually via proximity and contiguity—and scientifically meaningful, often by creating a series of areas where the population supports are the same, though this invites differences in rates of overall incidence.
CHAPTER 3

DATA AND METHODS

As stated earlier in this document, our goal is to develop a new method to (1) quantify population clustering introduced by kernel density estimation and (2) produce smoothed maps from spatial filters that are sensitive to underlying population characteristics. The hope is that these new tools will succeed in producing maps with a higher degree of both accuracy and utility when dealing with point-level data that may have a tendency to exhibit signs of spatial clustering; especially along boundaries of radical change in population density such as urban/rural transition zones. To that end, I produced a two-step process that calculates a Nearest Neighbor Index that is used to judge the severity of clustering present in uncompensated data and then—if the data is found to be sufficiently clustered—employs a process I call the disassociated aggregation and analysis method (DAAM).

This process of adjusting smoothed kernel density maps for difference in population density is divided into 4 broad steps:

1. Synthetic data – The process of creating the synthetic data set(s) used during the course of the project
2. Fishnet – Defining and creating the criteria-sensitive grid overlay that lies at the heart of the entire process
3. Cluster analysis – How to detect and classify any clustering that may appear in the data
4. Kernel density estimation – Performing the kernel density estimation on the partitioned data
Tools

We employed three major computational tools during the course of this thesis. Replicating the results shown elsewhere in this document will require a researcher to either use the tools listed here or else employ similar analogues that they have access to and that can be held to the same rigorous academic standards.

1. **ArcGIS 10.1** – Produced by ESRI, (c)1999-2012. ArcGIS is one of the industry-standard GIS programs and includes a wide array of mapping and statistical analysis tools, many of which are used throughout the processes outlined elsewhere in this document. Whenever a specific ArcGIS operation is employed, its name will appear in *italics*.

2. **Geospatial Modelling Environments (GME)** – Produced by Dr. Hawthorne L. Beyer, the GME provides additional statistical functionality for ArcGIS. For the purposes of this thesis, however, I only employ its *r.sample* function, which is used one wishes to take a set of points and randomly select a subset from those points. It is available for free online at [http://www.spatailecology.com/gme/](http://www.spatailecology.com/gme/).

3. **WebDMAP** – Produced by Dr. Chetan Tiwari, it is the primary Kernel Density Estimation software used for this thesis. It is available for free online at [http://webdmap.com/](http://webdmap.com/)

Synthetic Data

Being concerned primarily with the creation and testing of the classification system for spatial clustering and the optimization of context-sensitive kernel density estimations, this thesis uses multiple synthetic datasets. The primary reason for this is that the method was
originally conceived for use with point-level datasets, but the difficulties present in creating or obtaining real-world data in this format are beyond the scope of a project of this level. Additionally, synthetic datasets can be manipulated to produce points of a variety of spatial distributions, meaning that multiple scenarios can be tested without the need for expending the resources it would require to secure or produce such diverse real-world datasets. Because of this, it is also much easier to stress-test both the classification system and the KDE method, allowing the results to be analyzed in a more robust fashion.

The initial dataset, henceforth called Dataset 1, was created using 2010 Census data for Denton County, Texas. TIGER/Line files of Census Tract boundaries were used for the starting geography and this data was joined with 2010 population figures taken from the American Factfinder website. To serve as our Control Population, 66,262 points—10% of the total population—were assigned random locations using the ArcGIS Generate Random Points tool and the Census Tract boundaries to constrain population density. From these points, 10%—or 6,626 points—were randomly selected to be our Case Population using the r.sample function of the Geospatial Modelling Environment tools.

The exact process used to create Dataset 1 is outlined below. Words in **bold** represent specific files either imported from outside sources or created during the process. Words in *italics* are the names of the operations being performed:

1. Use **Census Tract** shapefile
2. Use **Census Population by Tract** csv file
3. Perform a *join* on the two files by Tract ID
4. Create a Control column with value of 10% of the total population
5. Use Create Random Points tool based on Control column to create Control file, the result of which can be seen in Figure 1

6. Run an r.sample on the Control file to select 10% of controls as cases

7. Open Control file's attribute table and select the randomly chosen points

8. In the right-click menu, choose Selection > Create Layer From Selected Features

9. In the right-click menu, use Data > Export Data... to create Cases file

10. In the Cases attribute table, Add Field twice: once for longitude and once for latitude

11. Calculate Geometry for both fields

12. Export Data as text file

13. In the Control attribute table, Add Field twice: longitude and latitude

14. Calculate Geometry for both fields

15. Export Data as text file

16. Convert both text files into .csv files without headings and with only four columns:

   ID, Longitude, Latitude, and Weight

Figure 1: Controls file.
Fishnet

At the heart of the process outlined in this thesis is the spatial partitioning of the data into discreet units which can be analyzed individually and in the context of other parcels that exhibit user-defined similarities. This is done by the creation of a spatial matrix called a 'fishnet' to overlay the map using a level of fineness determined by the individual or agency employing the method. This fishnet is employed to 'capture' the data from the points or regions contained within its boundaries. At its simplest—and in the method outlined below—it simply counts the number of cases within its boundaries, but remains capable of more complex calculations such as ratios or the incidence of additional factors of interest to the researcher.

For our primary example, the fishnet is created using Dataset 1 and the Cases file produced during the Synthetic Data process detailed in the section before this one. We will henceforth refer to this file as Fishnet 1.

The exact process used to create Fishnet 1 is outlined below. Words in **bold** represent specific files either imported from outside sources or created during the process. Words in *italics* are the names of the operations being performed:

1. The user determines the granularity of the fishnet they wish to produce, depending on the scale of the original data, the size of the spatial area to be analyzed, and any other relevant factors.

I think it is prudent to mention at this point that this step can be one of the most difficult in the entire process, because it is the point at which there begins to be a risk of introducing human error into the process. Because of the vast diversity of data types, map scales, analysis extents, and other issues, it is impossible to produce a single optimum setting
for fishnet granularity; indeed, I cannot even define a formula or other method for
guaranteeing the best results the very first time this process is followed. In Chapter 4, the
Results section of this document, I discuss techniques and future possibilities for refining this
and other user-controlled elements of the process, but for now, a researcher will be required to
rely on their understanding of the data with which they are working when deciding how to
define fishnet granularity.

2. Use the Fishnet tool, which will create the Initial file

3. Set the Output Feature Class to desired location

4. Set the Template Extent to the Census Tract shapefile

5. Set Cell Size Width to the desired granularity determined in Step 1
   a. You can ignore this step but you will need to fill in Number of Rows instead,
      creating a Grid with a set number of rows

6. Set Cell Size Height to the desired granularity determined in Step 1
   a. You can ignore this step but you will need to fill in Number of Columns instead,
      creating a Fishnet with a set number of columns

7. Disable Create Label Points

8. Set Geometry Type to Polygon

9. Select OK, which will produce results similar to those seen in Figure 2
10. In the Table of Contents, right-click on the Grid file and select Joins and Relates > Joins...

11. Under What do you want to join to this layer? select 'Join data from another layer based on spatial location'

12. Under Choose the layer to join to this layer select the Cases file

13. Under You are joining select 'Each polygon will be given a summary...'

14. Under The result of the join will be saved into a new layer, select the desired location to save your Fishnet file

Cluster Analysis

Our initial problem was the lack of a good standard by which to judge the degree or nature of oversmoothing that is taking place; that is, how individual variations in point density are translated to the overall density of a given aggregate area in a kernel density estimation.

Figure 2: Fishnet overlaid with cases.
Therefore, our first goal was to devise a method that is useful for defining the severity of clustering that can occur in non-aggregated data as well as a system for quickly and impartially ascertaining the severity when dealing with potentially tens or hundreds of thousands of points of data.

As shown in Table 1, the decision was made to classify clustering by the amount of clusters spatially within an analysis extent. This produces one of five possible situations: an aggregated area with a uniform distribution of points, a random distribution of points, or one of three instances of cluster severity based on the results of a nearest neighbor analysis. Originally, it was posited that a simple NNI analysis could be run on the entire map to determine clustering severity, but I decided that there was too much room for error in a process designed to be responsive to local changes in point density. To that end, I propose divvying up the entire map into a series of areas and calculating the degree of spatial clustering within each of these areas independently.

Table 1: Cluster intensity matrix.

<table>
<thead>
<tr>
<th>NNR</th>
<th>Uniform Distribution</th>
<th>Random Distribution</th>
<th>Minor Clustering</th>
<th>Moderate Clustering</th>
<th>Critical Clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥ 1.5 (to a maximum of 2.1491)</td>
<td>≥ 1.0 to &lt; 1.5</td>
<td>≥ 0.80 to &lt; 1.0</td>
<td>≥ 0.5 to &lt; 0.80</td>
<td>≥ 0.0 to &lt; 0.5</td>
<td></td>
</tr>
</tbody>
</table>

To this end, I begin by using WebDMap to perform a standard Kernel Density Estimation (KDE) on your dataset. When it is completed, you can take the resulting grid of points and use
ArcGIS to run the process outlined in Figure 3, which automates the process using the Model Builder application available in ArcGIS. The process outlined in Figure 1 uses the $d\text{distance}$ column in the WebDMap shapefile to draw a buffer around each grid point equal in size to the area being analyzed to form that point’s $\text{ratio}$ column. Then each individual buffer in the resulting file is selected one at a time and used as a mask to clip points from the dataset being analyzed. Then a Nearest Neighbor Index is run only on those points using ArcGIS's built-in $\text{Average Nearest Neighbor}$ function and all the results are provided as a single file in a separate window. Finally, those results are parsed to distill the results of each individual analysis, leaving you with the information in a format that can be easily read and analyzed.

Once the statistics for each grid buffer have been calculated, the nearest neighbor ratio (NNR) can be used to determine the severity of the clustering within that buffer, which I have classified as its $\text{intensity}$. As shown in Table 1, there are three scales of intensity—$\text{minor}$, $\text{moderate}$, and $\text{critical}$—which reflect an increasing degree of disparity between the average point density of an aggregated area and the expected point density of that same area.

![Figure 3: ArcGIS model builder for nearest neighbor index.](image-url)
If the final result is an NNR of 1.0 or greater, there is no tendency towards clustering; a result between 1.0 and 1.5 is a random distribution, while an NNR higher than 1.5—up to a maximum of 2.1491—exhibits uniformity. Only when the Nearest Neighbor Ratio is lower than 1 is there evidence of clustering, with severity increasing as the NNR approaches 0. Assignment of severity is based on the score's proximity to 0, with the scale being weighted for lower ratio values to be less severe. Thus, the difference between a score of 0.9 and 0.8 is considered more significant than the difference between 0.5 and 0.4. With an NNR that is less than 1.0 but greater than or equal to 0.8, the severity is considered *minor*: enough to be accounted for by common spatial variations and of no concern to further calculations. An NNR lower than 0.8 but greater than or equal to 0.5 is *moderate*; though not necessarily a cause for immediate condemnation of the traditional kernel density estimation method, some solution for the clustering issue—such as the one proposed in this thesis—may be worth considering. At an NNR of less than 0.5, the clustering is *critical* and, without some form of spatial compensation, information derived from the aggregated area cannot be considered quantifiably valid.

Once the kernels have been assigned their clustering severity, it falls to the researcher to decide what their threshold for action should be. In real-world data, certain amounts of clustering are inevitable, and it is therefore the responsibility of each individual involved in a spatial study to determine acceptable tolerances that will form the threshold between accepting a traditional kernel density estimation or employing the method for cluster compensation detailed in this thesis.

**Kernel Density Estimation**

Now that I have identified the problem, I am faced with the main purpose of this thesis:
solving it. To that end, I present a new method for using kernel density estimation to create aggregated areas that prioritize point density rather than physical adjacency: in short, to create surfaces that are statistically contiguous rather than spatially contiguous. In addition to point data, this method will be able to organize areas based on surface characteristics—such as income, population characteristics, or disease rates—to create aggregation areas that are sensitive to underlying factors that might prove important to a thorough analysis of data.

This process is known as the disassociated aggregation and analysis method (DAAM) and the creation of these non-contiguous aggregation surfaces is relatively simple. Continuing as I have in the previous sections, the exact process used to generate the cluster analysis is outlined below. Words in bold represent specific files either imported from outside sources or created during the process. Words in italics are the names of the operations being performed:

1. Begin by establishing how many statistical groups you wish to form and the thresholds between them

   As with defining the fishnet granularity, I am opening the process up to human error in a way that cannot be reliably compensated for at this stage in the process. The results of the analysis are going to depending heavily no only on the size of the individual cells of the fishnet, but on how many groups are going to be created and what thresholds you, as the researcher, are going to decide to employ. Once again, I discuss some ideas about how to ameliorate these difficulties in Chapter 4, but—initially, at least—the researcher will be required to exercise their own understanding of the data being analyzed to produce the results that they desire.

2. In the Table of Contents, right-click on the Fishnet file and select Attribute Table
3. Click on Select by Attributes

4. In the selection box, use the equation "Count" > 0 AND "Count" < X where X is the upper bound of your first group

5. Select Apply

6. In the Table of Contents, right-click on Fishnet file and select Data > Export Data...

7. Under Export, select Selected Features

8. Under Output Feature Class, select the desired location to save your Partition1 file

9. Select OK

10. Repeat steps 2 through 9 until you have created a Partition file for every statistical group you wish to create, which will produce a finished result similar to the one seen in Figure 4

The partitions you have just created are going to form the different statistical 'zones' into which you will be dividing the data. These zones lay at the heart of the DAAM, isolating the population into sections of comparable density so that each can be analyzed only in the context of other areas of similar population density. This is a temporary state, and the completed map will be reassembled later, but for the moment, you will be working with a number of separate maps equal to the number of partitions you have just created.
11. On the toolbar, click on Select > Select by Location

12. Under Selection method, use select features from

13. Under Target layer(s), select the Cases file

14. Under Source layer, select the Partition1 file

15. Under Spatial selection method, select intersect the source layer feature

16. Select OK

17. In the Table of Contents, right-click on the Cases file and select Data > Export Data...

18. Under Export, select Selected Features

19. Under Output Feature Class, select the desired location to save your CasePartition1 file

20. Select OK
21. Repeat steps 11 through 20 until you have created a CasePartition file for each Partition file you created earlier.

22. Once you have finished, repeat steps 11 through 21, except use the Controls file to produce ControlPartition files for each Partition file.

23. In the Table of Contents, right-click on the CasePartition1 file and select Attribute Table.

24. Select Table Options > Export.

25. Under Export, select All records.

26. Under Output table, select the desired location to save your CasePartitionTable1 file; make sure to select TXT file.

The next few steps are all part of the normal Kernel Density Estimation process using WebDMap, but they are included both for the sake of completeness and so that users that are inexperienced with the software can following along without having to consult outside documents. The important thing to remember about these steps is that all the files related to each partition should be handled separately from one-another; this division lies at the heart of the disassociated aggregation and analysis method. By forcing each section to analyzed only in the context of a relatively homogenous population density, I am able to minimize the effects of transition zones between areas of radically different densities. These compensations will then be reflected in the final product, even after the partitions are recombined.

27. Locate the CasePartitionTable1 file and change the file extension from .txt to .csv.

28. Open the file and delete the column headers as well as all columns except the FID,
Latitude, Longitude, and Weight

29. Repeat steps 23 through 28 until you have generated a clean CSV for all your 
    CasePartition and ControlPartition files

30. In WebDMAP, using the corresponding CasePartition and ControlPartition files, 
    create a series of Rate files

31. Employing a process similar to 11 through 20, use the individual Rate files and their 
    corresponding Partition files to create a series of RatePartition files

32. Use the Merge tool to combine all of the RatePartition files into a MasterRate file, 
    the effects of which can be seen in Figure 5

33. In WebDMAP, use MasterRate, Cases, and Controls to run the final analysis

Figure 5: KDE grids for standard KDE (red) and DAAM (blue).
CHAPTER 4

RESULTS AND DISCUSSIONS

Using the processes outlined the Methods section, a preliminary synthetic dataset was constructed with a total of 66,626 control points and 6,626 case points which can be seen in the map shown in Figure 6. Once the synthetic data set was generated, a traditional kernel density estimation was run on the data to give us a baseline for comparison; the map resulting

**Synthetic Dataset 1 - Denton, TX, 2010**

![Map of synthetic dataset 1.](image)

Figure 6: Map of synthetic dataset 1.
from this analysis can be seen in Figure 7. The next step was to generate the grid which would be used to break the data into the different analysis extents. It was decided—for ease of use with the GCS_WGS_1984 coordinate system which the data was projected in—that the area would be divided into squares .01 degrees in both height and width, which generated a fishnet of 56 squares by 44 squares, for a total of 2,464 individual partitions. When the fishnet was completed, I used the point grid produced by the KDE and the nearest neighbor index model to apply our cluster intensity method to the data, producing a grid with 1,485 individual points around which to build our buffers. The results of the cluster intensity survey are available in Table 2, where they indicate that a tremendous amount of minor, local-scale clustering is being exhibited, as well as a small amount of moderate clustering. While there is no critical clustering taking place, the sheer amount of minor clustering is such that our dataset is a viable candidate for some sort of compensation.

<table>
<thead>
<tr>
<th>Uniform Distribution</th>
<th>Random Distribution</th>
<th>Minor Clustering</th>
<th>Moderate Clustering</th>
<th>Critical Clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NNR</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>27</td>
<td>207</td>
<td>1,243</td>
<td>8</td>
</tr>
</tbody>
</table>

Once this spatial analysis was completed, I employed the disassociated aggregation and analysis method (DAAM) detailed in the Methods section to parse the map into four different partitions and analyzed them. The map produced by this method is available as Figure 8.
Figure 7: Standard kernel density estimation for synthetic dataset 1.

Figure 8: Disassociated aggregation and analysis method for synthetic dataset 1.
Though differences are apparent to even a cursory examination, I decided to help and compensate for the unreliability of the human eye at discerning the difference in pattern of multiple images by employing a simple difference calculation was performed on each set of raster maps, producing Figure 9. The calculation was formatted to have the disassociated aggregation and analysis method raster subtracted from the Standard Kernel Density Estimation. Thus, areas where the KDE resulted in higher numbers of cases are displayed in red, while areas where DAAM produced higher numbers of cases are displayed in blue.

This map clearly illustrates the effectiveness of the DAAM, with the blue areas tracing the population concentrations and preserving the urban-rural transition zone boundaries that are so important to correctly understanding how risk is spread. The red areas displaying the KDE concentrations tend to be clustered at the hearts of population groups, the characteristics

Figure 9: Difference map for synthetic dataset 1.
of the population being focused on a central area while the transition zones are smoothed out into the less-densely-populated areas, resulting in the loss of fidelity. What this means is that the KDE tends to result in a higher degree of spatial clustering, while the DAAM is more sensitive to the population characteristics of a given area, allowing us to see the extent of the simulated effect with a higher degree of fidelity.

Due to the nature of the raster calculations, the less-populous rural areas also exhibit some artifacting in the form of the yellow and teal circles that appear. This is due primarily to the offset between the center of the kernels in the standard KDE and the DAAM kernels. It has no significant bearing on the separate maps presented in Figure 7 and Figure 8; it is merely a data anomaly that appears in Figure 9, as well as appearing in all the other difference maps run on additional sets of synthetic data.

Table 3: Cluster Intensity Matrix for Dataset 1, Standard KDE vs DAAM

<table>
<thead>
<tr>
<th></th>
<th>Uniform Distribution</th>
<th>Random Distribution</th>
<th>Minor Clustering</th>
<th>Moderate Clustering</th>
<th>Critical Clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Standard KDE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NNR</td>
<td>27</td>
<td>207</td>
<td>1,243</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td><strong>DAAM</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NNR</td>
<td>0</td>
<td>1,025</td>
<td>501</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

An additional layer of confirmation in the effectiveness of the DAAM can be seen by running a second Nearest Neighbor Analysis, this time using the DAAM grid instead of the KDE grid. The results, which are seen in Table 3, clearly show the benefits of the DAAM, with the
instances of moderate clustering dropping from 8 down to a single instance while the minor clusters are reduced to a mere 40% of the KDE total. As an unexpected side effect, the instances of uniform distribution have also dropped to 0 from an original total of 27. A determined researcher can actually use multiple iterations of the DAAM and the Nearest Neighbor calculation to refine the user-defined variables in the DAAM, such as the granularity of the fishnet and the breaks used to create the partitions. By repeating the process with different variables and comparing the results of the NNI, it should be possible to find an optimum solution, or at least a solution close enough to optimum that it meets the user's criteria.

To test the this process, three more sets of synthetic data were produced, each using a unique new set of 6,626 points drawn from the original control file of 66,626 points. Each set of data was put through the same series of modifications via the DAAM as Set 1, producing 3 additional maps per dataset for a total of 9 more maps which are included below.

Dataset 2 shows the results of the standard KDE (Figure 10) and DAAM (Figure 11), reflecting profiles very similar to dataset 1. This impression is further advanced by the raster calculation in Figure 12; though not identical to dataset 1's Figure 9, the DAAM results still clearly outline the extent of the areas of denser case population while the KDE results remain much more highly clustered and the urban/rural transition boundaries are far less distinct where they are not completely absent.
Figure 10: Standard kernel density estimation for synthetic dataset 2.

Figure 11: Disassociated aggregation and analysis method for synthetic dataset 2.
Figure 13 shows the KDE for dataset 3, which remains consistent with those seen in datasets 1 and 2; a single concentration of red at the center of the county, with several additional concentrations located in the south. Meanwhile, Figure 14 displays the DAAM results that indicate a population that is spread out to more closely follow the distribution of the point data from which the maps were derived. This becomes much more starkly apparent with the raster calculation in Figure 15. Once again, the DAAM has produced a significantly more accurate picture of the relative population densities involved, including perhaps the most vivid rendition of the central population zone, with the visible 'horns' at the northern edge of the central population mass. The fidelity in the southern population zones is not quite as pronounced in this raster calculation as it is in the other datasets, which may indicate that the randomly selected points were more heavily drawn from the geographic center.

Figure 12: Difference map for synthetic dataset 2.
Figure 13: Standard kernel density estimation for synthetic dataset 3.

Figure 14: Disassociated aggregation and analysis method for synthetic dataset 3.
Finally, dataset 4 continues the trend; its KDE (Figure 16) and DAAM (Figure 17) maps continue to exhibit the characteristic patterns, which are only enhanced when seen using the raster calculation in Figure 18. Areas of dense population continue to be better-represented by the DAAM, while the KDE results remain much more intensely clustered. As with all the other raster calculations, it also retains the circular artifacting in the less-densely populated areas, where the lower population totals and the significant offset in the grid points between the KDE and the DAAM produce the circles that do not appear on Figures 16 or 17.

That is not to say that the method is completely without flaw, and there are numerous areas where additional research and development could be useful. While currently, many parts of the process—such as the granularity of the Fishnet and the thresholds which define the different partitions—are based solely on the personal input of the researcher conducting the
Figure 16: Standard kernel density estimation for synthetic dataset 4.

Figure 17: Disassociated aggregation and analysis method for synthetic dataset 4.
process, the newness of the method presented here means that there is no reliable body of literature for how to go about selecting those features. Employing the DAAM-to-NNI iteration method discussed briefly above, it would be possible—albeit time-consuming—for a researcher to run through the process multiple times until they found an optimal set of input for the variables, or at least input that reduced cluster instances to a degree sufficient for that user's purposes.

Additionally, many of the processes discussed elsewhere in this document would benefit from automation with tools or programs in a similar vein to WebDMAP, which would go a long way to eliminating the amount of time and labor required by the current system. Not only would this system speed up the process—and speed up the multiple iterations necessary to quality-check the variable inputs as discussed in the previous paragraph—but these tools could
be equipped with preselected settings for the variables such as fishnet size and partition breaks. These presets would enable the system to be made available to non-specialists, such as policymakers or researchers without a strong GIS background, allowing the method to find wider use while lessening the danger of inexperienced or unqualified users introducing too much error when making their selections about the variables.

Unfortunately, these tasks are beyond the scope of a single Master's-level thesis.

The development of both the cluster intensity scale (and the means of automating its application) and the disassociated aggregation and analysis method show tremendous promise with regards to helping geographers prepare for an age of precision. They will allow for a higher degree of control over how data is analyzed and how risk is calculated. Using the methods outlined in this document, everyone with an interest in the spatial relationships between risk, data, and population will have a new and powerful tool in the search for understanding.
CHAPTER 5

CONCLUSIONS

This thesis began with the central question of how to improve the standard kernel density estimation method used to create the maps seen every day in presentations, academic papers, and on the news. With a conviction that health and spatial data would only become more precise and more easily available, I felt that new methods which looked towards this future were an important area to be addressed and that advances in how data was handled could lead directly to breakthroughs in how risk was understood and how care and response were planned. To that end, our focus was to create a method of analysis that would allow for differences in spatial distribution and underlying population characteristics to play a larger role in how data was displayed.

One of the chief problems I hoped to combat was the idea of smoothing, where variations in the density of population such as those found in population clusters or along rural-urban transition zones were taken into account by the process, rather than having their unique characteristics smeared together with the characteristics of the populations around them. By understanding the factors that created risk—and by creating maps where those factors were given a role in determining how data was displayed—our hope was to produce a system that could take advantage of precise population information to eliminate the dangers of oversmoothing without losing the advantages that kernel density estimation already provides, such as its ability to compensate for issues like the small numbers problem or the MAUP.

The disassociated aggregation and analysis method has produced results in line with the original goal of this thesis, creating maps that show a measurable difference when compared to
the products of a traditional kernel density estimation. DAAM-produced maps allow the viewer to more accurately trace population concentrations and define borders of difference in population while KDE maps tend to aggregate populations into smaller clusters, which imply that DAAM maps possess a higher degree of sensitivity with regards to those qualities selected for during the map creation process outlined elsewhere in this document.

Furthermore, with the streamlined method for determining the local intensity of spatial clustering, it is now possible to analyze a map or series of maps and quantify the degree of spatial bias introduced by geography and population. Due to the time-consuming approach currently required for the disassociated aggregation and analysis method when compared to a standard kernel density estimation, the introduction of the cluster intensity metric can help save time by ensuring that the method is only employed on maps where clustering is both evident and critical. In situations where it would not produce results very different from traditional kernel density estimations, the intensity system is effective at reducing a researcher's workload, allowing them to use easier, automated methods like WebDMAP with confidence.
REFERENCES


