Journal of Geography and Geology
http://ccsenet.org/jgg

Vol. 1, No. 2
November 2009

Spatial Pattern of Uncertainties: An Accuracy Assessment of the TIGER Files

Byoungjae Lee (Corresponding author)
Department of Information Technology Leadership, Washington & Jefferson College
Washington, PA 15301, U.S.A
Tel: 1-724-503-1001 ext.3403 E-mail: blee@washjeff.edu

Abstract

This study discusses the ways how the positional accuracy of the TIGER files can be measured and spatially reported. Many people and companies use the address range of the TIGER files with the geocoding package within a Geographic Information Systems (GIS). However, the problem is that many people have little understanding of the inaccuracy of the TIGER files. This study examines the relationships between the distribution of inaccuracy and physical factors such as stream and urbanity. Next, the inaccuracy of the hydrography shape file of TIGER 2000 files is calculated by comparing it with the stream points data of United States Geological Survey (USGS)'s Geographic Names Information System. Finally, this study examines whether there are individual patterns in each spatial data by comparing the spatial pattern of the inaccuracies of the road and hydrography shape file.

Keywords: GIS, Spatial data quality, Inaccuracy, TIGER files

1. Introduction

As the TIGER files are used more widely, the necessity of making any inaccuracy of the TIGER files generally known increases. Because they are free, the TIGER files are widely used. Studies at the individual address level are now generally carried out. Improvements of the ability within GIS and the increase of storage capacity make this possible. However, compared with non-free data, the accuracy of TIGER files is of lesser quality. The TIGER files were built and have been continuously updated using a wide variety of source materials and techniques, including the GBF/DIME files, USGS 1:100,000-scale topographic maps, local and tribal maps, and enumerator updates of differing positional accuracy (O'Grady and Godwin, 2000). The varied update history has resulted in the inaccuracy of the TIGER files. Hence, an accuracy assessment of the TIGER files is necessary.

Ratcliffe (2001) shows a practical example. He performed an accuracy assessment of individual address locations in the form of high-resolution geocoded point data, by comparison with both cadastral records that delineate the individual target properties, and areal units. These studies concentrated on assessing the accuracy of the spatial data by using computer-graphical methods. They applied the same standards for each region. They assumed that the imperfections of the spatial data resulted from only the carelessness of the mapmaker. But, they didn't provide a statistical trend or spatial pattern for the inaccuracies. To find the reasons for the inaccuracies, a statistical approach is required.

For the GIS researcher using the TIGER files, the spatial pattern and causes of the inaccuracies and the statistical mapping of the inaccuracy can help to eliminate the imperfections of their projects. Moreover, people who use the information derived from the TIGER files can interpret the information correctly.

The objectives of this study are to find the spatial pattern and reasons of the inaccuracies by statistical methods. If the reason or pattern is revealed, it is very helpful to minimize the distortion when people perform the project by using the TIGER files. Although the TIGER files are very popular data, many people have little understanding of the inaccuracy. Even people who do have an understanding often don't know why the TIGER files are inaccurate.

The definition of accuracy is the degree to which information on a map or in a digital database matches true or accepted values. There are four types of accuracy. These are positional, attribute, conceptual, and logical accuracy (Goodchild and Gopal, 1991). This study focuses on the positional accuracy and uses mainly the TIGER 2000 file of Erie County, New York State. The positional accuracy of a spatial object or a digital representation of a feature is the measurement of the difference between the apparent location of the feature as recorded in the databases, and the true location

(Goodchild and Hunter, 1997). For the reference data, Geography Data Technology, Inc. (GDT) Dynamap/2000 Street network data and United States Geological Survey (USGS)'s Geographic Names Information System data are used. And, for the tested data, the roads and hydrography shape file of the TIGER 2000 files are used. According to GDT(Note 1), they used much more points than any other non-free data like TIGER 2000 files. They are continuously updating more often with new information from the USPS and many private sources. Furthermore, because GDT also participated in creating TIGER files, it can be minimized the undesired errors. The errors can result from the different creation process of the reference data and tested data. This study is divided into two parts. One is the comparison of the road shape file and GDT Dynamap/2000 Street network data by using geocoding method. The other is the comparison of the hydrography shape file and the stream points data of USGS's Geographic Names Information System. By using ArcGIS Spatial Analyst and Geostatistical Analyst, the spatial patterns of the inaccuracies in each shape file are found. The results will show that there are individual spatial pattern of the inaccuracies in each spatial data.

2. Background

2.1 Data quality issues in GIS

The computing saying 'garbage in, garbage out' applies to GIS since if you put poor quality data into your program, the quality of your output will be poor. The results of analysis are only as good as the data put into the GIS (Heywood et al., 1998). Concern for geospatial data quality has grown rapidly because of increased data production by the private sector, increased use of GIS as a decision-support tool, and increased reliance on secondary data sources. These trends have affected the responsibilities of data producers and consumers for data quality. The producer was responsible for only sanctifying databases meeting official quality thresholds (Veregin, 1999). Heywood et al. (1998) mentioned that two issues are important in addressing quality and error issues: first, the terminology used for describing problems, and second, the sources, propagation and management of errors. However, Duckham (2002) noted that an obvious criticism about many spatial data quality standards and research is that these focus only on the storage, management and propagation of data quality information rather than how to use such information. Moreover, he insisted on the importance of the error-sensitive GIS. The error-sensitive GIS can be characterized as comprising three distinct stages: first, deciding upon the core data quality concepts; second, developing and implementing an error-sensitive data model based on these concepts; and third, developing interfaces able to deliver the error-sensitive services and functionality to users.

At this point, most people's perspectives are generally in sympathy on the importance of the concept of 'fitness for use' about the spatial data quality. Responsibility for assessing whether a database is proper for the needs of a particular application has shifted to the data users (Veregin, 1999). The data providers should supply enough information about the quality of a data set to help a data user make a proper decision in a particular situation (Chrisman, 1991). To meet "fitness for use," the producer's role has shifted to data quality documentation or "truth-in-labeling." According to the truth-in-labeling paradigm, errors are inevitable and the data quality problem results from incomplete knowledge of data limitations (Veregin, 1999). Nevertheless, the fitness for uses of a data set cannot be assessed entirely objectively. Rather than a simple 'yes' or 'no' answer, 'fit' or 'unfit', the degree of fitness for use will be qualified subjectively (Duckham, 2002). This results in various demands of the different users on data quality issues. Even a single organization or person may perform many of the different roles. In spite of the importance of 'fitness for use', previous methods focused only on quantitative factors such as how close the point is to the real point. This is insufficient to meet the requirements for 'fitness for use'. To meet the various demands of different users, detailed characteristics of the errors are considered necessary. The purpose of this paper is to find the spatial pattern of the inaccuracy and to provide the possibility to maximize fitness for use. This view corresponds to the view of the truth-in-labeling paradigm. The errors are not just a bad thing, but an inevitable thing. The errors are another attribute of the spatial data. Thus, the characteristics of the errors in spatial data should be clarified so that good quality results and output can be produced.

2.2 Measuring the positional accuracy of spatial data

An assessment of positional accuracy is related to the quality of the final product after all transformations. The lineage part of the quality report deals with the information on transformations. In the description of positional accuracy, the date of the test should be included. Additional attributes of spatial objects or a quality overlay (reliability diagram) is needed for variations in positional accuracy (U.S. Geological Survey, 1997).

Drummond (1995) divides the determination of positional accuracy into two steps. One is to measure the error generated by the systems. The other is to estimate the error generated by the systems. 'Measure' means to only consider the final positional information and compare the tested data to a known higher standard. This approach requires the availability of checkpoints whose *x*, *y* and *z* values are known. 'Estimate' needs the associated contributing standard deviation in each step of the processing.

Open GIS Consortium (1999) has developed Drummond's idea. Open GIS Consortium breaks down error estimation methods into five groups (Note2): professional estimate, computed estimate, compared to similar quality data, tested similar quality data, and tested sample actual data.

Practically, accuracy testing is performed in terms of horizontal accuracy and vertical accuracy. FGDC (2002) provides the standards for accuracy testing and verification. "Map testing should be performed within a fixed time period after delivery. Horizontal accuracy is tested by comparing the planimetric coordinates of well-defined ground points with coordinates of the same points from an independent source of higher accuracy. Vertical accuracy is tested by comparing the elevations of well-defined points with elevations of the same points as determined from a source of higher accuracy."

U.S. Geological Survey (1997) made a synthesis of the methods for positional accuracy in Spatial Data Transfer Standards (SDTS) into four categories: deductive estimate, internal evidence, comparison to source, and independent source of higher accuracy.

The concepts of the U.S. Geological Survey (1997) have a connection with the previous views. *Deductive estimate* includes not only the *Estimate* concept of Drummond (1995), but also the professional estimate and computed estimate of Open GIS Consortium (1999). 'Independent Source of Higher Accuracy' indicates comparison of data with high quality data. Thus, it includes the 'Measure' concept of Drummond (1995), the three comparison methods of Open GIS Consortium (1999), and the horizontal/ vertical accuracy concepts of FGDC (2002).

2.3 Previous analyses on TIGER/Line files

According to the U.S. Census Bureau, TIGER 2000 Line files are designed to show only the relative positions of elements. In the 2000 TIGER/Line files technical documentation, the following statements appear about positional accuracy:

"Coordinates in the TIGER/Line files are in decimal degree and have six implied decimal places. The positional accuracy of these coordinates is not as great as the six decimal places suggest. The positional accuracy varies with the source materials used, but at best meets the established National Map Accuracy standards (approximately +/- 167 feet) where 1:100,000 scale maps from the USGS are the source. The U.S. Census Bureau cannot specify the accuracy of feature updates added by its field staff or of features derived from the GBF/DIME-Files or other map or digital sources."

Previous analyses on TIGER/Line files have focused on comparing it with an 'independent source of higher accuracy'. Zent (1996) compared the U.S. Census Bureau's TIGER/Line files with the U.S. Geological Survey's Digital Line Graph (DLG) files. The relative accuracy of TIGER/Line files and DLG were tested by using the content of USGS 1:12,000 scale digital orthophoto quarter-quadrangle backdrop images as the reference data for dataset accuracy. The results demonstrated that TIGER/Line files, in 4 out of the 6 study areas, were found to be more positionally accurate in their coordinates' location than the DLG dataset intersections.

Ratcliffe (2001) tested a TIGER-type geocoding process by using point-in-polygon methods. A study of over 20,000 addresses in Sydney, Australia showed that 5-7.5% of addresses may be misallocated to census tracts and more than 50% may be given coordinates within the land parcel of a different property.

After reviewing the positional accuracy information of TIGER/Line files, O'Grady et al. (2000) stated three needs to improve the positional accuracy of TIGER: internal needs, a desire to use local and tribal files for updates, and a desire to facilitate data exchange. Here, internal needs are related to a technological requirement. For example, Global Positioning Systems (GPS) technology is considered a powerful way to capture new coordinates for existing anchor points.

According to Liadis (2000), the Geography Division (GEO) of the U.S. Census Bureau uses the GPS to assess the spatial accuracy of the TIGER data base in its preparation for TIGER modernization. A tool called the GPS TIGER Accuracy Analysis Tool (GTAAT) is developed to evaluate the spatial accuracy of attributes derived from a variety of operations and sources. The GTAAT calculates the distance and azimuth difference between the GPS collected point and the equivalent TIGER 0-cell (point). By utilizing the GTAAT, it was revealed that there was a large variance in the mean distance difference from TIGER to ground truth based on the source code. It resulted from an inherent positional accuracy of each data source. Thus, the GEO concluded that the current accuracy of point and linear features in the TIGER system limits the ability to exchange data digitally through partnerships.

Moreover, O'Grady (2001) introduced a DOQ (Digital Orthophoto Quadrangles) test method to improve the TIGER. She stated that there are two components of TIGER improvement: Updating the data base by adding new features and spatially enhancing existing features. GPS technology is useful to test the updated TIGER data base, while, improving the positional accuracy of and spatially enhancing TIGER is tested by the DOQ. The DOQ test is composed of two parts. One is to capture the coordinates of certain TIGER feature intersections called "anchor points". The other is to transform all TIGER coordinates using the newly collected DOQ anchor point coordinate data.

To sum up, previous analyses on TIGER/Line files show that everyone agrees that the inaccuracy problem of TIGER/Line files limits the ability to exchange data. Thus they attempt to test TIGER/Line files with various methods and reference data. However, there are no attempts to maximize current fitness to use with detailed information of the inaccuracies. Therefore, the spatial characteristics of the inaccuracies are considered necessary for that.

3. Study area

According to Erie County Works (Note3), Erie County is a metropolitan area located in the western part of New York State. It covers 1,058 square miles. The County is bounded by Lake Erie to the west, Niagara County and Canada to the north, Genesee County and Wyoming County to the east, and Cattaraugus and Chautauqua Counties to the south. "More than half of the population in both countries, as well as 52 percent of the personal income ($1.4 trillion) created by the United States and Canada are within 500 miles of Erie County. In addition, three-quarters of Canada's manufacturing activity and 55 percent of the United States' manufacturing activity fall within that radius. Located within the County are three cities and 25 towns, including the City of Buffalo, the second largest city in New York State." The land use pattern has led to expansion in the suburban towns and a mixed pattern of stability, decline, and redevelopment in the City of Buffalo. The northern towns have grown relatively more. The eastern towns are beginning to develop, while the southern towns are developing at a slower pace.

4. Data

4.1 TIGER files

This study uses mainly the roads and hydrography shape file of the TIGER 2000 files. According to the U.S. Census Bureau (Note 4), most information in TIGER outside the urban centers was derived from the USGS 1:100,000-scale digital line graphs, which were vectorized from the digital scanning of the original artwork. The original artwork was in Universal Transverse Mercator (UTM) projection. After the map sheets were scanned, the coordinates were transformed from UTM into projectionless geographic coordinates of latitude and longitude. For most urban centers, the information in TIGER was derived from the GBF/DIME files produced for the 1980 census. There were a variety of other sources used in creating the Census TIGER data base. The features from those sources also were stored as latitude and longitude coordinates. Subsequent updates to the Census TIGER data base also came from a variety of sources, including paper maps annotated in the field and subsequently digitized without rigorous adherence to a projection or coordinate system.

4.2 Dynamap/2000 Street network data

This data is used for testing the accuracy of the road shape file of TIGER 2000 files. Geography Data Technology (GDT), Inc. built the Dynamap/2000 Street network data. According to GDT (Note 5), the boundary layers of the Dynamap/2000 Street network data, except for ZIPs, have not been generalized. Every polygon (area surrounded by boundary segments) and every feature (geographic unit formed by one or more polygons) has as many points as are required to draw its shape accurately. Hence, this data was used for the reference data. The version of the Dynamap/2000 files used in this study is 11.2 (July 2001). The scale of this data is 1:24,000. All coordinates are based on the 1983 North American Datum (NAD83), like the TIGER 2000 files.

4.3 Address data

This data is used for performing geocoding with the road shape file and GDT Dynamap/2000 street network data. For statistical analysis, randomly and independently selected address data is needed. To perform geocoding, the address data of the schools in Erie County was used. Compared with other kinds of data such as hotels, restaurants, and so on, schools are evenly distributed and each community has schools. The school address data is acquired from National Center for Education Statistics (NCES) website (Note 6).

4.4 Geographic Names Information System

This data is used for testing the accuracy of the hydrography shape file of TIGER 2000 files. According to the USGS (United States Geological Survey) (Note 7), "The Federally recognized name of each feature described in the data base is identified, and references are made to a feature's location by State, county, and geographic coordinates." In this study, the stream points data in Erie County, New York is used. According to the metadata, the accuracy of these data is based on the use of source graphics which are compiled to meet National Map Accuracy Standards. The main sources are 1:24,000-scale topographic maps, records of the U.S. BGN, and U.S. Forest Services 1:24,000-scale topographic maps. Because the TIGER files are based on 1:100,000-scale topographic map, these data can be used as more accurate reference data.

5. Methodology

5.1 Test of the road shape file of TIGER 2000 files

The purpose of this part is to test whether there is spatial pattern of the inaccuracies and whether there is a relationship between these spatial patterns and physical factors.

5.1.1 Data preparation

Using the road shape file of TIGER 2000 files as the tested data and the Dynamap/2000 Street network data as the reference data, geocoding is performed. The output of the geocoding is x, y coordinates. The street files option in

ArcGIS is checked. In geocoding, the suitable line segment is selected by using the target address, and then a location is interpolated between the ‘from node’ and the ‘to node’.

Ratcliffe (2001) showed the importance of the offset in geocoding. Moreover, he mentioned the potential problems with the geocoding. The problems are out-of-date street directories, abbreviations or misspelling, local name variations, address duplications, non-existent address, line simplification, noise in the address file, geocoding non-address locations, geocoding imprecision, and ambiguous or vague addresses. To overcome these problems, addresses that don’t score 100 or match a unique location are ruled out. Furthermore, by using statistical methods such as leverage values, outliers are excluded.

5.1.2 Inaccuracy mapping and finding the reasons of the inaccuracy

After the geocoding is performed, the distances between the coordinates from the tested and reference data are calculated. There are many ways to calculate the distance between two points on the earth's surface, defined by their latitude and longitude. In this study, the Great Circle Distance based on Spherical trigonometry is used. This method assumes that 1 minute of arc is 1 nautical mile and 1 nautical mile is 1.111 miles. The formula (Note 8) is as shown below.

$$D = 1.111 * 60 * \mathrm{ARCOS}\ (\mathrm{SIN}\ (L1) * \mathrm{SIN}\ (L2) + \mathrm{COS}\ (L1) * \mathrm{COS}\ (L2) * \mathrm{COS}\ (DG)) \tag{1}$$

L1 = latitude at the first point (degrees)

L2 = latitude at the second point (degrees)

DG = longitude of the second point minus longitude of the first point (degrees)

D = computed distance (mile)

The distances with the addresses are divided into 5 categories by natural break. Then, these are mapped within ArcGIS.

Here, to find the spatial pattern of these mapped points, interpolation is performed. Interpolation means to predict values at locations where data has not been observed. To do that, Kriging was used in the Geostatistical Analyst in ArcGIS is used. According to Johnston el al. (2001), the kriging is a statistical interpolation method that uses data from a single data type to predict values of that same type at unsampled locations. After looking over the result of the interpolation, the independent variables are selected for the distances as the dependent variables. For example, the length of the line and the width of the line can be independent variables for the inaccuracies. Finally, correlation values between independent variables and the distances are calculated.

5.2 Test of the hydrography shape file of TIGER 2000 files

The purpose of this part is to examine whether there is an individual spatial pattern of the inaccuracies in each spatial data set by testing another spatial data set.

5.2.1 Data preparation

The stream points data in Erie County, New York are obtained by querying the Geographic Names Information System (GNIS) online database. The fields in the result table are feature name, state, county, type such as stream, latitude, longitude, and related USGS 7.5’ map. The number of points is 73.

These points don’t have a specific pattern. In ArcGIS, the point data should have the decimal degree coordinates to create point coverage. Thus, in Microsoft Excel, the latitude and longitude of the stream points are converted to decimal degrees. By using converted decimal degrees, the point coverage is created. The hydrography shape file of TIGER 2000 files is also converted to arc coverage by utilizing ArcToolbox in ArcGIS. To calculate the distance between each stream points and hydrography shape file, these two data should be coverage data formats that have a topology.

5.2.2 Inaccuracy mapping and finding the reasons of the inaccuracy

The distance between the stream points and the hydrography shape file is calculated by using ArcToolbox’s near function in the Analysis category. The distances are divided into five categories by natural break. Then, ArcGIS performed mapping with this distance. To find the spatial pattern of these mapped points, interpolation is performed. To do that, the Geostatistical Analyst in ArcGIS is used. In the Geostatistical Analyst, the kriging method is selected.

6. Results

6.1 Inaccuracy of the road shape file

For the geocoding, the addresses of the 235 public schools in Erie County, New York are used. After the points that don’t score 100 or match a unique location are ruled out, 187 points remain. When the distances between the coordinates that come from the tested and reference data are calculated, the points whose distance is more than 1 mile are regarded as outliers. Now, 168 points remain (Figure 1). These 168 points are used for the comparison of the reference data and target data.

With x, y coordinates and the value of the distances, the Geostatistical Analyst creates the interpolation map based on the kriging method to show the spatial pattern of inaccuracies (Figure 2). The dark area is the area less accurate relatively. At this point, the spatial pattern appears. The map shows the less accurate area in rural area.

For the pilot study, urbanity and stream are selected as potential reasons for the inaccuracies. These are the independent variables. Any other factors can be the reasons. The dependent variable is distance. By using the Statistical Package for the Social Sciences (SPSS), the correlation values are calculated between the independent variables and dependent variable. By using the selection by location function, the points in the urban area are selected and scored 1. The other points scored 0 (Figure 3). It means that the points in the dark area in figure 3 are scored 1. Moreover, buffering makes it possible for the points within 1 mile from the stream to score 1. The other points scored 0 (Figure 4). And then, SPSS calculated the value of the correlations between the distances and these scores (Table 1 and 2).

In table 1, the significant (0.048) correlation value (-0.126) means that the points in the urban area are closer to the referencing points than outside the urban area. As mentioned, most information in TIGER outside the urban centers was derived from the USGS 1:100,000-scale digital line graphs, which were vectorized from the digital scanning of the original artwork. For most urban centers, the information in TIGER was derived from the GBF/DIME files produced for the 1980 census. This means there is basically a difference between the urban centers area and the area outside the urban centers. In table 2, the significant (0.037) correlation value (0.105) means that the points near the stream are less accurate than of those far from the stream. The roads near the stream cannot maintain a straight line. Because the stream is changeable, the shape of the roads near the stream is also changeable. When interpreting the results, the significance value is a little bit high. However, because the purpose of this study is not calculating an accurate spatial pattern of inaccuracies, but checking the existence of spatial pattern, it can be neglected.

6.2 Inaccuracy of the hydrography shape file

From USGS's Geographic Names Information System database, the coordinates of 73 stream points in Erie County, New York are obtained. After creating point coverage in ArcGIS, it can be seen that these points are distributed randomly. Thus, this data is proper for performing mapping of the inaccuracy and looking at the spatial pattern of the inaccuracies. Basically, these points should be placed on the hydrography shape line, but they are not. The distances between the stream points and hydrography shape lines are calculated by the near function in ArcToolbox. These distances indicate the degree of inaccuracies for each stream points. The results show that the distances are not uniform and showing spatial pattern of inaccuracies (Figure 5).

To see the spatial pattern of inaccuracies well, mapping of the inaccuracies is done by using the Geostatistical Analyst in ArcGIS. Interpolation is performed by kriging methods same as in the case of the road shape file.

The result of mapping shows that there is a spatial pattern to the inaccuracies (Figure 6). The dark area is the area that has relatively high inaccuracy. In this study, the dark areas are located near the downstream. Because the width of the stream is wider than the upper stream, it seems more difficult for the hydrography shape line to be accurately placed on the real center line of the stream.

7. Discussion and future work

In conclusion, it is certain that there is a spatial pattern and specific reasons for the inaccuracies about the TIGER 2000 files. The inaccuracies of the road shape file of the TIGER 2000 files are related to the urbanity and the distance from the stream line. The inaccuracies of the hydrology shape file are related to the width of the stream. Moreover, it is revealed that the spatial pattern of the inaccuracy exists individually in each spatial data set. After the interpolation of the inaccuracies of the road and hydrology shape file is performed, it can be easily observed that there is a big difference between the results of the interpolation. Thus, the spatial pattern of the inaccuracies in the spatial data set should be examined separately. Each spatial data have their reasons for the inaccuracies.

To get more significant correlation values, more factors should be tested as independent variables. In this study, only the urbanity, the distances from the stream line and the widths of the stream are considered reasons for the inaccuracies. For more correct estimates, more factors such as the elevation, the width of the road, and so on should be considered. The presence of the correlation between the factors also should be checked carefully.

Areas that have different characteristics should also be tested. Erie County is a relatively flat area. Thus, it is necessary that the area where the change of elevation is severe should be tested. The area where there is no stream, such as a desert, can have the different spatial pattern of the inaccuracies unlike the area where there are many streams.

Digital spatial data cannot involve everything in the real world. Thus, the inaccuracy of spatial data is absolutely natural. The only thing to do is to recognize the characteristics of the inaccuracy as another attribute of spatial data. For 'fitness to use', the metadata should report not only the general accuracy of data, but also the spatial pattern of the inaccuracy. If the users know the spatial characteristics of the inaccuracies in the spatial data set, they can cope with the situation of the inaccuracies.

In this study, only two shape files of the TIGER 2000 files are tested. The other files of the TIGER 2000 files should be tested. Moreover, the subjects of the study about the spatial characteristics of the inaccuracies involve not only vector data, also field-like data in raster format such as DEM. In future studies, these data should be tested. The ultimate purpose of these kinds of studies is to maximize the quality of the data and to help the user to use an imperfect spatial data set properly. Thus, the way to systematically increase the quality of the data with the spatial characteristics of the inaccuracies should be developed. Furthermore, the standard to avoid improper use of the spatial data which results from ignorance about the spatial characteristics of the inaccuracies remains for the further research.

References

Chrisman, N. (1991). The error component in spatial data. In Maguire, D. J. et al. *Geographical Information Systems*. New York: John Wiley & Sons, Vol. 1:165-174.

Drummond, J. (1995). Positional accuracy. In Guptill S. C., Morrison J. L. *Elements of spatial data quality.* New York: Elsevier Science: 31-58

Duckham, M. and Drummond, J. (2000). Assessment of error in digital vector data using fractal geometry. *International Journal of Geographic Information Science,* 14, 67-84.

Federal Geographic Data Committee (FGDC). (2002). Geospatial Positioning Accuracy Standards.

Goodchild, M.F. and Hunter, G.J. (1997). A simple positional accuracy measure for linear feature. *International Journal of Geographic Information Science,* 11, 299-306.

Goodchild, M.F. and Gopal, S. (1991). *Accuracy of Spatial Data Base* (New York: Talyor & Francis).

Heywood, I., Cornelius, S., and Carver, S. (1998). *An Introduction to Geographic Information Systems*. New York: Addison Wesley Longman.

Johnston, K., Ver Hoef, J.M., Krivoruchko, K., and Lucas, N. (2001). *Using ArcGIS Geostatistical Analyst*. Redlands: ESRI Press.

Kainz, W. (1995). Logical consistency. In Guptill S. C., Morrison J. L. *Elements of spatial data quality.* New York: Elsevier Science: 109-138.

Liadis, J.S. (2000). GPS TIGER Accuracy Analysis Tools (GTAAT) Evaluation and Test Results. TIGER Operation Branch, Geography Division in the U.S. Census Bureau. [Online] Available: http://www.census.gov/geo/www/tiger/gtaat2000.pdf

O'Grady, Kristen. (2001). A DOQ test project: collecting data to improve TIGER. U.S.Census Bureau. [Online] Available: http://www.census.gov/geo/mod/esri_paper.pdf

O'Grady, Kristen., and Godwin, Leslie. (2000). The Positional Accuracy of MAF/TIGER. U.S. Census Bureau. [Online] Available: http://www.census.gov/geo/mod/positional_accuracy.pdf

U.S. Geological Survey. (1997). Spatial Data Transfer Standard.

Veregin, H. (1999). Data qulity parameters. In Longley, P. A. et al. *Geographical Information Systems*. New York: John Wiley & Sons: 177-189.

Zent, C.J. (1996). TIGER vs. DLG Road Feature Data. SUNY at Buffalo: *Masters thesis of Geography*.

Notes

Note 1. Dynamap/2000 user manual (http://www.geographynetwork.com/data/download/gdt/gdt_dynamap_gn.pdf)
Note 2. Open GIS Consortium, 1999. The *OpenGIS Abstract Specification*. Topic 9. Quality: 17.

Note 3. Erie County Overview (http://www.erie.gov/overview/)

Note 4. TIGER/Line Metadata (http://www.census.gov/geo/www/tlmetadata/metadata.html)
Note 5. Dynamap/2000 user manual
(http://www.geographynetwork.com/data/download/gdt/gdt_dynamap_gn.pdf)
Note 6. Http://nces.ed.gov/

Note 7. Http://geonames.usgs.gov/gnishome.html

Note 8. Source: Geoscience Australia (http://www.auslig.gov.au/geodesy/datums/distance.htm)

Table 1. Correlation value between the inaccuracy of the road and urbanity

Correlations

		Distance	Urban
Distance	Pearson Correlation	1	- 0.126
	Sig.(2-tailed)	N/A	0.048
	N	168	168
Urban	Pearson Correlation	- 0.126	1
	Sig.(2-tailed)	0.048	N/A
	N	168	168

Table 2. Correlation value between the inaccuracy of the road and stream

Correlations

		Distance	Stream
Distance	Pearson Correlation	1	0.105
	Sig.(2-tailed)	N/A	0.037
	N	168	168
Stream	Pearson Correlation	0.105	1
	Sig.(2-tailed)	0.037	N/A
	N	168	168

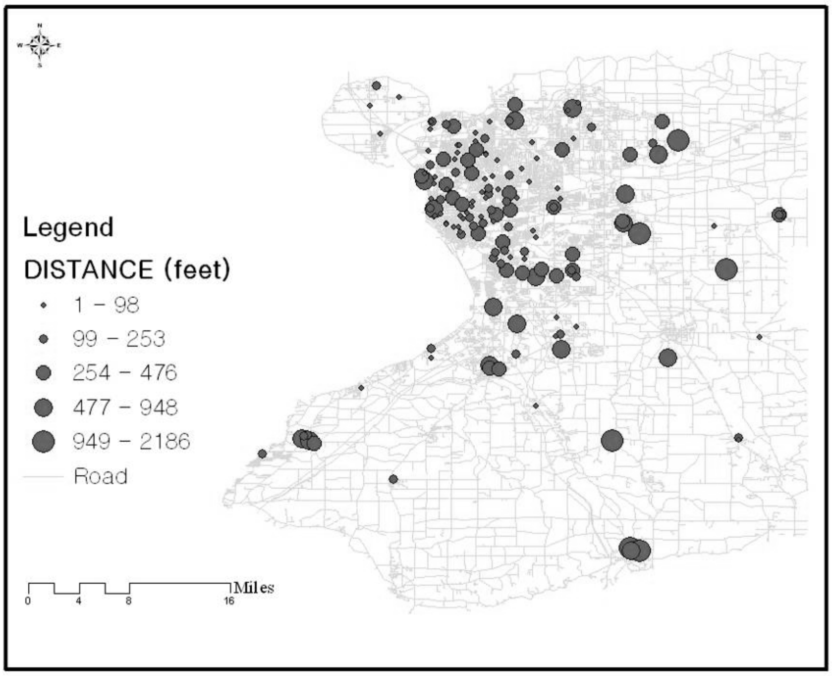

Figure 1. Inaccuracies of the road shape file of TIGER 2000 files based on the distances between geocoding results of the schools

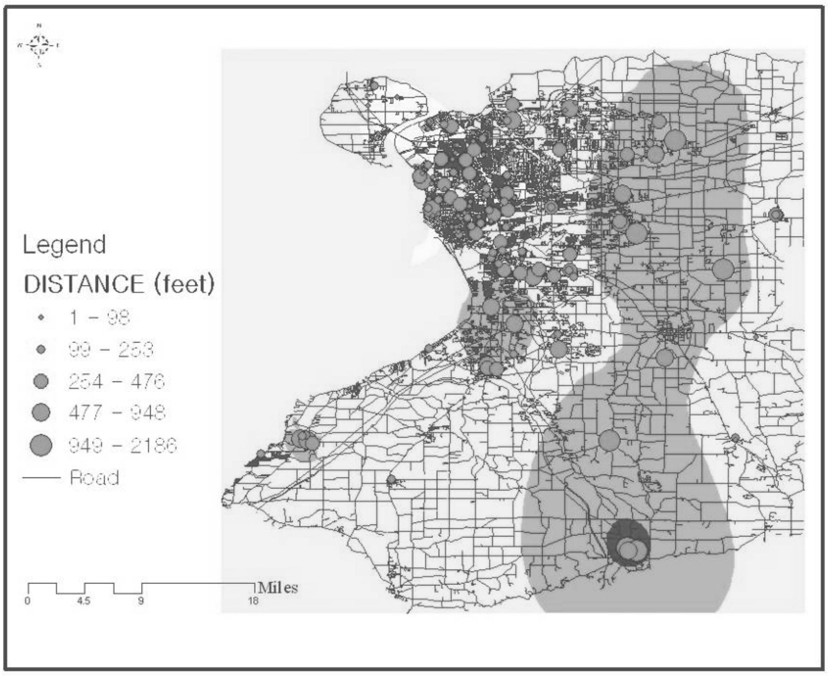

Figure 2. Mapping the inaccuracies of the road shape file by kriging methods

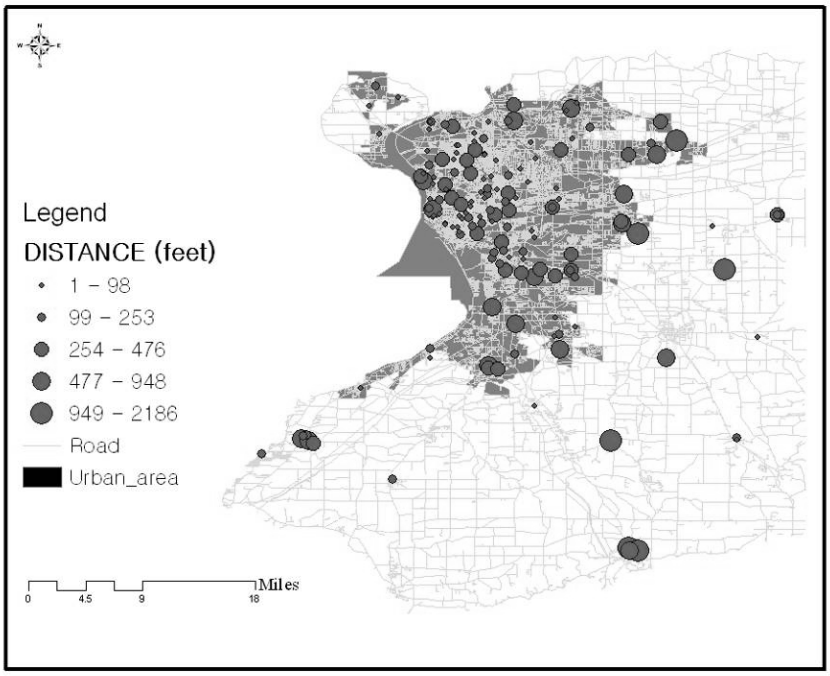

Figure 3. Factor 1 for the inaccuracies of the road shape file: Urbanity

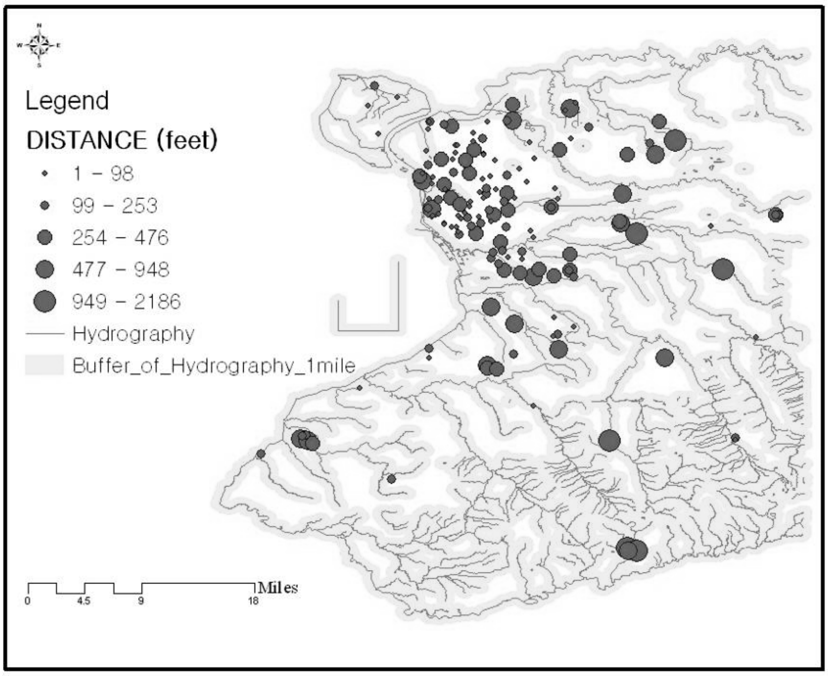

Figure 4. Factor 2 for the inaccuracies of the road shape file: Stream

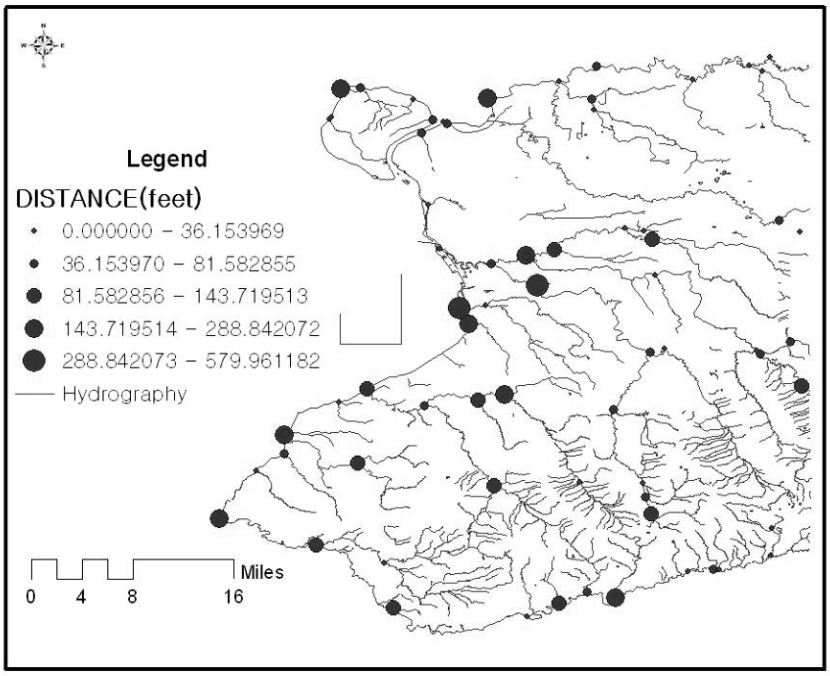

Figure 5. Inaccuracies of the hydrography shape file of TIGER 2000 files

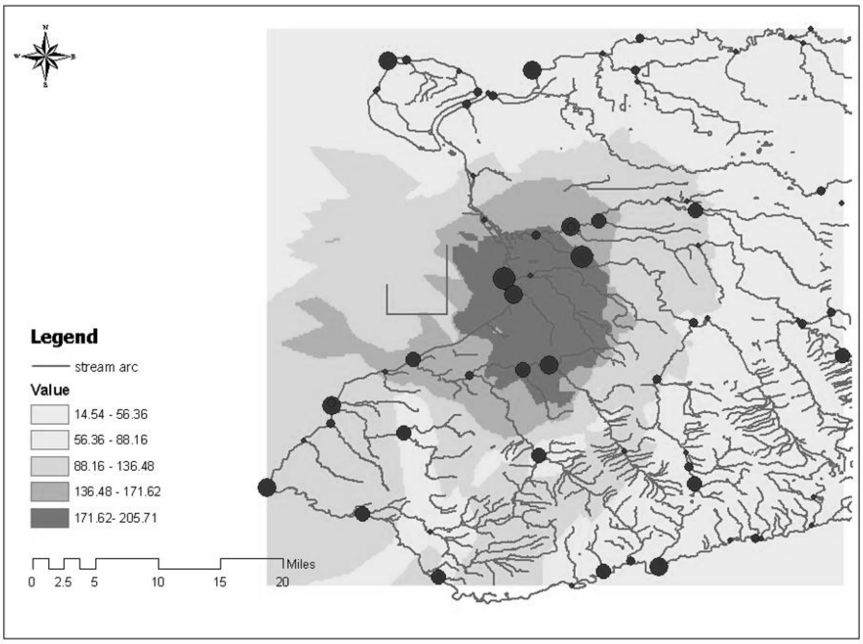

Figure 6. Mapping the inaccuracies of the hydrography shape file by kriging methods