IP ADDRESSES: EXPLORING THE NATURE OF THE GEOGRAPHIC
DATA AND THE PATTERNS THAT CAN BE EXTRACTED

A Thesis
Presented to the
Faculty of
San Diego State University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Geography
with a Concentration in
Geographic Information Science

by
Daniel Lewis Lusher

Spring 2013
SAN DIEGO STATE UNIVERSITY

The Undersigned Faculty Committee Approves the

Thesis of Daniel Lewis Lusher:

IP Addresses: Exploring the nature of the geographic data and the patterns that can be extracted

Ming-Hsiang Tsou, Chair
Department of Geography

Li An
Department of Geography

Dipak K. Gupta
Department of Political Science

3-22-2013
Approval Date
ABSTRACT OF THE THESIS

IP Addresses: Exploring the nature of the geographic data and the patterns that can be extracted
by
Daniel Lewis Lusher
Master of Science in Geography with a Concentration in
Geographic Information Science
San Diego State University, 2013

Internet Protocol (IP) addresses are frequently used as a method of locating web users by researchers in several different fields. However, there are competing reports concerning the accuracy of those locations, and little research has been done in manually comparing the IP geolocation databases and web page geographic information. Some research has been done using a variety of analytical techniques in order to estimate IP address accuracy, but there has not been a large scale manual investigation. Members of the GIScience community have developed techniques for visualizing point data, but there has not been much research in how to apply these practices to the phenomena of IP address point data. This thesis research worked in the intersection of GIScience, Internet research, and Geostatistics in order to examine IP accuracy (IPv4) to the city-level and to determine the better methods for visualizing IP address point values.

The first part of this thesis examined IP geolocation address accuracy in more depth, and processed several datasets for locational accuracy. Using a previously built custom searching tool that uses popular search engine APIs to extract a list of web pages for a keyword, six keywords were gathered. These keywords were “Mitt Romney”, “Rick Santorum”, “Michael McGinn”, “Jerry Sanders”, “Flu”, “HPV Vaccine”. When manually visiting each web page gathered by the searching tool, three types of data were gathered. First, I categorized each web page into one of twelve categories, ranging from “Blog” and “News” to “Education” and “Governmental”. Second, the slant of the web page was examined; answering whether is it supporting the subject at hand or attacking the subject. Third, and most important, this research looked to find the mailing or street address of the web page’s content creator and compare this address to the given IP address.

The second part of this thesis has attempted to answer the question of how to visualize IP address geolocation data, using the processed IP address data. Framing IP addresses as points and using a kernel density function in order to create a surface, how do different input parameters affect the final surface, and therefore pattern recognition? This study examined the kernel radius value, the method of calculating a population value, the method used to normalize the data, and the interactions of the different categories discussed above. By comparing the resulting maps side by side, the spatial patterns identified using IP addresses are better understood while recognizing better techniques to remove unwanted IP address spatial inaccuracies. The resulting spatial patterns with relevant properties of the keywords were also compared.
This research has attempted to find the optimal method for identifying the ‘signal’ of meaningful data and spatial patterns from the background ‘noise’ of insignificant results. With a better understanding of the signal, IP address information becomes more robust for statistical research and can be used to comprehend multiple sources of online data. By manually extracting the accurate locational information and then processing the results using a series of methods, a more proper technique for displaying IP address information in spatial analysis and GIScience has been shown.
# TABLE OF CONTENTS

| ABSTRACT | iv |
| LIST OF TABLES | viii |
| LIST OF FIGURES | ix |
| ACKNOWLEDGEMENTS | xi |

## CHAPTER

1. **1 INTRODUCTION**  
   1.1- Introduction  ................................................................. 1  
   1.2- Research Questions ........................................................... 2  

2. **2 LITERATURE REVIEW**  
   2.1- GIScience ........................................................................... 5  
   2.2 - The Internet, Search Engines, and Web Visualization .......... 8  
   2.3- Geostatistics .................................................................... 12  

3. **3 METHODS**  
   3.0- Overview of Methods .......................................................... 16  
   3.0.1- List of Software and Data .................................................. 18  
   3.1- Data Gathering- Searching Tool/Process .............................. 18  
   3.1.1- Map/Data Projection ......................................................... 19  
   3.2- Data Processing – Manual Classification Process ............... 20  
   3.2.1- Kernel Density Function .................................................. 23  
   3.2.2- Population Processes ....................................................... 24  
   3.2.2.1- No Population Value .................................................... 25  
   3.2.2.2- Inverse Rank Value ...................................................... 25  
   3.2.2.3- Logarithm of Rank Value .............................................. 25  
   3.2.2.4- Fuzzy Values .............................................................. 26  
   3.2.3- Normalization/Standardization Processes ......................... 28  
   3.2.3.1- Maximum Score ......................................................... 29  
   3.2.3.2- Score Range Procedure ............................................... 29
vii

3.2.4- Web Page Categories ................................................................. 30
3.3- Evaluation ..................................................................................... 31
4  4 RESULTS ........................................................................................ 32
   4.1- Manual Classification Results .................................................. 32
       4.1.1- Web Page Categories ......................................................... 32
       4.1.2- Relevancy/Intensity Codes ............................................... 34
       4.1.3- Geolocational Accuracy ..................................................... 35
       4.2- Kernel Density Maps .............................................................. 36
5  5 DISCUSSION .................................................................................. 40
   5.1- Manual Classification Results .................................................. 40
       5.1.1- Categories ........................................................................... 40
       5.1.2- Relevancy/Intensity ............................................................. 42
       5.1.3- Geolocation ........................................................................ 43
       5.1.4- Spatial Accuracy of Categories ............................................ 46
       5.2- Kernel Density Maps .............................................................. 47
           5.2.1- Population Differences ..................................................... 47
           5.2.2- Kernel Radius Sizes .......................................................... 51
           5.2.3- Normalization/Standardization Methods ............................ 54
           5.2.4- Category Kernel Densities ............................................... 57
           5.2.5- Public vs. Private Categories .......................................... 61
           5.2.6- Individual vs. Group Categories ...................................... 62
       5.3- Spatial Accuracy Correlations ................................................ 63
6  6 CONCLUSION ................................................................................. 67
REFERENCES ...................................................................................... 70
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1</td>
<td>List of Software</td>
<td>18</td>
</tr>
<tr>
<td>Table 2</td>
<td>List of Data</td>
<td>18</td>
</tr>
<tr>
<td>Table 3</td>
<td>Geolocation Summary Statistics</td>
<td>45</td>
</tr>
<tr>
<td>Table 4</td>
<td>Pearson’s r correlation coefficient for each set of datasets</td>
<td>65</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

PAGE

Figure 1: The KDD framework, condensed from nine into five steps. (Fayyed et al, 1996, p.41) .....................................................................................................................5

Figure 2: An overview of the Methods section .....................................................................................................................16

Figure 3: The VISION framework. (http://mappingideas.sdsu.edu/?page_id=136) .....................................................................................................................17

Figure 4: A blank map showing the base map data .....................................................................................................................20

Figure 5: Models of fuzzy classification from the GIS software IDRISI from Clark University .....................................................................................................................27

Figure 6: Chart displaying the sample population lines for a dataset that has 100 records .....................................................................................................................................................................................28

Figure 7: Pie charts showing the percentages of each dataset for each category of web page type .....................................................................................................................................................................................33

Figure 8: Pie charts showing the percentages of each dataset for each code for relevancy/intensity of the results .....................................................................................................................................................................................35

Figure 9: Pie charts showing the percentages of each dataset for each code for geolocation accuracy for the results. The number in parenthesis in the title reflects the number of records in each set of results .....................................................................................................................................................................................36

Figure 10: Examples of the processed keywords. All maps created using the logarithmic population, a kernel radius of 200,000 meters, and the maximum score normalization .....................................................................................................................................................................................38

Figure 11: Spatial accuracy rates by category, as a percentage of all of the records in a category .....................................................................................................................................................................................46

Figure 12: From the “Mitt Romney” dataset, all four population calculations at a kernel radius value of 150,000. The normalization is maximum score .....................................................................................................................................................................................48

Figure 13: From the “HPV Vaccine” dataset, all four population calculations at a kernel radius value of 150,000. The normalization is maximum score .....................................................................................................................................................................................49

Figure 14: From the “Michael McGinn” dataset, all four population calculations at a kernel radius value of 150,000. The normalization is maximum score .....................................................................................................................................................................................49

Figure 15: From the “Mitt Romney” dataset, each of the different kernel radius options. All maps are using the logarithmic population and the maximum score normalization .....................................................................................................................................................................................53
Figure 16: The six datasets, with maximum score in the left column and score range in the right column. All of the maps use a logarithmic population and a kernel radius at 150,000 meters. ..........................................................56

Figure 17: The kernel density results of the manual classification categories. Each dataset uses the logarithmic population, a kernel radius of 150,000 meters, and the score range normalization. .....................................................................................59

Figure 18: Public categories vs. Private categories. Both maps use the logarithmic population, a kernel radius of 150,000 meters, and the score range normalization. ..............................................................................................................61

Figure 19: Individual categories vs. Group categories. Both maps use the logarithmic population, a kernel radius of 150,000 meters, and the score range normalization. ..............................................................................................................62

Figure 20: Comparisons between the kernel density maps for the spatially accurate results (left) and the complete results (right). Each map was created using a logarithmic populations, a kernel radius of 150,000 meters, and the score range normalization. ....................................................................................................65
ACKNOWLEDGEMENTS

The writing of a thesis is not the work of one person. I would like to thank Dr. Ming-Hsiang Tsou for all of the help over the past two years. Dr. Ming was the only person to give me a chance when I first arrived at San Diego State University, and I cannot be grateful enough for the opportunity to work with him. I would also like to thank Dr. Li An and Dr. Dipak Gupta for their helpful suggestions throughout the framing and writing of this work. I would like to thank all of my friends, especially those in the department, who helped to keep me going after some long days (especially in our office). Perhaps most of all, I would like to thank my parents and my brother who have stood by me since I first decided to go to graduate school, and my girlfriend, who put up with the months of worrying and late night pacing around our apartments. It was only my fingers that typed this work; without everyone else, I would have never finished it.
1 INTRODUCTION

1.1- INTRODUCTION

Every computer with a network connection has an address, just like every house connected to the postal system has a street address. The address that places a computer is called an Internet Protocol (IP) address, and every computer, server, printer, or mobile device in a network has this IP address. Some addresses are static, where it is the same over time and some addresses are dynamic, where the addresses change as devices log on or off the network. Each IP address is broken into bit lengths, with IPv4 (Internet Protocol version 4) addresses consisting of 32 bits (as is used in this research), and IPv6 (Internet Protocol version 6) addresses containing 128 bits. Many researchers, both in geography and outside the field, use IP addresses as a method for obtaining locational data within cyberspace. The fact that IP addresses are unique, just like street addresses, makes them a prime system for storing web content in databases.

However, despite the wide use of IP addresses, there is some question about the accuracy of IP addresses. A spatially correct IP address would indicate that the server located via the IP address is in the same location as the street address that is on record, and IP address geolocation accuracy would be defined as the frequency that an IP address is correctly located. How do researchers know that the derived street addresses used in academic research actually reflect the physical locations of these devices? Every IP address has a registered location, but does this location mean what researchers are using it to mean (a proxy for an actual Internet user or web server)? These questions are part of what this research is designed to explore. In order to use IP addresses, validation of the IP address location must be done to ensure that patterns drawn from the cyberspatial data are relevant, meaningful, and in line with reality.

Within Geographic Information Science, or GIScience, there are several established techniques for processing data in order to draw out significant patterns. This thesis research has utilized and compared several of these methods in order to determine which spatial tools
are better suited for preprocessed IP address data, especially in the context of plotting the IP addresses.

The research and experimentation within this thesis resides firmly in the geographic sub-discipline of GIScience. The entire premise of the project is using computer information systems (IP addresses) to map out the location of other computer systems that fit a certain criteria (server probabilities based on keywords). In drawing upon GIS, the research rests on several decades of improvements in computer science and with research done to develop the Internet, in particular the World Wide Web, and all other networks of cyberspace. Within GIScience, the work of this project follows closely to the Oppenshaw (1991) view of the role of GIS. Using the power of these information systems, geographers can work to analyze topics within and outside of geography. Adhering to a postpositivist worldview, GIScience emphasizes determination, reductionism, empirical observation and measurement, and theory verification (Creswell, 2009, p.6). By using mathematical algorithms to analyze spatial patterns and spatial problems, quantitative data is processed in order to find potential causal links between events and realities. GIScience can also be described as part of a pragmatic worldview. As many of the applications for GIS are based in real-world situations, the methods and data that produce the right answer are the right methods and tools, regardless of the particular philosophical underpinnings. Truth and factuality is the key, whether derived from qualitative or quantitative mechanisms.

1.2- RESEARCH QUESTIONS

There are three major questions that I attempted to answer with this thesis research. All are related to better understanding the networks of cyberspace and the methodologies researchers use in studying it.

1. How accurate are the locations cited in IP address databases with respect to where the website content creators claim as their location?

In order to answer this question, I collected information from a large IP address geocoding database. The next step was manually visiting each website and recording
the stated mailing address or contact information for that website. Comparing these locations allowed statistics to be generated that express the accuracy level of IP address databases.

2. *Which geographic visualization techniques return the best result for generating patterns from the preprocessed IP address data?*

After gathering the manually processed geolocation data, I have pulled out the data that has matching locations. Using such data, I then processed the data with a series of different density, interpolation, and vector tools and compare the results to see which methods produce results that lead to a better understanding.

3. *What are key factors within the data visualization and processing of different categories of websites that can influence how well the ‘signal’ in the data shines through the ‘noise’ of inaccurate false positives?*

By looking at the type of website (be it a blog, a commercial site, a news site, etc.), I have been able to find a better method of pattern extraction. The use of different categories and different combinations of categories pointed to a theoretical difference in how different groups of people utilize cyberspace. The differences in whom the content creators are influence the rates of IP address spatial accuracy and the level of ‘signal’ that can be observed.
2 LITERATURE REVIEW

This research used the model of ‘Knowledge Discovery in Databases’ (KDD), as formulated in Fayyad, Piatetsky-Shapiro, and Smyth (1996) as the overall research framework. Fayyed et al outlined a system for KDD that goes through a series of steps beginning with a large database and moving toward knowledge (Figure 1). The process begins with an understanding of the domain and the data at hand and setting goals of knowledge discovery, and then moves into the selection of a target dataset. This dataset emphasizes a particular set of variables or samples that provide the base of research on the topic. The third step involves preprocessing the information to clean the noise from the data. With messy data, this step is invaluable. Following this, the fourth step is to reduce the data and transform it into something that can be used more effectively to accomplish the goals from the first step. The fifth step involves taking the goals and applying the optimal data-mining method, whether it be geographic in nature (hot-spot analysis, geographically-weighted regression) or more classically statistical. The sixth step focuses on exploratory analysis and choosing a proper hypothesis, including in the proper models, parameters, mining algorithms, etc. The seventh step actually accomplishes the data mining process, while the eighth step focuses on interpreting the previously mined patterns. This may involve moving back through the previous steps and reevaluating any choices. The ninth and final step requires acting on the discovered knowledge, whether it is direct use or writing and publishing the knowledge for the future. The steps of preprocessing and data mining are the most relevant to this research. The preprocessing work will be done in the manual classification step (section 3.1 in this proposal). As step three in KDD, data cleaning and preprocessing is important as it removes noise and handles missing data. The data mining step in KDD has been performed as the various visualization techniques outlined later in this proposal (Section 3). As step seven in KDD, this is the crucial step of actually looking for patterns in the data that can eventually lead to knowledge (Fayyed et al, 1996, p.42).
2.1- GISCIENCE

Geographic Information Science, or GIScience, is the study and theorization of the world and everything within it based on its spatial location, place, scale, and time; tightly bound with spatial analysis, geographic information systems (GIS) form the backbone of this thesis research. GIScience is the theoretical framework for GIS; examining the increasing complexity of questions regarding spatial data, especially related to data collection, data management, data capture, data handling, data structures, and data modeling (Goodchild, 1992). As one of the first to coin the term GIScience instead of GIS, Goodchild attempted to expand the discipline from technological advances into a sincere academic field that studies scientific questions (Goodchild, 1992). Expanding from Goodchild on how GIS is relevant in a world increasingly dependent on computer science, Chang states that the way to differentiate a GIS from any other type of information system (such as a file transfer system or a database management system) is its ability to process geospatial data (Chang, 2010, p.2).
For the rest of this thesis, GIScience will refer to geographic information science, and GIS will refer to geographic information systems.

Spatial Analysis can be broken down into four categories, as done in O’Sullivan and Unwin (2003). The first is spatial data manipulation, which includes basic data conversions and functions such as buffering and clipping of data. Second, spatial data analysis, refers to the “descriptive and exploratory” tasks that start a project, including identifying characteristics of the data and looking for early trends (O’Sullivan and Unwin, 2003, p.2). The third category, spatial statistical analysis, refers to applying models and techniques from classical statistics to geographic phenomena. By applying statistical models within a GIS, important information can be gleaned from the data. Fourth, spatial modeling is the method for establishing a framework in which predictions of spatial activities may occur. All together, these four categories constitute the overall idea of spatial analysis. The third category, spatial statistical analysis is the overall focus and purpose of this thesis.

The existence of GIScience itself is a topic that is under frequent debate. In particular, there is academic disagreement as to whether GIS should be more properly thought of as a tool or as a science. Wright, Goodchild, and Proctor (1997) summarized and clarified a long email thread that consisted of a running debate among GIScience experts about the formal definition of GIS. The discussion eventually led to three opinions, best thought of as points along a continuum: GIS is a tool, GIS is toolmaking, or GIS is a science (Wright, Goodchild, and Proctor, 1997, p. 354-355). GIS as a tool focuses on the technological aspects of the application, and sees GIS as a sub-discipline of geography primarily due to the geographic concepts used within GIS. GIS as toolmaking sees it as more than a simple technology, and focuses on the engineering work of extending GIS that involves deeper thinking than software manipulation. GIS as a science sees GIS as a full-fledged research discipline that examines significant questions that can be separated from the technology in a way that emphasizes the inquiry side of science, as opposed to only a tool (Wright, Goodchild, and Proctor, 1997).

Cartography is an important component of GIScience and research in cartography plays a crucial part of proper analysis within GIScience. In addition, Crampton (2009) expanded on the power of cartography and how cartography is applied in the new ‘Web 2.0’ of social interaction and input within geography and cartography. He explores the rise of the
`geoweb` with applications such as Google Earth, and onto the rise of using these applications for `mashups`. A mashup is “is the combination of geographic data from one source with a map from another source (eg, Google or Yahoo) using an application programming interface (API). An API is used to ‘hook’ data into Google or Yahoo maps.” (Crampton, 2009, p.93). These mashups allow non-professional users to create their own maps, frequently using free and open source software (FOSS), often leading to geocollaboration and crowdsourcing. An example of this would be using a service in Google Books that allows a user to input a book, and then the service will scrap all of the places from the book, and finally map all of the locations from the book. Another example would be Microsoft’s Photosynth, in which a user submits several photos of an object and the software stitches them together to create a 3D model of the building. These objects do not have a traditional cartography, but have a geographic sense that is built by the non-professional users who create them. Photosynth is a visualization technique that takes data that is not commonly viewed as geospatial, and orients it in a geographic fashion. In this thesis, web server information related to a particular keyword (which is not generally thought of as classically `spatial`) were visualized by their IP geolocations. The idea of using a visualization technique to expand on the gathered data is similar to Google Books and Microsoft’s Photosynth.

In addition to traditional data, collected by a specialized group or government entity, modern GIS can be built on volunteered geographic information, or VGI. However, is this data valid, or too inaccurate for professional use? Flanagin and Metzger (2008) discuss VGI and the credibility of the data. Credibility differs from accuracy as credibility is the perception of the data, not necessarily the quality of the data; credibility is much closer to the “subjective perception on the part of the information receiver.” (Flanagin and Metzger, 2008, p.141). Credibility can be influenced by the use of experts, whether part of creating the data or later validating the data. Alongside of experts, people also tend to trust data that has a larger number of contributors; users are discerning of data. One issue in the type of information is the presentation; people expect political blogs to be subjective, but tend to see maps as objective, possibly influencing the credibility. As a potential solution, Flanagin and Metzger note that user rankings may be the key to really finding what types of VGI are more credible to the general public. This thesis examines the data a step beyond credibility and
looking at accuracy. By doing a manual accuracy assessment, I will be able to see how frequently the CyberDiscovery tool is obtaining an IP address that matches the actual locations or places of the website content creators. Many view IP address information is credible, perhaps as it is highly technical and similar to maps, people tend to view the type of information as credible. The credibility will be tested by examining the geolocation accuracy. If the results of manual verification by relative experts show IP addresses to be fairly inaccurate, then perhaps the credibility may suffer.

The research presented here uses GIScience as the backbone of its theoretical framework. Utilizing spatial analysis, the registered locations versus the actual locations of web page servers were examined in order to identify patterns within the distribution. A GIS was needed to process the locational data, and to provide a framework for applying the geostatistical methods to the geospatial data. By applying the definition from Chang, and utilizing the third category of spatial statistics from O’Sullivan and Unwin, this research analyzed the accuracy of IP addresses, and examined different GIS techniques to pull patterns from the web server data. Potentially, this research will help us to better understand the causal mechanisms that underlie the diffusion of ideas and concepts on the World Wide Web.

2.2 - THE INTERNET, SEARCH ENGINES, AND WEB VISUALIZATION

One of the primary foundations of this research is the idea that real-world information diffusion could be detected and visualized by online search engines. While in the past, geographical information has been gleaned from methods such as surveys, field work, and remote sensing, the data from this project will revolve around web servers and the location of these servers, along with searches done via commercial search engines. There is an ample history of this proving relevant. Google Flu Trends is a tool that allows individuals to look at the frequency of searches, especially related to influenza-related materials. Using a set of particular searches, outbreaks and interest can be identified quicker than standard government reports (Ginsberg et al, 2009). The validity of using internet searches to gather
real world data has also been demonstrated in different sectors. The auto-industry, retail sales data, home sales, and personal travelling patterns can all be predicted faster than traditional reports by using search engine data (Choi and Varian, 2009). Since patterns of the World Wide Web exist due to real world people, those who exist beyond cyberspace, interpreting these patterns online can reflect previously unknown patterns of the real world. By expanding these basic trends into an entire geospatial cyberspace infrastructure, these patterns can become even more pronounced and identified. Cloud computing, the Semantic Web, and an expansion of network studies can help achieve this (Yang et al, 2010).

Validity is a key issue when embarking on a research project that is founded upon a method that has not been verified by an outside source, and has not been fully exposed to the greater scientific community. While search engine data has shown to be useful in the examples above, is there any reason to depend upon IP addresses as an appropriate method for locating web servers? In the absence of a non-automated research study on this topic, this thesis research will include a component of manual classification of IP addresses, comparing the IP address location to the web site stated location. The precise process will be outlined in the methods section.

There has been some research into the accuracy of IP address geolocations, especially into the accuracy of databases such as MaxMind (which is used in this study). Shavitt and Zilberman (2010) take an exhaustive look at several different IP address databases, and attempt to examine the accuracy of each. In order to assess, they use ‘PoPs’, or points of presence, as the ground truth for IP addresses, and check to see when multiple IP addresses are geolocated to the same PoP using each database. Their results for MaxMind show that the database tends to have a NULL classification for 36% of the IP addresses, which includes both situations where no results are returned and situations where MaxMind returns the coordinates of the country’s center (Shavitt and Zilberman, 2010, p.7). When looking at the MaxMind accuracy to a point, the MaxMind database has a “probability of 62% to 73% to place a IP within 40km from the PoP majority vote, with IPligence and MaxMind placing over 80% of the IP addresses within a 500km radius.” (Shavitt and Zilberman, 2010, p.9). This indicates that one should expect the IP address to be roughly within a city 62% to 73% of the time.
Other research has attempted to find more accurate methods for identifying IP addresses than using IP databases. Simple ‘ping’ methods involve timing the delay in sending packets of data between landmarks in order to get an idea of distance. More robust methods add a maximum distance value based on the maximum speed of data through fiber-optic cables and look for mechanical delays to refine the estimate (often called constraint-based geolocation). One method expands on this to create a statistical geolocation model that finds the delays between known landmarks and uses a kernel density map to estimate delays and refine geolocation, resulting in spatial accuracy levels as high as 77% of the IP addresses within 100km of the proper location. However, much of this work has not been implemented for the type of keyword association research used here, and the accuracy testing done in this thesis has been based off of a database of IP addresses. The visualization results will be relevant once the spatial accuracy of IP addresses has been improved, regardless of the method of improvement (Youn, Mark, and Richards, 2009).

Recognizing the potential errors inherent in IP address geolocation, Fink et al (2009) demonstrate a method for identifying Web log location information without IP addresses. Since a large number of the online blogs are hosted by a few companies, the IP address geolocation tends to place them all in a few locations. This misrepresents the actual spatial distribution of blogging activity. Fink et al developed a method for crawling the text of over 800 blogs and using the Stanford Named Entity Recognizer (NER) to pull all of the locations from the text. These locations were weighted, and the study was able to successfully geolocate 63% of surveyed blogs with only the written text (Fink et al, 2009).

The search engine is the primary vehicle for pinpointing web pages in this research. The best method for search engines to generate results is PageRank. PageRank is the method used by Google to rank the relevance of the search engine results given a particular search. By objectively ranking a page based on popularity, not on the subjective content, PageRank can replicate the probable needs of a user, along with the tendencies of a random web surfer. This ‘popularity’ is determined by hyperlinks, both hyperlinks pointing to a particular page, and hyperlinks by a page pointing to others (Page et al, 1998).

Now, the technology exists to go a step forward beyond simple popularity. By looking at the actual location of the site, spatial data is generated that informs on that particular city. People are located with their technology, and are embracing the wave of
location-based services (LBS). By searching for servers related to a particular subject matter, it is assumed that people nearby are interested in that subject matter (Yang et al, 2010).

PageRank methods are a modern method of looking at information diffusion. Early analyses were more focused on person-to-person transmission. This may have been either coincidental, as neighbors run into each other, or purposeful, as professional societies attempt to disseminate information to their constituents (Wolpert, 1966). Presently, these person-to-person interactions are often used for disease modeling, and modeling of how new transportation technologies affect a pandemic (Balcan et al, 2010). PageRank attempts to simulate the person-to-person transmission of ideas via hyperlinks (Page et al, 1998).

Chen et al (1996) discuss a way to improve search engine performance, in a way that will improve the speed of the search by cataloging certain concepts before running the search. Since the version of search engines in use in 1996 were very slow at examining each web page, Chen et al looked to build a system that would first categorize each homepage for a website, and then the search engine could begin by searching the categories. Borrowing from the self-organizing maps (SOM) concept (in particular, the Kohonen variant), the team looked to use machine learning techniques to appropriately categorize several concepts, moving one specialty down at a time. The technique was successful in small-scale applications and results suggest that it could be successful in larger-scale situations, but more refinement would be needed to fully implement this idea (Chen et al, 1996, p.100). As part of the manual classification process, the websites are categorized into one of twelve categories (such as blog, news, government, etc.). Within each category, the IP address accuracy and the overall spatial patterns in the visualizations is the focus. This categorization process is related to the idea from Chen et al in that this research involves taking a large problem and trying to break it into smaller pieces based on the semantic clues from the data.

Chow (2008) takes a look at how GIS analysts can use web and maps APIs (application programming interface) to create better Internet mapping applications. Defining a Maps API as “a source code interface that grants web developers access to a program library and to request services in generating a map over the Internet,” Chow emphasizes the services that are available via an API (Chow, 2008, p.180). With a Maps API, a GIS programmer can select an area of spatial data and embed it as a map, can control the viewing
angles, navigation, and in some APIs, even allow routing and 3D visualization. These services, along with functions such as spatial query and thematic mapping, allow a GIS developer to create a service faster and more efficiently than starting anew for every function. In order to display the advantages of a Maps API, Chow created a service that would convert a GIS database to GML (geography markup language), query the spatial data, overlay the data into a corresponding Maps API, and then display the map. At the end, while a Maps API could view the spatial data effectively, and XML was helpful to the process as it could call the GML-written geodatabase, the Maps API was more limited compared to what could be done with typical desktop GIS processing. Here, a Maps API will be used in order to gather the CyberDiscovery results. The Yahoo! API allows for me to obtain up to 1000 results (as opposed to Google’s limit of 64 results) and seems to produce better outcomes than the Bing API. However, the same limitations that Chow found also arise; in particular, the Maps API cannot be for complex spatial analysis as it does not have the functionality. While the Yahoo API can be used for gathering the IP address geolocation information, ArcGIS is used for the kernel density function. In addition to the Google API’s limitation on results, it also would just end with points on a map and limited control over complex visualizations.

There is some backlash in terms of this information being used by corporations and researchers. Generally, people see their activity online as semi-anonymous, given that they are not doing anything infamous and there are so many other people online. However, data on location can be gathered fairly easily via IP addresses, as is used in this proposal. Court cases have set precedent on the extent of privacy laws in the context of the internet, as in Macquarie Bank Limited & Anor v Berg and International League Against Racism & Anti-Semitism (LICRA) and the Union of French Jewish Students (UEJF) v. Yahoo! Inc. County Court of Paris. Issues surrounding privacy within web surfing need to be considered when approaching a proposal regarding mapping the locations of web servers. (Svantesson, 2005).

2.3- GEOSTATISTICS

The field of Geostatistics has existed before the rise in GIScience, and is considered to be in a different (though complementary) field altogether. It rose to prominence during the
spatial quantitative revolution of Geography in the 1950s-60s, and has had a profound effect on geographic scholarship since that point. Geostatistics finds the origin of many of its concepts based in classical statistics; for example, spatial autocorrelation began from a correlation coefficient, and was expanded into a spatial concept (Getis, 2008, p.300). Several geographers have expanded the mathematical formulations of different spatial statistics over the past 60 years, from Dacey’s spatial structures, Moran’s global autocorrelation measure, and Geary’s autocorrelation measure to Aldstadt and Getis’ spatial weights, and the work of Fotheringham, Brunsdon, and Charlton in geographically weighted regression (Getis, 2008).

There are several common spatial statistics within the discipline of Geography. Some examine patterns in attribute data, some create surfaces from point data, those that use basic arithmetic to create new types of data, such as buffering polylines, and many others. Spatial autocorrelation and cluster analysis are categories that look for patterns in the data. Spatial autocorrelation, whether a local or a global measure, is designed to find patterns of association due to spatial causes, and spatial causes alone (Anselin, 1995, p.94). Some implementations of cluster analysis are designed to look for hot-spots and cold-spots in the data, or areas where the data are clustered in high or low values. Within surface creation, some methods look to count frequency occurrences such as kernel density, and some look to interpolate the data values into a surface, such as inverse distance weighting and kriging. In addition, some methods look to create buffered lines for roads or rivers, or detect Boolean combinations of polygon datasets.

Cluster analysis has used a variety of algorithms in order to identify and understand clusters. Ng and Han (1994) discuss common clustering methods, and then present a clustering algorithm of their own. They review the PAM and CLARA methods, which work well for small and large datasets, respectively, and then present the CLARANS algorithm that uses randomized searching to identify the correct number of clusters and the members of a dataset that belong to a particular cluster. By using a randomized searching function, the system is able to perform clustering analysis much faster.

Geostatistics is not limited to just clustering equations; it also concerns the methods that surround cartography. Monmonier (1996) presents an entire book on how maps are frequently used to lie. By tweaking visual patterns with different classification methods and number of classes, vastly different patterns arise. In particular, Monmonier demonstrates
situations where political maps are used improperly. Without some sort of neutral
measurement, these patterns can lead to wrong conclusions. In one example, Monmonier
shows a German propaganda map from 1939 that has the German states in pure white and the
Allied states (Great Britain and France) in a solid black. This gives an impression of good
vs. evil that is completely at the hands of the map maker, and is completely visual
(Monmonier, 1996, p.100-101)

Spatial statistics are beginning to be seen in a new light as database management and
pattern extraction become key in a research dominated by large datasets. Computational
statistics are key to all data, and increasingly spatially-aware social media data. Lazer et al
(2009) discuss the change in computational social science. Instead of social scientists
interviewing individuals as data collection, there are now massive amounts of social data
being created each day, such as email interactions. There is also the potential for researchers
to take full advantage of social networking sites for delving into the day-to-day interactions
between friends and family members. However, Lazer et al also focus on the impediments to
fully taking advantage of this new world of data. In general, the database tools and analysis
techniques have not been fully realized. There is also a problem with needing access to the
personal information of actual people; a problem that is much less prevalent in sciences such
as physics and chemistry. Privacy will be an enormous problem when accessing this data,
but the potential gains of investigating this data will be worth the trouble of access (Lazer et

Nelson, in her recent article in The Professional Geographer (2012), outlined past trends
in spatial statistics, and some ideas for the future that she gathered from surveys taken by
prominent geographers and statisticians. There were some commonly held views, such as the
belief that GIS has caused a great change in the manner that spatial statistics are performed,
and that the digitization of data sources is important for sharing and processing data (Nelson,
2012, p.85). However, it is interesting that Nelson also points out the reversal in idea sharing
since the 1950s and 1960s; instead of geographers borrowing from physicists and
statisticians, geographers are now able to inform ecology, economics, social sciences, and
epidemiology (Nelson, 2012, p.86). Related to this research, one of the major future
opportunities in spatial statistics is to develop methods for communicating complex results in
a simple fashion, as this thesis looks to explain IP address locations in a meaningful manner (Nelson, 2012, p.88).
In order to complete the searching process, I am using the Visualizing Information Space In Ontological Networks (VISION) developed by Ming-Hsiang Tsou, Dipak Gupta, Jean Marc Gawron, Brian Spitzberg, and Li An (Figure 3). This system is designed to seek understanding of the spread of ideas by expanding current media models into cyberspace. While other methods for understanding human interactions and communications are limited to either person-to-person interactions, or newspaper/newsletter/television news communications, the VISION method is designed to understand how ideas spread through the medium of cyberspace. By focusing on the distribution of web servers supporting web pages containing the keywords of different ideas, the locational impacts of those keywords, and the spread of those ideas to different locations can be mapped (CDI Website, 2012).
While some previous projects have examined the spread of online information, the VISION method expands the use of advanced spatio-temporal analysis. By using commercial search engines such as Yahoo and Bing, the VISION prototype takes advantage of a previously-established tool for ranking websites based on popularity (and therefore, impact). Using GIS technology and real-world coordinate data, the information contained has a stronger link to how people are actually thinking and discussing ideas. The VISION model uses existing web pages as the source of information, as opposed to any user-submitted data. Within the VISION research framework is the assumption that people are creating web pages on their own (without being influenced by this project) and in areas that are nearby and relevant to them, and it follows that these web pages represent a more valid picture of the mindset of the population (CDI Website, 2012).

Above, in Figure 2, an outline of these methods is displayed. The procedure begins with the CyberDiscovery search results (discussed in Section 3.1), which are subsequently processed using the Manual Classification Process (discussed in Section 3.2). Each of four different data processing steps will be performed in order to draw out different visualization processes. After the maps for each process has been generated, the maps will be compared side by side in order to evaluate how helpful each result is in understanding the spatial signal.

Figure 3: The VISION framework. (http://mappingideas.sdsu.edu/?page_id=136)
3.0.1- List of Software and Data

For this thesis research, a few software packages were used. They are in Table 1 below.

**Table 1- List of Software**

<table>
<thead>
<tr>
<th>Software Name</th>
<th>Key Function</th>
<th>Cost</th>
<th>Developer/Company</th>
</tr>
</thead>
<tbody>
<tr>
<td>ArcGIS</td>
<td>Mapping/Spatial Analysis</td>
<td>Dept. License</td>
<td>ESRI</td>
</tr>
<tr>
<td>Excel</td>
<td>Table View – Cell Editing</td>
<td>Dept. License</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Firefox</td>
<td>Web Browser</td>
<td>Free</td>
<td>Mozilla</td>
</tr>
</tbody>
</table>

For this thesis research, several datasets were used, both for research purposes and base map data for visual context. They are in Table 2 below.

**Table 2- List of Data**

<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>Format</th>
<th>Data Source</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CyberDiscovery results</td>
<td>Varies</td>
<td>Excel Table</td>
<td>CyberDiscovery tool</td>
<td>Search tool results</td>
</tr>
<tr>
<td>Background datasets</td>
<td>8.53MB</td>
<td>Excel Table</td>
<td>CyberDiscovery tool</td>
<td>General Server Activity</td>
</tr>
<tr>
<td>County Shapefile</td>
<td>0.50MB</td>
<td>ESRI Shapefile</td>
<td>ESRI Data</td>
<td>World Polygon - 1994</td>
</tr>
<tr>
<td>US States Shapefile</td>
<td>0.20MB</td>
<td>ESRI Shapefile</td>
<td>ESRI Data</td>
<td>Contiguous US States</td>
</tr>
<tr>
<td>Top 30 Cities Shapefile</td>
<td>0.03MB</td>
<td>ESRI Shapefile</td>
<td>ESRI Data</td>
<td>US Population - 1990</td>
</tr>
</tbody>
</table>

3.1- Data Gathering- Searching Tool/Process

The backbone of the searching process is the CyberDiscovery searching tool. Using the Yahoo API, a python script takes the user-inputted keyword and mines the corresponding ranked search engine data. The top several hundred results are collected, and for each result a variety of information is gathered, including URL, IP Address, and Host Name. The raw
list is geocoded using another python script that utilizes the MaxMind geocoding database to match IP Addresses to the real-world location. MaxMind is a company that specializes in IP address geolocation and has developed a very large database of IP addresses with the corresponding geographic addresses. The final result of the CyberDiscovery tool is an Excel spreadsheet with web page data, locational data, and ranking data. This Excel spreadsheet can then be processed in a GIS. Using ESRI’s ArcMap, part of the ArcGIS software package, the spreadsheet data is plotted using the coordinate information. From this point, the kernel density function can be performed, creating an interpolated raster that displays the extent of the data, but also takes into account the rank of each web page. The statistical measures, listed below, are assessed on this raster.

3.1.1- Map/Data Projection

The kernel density function, and most of the spatial statistics discussed below are affected by the choice of projection for the data and map in use. The kernel density function uses a fixed, linear unit for determining the neighborhood for density calculation. Most surface generating techniques need some sort of linear measure as opposed to an angular measure. This means that it is more accurate to use a unit such as miles or meters which are more constant in a map projection than degrees which vary based on latitude. Together, this implies that the choice of projection, and therefore the linear unit, is important to the results of the analysis.

For this research, the Albers Equal Area Conic projection has been chosen. The primary reason for this choice is to maintain the principle of equal areas, implying that the map will display identical area measurements to earth measurements at the expense of distorting conformality. The Albers equal area projection is especially good at displaying areas that are longer east-west than north-south, as is the contiguous United States (Dent et al, 2009, p.55). By using an equal area projection, the kernel density will display area appropriately, and visual pattern examinations of the subsequent clusters will be comparable in different cities. The equal area projection is a better choice than an equidistant projection because in the latter projection, distance is only maintained from one point to all others, not all points to all points, which would increase overall distortion for the different clusters (Dent
et al, 2009, p.46). A blank map showing the base map data in the Albers equal area projection can be seen in Figure 4.

![Blank map showing the base map data.](image)

**Figure 4: A blank map showing the base map data.**

### 3.2- DATA PROCESSING – MANUAL CLASSIFICATION PROCESS

The first step within this process is to test the IP address geolocational accuracy manually. This has been a very labor-intensive process as each web page has been visited directly and examined for an address. In order to build a database that can compare the Maxmind registered IP address location and the web page stated location, the keyword has first been searched using the CyberDiscovery tool, and then every web page in the results has been visited.

For each web page, three different data points were gathered. First, the web page was categorized into one of twelve categories. By placing the web pages into certain categories, the web pages can be scrutinized to see if there are any patterns based on web page type. The categories are:

1. **Blog (blog + personal + group):** personal blogs or group blogs, usually ‘Blog’ has clear index about who are the writers and when the article was posted, along with personal web pages with strong opinions. ‘Blog’ is generally distinguished not only by the format, as described in the category descriptions, but by the existence of any
sort of publication. ‘Blog’ is also applied when it was a site that was produced by a popular blogging company- such as Wordpress or Blogspot.

2. **Commercial websites**: (e.g., selling products or selling services, or explaining information related to commercial products).

3. **Educational**: (e.g., schools, university, and educational institutes, usually within the EDU domain).

4. **Entertainment + video**: ‘Entertainment’ is used when there is only a video or series of videos available, not attached to an article, such as Youtube.com. ‘Entertainment’ is also used for obvious joke or humorous websites, such as The Onion and Uncyclopedia.

5. **Forum**: (websites with forum software and user accounts to post comments).

6. **Governmental websites**: (e.g., local, state and federal governments, usually associate with .gov, .us .org etc.).

7. **Informational**: (e.g., Wikipedia type, such as About.com, Wikipedia.org and other similar online Yellow pages).

8. **News**: ‘News’ generally indicates a certain website is part of an online news organization or a publication. (e.g., Local News site or National News - ABC, NBC, CNN, KUSI, SignonSandiego, etc.). Caveat: Some sites appear to be ‘News,’ but are coded as ‘Special Interest Group’ if it is less related to news reporting and more about postings about one specialized topic or subject.

9. **NGO**: (e.g., non-profit organizations, such as Red Cross, Rotary club).

10. **Social media website**: (e.g., such as twitter sites, Facebook, online forums), those contents are created by users directly. These are spatially accurate if the social media company is in the same location as the servers.

11. **Special Interest Groups**: (e.g., websites are created to promote specific concepts or items - such as political party, or issues).

12. **Offline**: (e.g., broken links, page not found).

The category has been assigned based on my exploration of the entire web site, not just the page that is linked in the results.

The second dataset collected is the relevancy/intensity of each site, along a pro vs. con continuum. If the web site was not related to the searched keyword, then the site received as score of ‘Not Applicable”. If the web site was related, then the site was coded based on a range from -2 (very negative) to 2 (very positive). The scores are:

- -2: Very negative web page regarding the keyword
- -1: Negative web page regarding the keyword
- 0: Objective, balanced web page regarding the keyword
- 1: Positive web page regarding the keyword
- 2: Very positive web page regarding the keyword
- N: (not relevant) Web page is not relevant to the keyword
- A: (ambiguous) Web page cannot be classified as positive or negative, but there is biased language

The key to this section is determining what constitutes a positive or a negative slant before beginning the classification process. In early testing, it seemed to be easier to do this for a person, or a partisan concept, than an object or less-inciting idea. This data is useful as it can compare the IP address accuracy for hyper-opinionated results with that of more neutral results.

The third dataset is the most important; the location comparison. Each site was visited and the mailing address of the headquarters recorded for every site. Generally, the first step was to look for the site information on the site itself first, such as on a ‘Contact Us’ page or in the ‘Privacy Policy/Terms of Service’ pages (an explicit-to-the-site address, street address is on the website). If this information could not be found, the process further investigated by using an outside source, such as Wikipedia (an implicit-to-the-site address, not actually on the website, but it is apparent that it is the correct street address). The websites location scores are:

- E-1: The content creator/writer, website company, and server are all in the same city; info found on website
- E-2: The website company and the server are in the same city; info found on website
  OR
  The content creator/writer and the server are in the same city; info found on website
- E-3: The website company and the server are not in the same city; info found on website
- I-1: The content creator/writer, website company, and server are all in the same city; info found outside website
- I-2: The website company and the server are in the same city; info found outside website
  OR
  The content creator/writer and the server are in the same city; info found outside website
• I-3: The website company and the server are not in the same city; info found outside website
• N/A: No geoidentifying information found, or server is at geographic center of United States, or the web page is offline

A similar location meant that the web server and the city listed in the IP address were within fifty miles of each other. Fifty miles is chosen as a cutoff point that should include all entities of a particular city. These codes allowed generalization statistics on the spatial accuracy of IP address databases compared to a manual classification of each website.

A total of six keywords were processed using this manual classification process:

• “Mitt Romney” – Search performed on January 4, 2012
• “Rick Santorum” – Search performed on January 4, 2012
• “Michael McGinn” – Search performed on November 17, 2011
• “Jerry Sanders” – Search performed on September 8, 2011
• “Flu” – Search performed on May 6, 2012
• “HPV Vaccine” – Search performed on February 3, 2012

These keywords were chosen in order to examine different domains, while still having a pair in order to look for domain-specific noise patterns. The first two keywords are political figures (both former US presidential candidates), the second two keywords are mayors in large US cities (Michael McGinn is the mayor of Seattle, WA; Jerry Sanders is the former mayor of San Diego, CA), and the next two key words are related to public health. By comparing the members of a pair, this research can better understand commonalities in the noise of a particular domain.

3.2.1- Kernel Density Function

The kernel density function is the method used to examine hotspots and the overall values in the data. While there are several options for creating a map from a series of points, the kernel density function is the primary one that was used. First, the kernel density function works well with overlapping points. Since there are several situations where
multiple points can overlay at the exact same location (multiple stories from a newspaper, multiple separate blogs from the same blog company), the kernel density function is a more proper function. Second, the function works well with the other types of analysis that will be performed on the dataset. The population field (a kernel density function parameter) can be changed, the number of records can be changed, or the post-density normalization can be changed. ArcGIS processes the kernel density function quickly, and the result is a data file that allows map algebra to be performed. Third, since the results of this research are to be applied to many outside of specialized spatial analysis, the resulting ‘hotspots’ allow for patterns to be identified quickly and easily. Finally, the kernel density allows for manipulation of both the radius of points that are calculated to create an expected density, and manipulation of the curve that predicts the significance of a point.

One of the major parameters in using the kernel density function is the radius of the searching kernel. The searching kernel is used to delineate an extent for the calculation of the density value. The value itself is determined by a specific curve, but the size of the extent is an easily manipulated variable. Different kernel sizes were tested to see which extent is proper for the scale of the web server data. Starting at 50,000 meters, every increment of 50,000 meters until 500,000 meters was tested. The value of the kernel radius is in map units, which means meters must be used.

3.2.2- Population Processes

Many of the tools and techniques used in this research have a common parameter: the ‘Population’. This generally refers to the frequency a point is counted when the algorithm is run. In practice, this parameter is generally used as a proxy for importance, popularity, emphasis, or a physically larger occurrence at this location. For the purposes of this study, population was used as a way of representing the rank order of the web pages based on search engine results. However, what is the best way to represent search engine results? How quickly do users stop clicking on web page links after the first result? This is important because if people do not see a website, then the opinions, views, and information contained in a website do not diffuse into the general society. In order to explore this question, several
different types of population representations were tested. A line graph comparison of all of
the methods can be found in Figure 6, assuming a sample number of 100 records.

3.2.2.1- No Population Value

This is the most simple method of dealing with population; simply ignoring it. Each
web page location was counted exactly once, with no concern to the rank of the web page.
The advantages of this method include ease of calculation and the results are not partial to
any popularity adjustment. Conversely, the disadvantages of this method include equating
the top result to a result buried dozens of pages in the rankings, which is unlikely to be
accurate.

Expression of each record: [1]

3.2.2.2- Inverse Rank Value

This method consists of taking the total number of records, assigning the value to the
top ranked record, and assigning values to subsequent records in a decreasing linear fashion.
In order for each record to have a value, I added one to the total number of records in
assigning the top value. For example, if there are 100 records, the top record will have a
value of 101, and the last records will have a value of 1. The advantage of this method is that
the rank of certain pages is taken into account. On the other hand, the disadvantage is that
this method is inflexible, and assumes the same drop off in value from the top record to the
second record and the 153rd record to the 154th record, which may not be true.

Expression of each record: ([Maximum Rank] + 1) – [Rank]

3.2.2.3- Logarithm of Rank Value

This method weighs each record based on the logarithmic value of each ranking and
the maximum ranking, and applies the value to each record. For example, if there are 100
records, the top and last records would have values of:
First record: \( \log(100 + 1) – \log(1) = \log(101) – \log(1) = 2.0043 – 0 = 2.0043 \)

Last record: \( \log(100 + 1) – \log(100) = \log(101) – \log(100) = 2.0043 – 2 = 0.0043 \)

The advantage of this method is that it causes the population value assignment curve to be gentler, especially later in the dataset, and increase the relative relevance of the first page. The disadvantage is that this model may not value the later pages enough, and there is no real justification to use the logarithm of the ranking as opposed to the natural logarithm or any other slope.

Expression of each record: \( \log(([\text{Maximum Rank}] + 1)) – \log([\text{Rank}]) \)

3.2.2.4- Fuzzy Values

A major problem with the previous population methods is that they do not account for the idea that most search engine users rarely leave the first page of search results. Even when users do leave the page, they do not generally go very deep into the results, and very rarely all the way through the results. In order to have a method that can partially represent this, fuzzier categories were attempted using a stair-step form. Records with ranks ranging from 1 – 10 were assigned a value of the maximum rank plus one. Records with ranks ranging from 11 – 50 were assigned a value of the maximum rank minus nine. Records with ranks ranging from 51 – End were assigned a value of the maximum rank minus forty-nine, minus the rank value by the fifty. For example, if there were 100 records, the values would be as follows:

First record: 100 + 1 = 101
Twelfth record: 100 – 9 = 91
Fifty-Second record: (100 – 49) – (52 – 50) = 51 – 2 = 49
Last record: (100 – 49) – (100 – 50) = 51 – 50 = 1

The advantage of this method is that it allows for special emphasis on the first page of results, and some extra emphasis on the top pages of results without adding undue influence
on the buried results. The disadvantage is that we do not know how much to emphasize the top page. Perhaps this score gives too much or too little weight to the first page.

The ideas of this sort of model are similar to the methods the GIS software IDRISI uses to calculate weights for its decision support system. In particular, the fuzzy calculation here is similar in theory to the second graph in each row in Figure 5.

Figure 5: Models of fuzzy classification from the GIS software IDRISI from Clark University
3.2.3-Normalization/Standardization Processes

The current distribution of web servers across the United States is not equal, as would be expected. Just as there are urban and rural variations from region to region, there are areas of dense server farms and areas where there is no server for miles. Early testing has shown San Francisco, CA as an area where there are a disproportionately high number of web servers, probably due to the proximity to ‘Silicon Valley’. When looking for patterns, these variations need to be accounted for and the data needs to be adjusted. Without normalizing the data, the significance of a result may not be easily identified. Two different
methods for normalization were tested. These have been previously studied in GIS Decision Support Systems literature, but there may be an application to cyberspace research.

3.2.3.1- Maximum Score

The maximum score method is a linear method that produces a proportional range. This is done by basing the significance of a web server probability on its relation to the maximum value in the dataset. The technique involves dividing the calculated surface by the maximum value in the surface, and then subtracting a created background surface that has been divided by its maximum value. The background surface has been developed by searching 306 common words, randomly generated, and creating the surface in the same manner as the search results. This method will create a range from a potential negative one to positive one, although the actual range is determined by the data. If the highest point the searched surface is also where there is zero background activity, then the maximum score will be one; however, this is unlikely, and the score tends to range from closer to negative 0.5 to positive 0.5. (Malczewski, 1999, p.117)

Expression: \[
\frac{( \text{Keyword Surface} - \text{Background Surface} )}{\text{Maximum Keyword Surface}} - \frac{\text{Maximum Background Surface}}{\text{Maximum Background Surface}}
\]

3.2.3.2- Score Range Procedure

The score range procedure looks similar to the maximum score method, but creates different results. Instead of a linear scale, the score range produces a non-linear scale that stretches the results to a range that is exactly negative one to positive one when the minimum is not zero. While this is generally more practical for data that has a non-zero minimum value, the procedure may create a new surface that is significant. The score range takes the same expression as the maximum score, but subtracts the surface from the maximum value from every location. (Malczewski, 1999, p.118)

Expression:
(Max Keyword Value – Keyword Surface) - (Max Background Value – Background Surface)
Max Keyword Value – Min Keyword Value) Max Background Value – Min Background Value)

3.2.4- Web Page Categories

As part of the manual classification process described above, the type of website being explored has been determined, placing it into one of twelve categories. This created an interesting dataset that allowed this research to examine each category individually, and groups of categories. The first step was to create a surface for each type, and compare the results. The results are compared to each other, and then to the surface of all of the records.

The next step was to decide which groupings of categories would be significant for seeing the underlying server location patterns. For this, a framework that separates ‘public’ sites from ‘private’ sites was proposed. Public sites are those that are created by public institutions, and hypothetically have the interest of the user or citizen as the top priority. This includes the categories: “Government”, “Education”, and “NGO”. Private sites are those that are created by either private companies or individuals, and hypothetically have either an interest for themselves, their reputation, or a clouded interest. This includes categories: “Commercial”, “News”, “Informational”, “Special Interest Groups”, “Blog”, “Social Media”, “Entertainment”, “Social Media”, and “Forum”. Each of these categories relates closer to the market than the public sector. For this framework, the websites in the category “Offline” have been discarded as this information is not useful.

The second framework attempted to look at differences between an individual and a group. Individual-based sites are those that are primarily posted by an individual, without concern or required affiliation to a larger group. This includes categories: “Social Media”, “Blog”, and “Forum”. Group-based sites are those that are sponsored by a larger group, or are inherently based around more than one person’s content creation. This includes categories: “Special Interest Groups”, “Government”, “Education”, “Commercial”, “News”, “NGO”, “Informational”, and “Entertainment”. This framework could be less robust than the first framework. Since some organizations sponsor blogs, and some news sites are the work
of a single person, there may be some situations where the website fits the category but not the framework.

3.3- **Evaluation**

The evaluation was done by comparing each of the resulting maps for a keyword with each other. By examining the differences between the maps, the spatial patterns can be seen more clearly. Since the particular keywords have certain dates, the keyword subject can be compared with the properties of the subject and the temporal relationships to the concept. For example, one of the keywords is the 2012 US Republican primary candidate Mitt Romney. The keyword search was performed on January 4, 2012, on the date of the Iowa Caucuses. In general, the pattern emphasizes Salt Lake City, Utah, as early testing has shown that the candidate has support here, and activity near Iowa can be shown, since the candidate narrowly lost (and was announced as the winner on the night of the caucus). There is also some activity near Massachusetts given that Mitt Romney was the Governor of Massachusetts from 2003 – 2007.

The maps were compared and the overall results of the manual classification process were explored. By pulling out the statistics for each category of web page, and by looking at the rates of positive and negative web pages, the properties of IP address geolocation accuracy can be better understood. It is interesting to know, for example, which categories tend to have higher accuracy rates, which could be used in future research to improve IP addresses geolocation practices.
4 RESULTS

4.1- MANUAL CLASSIFICATION RESULTS

4.1.1- Web Page Categories

As discussed in Section 3.2, the Manual Classification Process was used to take a more in-depth look at the CyberDiscovery results. The first part of the process was to assign a category to the web page from a domain list of twelve categories. The results from this can be seen in Figure 7.
Figure 7: Pie charts showing the percentages of each dataset for each category of web page type.
As can be seen above, “News” websites tended to dominate the web pages, with “News” occupying the highest percentage in four of the six keyword choices. For the two presidential candidates, “Mitt Romney” and “Rick Santorum”, the category percentages were similar. However, this did not hold true for the other pairs. More on this subject will be discussed in Section 5.1.1 of this thesis.

4.1.2- Relevancy/Intensity Codes

As discussed in Section 3.2, the Manual Classification Process was used to take a more in-depth look at the CyberDiscovery results. The second part of the process was to assign a code related to the relevancy/intensity of the web page in relation to the keyword. The results from this can be seen in Figure 8.
Overall, most of the results pages had a high number of “0” or even levels of intensity toward the keyword. This could be due to the intended impartiality of the news media. However, the two mayors’ keywords, “Michael McGinn” and “Jerry Sanders”, had a high number of irrelevant web pages, perhaps reflecting a lack of presence in the national consciousness. More on this subject will be discussed in Section 5.1.2 of this thesis.

4.1.3- Geolocational Accuracy

As discussed in Section 3.2, the Manual Classification Process was used to take a more in-depth look at the CyberDiscovery results. The third part of the process was to assign a code related to the geolocational accuracy of the web page relative to the IP address geocoded result. The results from this can be seen in Figure 9.
There is a startling consistency among the six keywords, with no keyword having a geolocational accuracy above one-third of the total number of webpages. This is in stark contrast to the earlier estimates near 62% to 73% from the literature (Shavitt and Zilberman, 2010, p.9). The percentage of spatially accurate web pages for the two presidential candidates is nearly identical, near 20%. More on this subject will be discussed in Section 5.1.3 of this thesis.

4.2- Kernel Density Maps

By taking the manual classification results and selecting out only those records that had a score of “1” or “2”, a dataset of only spatially accurate results was created. This was done with all six keywords. In all, each dataset was processed with four different population
calculations, and run through the kernel density function with twelve different kernel radii, and then normalized with two different standardization processes. Since there were six datasets that were put through this process, it resulted in 864 different maps. In the interest in the length of this paper, a sample of the kernel density maps will be shown here, and more will be shown in Section 5 to illustrate particular points. Red areas have a higher than average probability of server activity, and blue areas have a lower than average probability for server activity.
Figure 10: Examples of the processed keywords. All maps created using the logarithmic population, a kernel radius of 200,000 meters, and the maximum score normalization.

As can be seen, each of the six keywords has an individual overall pattern to them. This can be thought of as a particular geospatial fingerprint, as different search terms can result in different patterns. While the fingerprint (cyberspace activity) for a keyword is what
it is, the different kernel density parameters can be thought of as the different materials used to record a fingerprint, whether it is graphite, iodine, wax, or charcoal. Part of the goal of this research is to identify an ideal method for recording and visualizing the cyberspatial patterns.

In Figure 10, some cities tend to show up more than others. Cities such as Washington, DC and New York, NY tend to be higher than average. This may be due to a higher number of spatially accurate servers in these cities than average. For instance, servers from the “Government” category are more likely to be in the location assumed, as governments (federal and local) tend to maintain control over their servers as opposed to leaving them with a private company. Therefore, Washington, DC, the center of the federal government, would tend to have more servers than expected for the area. This phenomena also works in the opposite direction, as San Francisco, CA/San Jose, CA tend to be lower than average on the maps. This could be due to the large number of blogs that reside in the cities. Both Wordpress and Blogspot, two of the largest online blog communities, are based in San Francisco/San Jose. This means that IP addresses track all blogs to this area, despite the blog authors living all over the United States or the world. More on these subjects will be discussed in Section 5 of the thesis.
5 DISCUSSION

5.1- MANUAL CLASSIFICATION RESULTS

5.1.1- Categories

For the "Mitt Romney" dataset the category with the highest number of records was "News", with 42.66%. The second highest category was "Blog", with 24.80%, the highest fraction for the "Blog" category of all six keywords. This is not surprising as at the time of data collection (January 4, 2012), Mitt Romney was a well-known Republican primary candidate with a high level of publicity. It would follow that a large number of news stories would be written about the candidate and that a large number of blog authors would post entries related to the favorite to win the Republican nomination.

For the "Rick Santorum" dataset, the category with the highest number of records was "News", with 47.34%, the highest fraction for the "News" category. The second highest category was "Blog", with 20.72%. As with Mitt Romney, this could be expected as Rick Santorum was also a well-known Republican primary candidate. Interestingly, the category breakdown for the two Republican candidates was very similar. They were both dominated by "News" and "Blog" web pages, with smaller groups of "Special Interest Group" and "Informational" web pages.

For the "Michael McGinn" dataset, the categories were much more evenly split. The category with the highest number of records was "Special Interest Group", with the second highest category being "Informational". The best explanation for the category breakdown may be related to the relevancy/intensity section. Since Michael McGinn is only regionally known, many of the results were for others who shared the same name. Since the category breakdown does not take this into account, the results may not be meaningful.

For the "Jerry Sanders" dataset, the category with the highest number of records was "News", with 31.37%. The second highest category was "Blog", with 23.62%. The breakdown was similar to the Mitt Romney dataset, although with a higher number of "Commercial Websites". This may have a similar explanation as the McGinn dataset, as the Sanders dataset also collected web pages that are not related to mayor (albeit on a smaller scale).
For the "Flu" dataset, the category with the highest number of records was the "Governmental", with the second highest category being "Commercial Websites". As many of the top web pages related to the flu are part of the US CDC (United States Centers for Disease Control), the "Governmental" web pages are the top result. Regarding the high number of "Commercial Websites", there are a large number of private companies that provide cold and flu remedies, whether actually curative or not.

For the "HPV Vaccine" dataset, the category with the highest number of records was the "News" category, with "Blog" narrowly holding the second highest spot. "Special Interest Group", "Informational", and "Offline" are all close to "Blog", each containing approximately a tenth of the dataset. While it would be expected that the HPV vaccine results would be similar to the flu results, the increase in "News" web pages could be related to the fact that the HPV vaccine was very controversial when released. This controversy created a large amount of media attention that overshadowed the government affirmations of the vaccine.

Of all three pairs of keywords, there was the most consistency between the "Mitt Romney" and "Rick Santorum" datasets. There could be several explanations for this phenomenon. First, they may be the most similar of the sets of pairs. Both candidates are national figures, former elected officials, belong to the same party, are relevant to the same areas, and are frequently associated together. This is opposed to Michael McGinn and Jerry Sanders who only really share the same office; they have different policies, locations, and levels of notoriety. Similarly, while the flu and the HPV vaccine are both public health concerns, they are fundamentally different entities; the flu is a disease while the HPV vaccine is a vaccine to prevent a disease. Second, the similarity could be related to the level of publicity. The two candidates are nationally relevant, while the two mayors are primarily known in only their own cities. This explanation does not account for the public health keywords as they are both well known. Finally, this may be due to bias within this research. The two candidate searches were performed on the same day, while the two mayor searches were over two months apart and the two public health searches were over three months apart.
5.1.2- Relevancy/Intensity

The results related to relevancy/intensity do not appear to have the same impact as the results of the categories or the georelations. Since these results are only the particular slant of the issue, they will most likely only be useful in explaining as an accessory to the categories or the georelations.

For the "Mitt Romney" dataset, the overwhelming majority of the web pages were "0", or neutral. There were more "-1" results than the "1" and "2" classes combined, slanting the overall dataset negative, but this is fairly slight since the dataset is primarily even. The "Rick Santorum" dataset has a very similar distribution. The top category is "0" and there are more "-1" results than "1" and "2" combined. Perhaps this is the result of these two candidates being dominated by "News" web pages. Overall, there seemed to be only a few extreme results, with the combination of "-2" and "2" classes resulting in only 10% - 15% of the results.

For the "Michael McGinn" dataset, the most striking feature is the high level of "N" websites. This implies that very few of the collected web pages are even related to the mayor. While the name may be part of the web page, the reference is to a different Michael McGinn. Once the "N" class is set aside, the "0" class again dominates. When looking at the "Jerry Sanders" dataset, the "N" class is also much higher than normal. While not as high as the McGinn dataset, it still accounts for nearly 40% of the distribution, nearly four times as much as the next largest "N" class (10.96% for "HPV Vaccine"). Interestingly, the next highest category for the Sanders dataset is the "2" category, or strongly positive toward the mayor.

For the "Flu" dataset, the highest category is "0", just as it was for the Romney and Santorum datasets. There is also a large number of "2" positive web pages. Perhaps this is due to the categories that make up the flu dataset. The flu dataset had the highest number of "Commercial Websites" and "Governmental" websites. If these web pages tended to back commonly accepted scientific ideas on the flu, whether to educate the public ("Governmental") or to sell the public successful remedies and doctors to visit ("Commercial Websites"), then this could explain the high number of positive results. For the "HPV Vaccine" dataset, the numbers are much more similar to the presidential candidates than the flu dataset, with "0" dominating and the other classes fairly split. Since there was a lot of media controversy over the HPV vaccine, the web pages may reflect more of a debate that
resembles a political election. While the uncontroversial flu data allows for a higher number of positive results, the backlash against the HPV vaccine leads to a split situation.

Overall, the biggest impact on the relevancy/intensity data seems to be the amount of national saturation. The four topics that are nationally relevant ("Mitt Romney", "Rick Santorum", "Flu", and "HPV Vaccine") have similar distributions to each other while the local topics ("Michael McGinn" and "Jerry Sanders") have similar distributions to each other different than the national group. This tie to publicity would need to be further researched by examining local phenomena that are not mayors, perhaps laws particular to a city or state or a college/university that does not have a nationally-known sports team. When first accounting for publicity, there also could be differences based on controversy. Disputed issues have a similar plot after accounting for publicity. This would need further research by performing the manual classification process on low-debate terms.

5.1.3- Geolocation

For the "Mitt Romney" dataset, the results clearly indicate that there is a large disconnect between the percentage of spatially accurate results and the overall dataset. Only about 20% (21.43%) of the web page IP address registrations were actually within 50 miles of the web page content creators. 46.62% were definitively in the wrong location, and 31.94% are unknown. The low spatial accuracy could be partially the result of a high number of “Blogs” and “Special Interest Groups”, as discussed in Section 5.1.4.

For the "Rick Santorum" dataset, the results are similar to the Romney dataset as the spatially accurate results are just over 20% (21.29%). Again, there is a large disconnect between the spatially accurate results and the complete set of IP addresses. However, the unknown percentage is lower (25.86%) than the Romney dataset, with virtually all of the records moving to the spatially inaccurate class (52.85%). Over half of the records that would typically be used as data points are ultimately wrong. As with the Romney dataset, the low spatial accuracy rate could partially be the result of a high number of “Blogs” and “Special Interest Groups”, as discussed in Section 5.1.4.

For the "Michael McGinn" dataset, the results show the highest percentage of spatially accurate web pages of all of the six search terms. The spatially accurate class is at
33.57\%, the only dataset to have at least one third of the records classified as accurate. However, 48.56\% of the records are spatially inaccurate, a higher percentage than the Romney or Flu datasets. With 554 total records classified, the second largest dataset, and an unknown percentage of 17.87\%, the second lowest unknown, the McGinn dataset may be the best results proponents of IP address geolocation can point toward. The explanation for the higher spatially accurate percentage may be related to the relevancy/intensity and categories percentages. Since many of the web pages in the McGinn dataset were not related to the mayor and instead to other Michael McGinns on other businesses, the percentages of category spatial accuracy need to be taken into account (Section 5.1.4). In particular, the low number of “Blog” results may have led to a higher spatial accuracy.

For the "Jerry Sanders" dataset, the spatially accurate class is similar to the Romney and Santorum datasets at 23.25\%. However, the spatially inaccurate class is much larger, at 60.33\%. As the unknown class (16.42\%) is the lowest of all six datasets, the Sanders dataset may provide the best approximation for spatial accuracy. Relative to the McGinn dataset, there does not seem to be a spatial accuracy connection between local topics as there was with the relevancy/intensity subject.

For the "Flu" dataset, the spatially accurate percentage is near the others at 26.25\%. This is slightly higher than the presidential candidates and Sanders, but lower than McGinn. With the largest unknown class of all six datasets, at 35.52\%, conclusions about the lowest spatially inaccurate class (38.22\%) are not robust. As can be seen in Table 3, there is a lower standard deviation in the spatially accurate class than the spatially inaccurate class, implying that the spatially accurate class is more compact.

For the "HPV Vaccine" dataset, the spatially accurate percentage is in line with the other results at 27.47\%. This is the second highest result, though still just over one fourth of the entire dataset. The spatially inaccurate class is large, reaching over half of the dataset at 52.01\%. These results are more robust than the flu results as the unknown class is much smaller at 20.52\%. 
Table 3: Geolocation Summary Statistics

<table>
<thead>
<tr>
<th></th>
<th>Average</th>
<th>Median</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Spatially Accurate</strong></td>
<td>25.54%</td>
<td>24.75%</td>
<td>4.26%</td>
</tr>
<tr>
<td><strong>Spatially Inaccurate</strong></td>
<td>49.77%</td>
<td>50.29%</td>
<td>6.72%</td>
</tr>
<tr>
<td><strong>Unknown</strong></td>
<td>24.69%</td>
<td>23.19%</td>
<td>7.11%</td>
</tr>
</tbody>
</table>

For the entire six datasets, the average spatially accurate percentage is 25.54%. The average spatially inaccurate percentage nearly doubles this value at 49.77%. However, all conclusions must be noted with the unknown average of 24.69%. These values are not much different than the median values, with the spatially accurate median at 24.75%, the spatially inaccurate median at 50.29%, and the unknown class median at 24.69%.

Overall, there are a few patterns than can be seen from the dataset. First, the consistency between the presidential candidates remains. The spatially accurate class differs by only 0.14%. While there is a smaller amount of unknown web pages with the Santorum dataset, the change happens almost completely within the spatially inaccurate class. If the manual classification process were to be performed on another candidate from the time period, such as Ron Paul or Newt Gingrich, the results would probably be the same. Second, other than the McGinn dataset, there seems to be a cap on the spatially accurate category near one fourth of the dataset. While the unknown category may increase or decrease, the change only happens in the spatially inaccurate class, indicating that there may be a large percentage of spatially inaccurate web pages in the unknown category. Third, there is not a large difference between the mean and the median of the geolocation results. This implies that there is a high level of consistency within the results, and rejects any conclusion that these results are the influenced by a few very high or very low values.
5.1.4- Spatial Accuracy of Categories

As would be expected, different categories have different rates of spatial accuracy. A full breakdown can be seen below in Figure 11. The percentage in Figure 10 for “Offline” is zero as all offline sites were categorized as “N/A”.

The most spatially accurate categories are “Educational” (73.86%), “Social Media website” (68.97%), and “Governmental” (60.98%). This generally makes sense, as these are categories that tend to emphasize data security over convenience. Most major universities keep server management in-house, which would lead to IP address registrations that are at the university. For social media websites, most of the social media websites we picked up were major companies that manage their own servers (Facebook, Twitter, LinkedIn) as opposed to renting server space somewhere offsite. Governments would seem to be the most concerned with data integrity as they manage large amounts of very personal data (social security numbers, tax information), both at the state and local level.

The least accurate categories are “Blog” (10.81%), “Special Interest Group” (12.81%) and “NGO” (20.93%). This also is intuitive, as blogs are one of the easiest web pages to...
purchase. Most blogs that were gathered are hosted by WordPress or Blogspot, which are both based in the San Francisco, CA/San Jose, CA area. Since only a small percentage of the entire country lives in this area, only a small number of blogs can be spatially accurate. Special interest groups are similar as domain names are relatively easy to buy, allowing any group with a goal to set up a basic website that is hosted by another major company. Since many of the special interests groups gathered here were only of moderate size/influence, they did not have the need to manage their own servers. The NGO percentage is more surprising. The explanation for this may lie in methodological bias, as the delineation between “NGO” and “Special Interest Group” is relatively vague. The result may also be due to intermediate sized NGOs that are large enough to function as an NGO abroad, but are small enough that server maintenance is out of the budget.

5.2- **KERNEL DENSITY MAPS**

5.2.1- **Population Differences**

As discussed in Section 3, four different calculations were used for the population variable as part of the kernel density function. Each of these four methods attempts to establish a value for search engine rank in a different way. This results in moderate differences in the locations of hot spots as value of a server in a particular city may vary. This can be seen in Figures 12, 13, and 14, where the inverse population is in the top-left, the no population is in the top-right, the logarithmic population is in the bottom-left, and the fuzzy population is in the bottom-right. As discussed above in section 3, the graph colorings are based on the relation of the dataset to the background normalization. Red areas are where there is a higher probability of a server related to the dataset keyword than the background average, and blue areas are where there is a lower probability of a server related to the dataset keyword than the background average.
Figure 7: From the “Mitt Romney” dataset, all four population calculations at a kernel radius value of 150,000 meters. The normalization is maximum score.
Figure 9: From the “Michael McGinn” dataset, all four population calculations at a kernel radius value of 150,000 meters. The normalization is maximum score.

Figure 8: From the “HPV Vaccine” dataset, all four population calculations at a kernel radius value of 150,000 meters. The normalization is maximum score.
Overall, the differences between the four population results are minor. In each of the three figures above, the two cities with the highest probabilities for each keyword are still the highest in each of the four population values. Any of these population values will return the most important results from the kernel density function. However, there are some differences outside of the top two cities. While these locations may have some relevance, it must be remembered that there can be some randomness to IP addresses, even those that are spatially accurate. Given this, the main focus should be the major hot spots that appear in the kernel density surface, and in the interest of clarity, an optimum method would reduce these unnecessary results.

The maps with no population have the highest amount of spurious results. In both the McGinn results and the HPV vaccine results there are a series of hot spots along the border of the United States and Canada, especially near Toronto and Montreal. These results are due to having a few search engine results that are low down on the list. With the no population maps, these results have as much importance as the top results in major cities, as opposed to the other population methods that properly devalue these lower results.

The maps with inverse population and fuzzy population tend to have less spurious hot spots, but there are still some locations that have faint results. In the HPV vaccine dataset, the inverse population map has some minor results near Milwaukee, MN and Albany, NY that could be confusing. Similarly, in the McGinn dataset, there are a lot of mixed colorings in the Midwest region that can be confusing.

For the logarithmic population maps, many of these partial or indistinct patterns have disappeared. The result is a much cleaner map that conveys the clearest patterns in the dataset. The deepest hot spots and cold spots shine without interference. This can best be seen in the McGinn dataset as the logarithmic population map eliminates most of the confusion in the Midwest while also eliminating the noise near the border of Canada. When compared to the other population methods, the logarithmic population is the optimum method to create a clear, easily interpreted map. This remains true for all domains- not just public health or city mayors; it could work for searches related to climate change.
5.2.2- Kernel Radius Sizes

As discussed in Section 3, twelve different kernel radius sizes were used for each dataset. The searching kernel is used to delineate an extent for the calculation of the density value. I tested several different kernel sizes to see which extent is proper for the scale of the web server data. Starting at 50,000 meters, I tested every increment of 50,000 meters until I reach 500,000 meters. The main balance in the kernel radius size is between having a kernel large enough that the phenomena can be seen and intensity levels can be viewed, but small enough to be able to visually distinguish the differences between nearby cities. As an example, Figure 15 is the complete set of kernel radii for the “Mitt Romney” dataset, all with logarithmic population.
As stated above, the first step is to establish a minimum kernel size that can be easily seen with intensity variations that can be viewed. Looking at Figure 15, the first map (row one, column one) does not really show the phenomena. While there are a couple of red dots, the blue areas do not show up. The second map (row one, column two) shows both red and blue areas, and shows intensity levels. Therefore, the minimum size for a kernel radius should be 100,000 meters, using this projection.

The second step is to establish a maximum kernel size that does not obscure cities. Overall, the eastern seaboard from Washington, DC through New York, NY provides a range of four cities that are only about 225 miles long, allowing the best comparison on the map.

Figure 10: From the “Mitt Romney” dataset, each of the different kernel radius options. All maps are using the logarithmic population and the maximum score normalization.
Also, the five Texas cities of San Antonio, Austin, Houston, Fort Worth, and Dallas create a triangle in which the farthest cities of Dallas to San Antonio are only about 275 miles. If the surface colorings are indistinguishable between these groups of cities, then the kernel radius is too high. Looking at Figure 15, the eastern seaboard cities start to get muddled around 250,000-300,000 meters, and the Texas cities are all covered in blue by 300,000 meters. This indicates that the maximum kernel radius is 200,000.

At this point, the optimum kernel radius size is between 100,000 and 200,000 meters. Of the remaining three options (100,000; 150,000; 200,000), the radius of 100,000 leaves the smallest room for delineating different intensity levels, so it should be eliminated. However, these visual methods cannot go any farther than this point. When examining all six datasets, there was no consistency in whether the 150,000 or the 200,000 was a better representation. The optimum choice between these two is situational.

5.2.3- Normalization/Standardization Methods

As discussed in Section 3, there are two normalization techniques examined in this research: the maximum score and the score range. Figure 16 shows the direct comparison between the maximum score and score rank procedures.
As can be seen in Figure 16, the differences between the two techniques are more minor than the kernel radius differences. However, as can be seen in the Romney and Santorum datasets, there are differences that can be seen. For the Romney dataset, the hot spot near Washington, DC is less concentrated, revealing a much larger distinction between Washington, DC and New York, NY. This would make it easier for a map user to quickly distinguish the “hottest” spot in the map. Similarly, in the Santorum dataset, the score range procedure reduces the intensity of Atlanta, GA and Miami, FL, drawing the map user’s eyes to the “hotter” spots of Washington, DC and New York, NY.
In three other datasets, the score range procedure just reduced the strength of both hot and cold spots throughout the entire dataset. For the McGinn, Sanders, and HPV vaccine datasets, the overall pattern does not become any more clear or less clear, but the vitality of the visualization is lessened. This is a neutral result for the method, as a user can still see the most important spots on the map. For the flu dataset, there is no perceptible change between the two methods.

Generally, the difference can be summarized as a reduction in color intensity with the score range procedure as opposed to the maximum score procedure. This makes sense, as the score rank stretches the distribution as it normalizes against a background. The stretching could cause a larger relative space between the intensity of hot spots/cold spots, creating the less forceful visualization. With the six examples above, using the score range procedure improved clarity for two of the keywords (Romney and Santorum) and had a neutral impact on the other four datasets (McGinn, Sanders, flu, and HPV vaccine). The score range procedure is the optimal normalization technique with this type of data.

5.2.4- Category Kernel Densities

As discussed in section 3, each category is separated and a kernel density map is created from the spatially accurate results. Figure 17 displays all of the 12 categories, except “Offline”, as offline websites were assigned a geolocation code of “N/A”.

![Map of Blog and Commercial Websites](image-url)
The most interesting part of comparing these categories is that there is an actual difference between all of the maps. This points to a spatial connection within the type of web page. While this would have to be tested using a new set of manual classification results, there are some intuitive reasons for why these maps may be reflecting the real world.

First, the map displaying “Governmental” web pages has hot spots in Washington, DC and Seattle, WA. As Washington, DC is the center of the United States federal government, it would be expected that there are a lot of government-related servers near the city. Since one of the keywords is the Seattle mayor, there is also a large number of government-related web pages included in the McGinn dataset.
Second, the focus on San Jose, CA for both the “Social Media” map and the “Entertainment” map is probably linked to Silicon Valley. Most of the major social media companies (Facebook, Twitter, and LinkedIn) are based in this area, leading all of the social media IP addresses to these servers. Google, who is based in Palo Alto, CA, owns Youtube.com, the most common web page for the “Entertainment” category.

Third, the “Educational” and the “Commercial Websites” maps appear to have a large number of hot (or “warm”) spots throughout the entire country. This would make sense as both of these categories are spread about evenly though the country. As opposed to the “Governmental” web pages that have a real world hub (Washington, DC), there are schools, colleges, and universities throughout the entire country. In the same vein, there are medium to large businesses with websites throughout the entire country. The “Special Interest Group” map also has a large presence throughout the south and the east, which could be reflective of the wide variety of interests among people throughout the country. It does not take a lobbying campaign in Washington, DC to organize a group of people and create a website; however, there is a cluster around Washington, DC, as it is the top result in the map.

There are also some patterns that are harder to explain. The east coast spread of spatially accurate “News” web pages does not seem to follow any reasoning. This is especially true as New York, NY, the most concentrated hot spot, does not benefit from The New York Times as the Times has several spatially inaccurate servers (often clustering in Denver, CO). A similar problem occurs with the “Informational” map, as there is no reasoning to have such a high cluster in New York, NY and a cold spot in Washington, DC.

A few of the explanations for these maps could be related to inadequacies of this research. The “Forum” map has a cluster in Salt Lake City, UT. This is due to the fact that there a large number of results from the website Ancestry.com located in this area. Many forum postings were about the last name “McGinn”; completely unrelated to the Seattle, WA mayor. If there were more manual classification results, this pattern would likely disappear.

For the “NGO” map, the map extent does not go farther south or west of Washington, DC. This is due to the kernel density map not going farther than the data extent. As there were no spatially accurate NGO servers south or west of Washington, DC, the map is incomplete.

The “Blog” map is highly influenced by the lack of blogs from WordPress and Blogspot. These companies are two of the largest blogging companies in our results and both are based
near San Jose, CA. However, since bloggers are spread throughout the country, very few of these results were spatially accurate. When this knowledge is paired with the number in Figure 10 for “Blog” spatial accuracy, 10.81%, the “Blog” map in Figure 16 is probably not very useful in looking at the spatial patterns of blogging.

5.2.5- Public vs. Private Categories

As discussed in section 3, there are two frameworks related to web page categories that may have some sort of meaning. The first framework, Public vs. Private, uses the categories “Government”, “Education”, and “NGO” for Public and the categories “Commercial”, “News”, “Informational”, “Special Interest Groups”, “Blog”, “Social Media”, “Entertainment”, “Social Media”, and “Forum” for Private. Figure 18 displays the two maps.

![Figure 13: Public categories vs. Private categories. Both maps use the logarithmic population, a kernel radius of 150,000 meters, and the score range normalization.](image)

Looking at Figure 18, the Public web pages tend to follow a similar pattern to the “Governmental” map from Figure 17. The two cities with the highest intensity are Washington, DC and Seattle, WA. The two regions with the lowest concentrations are San Francisco, CA/San Jose, CA and New York, NY. There is also some red activity in the Midwest, perhaps as part of the “Educational” influence.
For the Private map, the influence of the “Social Media” and “Entertainment” maps can be seen in the San Jose, CA area. There are also large cold spots in Washington, DC and Seattle, WA as the “Governmental” web pages are in the Public map. It is surprising that the full country influence of the “Commercial Websites” does not seem to have a larger influence.

It would require further research to see if this cyberspace pattern is reflected in the real world. Are there more Public servers in the north and Midwest areas? Are there really as fewer Private servers in the Washington, DC and Seattle, WA areas as shown on the map? The maps in Figure 18 can give one hypothesis, but the results are very preliminary until further research can be done.

5.2.6- Individual vs. Group Categories

As discussed in section 3, there are two frameworks related to web page categories that may have some sort of meaning. The second framework, Individual vs. Group, uses the categories “Social Media”, “Blog”, and “Forum” for Individual and the categories “Special Interest Groups”, “Government”, “Education”, “Commercial”, “News”, “NGO”, “Informational”, and “Entertainment” for Group. Figure 19 displays the two maps.

![Figure 14: Individual categories vs. Group categories. Both maps use the logarithmic population, a kernel radius of 150,000 meters, and the score range normalization.](image-url)
Looking at Figure 19 above, there see, to be similar results to the above comparison from section 5.2.5 in that the components are very apparent in the Individual map, just as the components were very apparent in the Public map. The only two major hot spots in the Individual map are Salt Lake City, UT, and San Jose, CA, reflecting the categories “Forum” and “Social Media”, respectively. The absence of the other categories presents a large number of cold spots on the map.

On the other side, the Group map seems to be more moderated, with no major cold spot (only one light blue area, in San Jose, CA). This indicates that there are a large number of group-oriented spatially accurate servers throughout the country, especially in regions such as Washington, DC and New York, NY.

As with the above comparison, it would require further research to see if this cyberspace pattern is reflected in the real world. Are there more Individual servers in Utah and San Jose, CA? Are there really as fewer Individual servers in the most of the country as shown on the map? The maps in Figure 19 can give a rough idea, but the results are very much an estimate until this issue is further explored.

5.3- Spatial Accuracy Correlations

As discussed in section 3, in order to identify the impact of spatial accuracy, the correlation between the maps created using only the spatially accurate results and the maps created using all of the results has been examined. Figure 20 shows each of the maps, with the spatially accurate results in the left column and the complete results in the right column. Table 4 shows the correlation results from the two maps.
Figure 20: Comparisons between the kernel density maps for the spatially accurate results (left) and the complete results (right). Each map was created using a logarithmic populations, a kernel radius of 150,000 meters, and the score range normalization.

Table 4: Pearson’s r correlation coefficient for each set of datasets.

<table>
<thead>
<tr>
<th></th>
<th>Pearson's r</th>
<th>t-Test Score</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Romney</td>
<td>0.504337</td>
<td>386.312</td>
<td>p &lt; 0.01</td>
</tr>
<tr>
<td>Santorum</td>
<td>0.195205</td>
<td>148.267</td>
<td>p &lt; 0.01</td>
</tr>
<tr>
<td>McGinn</td>
<td>0.985409</td>
<td>6721.92</td>
<td>p &lt; 0.01</td>
</tr>
<tr>
<td>Sanders</td>
<td>0.589415</td>
<td>1184.81</td>
<td>p &lt; 0.01</td>
</tr>
<tr>
<td>Flu</td>
<td>0.484688</td>
<td>877.580</td>
<td>p &lt; 0.01</td>
</tr>
<tr>
<td>HPV Vaccine</td>
<td>0.686586</td>
<td>1294.74</td>
<td>p &lt; 0.01</td>
</tr>
</tbody>
</table>
There is a difference in results depending on a visual comparison as opposed to a statistical comparison. Visually, the maps above in Figure 20 appear to be very different. Often, the spatially accurate maps are blue, or have a lower server probability throughout much of the western half of the country that disappears or turns to red in the complete maps. Also, four of the six datasets have a different top city in the spatially accurate dataset as opposed to the complete dataset. Often, Los Angeles, CA appears prominently in the complete datasets as opposed to blending in in the spatially accurate datasets, as in the Romney, Santorum, and HPV Vaccine datasets. The mayor datasets were less changed, perhaps reflecting the local impacts discussed in section 5.1.2.

As can be seen in Table 4, conversely, the Pearson r values show a high correlation between the sets of datasets. Only one dataset (Santorum) has an r value below 0.5, and all of the datasets have a positive correlation that is significant past the p < 0.01 level. This would typically imply that the datasets are correlated, and perhaps are not as different as they may appear above. However, this may be due to the urban bias in the data. Since there is obviously more activity near cities, there tend to be more servers located in cities as opposed to rural areas. This means that a large amount of all of the maps have a raster value of “0”. When the Pearson’s r calculation is done, these large areas of “0” values are identical, regardless of the changes in the cities. In essence, even if all of the cities change values, they are only a portion of the entire map that is dominated by “0” values. This urban bias may be skewing the correlation results.
6 Conclusion

The overall goal of this research was to take an in-depth look at IP addresses, particularly focusing on using IP addresses to geolocate servers. By combining geographic information science, Internet research, and geostatistics, this research has moved the understanding of IP address spatial analysis forward. This methodological examination has taken IP addresses related to search engine results, extracted the spatially accurate results, and attempted to examine what separates these results from the complete dataset. This research has also examined the best methods for using and processing these points, and any connection between the properties of the web pages (categories) and the visualization technique. In this conclusion, each research question is reexamined and answered in full using the results from above.

1. How accurate are the locations cited in IP address databases with respect to where the website content creators claim as their location?

The spatial accuracy percentages are in Table 3. The average spatial accuracy from the six datasets studied is 25.54% and the median spatial accuracy of the six datasets studied is 24.75%. These numbers are far smaller than the current cyberspace research, and are significant as this is the first time that IP address geolocation has been estimated using a manual technique. As can be seen in section 5.3.1, this poor spatial accuracy has an effect on the visual comparisons of the data.

2. Which geographic visualization techniques return the best result for generating patterns from the preprocessed IP address data?

There are several parts to this answer. When using a population method to represent a top-heavy ranking (such as search engine results), the logarithmic population method created the best results. For the kernel radius in the kernel density function, the kernel radius should be between 150,000 and 200,000 meters, or a distance that does not confuse important areas on the map (in this case, nearby cities). When using
a normalization technique, the score range procedure can create a cleaner map than the maximum score technique.

3. **What are key factors within the data visualization processing of different categories of websites that can influence how well the ‘signal’ in the data shines through the ‘noise’ of inaccurate false positives?**

There are a large number of factors that seem to impact the level of “signal” that can be viewed through the “noise” of obfuscating results. There are variations in the type of web pages and the intensity of the web pages. Higher numbers of “Educational” and “Governmental” websites and lower numbers of “Blog” and “News” web pages would increase the amount of spatially accurate results within a dataset. This can be seen in the McGinn dataset, and to a lesser extent, the Sanders dataset. Overall, the best summation of these results may to focus on the extent of attention. Locally known phenomena, such as the mayors, resulted in maps that were more resilient to spatial inaccuracies. They had enough of the above categories to present the pattern, but as can be seen in Figure 19, the images do not change much when the spatially accurate results are extracted. Also, from Table 4, the two mayors had two of the top three highest correlations between spatially accurate and complete results. While groups of categories (such as Private vs. Public) were examined, the results were too inconclusive to be relevant.

Using IP addresses to better understand cyberspace phenomena is a complex task with many problematic issues. With a spatially accuracy rate that is only one fourth of the entire set of results, IP addresses are not a viable method for understanding cyberspace patterns in the context of real space. Even when there is a correlation between the spatially accurate results and the complete results, there is no way to visually understand the pattern, as the cities with the highest and lowest server probabilities change, and the overall patterns within the country change. While performing a manual classification process will allow a user to extract the accurate results, this process is very time consuming and not generally
practical for a researcher attempting to look at large amounts of data. Locally-focused keywords will tend to have better results, but even these are problematic.
REFERENCES


