4-8-2008

The Spatial Distribution of Geoprivacy Concerns in Florida: A County Level Analysis

Joshua W. House
University of South Florida

The Spatial Distribution of Geoprivacy Concerns in Florida:

A County Level Analysis

by

Joshua W. House

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Arts
Department of Geography
College of Arts and Sciences
University of South Florida

Major Professor: Jayajit Chakraborty, Ph.D.
Pratyusha Basu, Ph.D.
Steven Reader, Ph.D.

Date of Approval:
April 8, 2008

Keywords: GIS, privacy, geocoding, reverse geocoding, masking

Table of Contents

List of Tables

List of Figures

The Spatial Distribution of Geoprivacy Concerns in Florida:
A County Level Analysis

Joshua W. House

ABSTRACT

Certain types of spatial data maintained and distributed by counties at taxpayer expense can be used with powerful mapping and analysis software, called Geographic Information Systems (GIS), to compromise an indvidual's locational privacy. The kind of privacy at threat here is referred to as *geoprivacy*, which is concerned with the rights to prevent disclosure of the location of one's home, workplace, or daily activities. While the availability of accessible and accurate geospatial data has increased geoprivacy concerns in recent years, this threat remains virtually unknown to the general public.

Although previous research has explored various components of the geoprivacy debate, the fragmented and localized nature of this work does not adequately address the threat on a large scale or lend itself for use in multi-level policy discussions. This thesis fills the need for a comprehensive and systematic geoprivacy study by examining county data availability in the entire state of Florida.

Ultimately, the success of geoprivacy violation attempts is determined by the availability and quality of the data being used. In order to evaluate this threat, a statewide inventory of the data necessary for a reverse geocoding operation, defined here as geoprivacy data elements, was created. A specific county (Bay County) with complete data availability was then selected and its geoprivacy data elements, specifically street, parcel, and address point layers were evaluated for their reverse geocoding and subsequent identity disclosure success. These findings were then compared with the results of the statewide inventory to determine the level of exposure that the state's residents are subjected to, based on their county's data offerings.

The statewide data inventory indicated substantial variation in county availability, quality, and delivery methods of the desired geoprivacy data elements. The results of the reverse geocoding operation performed with Bay County's geoprivacy data elements revealed that both property parcels and address points in conjunction with ownership information have a high rate of identity disclosure success. Geocodable streets were found to have a low rate of identity disclosure success and their results were comparable to a non-county maintained street layer that was used for control purposes. Although the street layers had a low rate of identity disclosure success, they could be used to identify a narrow range of address possibilities and still pose a geoprivacy threat. Forty-two counties in which approximately 13 million people reside make parcel data with ownership information available for free or purchase. Given the high success rate of the parcel data to disclose individual identity, this research suggests that the majority of the state's residents are vulnerable to potential geoprivacy violations.

1. INTRODUCTION AND OBJECTIVES

1.1 Background

Advancements in the field of information technology have greatly enhanced the ability to acquire, analyze, and distribute information of varied content. Although there are many benefits associated with such progress, it is important to also consider the risks, such as the potential for privacy infringement as there are "enhanced possibilities presented by information technology for collecting data about individuals without their consent" (Olvingson 2003; p. 183). The rapid pace of technological evolution can make it difficult to comprehend and effectively manage its collective impact prior to implementation, thus complicating privacy issues. This makes the development of effective information technology privacy protection efforts and risk management strategies challenging, and in many cases, reactive.

One of the benefactors from developments in information technology is the field of Geography, specifically a branch called Geographic Information Systems (GIS). GIS is described as "a computer-assisted process designed to acquire, store, analyze, and display spatial data and their attributes" (Dent 1999, p. 111). A small sample of the many applications of GIS includes environmental modeling, epidemiology, urban planning, and emergency response. Any subject that has a spatial component can somehow be served, or at least conveyed, in a

GIS.

Riding the wave of information technology, the rapid advancement of GIS and locational capabilities such as Global Positioning Systems (GPS) has resulted in the creation of vast amounts of accurate, accessible spatial data and analytical tools that did not exist until only a few years ago. Although these advancements have helped society discover and analyze spatial phenomena, their inherent power has also raised concern over an individual’s right to locational privacy because some of the information available for use in a GIS can be used to disclose an individual’s identity. This privacy subset is called “geoprivacy” and “refers to individual rights to prevent disclosure of the location of one's home, workplace, daily activities, or trips” (Kwan 2004, p. 15). Because of its ability to “integrate and analyze a large amount of geospatial data,” GIS is at the forefront of the geoprivacy debate. According to Kwan (2004; p. 15), “the potential of GIS to be far more invasive of personal privacy than many other information technologies has caused serious concern among GIS critics and the public.” When the physical location of an event is tied to its descriptive information, the potential for privacy breach is exponentially increased because of its interaction with other spatially located phenomena (VanWey 2005; p. 15339). GIS data such as roads, addresses, and property parcel boundaries with ownership information provide the means to link the location of events, often conveyed through other, seemingly benign means, to an actual individual. Such underlying data is commonly developed and maintained by government entities (at taxpayer expense) and made available for free download via the internet.

Contributing to the advancement of the geoprivacy threat is a lack of public awareness regarding both sensitive data availability and how it can be used within a GIS to disclose an individual's identity.   Therefore, the geoprivacy debate is occurring within a relatively small arena by only those who fully understand the gravity of the issue. If made known, however, the full scope of available information as well as what could be done with it by someone skilled in the spatial sciences, the issue of geoprivacy would likely garner more attention and concern from the general public.

Although it is difficult to comprehensively assess the sensitivity of any piece of information, an example of that which would be deemed "private" is the location and identity of an individual with a certain disease. Maps depicting locations such as points are commonly produced for medical studies aiming to discover spatial relationships among the afflicted. However, "it is not widely known that such maps can be "hacked" to allow individual-level information to be recovered" (Armstrong 2005; p. 67). If there is sufficient detail and fidelity in the map, the locations of the individuals can be extracted via GIS and spatially cross referenced with other data, such as property boundaries with ownership information obtained through a county website, to disclose identity.

It is certainly not the goal of the research community to compromise their subject's identity. The assurance of the preservation of confidentiality is not only consistent with ethical research guidelines as defined by the American Association of Geographers, American Psychological Association, American Political Science Association, and American Sociological Association, but also “necessary to guarantee the continued participation of the public in censuses and social surveys” (VanWey 2005; p. 15337). If the public perceives that their privacy is being breached, they will be less likely to participate in locational studies, effectively minimizing the potential public benefits that the study could provide. A lack of effective research can lead to a lack of researchers so, for many reasons, it is in the research community’s best interest to maintain their subject’s confidentiality and trust.

Further complicating the geoprivacy issue is that the transparent nature of research (e.g., disclosing methods, sources, data) presents an additional source of vulnerability. Research and the progression of knowledge requires outside parties validating, recreating, and building upon prior methods, data, and findings. However, such efforts can compromise confidentiality as disclosure risk ultimately increases with access. Even if access to sensitive research material was prevented, merely knowing which research entities were involved with its production represents a potential geoprivacy leak, as institutional knowledge can lend insight or provide an unprotected gateway to the private information (Van Wey 2005).

Guidelines do exist for disclosure of non-spatial medical and financial records, but universally accepted or effective rules have not been developed for spatial records. Where there are rules for spatial data (HIPAA) there is opportunity for disclosure as “the rule (HIPAA) creates an exception permitting disclosure of personal health information to public health authorities for public health purposes without such authorization” (Rushton 2006; p. S19). Disclosure is essentially governed by an individual privacy versus public benefit debate. However, with little public knowledge that these discussions are being held, the debate is one sided and can too easily conclude with a decision to compromise individual privacy for public benefit.

Limiting the jurisdictional power of research guidelines is that the guidelines only have meaning to those who aim to be accountable to the overseeing organizations. “According to the concept of confidentiality, it is only possible to share data with others who are obliged to the same confidentiality concept and need the information in their profession” (Olvingson 2003; p. 181). It is likely that a great amount of geospatial analysis with sensitive data occurs outside of these organizational guidelines and is thus unregulated. Media outlets, for example, are currently not subjected to the same guidelines which govern the presentation of locational information of health study publications (Olvingson 2003), yet their work (e.g., newspapers, newscasts) could conceivably reach a greater audience than an academic journal article and present a greater geoprivacy threat.

In an effort to mitigate the disclosure threat, several mechanisms and procedures have been developed. These are referred to as masking techniques and they aim to provide adequate analytical capabilities while preserving individual privacy. Given the subjectivities involved with determining what constitutes an adequate analysis, however, this is not an easy task. What may be a suitable masking technique for one purpose may not be for another, because an “adequately masked” data set could be combined with additional information or knowledge by a third party resulting in disclosure. With technology making the distribution and acquisition of information easier and effortless, it is difficult to predict the intended and unintended uses of a mapping product as there are “many unforeseeable downstream users and uses” (Olvingson 2003; p. 183).

In summary, there is an abundance of accessible, unregulated spatial data that can be used with powerful mapping and analysis software to disclose individual identity. This threat exists and is virtually unknown to the general public. Ultimately, the success of geoprivacy violation attempts is determined by the availability and quality of the underlying geospatial data as well as the ability to use such data in conjunction with GIS software. With the increasing availability of high quality data and the advancement and pervasiveness of the software used to engineer geoprivacy violations, the geoprivacy threat will continue to grow if left unchecked.

1.2 Goals and Objectives

Although there is a growing body of literature on geoprivacy and its various components, most of it is field-specific and carries a technical tone that might be abstract, irrelevant, and inaccessible to someone unfamiliar with modern spatial technologies and venues for accessing related material. Moreover, the existing body of research is fragmented and difficult to be used “as is” to raise awareness of the issue and serve as a springboard for widespread discussion. In addition, previous research has largely been localized, focusing on specific towns, census tracts, or individual counties. These delimiters, while certainly valid for their respective purposes, do not fully explore the nature, extent, and magnitude of the geoprivacy threat.

To address the need for a more comprehensive, systematic, and tangible assessment of geoprivacy, the state of Florida and its counties serve as the study area for this thesis project. Florida is an appropriate setting for a geoprivacy study of this scope because of two reasons:

1) Florida, its counties, and estimated 18 million residents (US Census Bureau 2006) provides a geographical context that is well-known, has jurisdictional significance, and appeals to a large audience.
2) Florida's Public Records Law states that government records, including computer records and subsequently GIS data, are public information (Florida Statues, Ch. 119, AGO 89-39). Although a public record preparation fee can be assessed (Florida Statutes, Section 119.07(4)(d)), several counties make their GIS data available for free download via the internet as "providing access to public records is a duty of each agency" (Florida Statues, Section 119.01(1). Some of the GIS data that is made available by Florida government entities is suitable for use in a reverse geocoding / map hacking process.

These two factors provide an important basis for investigating the following research questions:

1) In what manner does the availability of certain types of information necessary to engineer geoprivacy violations influence its success?

2) To what extent are Florida counties and its inhabitants at risk for geoprivacy violations?

By investigating these questions, the thesis documents and analyzes the geoprivacy threat in a systematic manner that is easily understood yet grounded in sound research practices; ultimately lending itself for use in policy discussions at all levels throughout the state of Florida.

2. BACKGROUND AND LITERATURE REVIEW

This chapter provides a detailed summary of several key aspects of geoprivacy and associated methodologies. It is important to consider that the range of geoprivacy research is rather broad and encompasses several techniques, issues, and practices. While practices such as mobile phone tracking and video monitoring are relevant, this thesis focuses primarily on the risks associated with the display of point data. The scope of the geoprivacy threat, however, is not limited to what is presented in this specific study and accompanying methodology.

2.1 Geocoding

Locational studies typically aim to determine if there is an association between an entity and its proximity to an event. The questions researchers try to answer is: where do the subjects live, and is there anything acting upon them that is related to the factor(s) being investigated? These questions can be explored using the powerful mapping and spatial analysis capabilities of GIS (Geographic Information Systems). One of the many functions available in GIS is a process called Geocoding, which can best be described as “the practice of assigning a geographic identifier to a computer record that lacks it, thereby tying information to geographic space.” (Rushton 2006; p. S16) This process is widely used in locational research (Brownstein 2006; p. 2) as the subjects’ need to first be located, typically from address information obtained by the researcher(s), before analysis with other spatial phenomena can begin.

The geocoding process is accomplished by using GIS software in conjunction with reference data that serves as an underlying framework for the assignment of a real world geographic identifier. Such reference data is available from a variety of sources and exists as either a line, polygon, or point. A further explanation of this reference data and how it is used in the geocoding process to assign addresses is provided below:

Line (street network based) – a spatially referenced GIS “layer” which depicts streets as individual line segments. Information such as the street name, address range, etc. are assigned to each line segment in the street layer’s attribute table. Addresses are identified by using geocoding algorithms that attempt to locate the address(es) of interest on the underlying street network. This is conceptually performed by searching for the components of the desired address (Street Number, Street Name, City, State, Zip) and then using the address range information inherent to the line to locate the desired street number by means of linear interpolation along that street segment.

Polygon (parcel based) – a spatially referenced GIS “layer” depicting the boundaries of property parcels. These parcels correspond to ownership boundaries and have the address(es) assigned in the layer attributes of each parcel. This information, in turn, can be used by the mapping software to locate or assign an address.

Point (address point based) – a spatially referenced GIS “layer” which denotes an address as a discrete point location. This represents the highest level of accuracy for address information. Addresses are encoded into the layer attributes which is used by the mapping software to locate or assign an address.

Once geocoded, the subjects can be viewed and analyzed in conjunction with other spatial data. This is the power of GIS; locating, integrating, and analyzing spatial data of varied themes. While of great benefit to a researcher who is looking to determine if high rates of cancer are related to residential proximity to toxic waste facility, the very same tools used to answer such questions can be used in conjunction with data containing personal information to disclose identity. Evaluations of geocoding methods and their effectiveness have provided mixed results. An assessment of the capabilities of firms that offer geocoding services (line / street network based) resulted in high variability among their products, pointing to the quality of the reference information used as well as the tolerances used for determining what constitutes a match (Whitsel 2006; p. 8). In addition, the geographic characteristics of the population being geocoded plays a role in geocoding success. Cayo’s 2003 study, “Positional error in automated geocoding of residential addresses,“ examined the effect of population density on geocoding accuracy (line / street network based) and observed that rural addresses were less accurately located than more urban areas. Sources of geocoding error include inaccurate geometry, inaccurate attributes, and inaccurate ranging and there is also an accuracy tradeoff due to the matching tolerances applied during the geocoding process (Rushton 2006; p. S17-S18).

Despite the limitations of geocoding, it is possible to locate addresses in accurate manner. Prior studies such as Cayo (2003), Whitsel (2006), and Rushton (2006) cite the significance of accurate base data and consistent address formatting of the input and reference data as determinants of geocoding success. Both Whitsel (2006) and Cayo (2003) proclaim the increased accuracy of using polygon / parcel based geocoding which "is expected to grow over time as high quality, parcel-level databases become more uniformly available across larger study areas" (Whitsel 2006; p. 10).

2.2 Reverse Geocoding (also known asinverse geocoding or map hacking)

The functionality of geocoding, which spatially locates addresses using underlying reference information such as street lines, parcel polygons, and address points is a critical component of many locational studies. In addition to providing geocoding capabilities, GIS software can also be used to determine the location of a feature that has been mapped (in hardcopy or other form outside of a GIS environment) but does not have a real world location or assigned address. This process is known as reverse geocoding or map hacking (Rushton 2006; p. S19).

Curtis' 2006 study, "Spatial confidentiality and GIS: re-engineering mortality locations from published maps about Hurricane Katrina" portrays a common reverse geocoding scenario whereby a published map is scanned, georeferenced, and the features of interest digitized to determine real world locations. Once real world locations, or coordinates, have been assigned, these features can then be linked and analyzed in a variety of ways to and with other data; both geographic and non-geographic.

In Curtis' study, reverse geocoding was performed on a map that was published in the Baton Rouge Advocate newspaper. This map displayed Hurricane Katrina mortalities as point locations and used census tract boundaries as a background reference theme to provide the reader with an idea of where the mortalities occurred. This map was clipped from the newspaper, scanned, and then georeferenced to an existing GIS layer of census tract boundaries. The point locations were then digitized from the scanned and georeferenced map, which gave the features that were once merely dots on a map in a newspaper real world coordinates.

To determine the accuracy of the reverse geocoding process, the real world coordinates of the digitized features were then compared with GPS measurements of homes in which mortalities occurred. Given that the published map and the georeferencing target were of a common theme, (census boundaries), there was a high success rate of reverse geocoding them to their true location. The goal of Curtis' research was to evaluate the accuracies of the reverse geocoding process, which proved to be high. Adding to the success of reverse geocoding is the presence of additional themes on the map such as political boundaries and roads. These greatly assist in the reverse geocoding process as they provide a common link for georeferencing. "The general point is that layers or themes potentially displayable on a map add to the security threat" (VanWey 2005; p. 15540). These themes lend the location of the study to being vulnerable to general geographic knowledge of the area as physical indicators such as coastlines, rivers, streets, and topography could give away the location (Armstrong 1999).

Of course, not every single map lends itself to successful reverse geocoding: "contributing factors in the successful re-engineering of information from a cartographic display is the published map's scale, the size (and quality) of the published map, the projection used, and the accuracy (or error) in the initial mapping of the points" (Curtis 2006; p. 2). These items, in conjunction with other information displayed on the map as well as the availability of the information that is used in the georeferencing process ultimately govern a map's hacking potential (VanWey 2005). However, if the maps used in reverse geocoding "accurately depict locations, they can be used to recover individual-level information such as an address" (Rushton 2006; p. S19). It is the map's accuracy that is paramount; factors such as resolution are not as significant assuming the map also has moderate visual clarity. Brownstein's 2006 study of the effect of map resolution on reverse geocoding success determined that "the home addresses of many of these patients could be discovered, despite the low resolution of the disease maps" (Brownstein 2006; p. 2).

The significance of these findings is that it is possible to use GIS to tie these reverse geocoded or hacked locations to other spatial data, such as property ownership parcels. This type of information is commonly distributed free of charge from county maintained websites. If this data contains ownership information it can be spatially cross-referenced to the reverse geocoded features, resulting in identity disclosure.

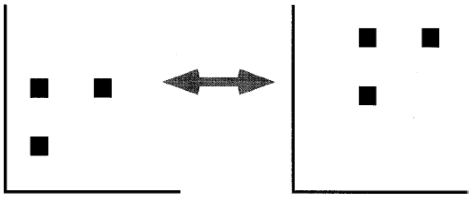
2.3 Masking

Given the vulnerabilities associated with mapping individual locations, researchers have worked to develop methods to protect individual locational privacy while at the same time allowing valid spatial analysis to be performed. These methods are referred to as geographic masks or masking. "The goal of these masks is to modify the geographic information sufficiently to prevent disclosure of individual identities, while retaining enough spatial accuracy for geographic trends, clusters, or other patterns to be detected" (Rushton 2006; p. S20).

Armstrong provided a comprehensive summary of masking techniques in his 1999 work, "Geographically Masking Health Data to Preserve Confidentiality." A description of these masks and how they affect data are described below. Where applicable, a graphic is provided to assist with understanding the masking concept:

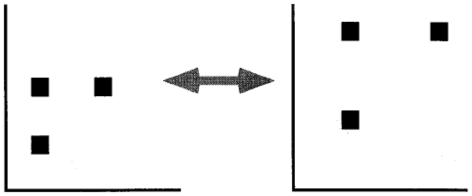
Displacement using translation (Figure 1) – moves "each point from its original location by a fixed increment." (Armstrong 1999; p. 502) This results in a uniform shift of the entire data set.

Figure 1. Displacement Using Translation



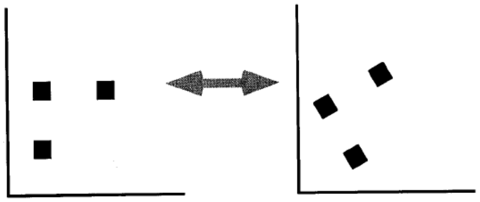
Scaling (Figure 2) – this mask “changes both the distance from the origin of the co-ordinate system (thus executing a displacement) as well as the distances between point locations.” (Armstrong 1999; p. 502) This results in a uniform shift of the entire data set as well as a fixed increase or decrease in the distance between each feature.

Figure 2. Scaling



Rotation (Figure 3)– simply rotates “each point by a fixed angle about a pivot point.” (Armstrong 1999; p. 502) This results in a uniform “twist” of the data at a specified rotation point.

Figure 3. Rotation

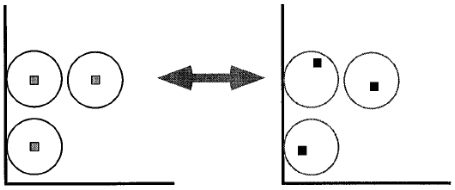


Concatenated Mask – using any combination of displacement, scaling, and rotation masks in conjunction with one another. (Armstrong 1999; p.

503)

Random Perturbation (Figure 4) – displaces “each point by a randomly determined amount, and in a randomly determined direction, specific to its original location.” (Armstrong 1999; p. 504)

Figure 4. Random Perturbation



Point Aggregation – this technique “uses a single location to represent a defined subset of the original locations.” (Armstrong 1999; p. 506) An example would be to use one point to depict that several incidences of cancer occurred within the greater vicinity of that point, but not at that discrete location.

Areal Aggregation – protects against disclosure by “enumerating the total that exists within a region.” (Armstrong 1999; p. 506) An example would be to show the total amount of incidences of cancer that occurred within a census block.

The previously discussed masking methods have dealt with altering the physical location or amount of information that is shown. There are, however, other ways of protecting spatial data confidentiality that are based on some form of data access control, agreements among the parties involved, or alternate forms of display. These include:

Enclaves and Cold Rooms – where data is made available for analysis at a physical location. No data is permitted to leave the premises, and access can be restricted to certain individuals. (VanWey 2005; p. 15338)

Virtual Enclaves - a computer network accessed environment where “restricted access to data can take place, without requiring travel, access fees, or delays before the results are available to the researcher” (VanWey 2005; p. 15341). This is similar to the concept of an Enclave / Cold Room but the data user can obtain remote access to the data and / or results of the object in study.

Software Agents – this masking technique involves remote access by using software to formulate data requests which are “sent to the original data repository, so the analysis could be done inside the original data repository and then a summary aggregate report sent back to the researcher” (Kamel 2006; p. 165). In 1999, Armstrong discussed software agents in that “users would not be required to have access to confidential health records. Rather, they would submit a request to an intelligent analysis agent that would assess the request, and if found appropriate, would complete the analysis and return a result to the data user without exposing any individual-level health data (Kamel 2006).

Virtual Institutions and Virtual Organizations – building on the concept of virtual enclaves, these are very generally described as a combination of data distribution and analysis services whose access is governed by means of pre-defined agreements. These entities can function independently or in conjunction with others (including software agents) to judiciously serve data needs (Kamel 2006).

Privacy, Access, and Usage Agreements – specific agreements outlining what can and cannot be done with the data.

Reduction of Basemap Detail and Contextual Information – reducing the amount ancillary information displayed on a map (political boundaries, roads, descriptions) so that the possibility for identification of the study area is minimized.

Abstract Methods (spider plots, graphs) – displays spatial information in a manner whereby geographical relationships are effectively communicated but not explicitly presented in a conventional mapping format.

Omission - not using or limiting the use of maps for publication purposes.

The implementation of any masking technique will ultimately result in some type of information loss when compared to the original data. However, the information which is lost at the hands of masking might not be necessary for the objectives and subsequent accuracy needs of an analysis (Rushton 2006; p. S20). Kwan's 2004 study, "Protection of Geoprivacy and Accuracy of Spatial Information: How Effective Are Geographical Masks?" evaluated how the degree of random perturbation affected accuracy. A masking threshold value was discovered at which the masked results differ substantially from the non-masked data. Results below the threshold were deemed adequate for the analysis. This suggests that it is possible to use masking and find balance between information loss and accuracy needs.

Despite the possibility for masking success, which is certainly a subjective decree, a universally accepted solution that could be implemented on a large scale (a scale which matches the amount of sensitive data that is easily accessible) has not been developed. An evaluation of the masking techniques which alter the position of the original data (displacement, scaling, rotation, and random perturbation) suggest that “random perturbation is superior from a comprehensive information preservation standpoint” (Armstrong 1999; p. 512). Virtual solutions (enclaves, agents) are theoretically strong yet mostly conceptual in nature and implementation is complex. Real enclaves are believed to carry the lowest risk of confidentiality breaches (VanWey 2005) yet this may not be something that would be possible for all data producers to implement as the startup costs and continued management is prohibitive.

Even with the application of a masking technique, a dataset is still vulnerable to being “hacked” if the masking method employed is discovered, if alternate masked versions of the same dataset are obtained, or if ancillary information (that may have been produced or disclosed afterwards by a different party) is used in conjunction with the masked data to reveal the original locations. Unfortunately, “there are relatively few simple cases or simple solutions” when it comes to managing the geoprivacy threat (VanWey 2005; p. 15338).

Despite the uncertainties involved with masking and effective implementation on a large scale, if disclosure risk is to be minimized it is “vital that some masking occurs of the original point data." (Curtis 2006 p. 10) This belief was echoed at a recent symposium hosted in part by the Association of American Geographers as "there was also a general concern expressed about preserving individual confidentiality within spatial displays. This concern is justified as map making, and the ability to deliver maps to a mass audience through the Internet becomes steadily easier [5-8]" (Curtis 2006; p. 2). Providing further grounds for masking is that "administrative records and other information, sometimes obtained as an adjunct of newly emerging location based services, can be mapped and cross-referenced to reveal the identities and characteristics of individuals from information that is often available on-line" (Armstrong 2005; p. 64).

Lending further support for the need to mask sensitive data are the uncertainties associated with other information that is currently or will become available:

> "an experiment using 1990 U.S. Census summary data surprised the public health community by showing that datasets previously thought to be adequately de-identified, containing only 5-digit ZIP code, gender and date of birth, could be linked with other publicly available data (e.g., voting records) and used to uniquely identify 87 percent of the population of the United States [15]" (Brownstein 2006; p. 4).

A seemingly sound masking technique can be foiled by something that is difficult to prepare for: the unknown.

3. METHODOLOGY

The geoprivacy threat is real and it requires substantial research attention. Understanding the nature and magnitude of the threat, however, is difficult given the fragmented and intangible nature of the existing body of geoprivacy research. This study hopes to address this need by exploring the following research questions:

1) In what manner does the availability of certain types of information necessary to engineer geoprivacy violations influence its success?
2) To what extent are Florida counties and their inhabitants at risk for geoprivacy violations?

These questions are explored by emulating how a “map hacker” could attempt to disclose the identity of non-masked, accurately mapped individuals using public domain data via a reverse geocoding operation. This process was performed and evaluated based on the following steps:

1) For all counties in Florida, determine the availability and quality of the geospatial data that can be used for identity disclosure to occur.
2) Perform a reverse geocoding procedure to evaluate the capability of available county data to disclose identity.
3) Investigate the relationship between population density and reverse geocoding success.

4) Explore the statewide implications of these findings by determining population exposure with respect to the availability and reverse geocoding success of the geoprivacy data elements.

A more detailed description of this process and each individual step is described in the remaining sections of this chapter.

3.1 Geoprivacy Data Availability

Since geocoding and reverse geocoding require appropriate data, the first step was to perform a county-by-county inventory of the information that is necessary and typically used in these operations. The data must have been made available through a county conveyance so as to be considered public domain. The availability of the following county maintained data, referred to as “geoprivacy data elements”, was determined:

A. Geocodable Street Layer

B. Property Ownership Parcels

C. Address Points

Given that there were some counties which did not have any of these data elements, a non-County source geocodable street layer (Census TIGER/Line file) was used to establish a baseline to which the other elements were compared.

Another critical component of geocoding, reverse geocoding, and mapping in general is the availability of current and accurate aerial photography. Although many counties acquire this imagery on their own and make it available for purchase or download, at the time of this study the US Department of Agriculture released imagery for the entire state of Florida whose acquisition date (2007) and quality rivaled or exceeded what most counties typically offer. Since a set of high quality aerials for every county in the state are now available for free download, the need to evaluate variability across counties was eliminated.

In addition to the mere existence of the geoprivacy data elements, factors such as accessibility, ease of use, completeness, and cost were assumed to influence the ability of the layers to be used in a reverse geocoding operation. These characteristics were used to develop a classification scheme that represented the final availability code for each geoprivacy data element. The six categories in this classification scheme are described in Table 1.

Table 1. Geoprivacy Data Availability Codes

Data Availability Code	**Description**
Yes	Layer is available and can be obtained anonymously and without charge (includes parcels that must be joined to ownership table)
Purchase	Layer is available for purchase therefore payment information and in some cases a usage agreement is required (includes parcels that must be joined to table) and cannot be obtained anonymously.
Indirect	Layer is available but difficult to obtain, use, or contact information and / or a usage agreement is required.
Ineffective	Layer is available but not able to be used for reverse geocoding and identity disclosure purposes due to a lack of addresses and / or ownership information.
No	Layer confirmed unavailable (includes layers in process).
Inconclusive	Layer availability unable to be determined.

Based on the classification scheme depicted in Table 1, the county geoprivacy data elements made available for each county in the State of Florida was compiled into: (a) a table that documents each county’s offering; and (b) a series of county-level maps depicting the spatial distribution of geoprivacy data element availability.

3.2 Evaluating Geoprivacy Risk

In order to evaluate whether or not the aforementioned geoprivacy data elements contribute to geoprivacy risk, one county (Bay County) which has complete data availability (geocodable street layer, property ownership parcels, and address points) was selected and each geoprivacy data element was successively evaluated for its identity disclosure success. The subjects whose identity was targeted for disclosure were identified by randomly selecting 100 address points which represent owner / occupiers of the property. The address point layer is the most accurate geoprivacy data element and best represented an individual's discrete location.

With the test population identified, it was then mapped in a GIS environment, specifically ESRI's ArcGIS, to provide the source material on which the reverse geocoding operation was performed. A county scale map was then produced showing only the county boundary, major roads, and the residences of the mapped individuals (un-masked, of course) displayed as point locations. The map conformed to a page size of 8.5 inches by 11 inches (letter size) and was printed on a conventional laser printer.

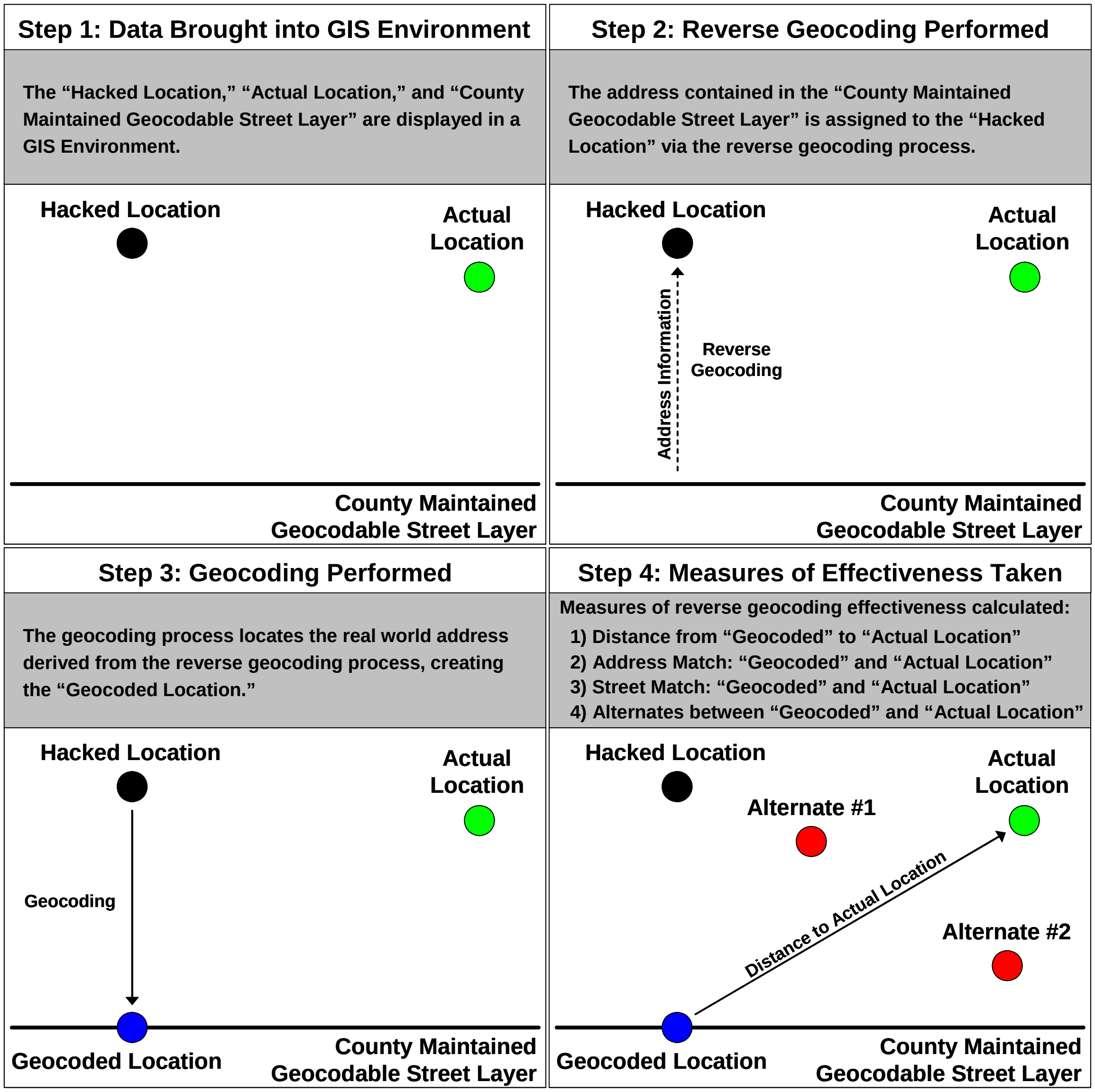
Consistent with routine "map hacking" practices, the "published" map was then scanned at a resolution of 200 dots per inch (dpi) and geo-referenced in ArcGIS using the available source data (county boundary and roads) as registration points. To determine the location of the residences as depicted on the scanned and georeferenced map, the point locations were then determined by first creating a vector polygon circle that conformed to the areal extent of the circle representing the boundary of the point on the georeferenced map. Maintaining a constant capture scale, the boundary of each mapped residence (displayed as points) was determined in this manner, and resulted in one vector polygon for each mapped residence. The discrete location of the mapped individuals was then determined by calculating the centroid of the digitized vector polygon circles. This location represents the “hacked” location of the mapped individuals and was used as the common starting point for each identity disclosure effort.

At this point the actual location of the individual’s residence as well as the hacked location had been established. Given that no personal information such as name or address has been introduced, the possibility for identity disclosure is minimal assuming that the “map hacker” does not have any knowledge of the study area nor its residents. To establish a control value, the distance between the hacked and actual locations was determined.

The real-world address of the hacked points was then determined by using each of the data elements (geocodable street layer, property ownership parcels, and address points) and their corresponding reverse geocoding method in ArcGIS. The assumption here is that a street address is needed for identity disclosure. For example, a hacked location may place the point in the middle of a pasture. This location is deemed to be of little value until a street address is assigned. Since the real-world address is critical, the hacked point was then moved via geocoding to the point on the reference data that represents the location of the address obtained through the reverse geocoding operation, and the distance the hacked point moved was recorded. Finally, the distance from the hacked, reverse geocoded, and geocoded position to the actual residence was determined.

After determining the distance from each hacked, reverse geocoded, and geocoded location to its actual location, summary statistics were calculated to analyze the various geoprivacy data elements and their corresponding effectiveness (measured in distance from actual location). For all data elements, the number of alternate residences within the distance to the actual location was determined as well as an indication of whether or not the reverse geocoded point matched the actual address, actual street, or neither. It should be noted that when ownership information is available, an address match reveals an individual's personal identity. A graphical depiction of this process, using the County Maintained Geocodable Street Layer Geoprivacy Data Element, is offered in Figure 5, Reverse Geocoding Workflow.

Figure 5. Reverse Geocoding Workflow



Step 1: Data Brought into GIS Environment

The “Hacked Location,” “Actual Location,” and “County Maintained Geocodable Street Layer” are displayed in a GIS Environment.

Step 2: Reverse Geocoding Performed

The address contained in the “County Maintained Geocodable Street Layer” is assigned to the “Hacked Location” via the reverse geocoding process.

Step 3: Geocoding Performed

The geocoding process locates the real world address derived from the reverse geocoding process, creating the “Geocoded Location.”

Step 4: Measures of Effectiveness Taken

Measures of reverse geocoding effectiveness calculated:

1) Distance from “Geocoded” to “Actual Location”
2) Address Match: “Geocoded” and “Actual Location”
3) Street Match: “Geocoded” and “Actual Location”
4) Alternates between “Geocoded” and “Actual Location”

3.3 Population Characteristics and Reverse Geocoding Success

Previous studies have indicated that population density has a positive influence on geocoding success when using line based, interpolation methods (Cayo 2003). Highly populated areas, such as cities, typically have shorter streets and a more uniform distribution of addresses. These two factors allow for the line based geocoding process, which utilizes linear interpolation methods, to more accurately predict address placement in urban areas than in rural areas where streets are typically longer and address distribution less uniform. With respect to reverse geocoding, however, higher population densities offer an ambient level of masking as there are simply more possibilities (people) shielding the targeted individual(s) due to closer residential proximity. It was thus necessary to look at the opposing influences of population densities, as this experiment involved both line based geocoding and reverse geocoding. This was accomplished by comparing the population densities for each geoprivacy data element's reverse geocoded point as documented by its corresponding 2000 US Census Block Group value with its reverse geocoding success as measured by distance to actual location. The results for each geoprivacy data element were displayed on a scatter plot and included R-squared values as an indication of linear association.

After the determination of data availability, reverse geocoding success, and influence of population density, these findings were examined with respect to county population totals. This comparison was used to obtain an understanding of potential exposure to geoprivacy in the state of Florida.

4. RESULTS

4.1 Geoprivacy Data Availability

The initial effort to determine data availability in Florida consisted of an internet search utilizing the Google engine (http://www.google.com). For each county, three separate searches were performed using the following key words:

1) Desired County Property Appraiser
2) Desired County GIS Department
3) Desired County GIS Data

In addition to revealing the sought after data and contact information, these search criteria quickly returned a web site that catalogs links to Florida county GIS websites. The information provided by this website was used to supplement the existing search criteria and help determine data availability. The websites retrieved from the search were examined for the presence of the three geoprivacy data elements: county maintained roads, property parcels, and address points. Where data elements were found and freely available, they were downloaded and examined for their ability to be used in a reverse geocoding operation. This examination was not only for the existence of address and ownership information, but also for any characteristic which impacted the usability of the data. Traits such as difficult access, fragmented files, or the need to perform additional processing steps such as joining ownership tables to the

GIS layer, was documented. The online search revealed a substantial amount of variability among all counties as to what information was available, how it was offered, and how it was described.

For counties whose availability could not be determined from the online search, the best contact information (email, phone number) offered by the website was obtained. For example, if there was a specific GIS or mapping contact listed, that information was determined to be the best contact and was pursued. If there was no GIS or mapping contact, the general email address or phone number was used. Keeping with the desire to remain anonymous, email contacts took priority over phone numbers. Only one contact was obtained for each county mapping entity, which typically was the property appraiser and county GIS department.

To fill in the gaps for counties whose data availability could not be determined from the initial web search as well as insure the findings of the online search, an anonymous email was sent to all counties which offered an email contact inquiring about the existence of the geoprivacy data elements. The following is a transcript of the email that was sent:

Hello. Could you please inform me as to how I can obtain GIS information for your county? I am specifically looking for the following layers:

- Streets with Address Ranges (geocodable)
- Property Parcels with Ownership Information
- Address Points
- Recent Aerial Photography*

Is there a site (web, ftp) from which I can directly access any of this information? Any assistance you could provide would be greatly appreciated. Thanks.

*Aerial photography was deemed irrelevant due to the release of the USDA imagery, however the email request was sent prior to this conclusion.

For counties whose best or only contact was a phone number, an anonymous phone call (*67) was made to determine data availability and consisted of the same verbiage as the email message.

The data availability effort required 201 unique web searches, 68 emails, and 35 phone calls. It should be noted that only one knowledgeable contact for each county entity, which in most cases was the County Property Appraiser and County GIS Department, was pursued. With 67 counties in the state and typically two departments being responsible for the desired GIS information, contacting every conceivable entity to achieve absolute certainty would be a monumental task. For the purposes of this thesis, it was determined that if the information could not be located by a thorough web search, email, or phone call to a knowledgeable source then the information is presumed to be difficult to obtain which provides some level of protection, intentional or not.

Unexpectedly, there were a few referrals to county Emergency Services /

Management department(s) for street and addressing information. The referring county staff did not give an indication of whether or not these departments would actually provide this information, just that they were the caretakers and to contact them for availability. Given that these departments focus on providing emergency services to individuals in need and every moment of their time is critical, these contacts were not pursued.

The outcome of the county data search revealed significant variability in both the availability and accessibility of the county maintained geoprivacy data elements (Table 2). A review of the detailed county availability (Table 3) portrays this variability in greater depth. Only seven counties make all three layers available for free, and only 12 have all three available for free or purchase. However, 24 counties make at least one county maintained geoprivacy data element available for free and 45 counties make at least one available for free or purchase. Therefore, at least one county maintained geoprivacy data element

Table 2. County Availability Summary

Data Availability Code	Description	Streets	Parcels	Address Points	Total
Yes	Layer is available and can be obtained anonymously and without charge (includes parcels that must be joined to ownership table)	18	15	12	**45**
Purchase	Layer is available for purchase therefore payment information and in some cases a usage agreement is required (includes parcels that must be joined to table) and cannot be obtained anonymously.	6	27	2	**35**
Indirect	Layer is available but difficult to obtain, use, or contact information and / or a usage agreement is required.	7	5	10	**22**
Ineffective	Layer is available but not able to be used for reverse geocoding and identity disclosure purposes due to a lack of addresses and / or ownership information.	10	2	0	**12**
No	Layer confirmed unavailable (includes layers in process).	1	0	9	**10**
Inconclusive	Layer availability unable to be determined.	25	18	34	**77**
Total		**67**	**67**	**67**	**201**

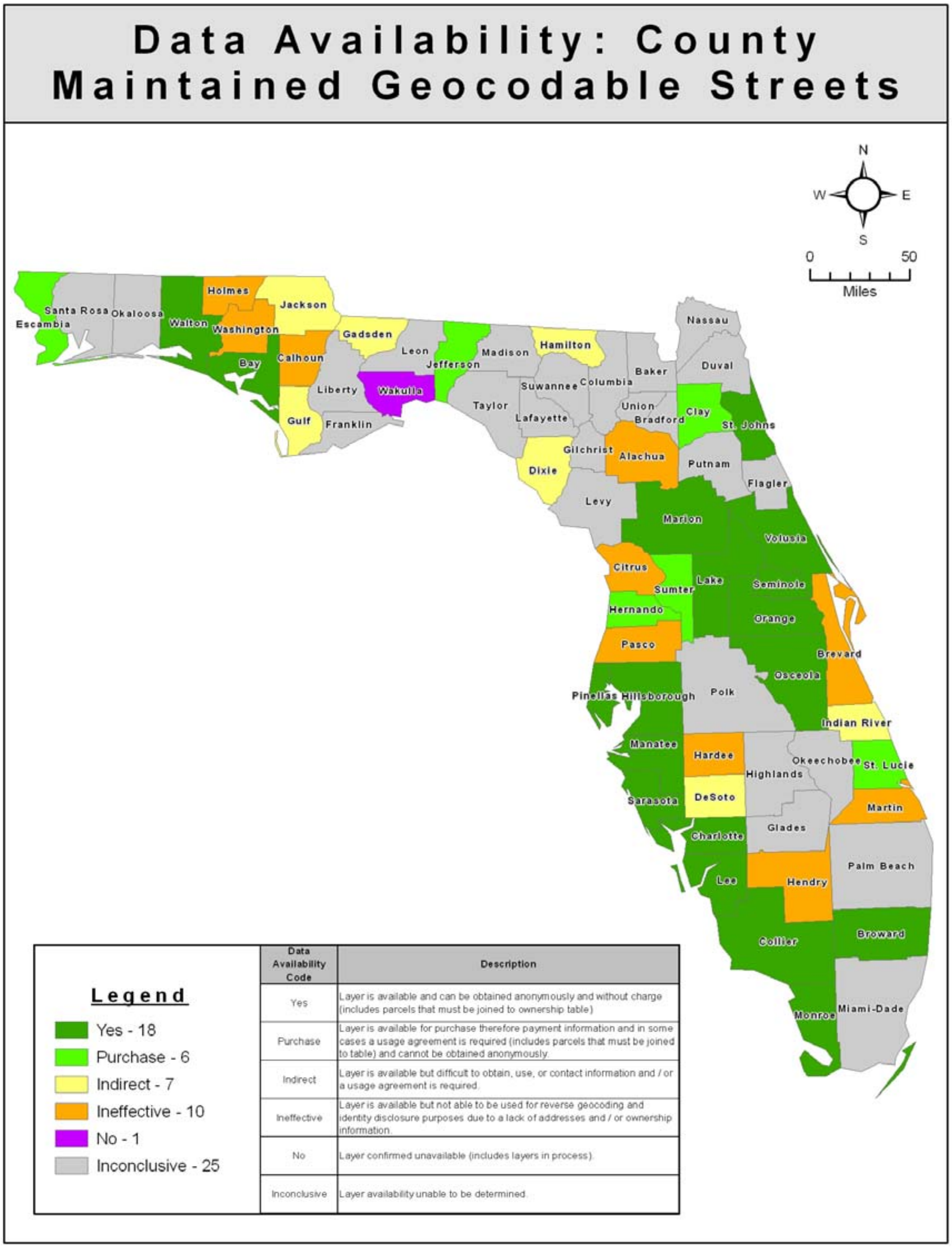
can be obtained for the majority (67 percent) of the state.

Table 3. Detailed County Availability

Detailed County Availability							
County	**Roads**	**Parcels**	**Address Points**	**County**	**Roads**	**Parcels**	**Address Points**
Alachua	Ineffective	Indirect	Indirect	**Lee**	Yes	Yes	Inconclusive
Baker	Inconclusive	Inconclusive	Inconclusive	**Leon**	Inconclusive	Inconclusive	Inconclusive
Bay	Yes	Yes	Yes	**Levy**	Inconclusive	Inconclusive	Inconclusive
Bradford	Inconclusive	Inconclusive	Inconclusive	**Liberty**	Inconclusive	Purchase	Inconclusive
Brevard	Ineffective	Indirect	Yes	**Madison**	Inconclusive	Inconclusive	Inconclusive
Broward	Yes	Purchase	No	**Manatee**	Yes	Yes	Inconclusive
Calhoun	Ineffective	Purchase	No	**Marion**	Yes	Purchase	Inconclusive
Charlotte	Yes	Yes	Yes	**Martin**	Ineffective	Yes	Inconclusive
Citrus	Ineffective	Indirect	Inconclusive	**Miami-Dade**	Inconclusive	Purchase	Inconclusive
Clay	Purchase	Purchase	Inconclusive	**Monroe**	Yes	Yes	No
Collier	Yes	Purchase	Yes	**Nassau**	Inconclusive	Purchase	Inconclusive
Columbia	Inconclusive	Purchase	Inconclusive	**Okaloosa**	Inconclusive	Inconclusive	Inconclusive
DeSoto	Indirect	Purchase	Indirect	**Okeechobee**	Inconclusive	Inconclusive	Inconclusive
Dixie	Indirect	Purchase	Indirect	**Orange**	Yes	Purchase	Yes
Duval	Inconclusive	Inconclusive	Inconclusive	**Osceola**	Yes	Purchase	Yes
Escambia	Purchase	Indirect	No	**Palm Beach**	Inconclusive	Inconclusive	Inconclusive
Flagler	Inconclusive	Purchase	Inconclusive	**Pasco**	Ineffective	Yes	No
Franklin	Inconclusive	Inconclusive	Inconclusive	**Pinellas**	Yes	Yes	No
Gadsden	Indirect	Purchase	Indirect	**Polk**	Inconclusive	Yes	Indirect
Gilchrist	Inconclusive	Inconclusive	Inconclusive	**Putnam**	Inconclusive	Ineffective	Yes
Glades	Inconclusive	Inconclusive	Inconclusive	**Santa Rosa**	Inconclusive	Inconclusive	Inconclusive
Gulf	Indirect	Purchase	No	**Sarasota**	Yes	Purchase	Inconclusive
Hamilton	Indirect	Purchase	Indirect	**Seminole**	Yes	Yes	Yes
Hardee	Ineffective	Purchase	No	**St. Johns**	Yes	Purchase	No
Hendry	Ineffective	Ineffective	Inconclusive	**St. Lucie**	Purchase	Purchase	Inconclusive
Hernando	Purchase	Purchase	Purchase	**Sumter**	Purchase	Purchase	Purchase
Highlands	Inconclusive	Inconclusive	Inconclusive	**Suwannee**	Inconclusive	Inconclusive	Inconclusive
Hillsborough	Yes	Yes	Yes	**Taylor**	Inconclusive	Yes	Inconclusive
Holmes	Ineffective	Purchase	Inconclusive	**Union**	Inconclusive	Inconclusive	Inconclusive
Indian River	Indirect	Indirect	Indirect	**Volusia**	Yes	Yes	Yes
Jackson	Indirect	Purchase	Indirect	**Wakulla**	No	Inconclusive	Inconclusive
Jefferson	Purchase	Purchase	Indirect	**Walton**	Yes	Yes	Yes
Lafayette	Inconclusive	Inconclusive	Inconclusive	**Washington**	Ineffective	Purchase	Indirect
Lake	Yes	Yes	Yes				

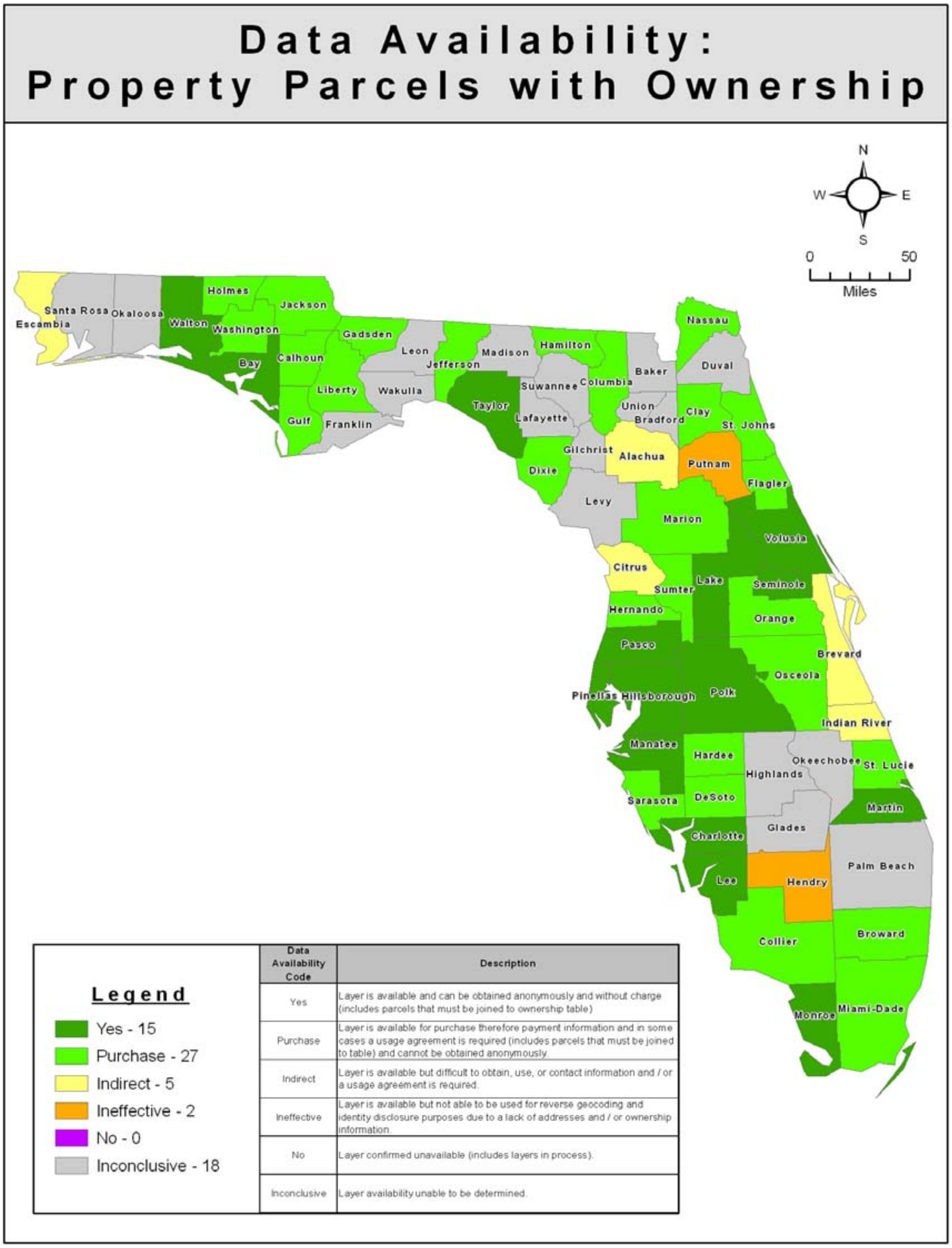
Despite performing an online search and attempting to reach the best contact for each county entity, the availability of elements for several counties was unable to be determined and were therefore deemed “Inconclusive.” The availability of all three county geoprivacy data elements could not be determined in 17 counties and there were 36 counties with at least one data element whose availability could not be determined. Despite an inability to determine what, if any, county maintained information is available for these areas, it is important to remember that non-county maintained geocodable streets are available for the entire state. The following map series (Figures 6 – 8) provides a graphical representation of statewide data availability.

Figure 6. Statewide Availability of County Maintained, Geocodable Streets



Data Availability Code	Description
Yes	Layer is available and can be obtained anonymously and without charge (includes parcels that must be joined to ownership table)
Purchase	Layer is available for purchase therefore payment information and in some cases a usage agreement is required (includes parcels that must be joined to table) and cannot be obtained anonymously.
Indirect	Layer is available but difficult to obtain, use, or contact information and / or a usage agreement is required.
Ineffective	Layer is available but not able to be used for reverse geocoding and identity disclosure purposes due to a lack of addresses and / or ownership information.
No	Layer confirmed unavailable (includes layers in process).
Inconclusive	Layer availability unable to be determined.

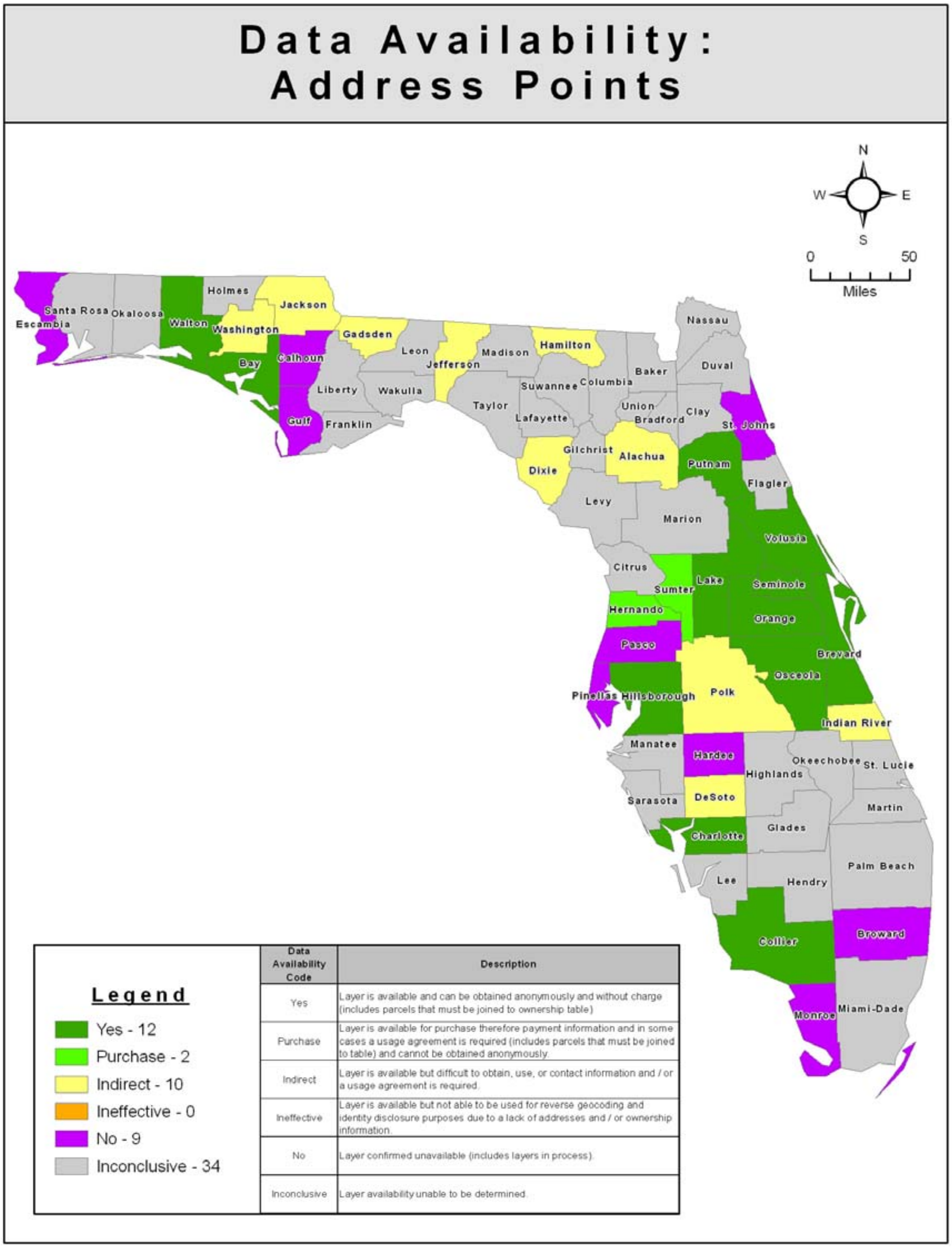
Figure 7. Statewide Availability of Parcel Data with Ownership Information



Data Availability:
Property Parcels with Ownership

Data Availability Code	Description
Yes	Layer is available and can be obtained anonymously and without charge (includes parcels that must be joined to ownership table)
Purchase	Layer is available for purchase therefore payment information and in some cases a usage agreement is required (includes parcels that must be joined to table) and cannot be obtained anonymously.
Indirect	Layer is available but difficult to obtain, use, or contact information and / or a usage agreement is required.
Ineffective	Layer is available but not able to be used for reverse geocoding and identity disclosure purposes due to a lack of addresses and / or ownership information.
No	Layer confirmed unavailable (includes layers in process).
Inconclusive	Layer availability unable to be determined.

Figure 8. Statewide Availability of Address Points



Data Availability Code	Description
Yes	Layer is available and can be obtained anonymously and without charge (includes parcels that must be joined to ownership table)
Purchase	Layer is available for purchase therefore payment information and in some cases a usage agreement is required (includes parcels that must be joined to table) and cannot be obtained anonymously.
Indirect	Layer is available but difficult to obtain, use, or contact information and / or a usage agreement is required.
Ineffective	Layer is available but not able to be used for reverse geocoding and identity disclosure purposes due to a lack of addresses and / or ownership information.
No	Layer confirmed unavailable (includes layers in process).
Inconclusive	Layer availability unable to be determined.

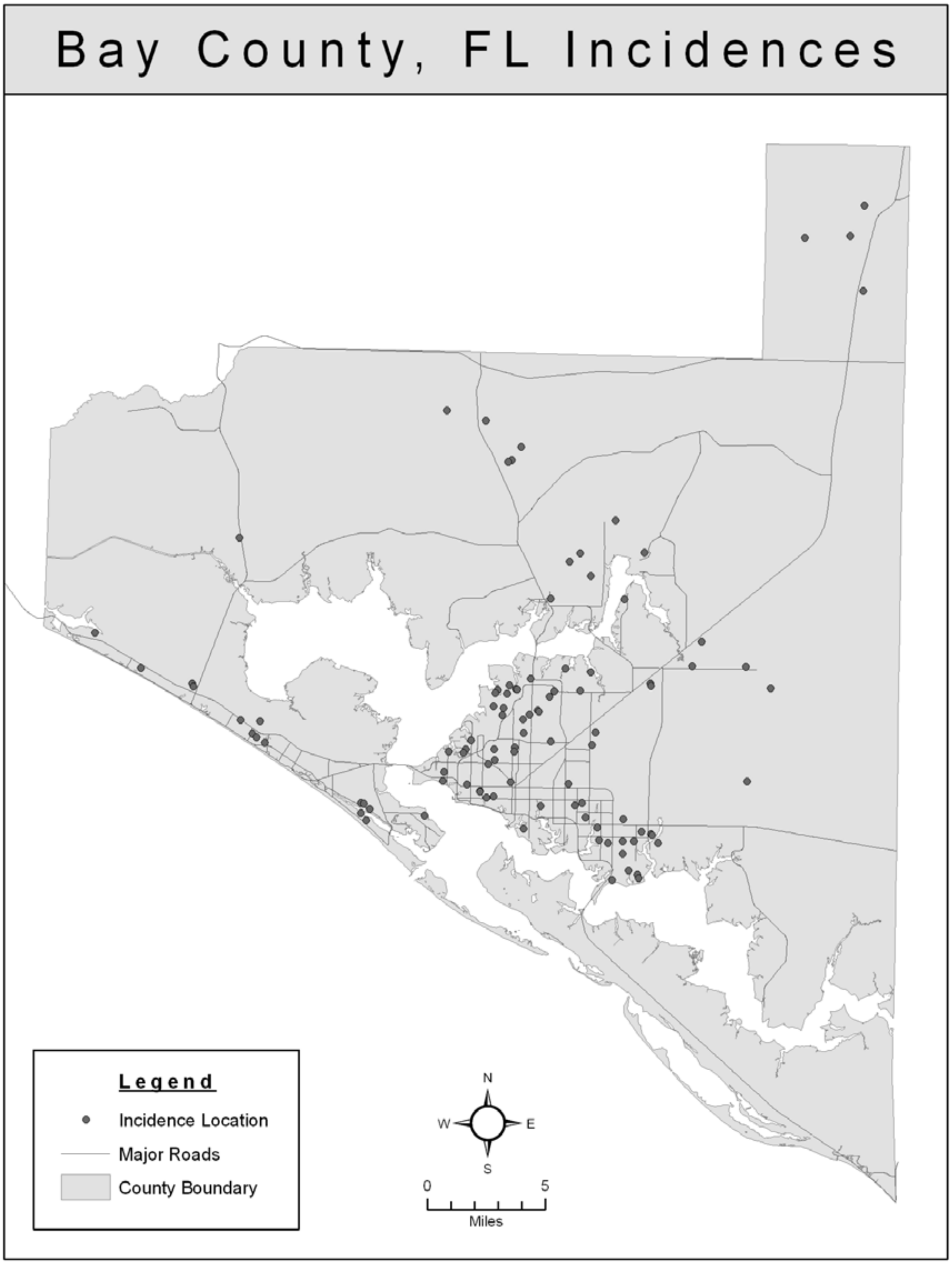
4.2 Evaluating Geoprivacy Risk

With the statewide geoprivacy data element availability determined, the next step was to select a pilot county for the reverse geocoding evaluation. After the initial online search, only two counties had complete geoprivacy data element availability: Bay County and Charlotte County. Bay County was selected as the pilot county because it offered a more robust suite of data beyond the geoprivacy data elements, included metadata, and had population characteristics that were more similar to statewide averages. (USCB 2006 Estimates)

Following the county selection, the subjects of the map hacking effort needed to be identified. The location of the subjects was conveyed by the county address point layer which, in the case of Bay County, is the most accurate geoprivacy data element. To best emulate a scenario which targets residents, the county property ownership database was filtered to contain only those individuals who were listed as owner / occupiers of the property. This reduced the eligible population from 78,090 to 22,755. Of this subset, a simple random sample of 100 individuals were selected as the test group. A limitation of a tool used later in this experiment influenced the decision to use a sample size of 100 individuals.

A county scale (1:316,800) monochromatic map was then developed in ArcGIS showing only the county boundary as a polygon, major roads as lines, and subjects as point locations (Figure 9). This map was then printed with a conventional black and white laser printer on standard copy paper. The 100 point locations, or “incidences,” were then counted to ensure that none were obscured by the other map elements and therefore able to be hacked.

Figure 9. Published Map: Bay County, FL Incidences



The “Bay County, FL Incidences” map was then scanned at 200 dpi to a Tagged Image Format (.tif) file, a lossless file format, and georeferenced in ArcGIS using four road intersections as control points. The incident locations were then “hacked” by digitizing polygon circles at a capture scale of 1:4800 and then determining the polygon centroid using the ArcGIS “Feature to Point” tool. The polygon centroid locations represented the starting points for all subsequent reverse geocoding operations and are referred to as the “hacked” locations or points.

The different types of reference data used in this experiment (point, line, polygon) warranted the use of a different process to assign addresses (reverse geocoding) to the hacked points. For the parcels (polygon layer) and address points (point layer) an ArcGIS tool called “Spatial Join” was used. This process assigns the attributes (the address information) of the closest feature in another layer (the geoprivacy data element) to each feature in the target layer (hacked locations) and calculates the distance between the two.

For assigning the address represented by the line features to the hacked locations, however, an ArcGIS add-in called ET Geowizards developed by ET Spatial Techniques was used. Several internal ArcGIS tools and code samples were explored prior to making the decision to use this utility, but the ease of use, low cost ($195), and effectiveness of ET Geowizards made it an appropriate choice for this experiment. This software utilized all available address components of the non-county and county maintained geocodable street layers to assign an address to the target point layer. The hacked and reverse geocoded points were then geocoded using the address information that was assigned from its corresponding reverse geocoding method. The distance from these points to the actual point, or incident, and the number of alternates was determined using a free ArcGIS add-in (Hawth's Analysis Tools for ArcGIS).

The reverse geocoding analysis with the non-county maintained geocodable streets produced eight points that were not geocodable and one point which was an extreme observation at 42,419 feet with over 100 alternates (a limitation of Hawth's Analysis Tools). The county maintained geocodable streets produced one point that was not geocodable and three with over 100 alternates. The parcels produced only one point that was not geocodable and no points with over 100 alternates. All of the address points were geocodable and had less than or equal to 100 alternates. The points that were not geocodable or had over 100 alternates did not produce results which could be compared quantitatively with the corresponding results for the other elements. To allow for a quantitative comparison based on comparable sample size across all four geoprivacy data elements, the most extreme ten percent (10 points) associated with each element were excluded.

Table 4 represents the summary statistics for the reverse geocoding effectiveness of the four layers and the initial hacked locations which serve as control values, and thus documents the accuracy of the initial map hacking process, prior to the reverse geocoding and geocoding steps. In addition to standard descriptive statistics measures, the Root Mean Square Error (RMSE) was used to evaluate reverse geocoding effectiveness. This computation measures the average magnitude of the error, giving a progressively higher weight to larger error values. RMSE provides an indication of the consistency of the process being measured and has been utilized in previous studies of spatial proximity and accuracy (e.g., Zandbergen and Green 2007).

Table 4. Reverse Geocoding Results

	Control	**Geoprivacy Data Elements**			
Summary Statistic (excludes 10% extreme observations)	Hacked Locations (no geocoding)	Non-county Maintained Roads	County Maintained Roads	Property Parcels	Address Points
Minimum (feet)	6.6	13.7	63.1	0.0	0.0
Maximum (feet)	81.3	503.5	357.1	93.5	51.3
Mean (feet)	39.8	144.1	132.8	19.4	0.6
Median (feet)	41.1	119.1	115.2	9.8	0.0
Standard Deviation (feet)	17.3	94.9	59.7	22.4	5.4
Root Mean Square Error	43.4	172.2	145.5	29.5	5.4
% Match Address*	Not Applicable**	3.3%	7.8%	87.8%	98.9%
% Match Street	Not Applicable**	66.7%	86.7%	98.9%	100.0%
Mean # Alternates	0.1	3.3	1.8	0.2	0.0

* When ownership information is present, an address match also reveals identity.

** The control value represents the initial map hacking effort, prior to address determination. Geocoding was not performed therefore these measures do not apply.

Table 4 shows that initial map hacking, as depicted by the values for the control “Hacked Locations”, was very accurate. The RMSE from hacked to actual locations was 43.4 feet with the least accurate point being an extreme observation at only 81.3 feet away from its actual location. Lending further support to the claim of high map hacking accuracy is that the mean number of alternates for the hacked locations was 0.1. These values are surprising in light of the map’s small scale (1:316,800) and limited detail. Furthermore, the map was reproduced twice (printing and scanning) prior to georeferencing and each reproduction presents an opportunity for errors to be introduced. No geocoding was performed for the hacked locations layer as it served as the starting point for the reverse geocoding process and was a control value to which the other elements were compared.

The Address Points layer produced the most accurate results, with the RMSE being 5.4 feet. This value is nine times greater than the mean of 0.6 feet and this disparity, although small in terms of real world distance, was influenced by the maximum value of 51.3 feet. The address and street address match rates were 98.9 percent and 100 percent, respectively, rendering the address point layer extremely effective at identity disclosure.

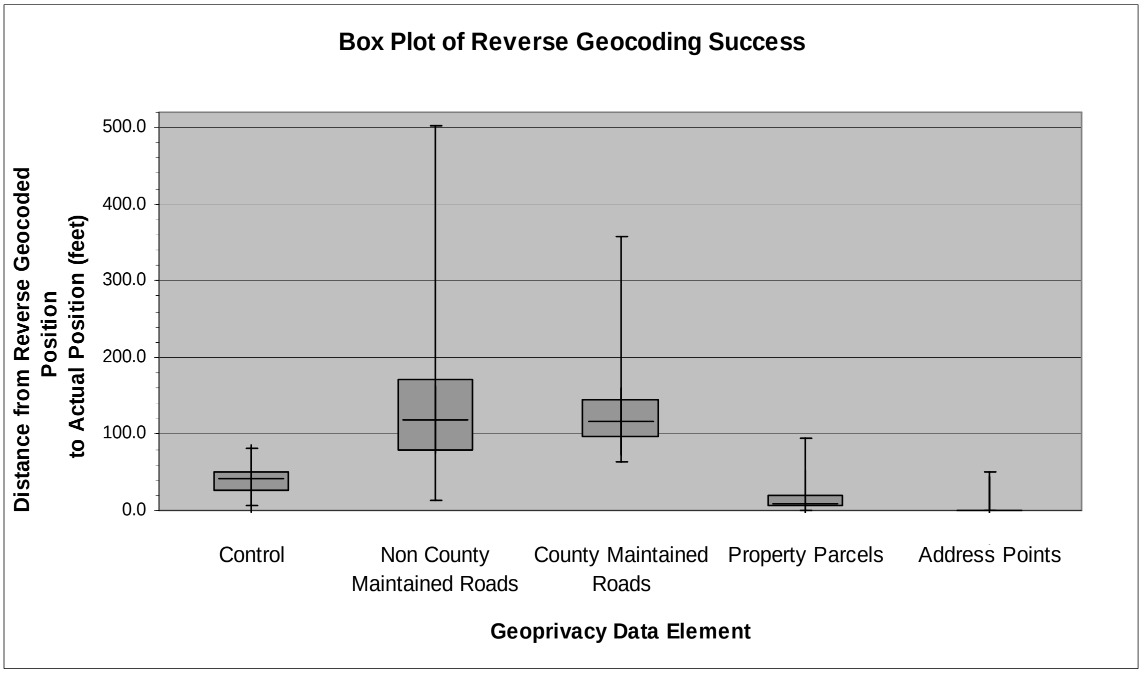
While the results for Property Parcels were also very accurate, these did not approach the accuracy of the address points, at least in a statistical sense. RMSE was determined to be 29.5 feet and match rates were 87.8 percent for the target's address and 98.9 percent for target's street. The maximum distance to actual was 93.5 feet, a relatively short distance for an extreme observation, but the majority of the points were very accurate as the mean for all points was only 19.4 feet and the median 9.8 feet.

When examining the results for the Non-County and County Maintained Roads layers, it is apparent that these geoprivacy data elements do not approach the high accuracy values of the Address Points and Property Parcels. RMSE values for the County Maintained Roads layer was 145.5 feet with an accompanying address match rate of 7.8 percent and street match rate of 86.7 percent. Results for the Non-County Maintained Roads layer were less accurate with a RMSE of 172.2 feet and address and street match rates of 3.3 percent and 66.7 percent, respectively.

Although many of the values for the road layers are several times greater than that of the Address Points and Parcels, it is important to process these values in their real world context. For example, although the RMSE value for Non-county Maintained Roads was nearly six times greater than that of the property parcels, the resulting difference in distance is only 142.7 feet; which is only a little less than half the length of a football field. The values for mean number of alternates for the roads layers were nine times higher than that of the parcels, but determining location to within an average of 3.3 alternates for Non-county maintained roads and 1.8 for county maintained roads still puts them reasonably close to the target.

The distribution of values representing reverse geocoding success (distance to the actual location) are depicted by a box plot in Figure 10. The extents of the vertically oriented lines represent the minimum and maximum distance to actual values, the extents of the grey box represent the upper and lower quartiles, and the horizontal black line which crosses the vertical line within the grey box represents the median value. This was prepared to visually convey and compare the aforementioned statistical parameters.

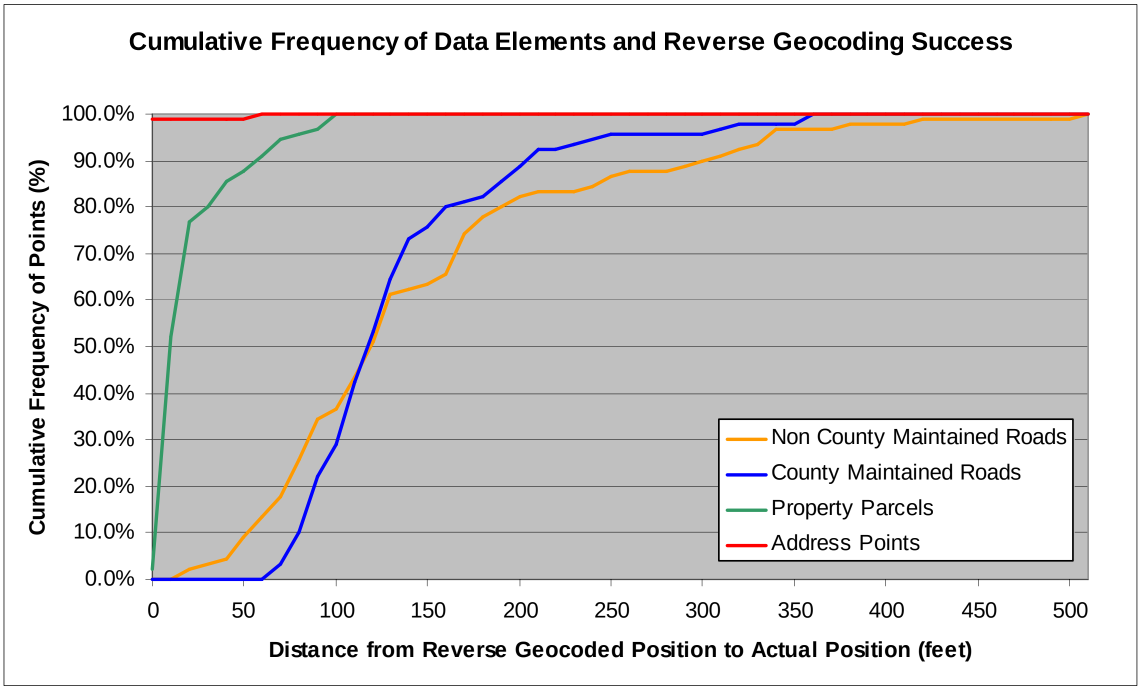
Figure 10. Box Plot of Reverse Geocoding Success



When examining the overall effectiveness of each element to disclose address and identity information, the summary statistics reveal a distinction between the road elements (non-county and county maintained) and the parcels and address points. When looking at the statistics alone, an argument could also be made that there is a similar disparity between the parcels and address points, but when these numbers are considered with respect to their real world significance the differences are marginal as they are both highly accurate.

Therefore, the parcels and address points were superior and similar in their reverse geocoding capabilities and the roads layers were inferior, yet also similar. A cumulative frequency distribution (Figure 11) of the data elements and their reverse geocoding success illustrates the pairing between the data elements. Whereas all points for the Address Point and Property Parcel layers are within 100 feet or less of the actual location, and most within 50 feet, both road layers exhibit a much more gradual, yet similar trend.

Figure 11. Cumulative Frequency Distribution of Data Elements and Reverse Geocoding Success



Although the roads were less effective in disclosing the actual address and do not have the capability to inherently reveal identity as do parcels, it should be noted that the mean number of alternates for both road data elements is still a remarkably small number: 3.3 for non-county maintained roads and 1.8 for county maintained roads. Despite not being able to provide the certainty of the parcels and address points, these elements did produce very few alternates and should be thought of as an effective means of narrowing down address possibilities to a particular street with few alternates.

4.3 Population Characteristics and Reverse Geocoding Success

Located in northwest Florida (aka the Panhandle) on the Gulf of Mexico, Bay County ranks 25th in the state with respect to population at 148,217 and is 30th in land area at 763.7 square miles. (USCB 2000) This puts the overall population density of the county at 194 people per square mile. Nearly 25 percent of Bay County's residents live within its largest municipality and most densely populated area, Panama City.

To investigate whether population density influenced reverse geocoding success within the county, a greater level of geographic detail was needed than an overall county population density calculation. Therefore, the county's 87 Census Block Group boundaries and associated population densities (Figure 12) were used to better reflect variability across the county. Summary statistics for Bay County's Census Block Groups are depicted in Table 5.

Table 5. Population Density Bay County

Population Density - US Census Block Groups (per 2000 Census)	
Summary Statistic	People per Square Mile
Minimum	9
Maximum	5,338
Mean	1,686
Median	1,525
Standard Deviation	1,192

Total Population of Bay County: 148,217

The population density value of the underlying the census block group was then assigned to each of the point locations of the test population (incidents) that were the target of this experiment. This value was then analyzed with respect to the distance that each geoprivacy data element and associated reverse geocoding method produced. This correlation analysis is depicted on the following scatterplots (Figures 13 - 16).

Figure 12. Bay County Population Density

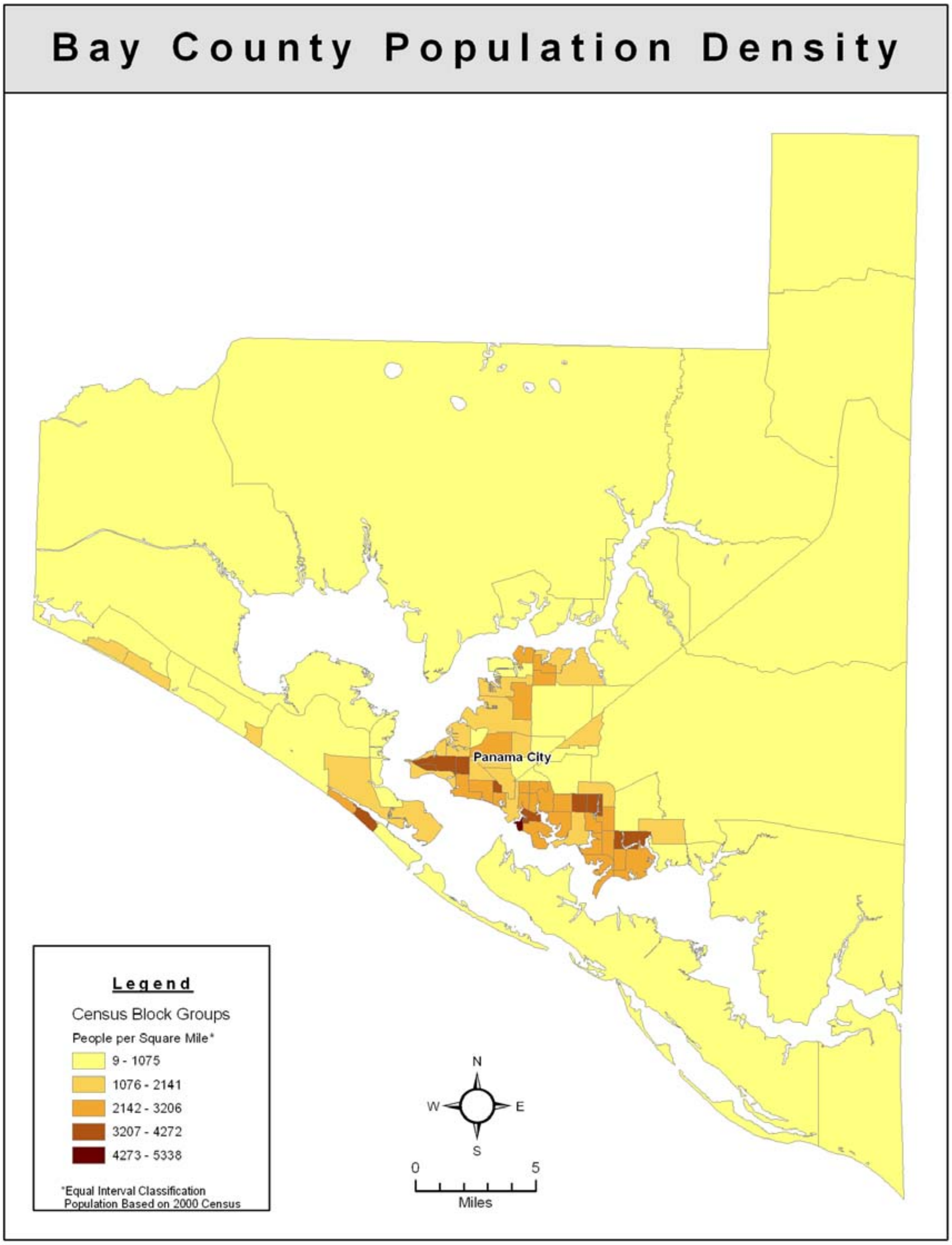


Figure 13. Scatter Plot: Non County Maintained Roads

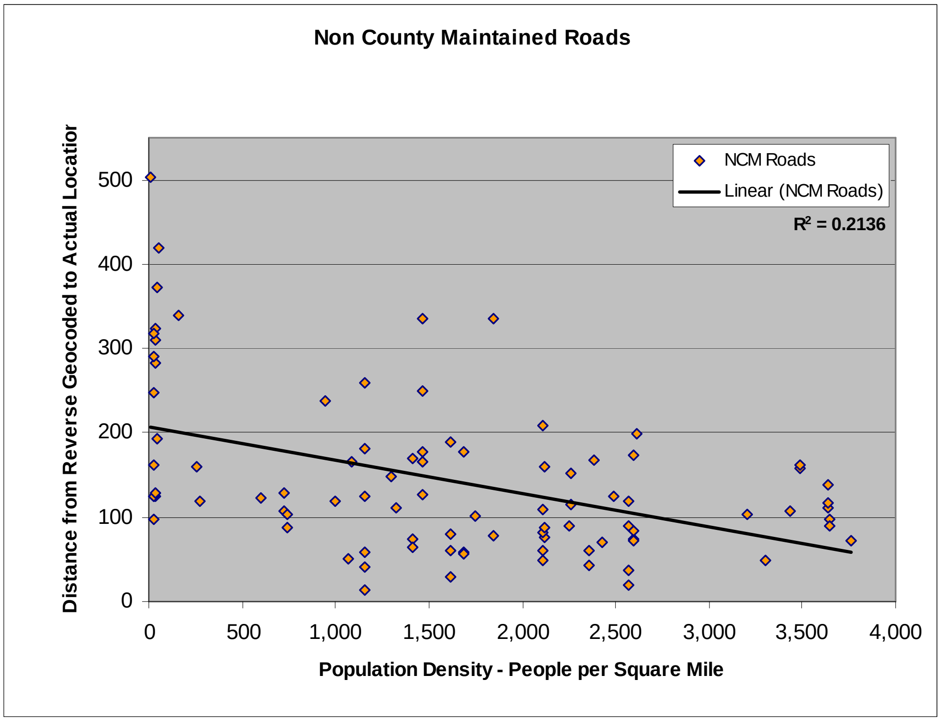


Figure 14. Scatter Plot: County Maintained Roads

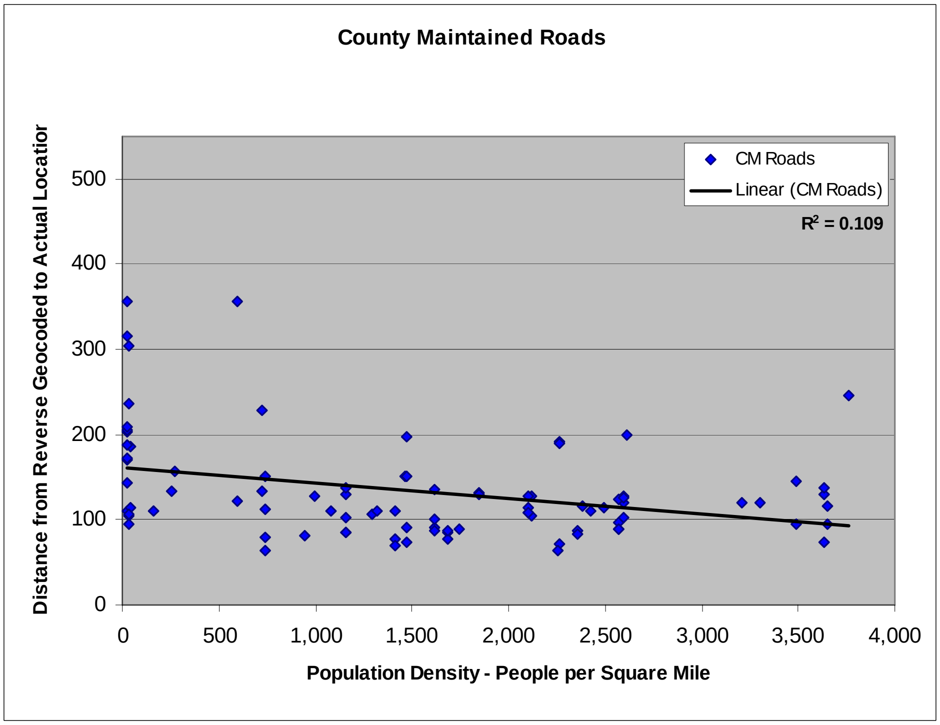


Figure 15. Scatter Plot: County Maintained Parcels

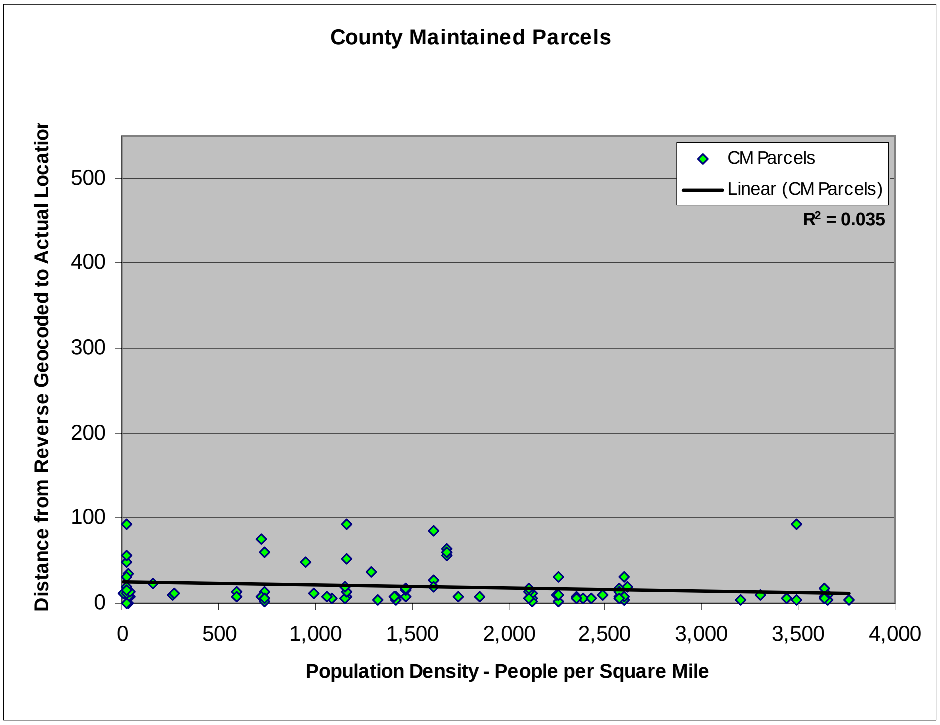
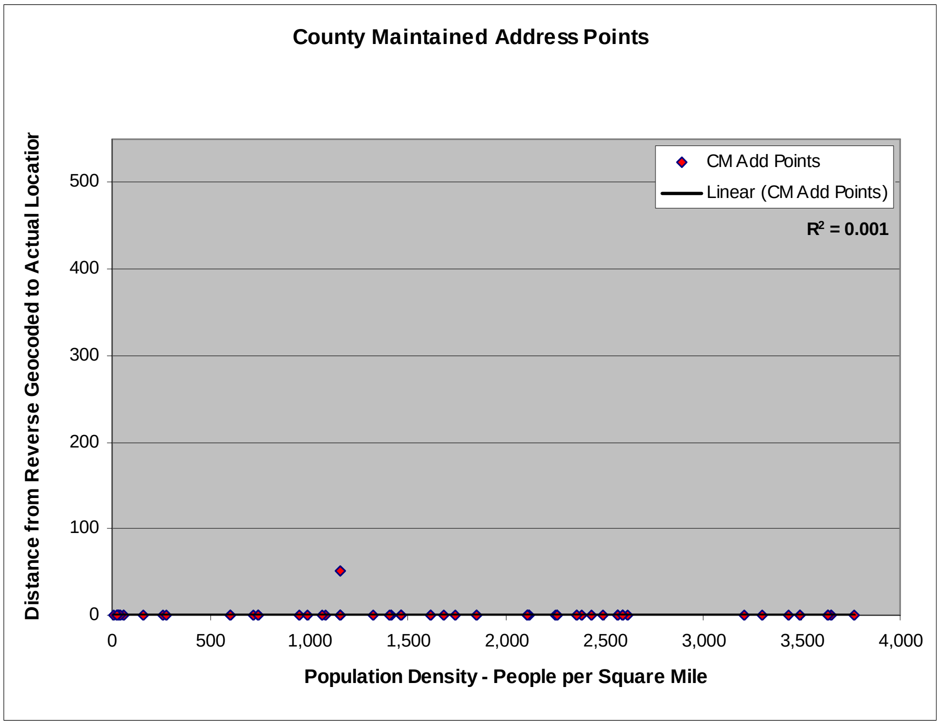


Figure 16. Scatter Plot: County Maintained Address Points



The results of the correlation analysis do not reveal any significant linear relationship between population density and reverse geocoding success for any data element. This is a probable consequence of the success of the initial map hacking effort, which placed points on average 39.8 feet from their actual location with the least accurate point registering 81.3 feet. Even in high density areas there can only be few, if any, alternates within these short distances. An exception would be high-rise residential structures, but these are not prevalent in Bay County.

Although low population density has been shown to negatively influence linear based geocoding, linear based *reverse* geocoding is a different process as it uses the street layer as a reference for determining the closest address to the target feature. Linear based geocoding was used, but it was based on an address derived from the source layer based on proximity, not an interpolation of a known address along a line segment. The address assigned via reverse geocoding may be incorrect, as is demonstrated by the low address match rates of both street based geoprivacy data elements, but it is an address that can reliably be used to locate the feature back to the closest address on the reference layer.   In summary, linear based reverse geocoding is not influenced in the same manner by population density as linear based geocoding, and in the case of Bay County population density does not influence line, point, or polygon based reverse geocoding success.

Data availability, the effectiveness of each geoprivacy data element, and the influence of population density has been determined; but what does this mean for the state of Florida? As stated before, Bay County has population characteristics (total, density) similar to Florida county averages (Figures 17 & 18). This likeness along with the finding that population density does not play a significant role in reverse geocoding success permits the results for Bay County to be applied to all counties and suggests statewide vulnerability.

Figure 17. County Population

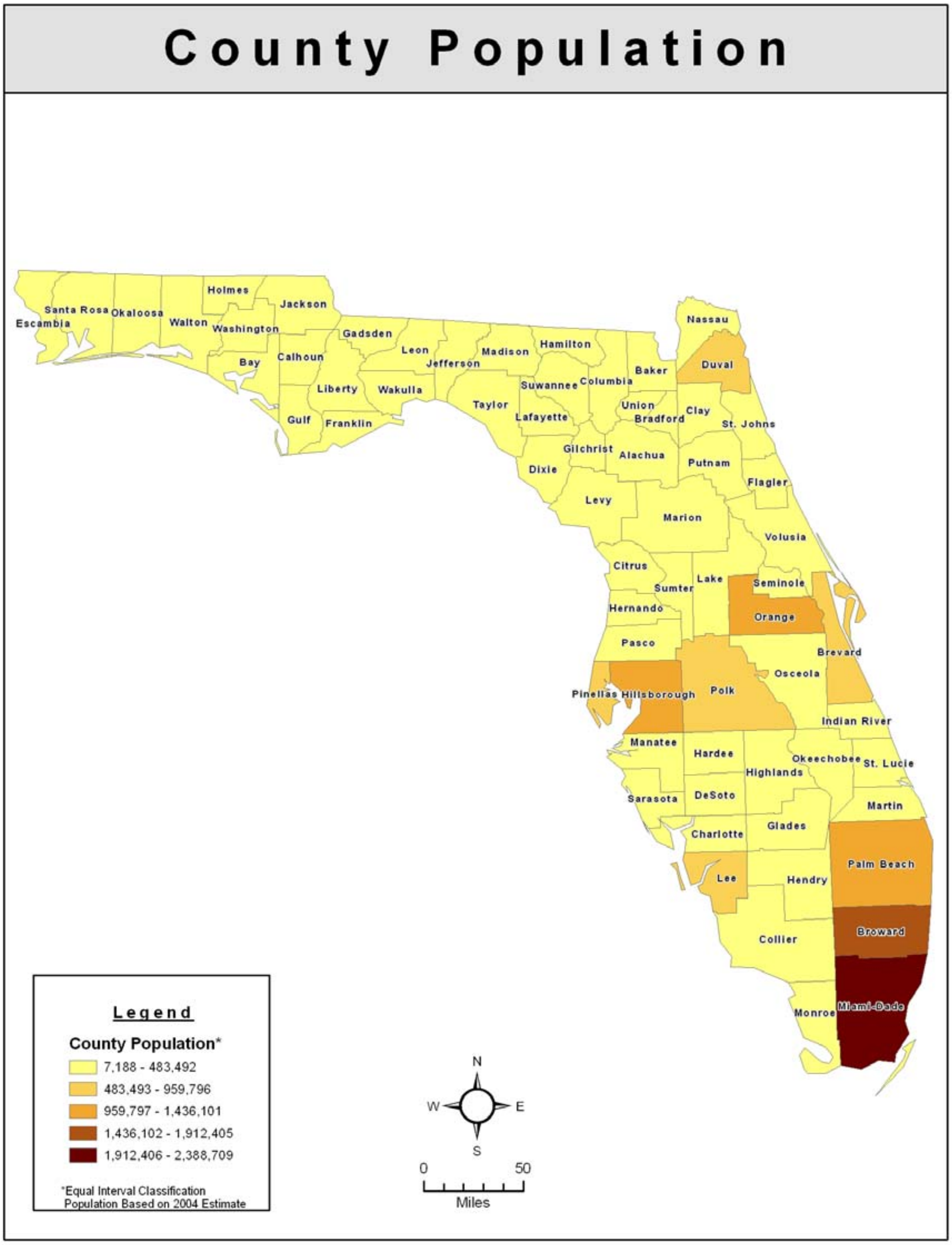
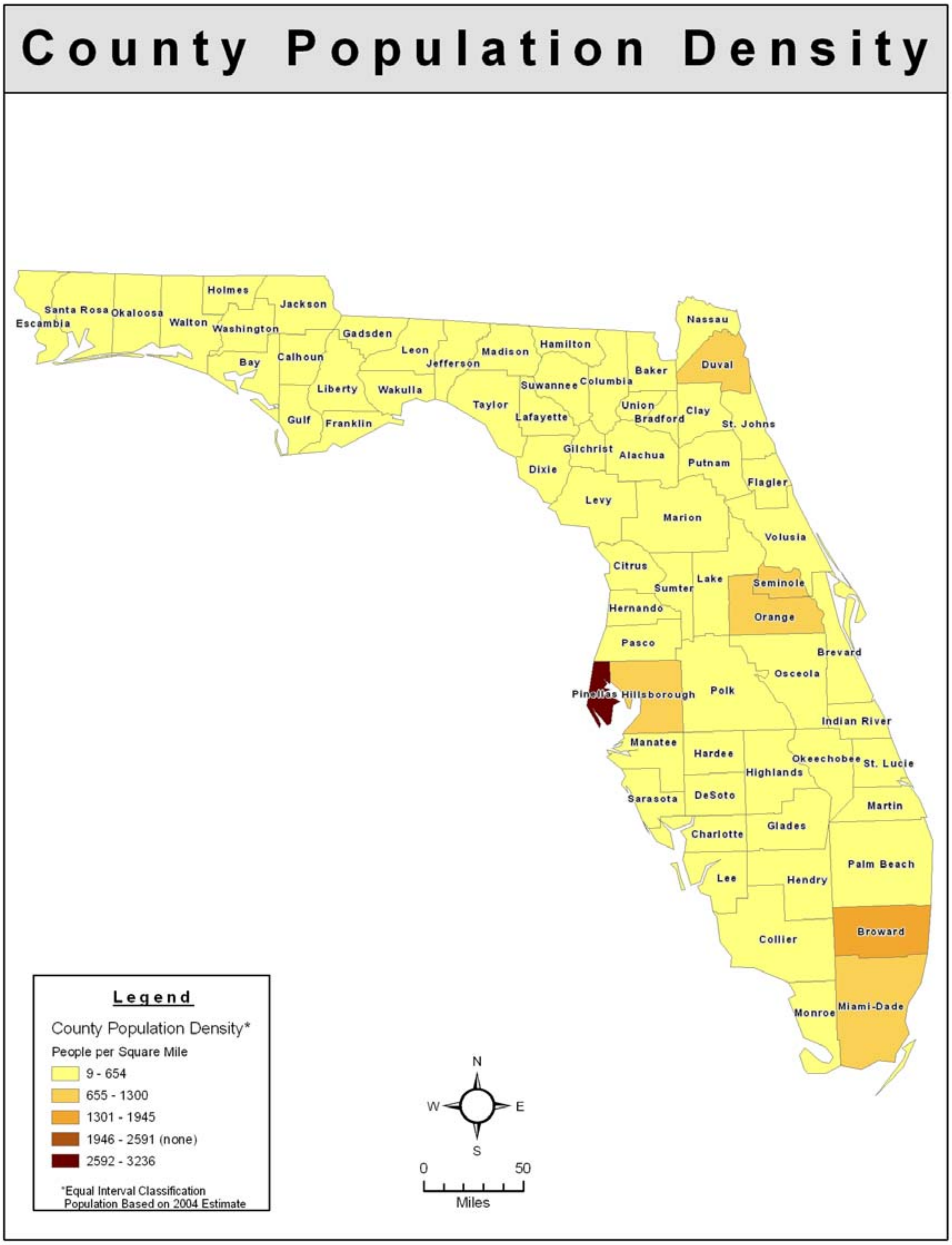


Figure 18. County Population Density



With population size and density not significantly influencing reverse geocoding success, it is apparent that data availability is the key component of geoprivacy risk. Considering that address points and parcels can provide for successful identity disclosure, the state's population (United States Census Bureau, 2004 Estimate) was totaled for each geoprivacy data element and corresponding availability code (Table 6). This was developed to provide some insight as to the amount of people who, due to the data that their county of residence provides, are potentially exposed to geoprivacy threats. As Table 6 indicates, approximately two thirds of Florida's residents live in counties that make parcel data available, and nearly half of those counties make it available for free. Given the high success rate of parcels to disclose both address and identity, this is reason for concern as it indicates that majority of the state is exposed to a potential geoprivacy risk.

Table 6. Population Facing Geoprivacy Risk

Population Facing Geoprivacy Risk			
	Streets	**Parcels**	**Address Points**
Yes	8,514,917	5,513,128	4,716,269
Purchase	912,839	7,519,229	211,446
Indirect	299,712	1,310,939	1,077,635
Ineffective	1,526,169	110,816	0
No	26,235	0	3,676,652
Inconclusive	6,183,176	3,008,936	7,781,046
Total	**17,463,048**	**17,463,048**	**17,463,048**

Population based on USCB 2004 Estimates

5. CONCLUSIONS

This thesis explored a very imminent, yet relatively unknown issue affecting every citizen in the state of Florida, and perhaps elsewhere; geoprivacy. Two very important questions were explored, the first of which being the manner in which certain types of information necessary to engineer geoprivacy violations influence success. The results indicate that if no masking of sensitive data occurs and reference material displayed on a map can be discovered or accurately replicated, reverse geocoding can be very accurate; even when working from small scale material that has been reproduced. All geoprivacy data elements evaluated could be used to assign an address to a point which lacks one and can subsequently serve as a reference layer to locate that address accurately relative to itself. The existence of parcel data or address points allow for a high risk of identity disclosure, assuming that address and ownership information is inherent to the data. Non-county and county maintained roads exhibit a moderate to high success rate of identifying the target's street and produced few alternates. Although less successful than parcels and address points, both street layers pose a threat for address disclosure; albeit a less certain one. The existence of county maintained roads does not greatly increase the chance for address disclosure when compared to non-county maintained roads and neither road layers suggest a risk of personal identity disclosure as ownership information is not inherent.

The second question addressed by this research was regarding the extent to which Florida counties and their inhabitants at risk for geoprivacy violations. Although the entire state is subject to the Public Records Law, there is substantial variability in the amount, quality, accessibility, and delivery methods of county maintained spatial data. Despite these inconsistencies, this study suggests that current data availability and associated reverse geocoding success makes the majority of the state's residents vulnerable to geoprivacy violations. The widespread accessibility of parcel data, its associated reverse geocoding success, and its lack of statistical association with population density support this claim.

With these findings, this thesis makes a unique contribution to the existing body of geoprivacy research by creating a comprehensive forum that exposes the geoprivacy threat to a greater audience while retaining its academic significance. It is anticipated that parties from various social, economic, and professional backgrounds will use this research to engage in geoprivacy policy discussions. State and local government should consider the privacy risks that spatial information can create for its citizens and a debate which involves all parties should be initiated to address geoprivacy concerns. Successful collaboration will result in a more effectively managed threat; allowing geospatial research and services to provide continued public benefit while simultaneously protecting the privacy of its subjects, which has been proven to be threatened by current practices.

It is important to consider some of the limitations of this study and explore related avenues for future research. An important limitation of this research was that statewide data availability and quality could not be determined with absolute certainty. Several geoprivacy data elements (77 of 201) were marked as “Inconclusive” for several counties. A more exhaustive effort to determine data availability could resolve these uncertainties. In addition to resolving the “Inconclusives”, an examination of the geoprivacy data elements that were available only for purchase would be needed to accurately determine their suitability for use in a reverse geocoding operation. Although this was attempted to be resolved through careful wording of the data requests (with ownership information, geocodable, etc.), without first-hand knowledge the county representative had to be relied upon to accurately convey the layer’s capabilities. Another opportunity to add to this research effort would be to modify the testing population and sampling techniques. One random sample of only 100 individuals from one county was used. Additional experiments conducted with larger samples and/or samples from other counties using their geoprivacy data elements could extend the findings presented in this study. Within these limitations, however, the methods used for this experiment are believed to be appropriate and the conclusion regarding statewide vulnerability conceptually and methodologically valid.

Future research endeavors should approach the geoprivacy threat in a holistic manner and carefully explore policy alternatives for managing the vulnerabilities associated with making private data publicly available. The geoprivacy data elements used for the purposes of this thesis exist because they provide a service to the entities who commissioned their development. By allowing public availability, this data also provides a benefit to many non-governmental entities that use this information for purposes which are part of the state's intellectual and economic engines. Widespread data availability, fostered by the state's Public Records Law, has created a complex scenario where there are concurrent positive and negative outcomes. While personal privacy is the key concern addressed by this thesis, it is important to include these and other external, yet important, factors when addressing policy alternatives. Such factors can only be identified, however, if all parties contributing to, knowledgable of, and influenced by geoprivacy concerns are involved. This will require extensive engagement and collaboration between public officials, citizens, GIS professionals, corporate representatives, and the academic community.

6. REFERENCES

Armstrong MP, Ruggles AJ. Geographic Information Technologies and Personal Privacy. *Cartographica* 2005;40:4:63-73

Armstrong MP, Rushton G, Zimmerman DL. Geographically masking health data to preserve confidentiality. *Statistics in Medicine* 1999;18:497–525.

Brownstein JS, Cassa CA, Kohane IS, Mandl KD. An unsupervised classification method for inferring original case locations from low-resolution disease maps. *International Journal of Health Geographics* 2006;5:56:1-7

Cayo MR, Talbot TO. Positional error in automated geocoding of residential addresses. *International Journal of Health Geographics* 2003;2:10.

Curtis AJ, Mills JW, Leitner M. Spatial confidentiality and GIS: re-engineering mortality locations from published maps about Hurricane Katrina. *International Journal of Health Geographics* 2006;5:44:1-12

Dent, BD. *Cartography Thematic Map Design* (5th Edition). 1999. Boston, MA: McGraw-Hill.

Kamel Boulos MN, Cai Q, Padget JA, Rushton G. Using software agents to preserve individual health data confidentiality in micro-scale geographical analyses. *Journal of Biomedical Informatics* 2006;39:160-170

Kwan M, Casas I, Schmitz BC. Protection of geoprivacy and accuracy of spatial information: how effective are geographical masks? *Cartographica* 2004;

39:2:15-28

Olvingson C, Hallberg J, Timpka T, Lindqvist K. Ethical issues in public health informatics: implications for system design when sharing geographic information. *Journal of Biomedical Informatics* 2003;35:178-185

Rushton G, Armstrong MP, Gittler J, Greene BR, Pavlik CE, West MM, Zimmerman DL. Geocoding in cancer research. *American Journal of Preventative Medicine* 2006; 30(2S):S16-S24

State of Florida Statutes 2007 http://www.leg.state.fl.us/Statutes/index.cfm (Last Accessed 11/19/2007)

United States Census Bureau, State and County Quick Facts http://quickfacts.census.gov/qfd/states/12000.html (Last Accessed 11/13/2007)

VanWey LK, Rindfuss RR, Gutmann MP, Entwisle B, Balk DL. Spatial Demography Special Feature: Confidentiality and spatially explicit data: Concerns and challenges. *Proceedings of the National Academy of Sciences* 2005;102:15337-15342

Whitsel EA, Quibrera PM, Smith RL, Catellier DJ, Liao D, Henley AC, Heiss G. Accuracy of commercial geocoding: assessment and implications*.* *Epidemiologic Perspectives and Innovations* 2006;3:8:1-12

Zandbergen PA, Green JW. Error and Bias in Determining Exposure Potential of Children at School Locations Using Proximity-Based GIS Techniques. *Environmental Health Perspectives* 2007;115:9:1363-1370.