International Journal of Health Geographics



Methodology



# Positional error in automated geocoding of residential addresses

Michael R Cayo* and Thomas O Talbot

Address: Geographic Research and Analysis Section, Bureau of Environmental and Occupational Epidemiology, New York State Department of Health, 547 River Street, Room 200, Troy, NY 12180-2216, USA

Email: Michael R Cayo* - mrc02@health.state.ny.us; Thomas O Talbot - tot01@health.state.ny.us

* Corresponding author

Published: 19 December 2003

*International Journal of Health Geographics* 2003, **2**:10

Received: 10 September 2003
Accepted: 19 December 2003

This article is available from: http://www.ij-healthgeographics.com/content/2/1/10



## Abstract

**Background:** Public health applications using geographic information system (GIS) technology are steadily increasing. Many of these rely on the ability to locate where people live with respect to areas of exposure from environmental contaminants. Automated geocoding is a method used to assign geographic coordinates to an individual based on their street address. This method often relies on street centerline files as a geographic reference. Such a process introduces positional error in the geocoded point. Our study evaluated the positional error caused during automated geocoding of residential addresses and how this error varies between population densities. We also evaluated an alternative method of geocoding using residential property parcel data.

**Results:** Positional error was determined for 3,000 residential addresses using the distance between each geocoded point and its true location as determined with aerial imagery. Error was found to increase as population density decreased. In rural areas of an upstate New York study area, 95 percent of the addresses geocoded to within 2,872 m of their true location. Suburban areas revealed less error where 95 percent of the addresses geocoded to within 421 m. Urban areas demonstrated the least error where 95 percent of the addresses geocoded to within 152 m of their true location. As an alternative to using street centerline files for geocoding, we used residential property parcel points to locate the addresses. In the rural areas, 95 percent of the parcel points were within 195 m of the true location. In suburban areas, this distance was 39 m while in urban areas 95 percent of the parcel points were within 21 m of the true location.

**Conclusion:** Researchers need to determine if the level of error caused by a chosen method of geocoding may affect the results of their project. As an alternative method, property data can be used for geocoding addresses if the error caused by traditional methods is found to be unacceptable.

## Background

There has been a dramatic increase in the number of public health applications using GIS. Software and hardware are now more accessible, affordable, and easier to use. Environmental, health and socio-demographic data are readily available through the Internet and optical disk media. Many colleges and universities now offer courses in GIS and spatial analysis. An increase in public awareness of these advances has led to increased demand for studies and maps investigating spatial relationships between health outcome, environmental risk factors and exposure.

Environmental health studies often rely on GIS and geocoding software to help delineate areas of potential exposure and to locate where people live in relation to these areas. A number of studies have used residential locations to determine whether individuals live within defined zones of exposure. Geschwind et al. [1] geocoded congenital malformation cases and controls to estimate an increased risk of living within 1 mile of hazardous waste sites. English et al. [2] used geocoded residential addresses to assess whether there was an elevated odds ratio of childhood asthma hospitalizations in children living within 550 feet (168 m) of roads with heavy traffic in San Diego, California. In a study of breast cancer on Long Island, New York, one-kilometer grid cells were created and cases, controls and chemical facilities were assigned to individual cells through automated geocoding methods. The risk of developing postmenopausal breast cancer was found to increase as the number of chemical facilities sharing the same cells as study subjects increased [3]. More recently, Reynolds et al. [4] geocoded childhood cancer cases to census tracts in California and used USEPA data to assign hazardous air pollutant scores to each tract. There was an increased risk of developing childhood cancer as the exposure level increased. Kitto et al. [5] used GIS to locate nearly 45,000 residential radon screening measurements, which were then associated with surficial geology. The association between surficial geology types and radon measures were used to predict radon levels in towns across New York.

Geocoded health data are also used to map rates of disease in order to determine areas of high or low incidence [6,7]. Rate maps can be used in conjunction with spatial statistics such as the local Moran's I [8] or the Spatial Scan Statistic [9] to locate the general areas where the rates are unlikely due to chance. Further investigations or more rigorous epidemiology studies are often needed to clarify the association of risk factors and adverse health outcomes when high rates are detected.

Many GIS software packages provide for street level geocoding. Geocoding software matches residential addresses to street reference files containing geographic centerline coordinates, street numbers, street names and postal codes. Researchers undertaking projects having a geocoding component should be concerned that positional error can be introduced by commonly used algorithms. Of more concern, they need to understand if this error could impact study results. Nondifferential errors with respect to exposure classification will bias the association between a risk factor and the health outcome towards the null, limiting the ability to detect true effects. The capability to detect an association thus depends on the magnitude of this error. However, if the positional error is systematic, it is possible an association may be found between a health outcome and an exposure where none actually exists. In the case of disease surveillance activities, localized high or low rates of disease may appear as an artifact of geocoding errors [10].

The percentage of addresses that geocode is commonly referred to as a match rate. The inability to geocode addresses can lead to a loss of study population causing sample bias and reduced statistical power in detecting important associations. Several investigators have provided statistics related to match rates [11-17]. Researchers have found that differences in these rates are dependent on population density [18,19]. This is because street reference files, such as the U.S. Census TIGER (Topologically Intergrated Geographic Encoding and Referencing) files or commercially enhanced TIGER files, often contain more complete address information in more densely populated areas. Gregorio et al. [20] analyzed the match rates of breast cancer cases from the Connecticut Tumor Registry and found that women of color and women living in low income neighborhoods were more likely to be successfully geocoded compared to white women and women living in higher income areas. Investigators should be aware that the geographic differences in match rates can alter study results. For example, if more cases are matched in inner city minority neighborhoods, the disease incidence may appear higher in these areas due to larger subject loss in other areas.

Achieving high match rates is dependent on accurate and complete address information of the study subjects and the street reference files. Many types of problems can occur in both, such as: spelling errors; street suffix, prefix and abbreviation inconsistencies; and erroneous ZIP code information. Reference files also contain errors such as missing, incomplete, and incorrect street segments and address ranges. The North American Association of Central Cancer Registries provide an extensive overview and guideline of the standardization and geocoding of patient addresses, problems encountered, and recommendations for improving the geocoding process of disease registries [21]. Match rates alone are not sufficient to evaluate a geocoding result. Some investigators have also provided statistics related to the percent of geocoded addresses misclassified to the correct town [13], census area [22,16], and land parcel [22]. The level of misclassification will change depending on the geographic scale of the regions used.

Very limited published information exists on positional error in automated street level geocoding. Hertz, of the California Department of Health Services, conducted a pilot study to assess geocoding accuracy of 70 addresses (A. Hertz, personal correspondence, 2002). In his study, Hertz used aerial photos to determine the true location of

each address and geocoded the same group using three different commercial products. Hertz found positional error to be in a range of 20–80 m depending on the product used and had some extreme outliers over 250 m. Researchers at the University of Connecticut compared the locations of addresses geocoded using the U.S. Census Bureau's TIGER [23] files to ground truth locations of approximately 536 addresses in Stratford, Connecticut. Four of these addresses were located more than 500 feet (152 m) from the correct location (E. Cromley, unpublished manuscript, 1997). In a recently published study, Bonner et al. [24] found differences between urban and non-urban addresses when examining distances between the geocoded and GPS determined locations. They found 89 percent of the addresses were within 100 meters in urban areas of Erie and Niagara Counties, NY, while in the non-urban areas 69 percent were within 100 meters.

This study had several objectives. The primary objective was to evaluate positional error in automated geocoding of residential addresses. We measured positional error by calculating the distance between geocoded locations provided by a commonly available off-the-shelf product and their corresponding true locations. This commercial product uses a proprietary enhanced version of the TIGER files. The second objective was to evaluate how this error varies between urban, suburban, and rural population densities. A third objective was to determine if the error can be reduced by adjusting default settings in the geocoding software. The street offset setting allows the user to change the default for how far a geocoded address is placed from a street centerline while the corner inset setting determines how far a geocoded address is placed along a street from an intersection. The final objective was to compare the error observed in the traditional geocoding method, which relies on linear interpolation, to a point geocoding method using property parcel data.

## Methods

### *Data*

We acquired residential addresses from the New York State Office of Real Property Services (NYSORPS). These represent the types of addresses we frequently geocode in health studies. The data included 1999 property parcel records for the New York State Capital District counties of Albany, Rensselaer, Saratoga, and Schenectady [25]. The City of Watervliet and the Town of Westerlo were excluded as they were not available at the time we processed the data. Local governments compile these data to assess town and school property taxes. Each record contains parcel level information, such as: street number and name, property use, billing information, and coordinates for the approximate centroid of the parcel. Towns vary in their method of determining the location of the parcel centroids. Some are derived visually from paper maps, which have a horizontal positional accuracy of ± 10 feet (3 m) [25]. Figure 1 provides an example of property parcel points overlaid on high resolution aerial orthoimagery. We further restricted analysis to residential properties classified as single, two or three-family houses. This selection included 215,007 addresses.

Using the property centroid, each address was assigned a population density class of urban, suburban, or rural. The most densely populated cities in the Capital District were classified as urban and included Albany, Schenectady, Troy, Rensselaer, Cohoes, Mechanicville and Green Island. The population density of these cities ranged from 1,059 persons/km$^2$ to 2,490 persons/km$^2$ [26]. The remaining areas were partitioned into suburban and rural areas by census tract. Suburban areas consisted of census tracts having greater than 250 persons/km$^2$, while rural areas contained census tracts with less than 250 persons/km$^2$. The four county study area and associated population density assignments are shown in Figure 2.

### *Geocoding*

In order to successfully geocode a residential address, a valid street number, street name and ZIP code is required. NYSORPS property data contain parcel specific street number and name information, but lack the ZIP code of the parcel address. We could not reliably assign parcel ZIP codes to 3,145 addresses (1.5%) in our residential property file. This group was excluded from further analysis and should have no effect on the overall results since they represent a very small portion of the addresses in the four county area.

MapMarker Plus Version 6.0 [27] was used to match the residential address records to the software's street reference files. The reference files used in this product were dated July-August 2000 and enhanced by Geographic Data Technology of Lebanon, New Hampshire. We did not include any record for further analyses unless it could be matched to the street reference files by exact house number, street name, and ZIP code. Using these criteria, MapMarker successfully geocoded 170,819 (81%) addresses. Match rates by population density class are summarized in Table 1.

In order to measure positional error, we determined the true location for a random sample of 1,000 addresses from each of the three population density classes for a total sample size of 3,000 addresses. This selection was drawn from the group that matched exactly on street number, street name, and ZIP code. We define the "true" location of each address as the point that visually represented the approximate center of the house using 1 m resolution digitally enhanced aerial orthoimagery. The orthoimagery was flown from 1994–98 and has a


**Figure 1**
**Residential parcel points and high resolution aerial imagery.** Property parcel points from the NYSORPS. Each circle represents the approximate centroid of a property parcel classified as residential.

horizontal accuracy of 10 m [28]. Through this method, 2,674 (89%) addresses of the study group were assigned true locations.

Closely spaced homes were the most common problem in identifying the true location in urban areas. One meter resolution orthoimagery made it difficult to delineate some of the building rooftops. A more common problem in suburban addresses was dark rooftops surrounded by dense canopy cover from trees. In rural areas, detached garages, barns, and other large outbuildings made it difficult to distinguish the actual house.

Overall, we found only small differences in our ability to assign a true location between the three density classes. Four individuals were involved in creating true locations for the study sample addresses. A QA/QC assessment was performed on a random sample of 100 addresses to compare their decisions of where to place the true location. Results showed that discrepancies between all individuals were minimal, averaging only 3.3 m.

Fieldwork was completed in the summer of 2001 for the remaining 326 addresses which could not be confidently identified using in-house techniques. Staff used real time Global Positioning System technology and mapping software as a navigational aid to locate the address and

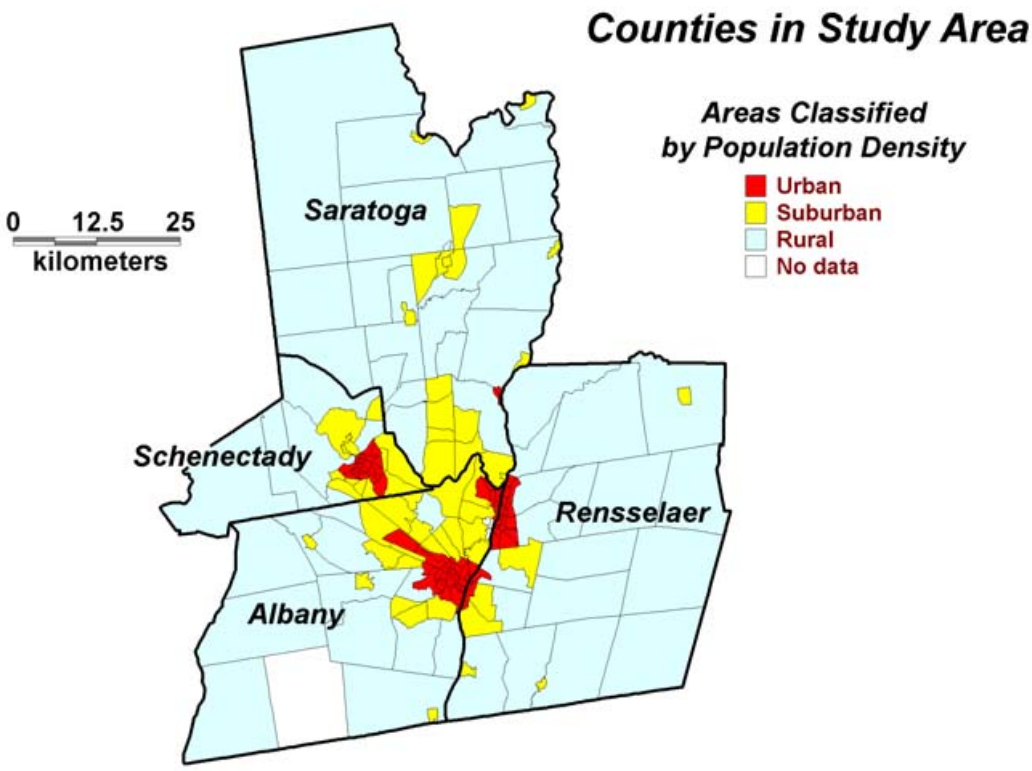


**Figure 2**
**Study area by population density class.** Densely populated cities were classified as urban. Census tracts were used to partition the remaining areas into suburban (>250 persons/km2) and rural areas (<250 persons/km2).

**Table 1: Geocoding match rates by population density. Values are based on exact matching of house number, street name, and ZIP code.**

| Population Density | Number of Residential Addresses | Number Exact Matched | Percent Exact Matched |
|---|---|---|---|
| Urban | 53,602 | 50,291 | 93.8% |
| Suburban | 90,759 | 78,706 | 86.7% |
| Rural | 67,501 | 41,822 | 61.9% |
| Total | 211,862 | 170,819 | 80.6% |

identify the correct structure for that address. As with the in-house procedure, the point was then manually placed in the center of the correct structure using aerial imagery.

### *Analyses*

Once true locations of all 3,000 addresses in our sample were determined, we calculated the straight-line distance between coordinates of the true locations and the automated geocoded points. This allowed us to compute the positional error, by population density, from traditional automated geocoding.

Without knowing the optimal settings, geocoding was initially performed using a street offset and corner inset of

zero. We adjusted the default offset and inset settings in MapMarker to see if the positional error in the geocoded addresses could be reduced. The sample address file was re-geocoded using 5 m iterations of these values and compared to true locations to determine the optimal combination.

As an alternative to traditional methods relying on street centerline files, we calculated the distance between the true locations and the property parcel centroids assigned by local governments. This allowed us to compare the positional error between automated geocoding and using parcel data to locate addresses. An example of the three locations for each address is shown in Figure 3.

We also investigated whether directional bias in the error could be introduced by data conversion issues, such as inconsistent projections or datums in the various GIS layers. A rose diagram was constructed using the directional error of the 3,000 addresses. We also calculated the angle of the errors to determine if the direction of the errors were uniform for both the automated geocoded points and the property parcel points using the modified Rayleigh test [29].

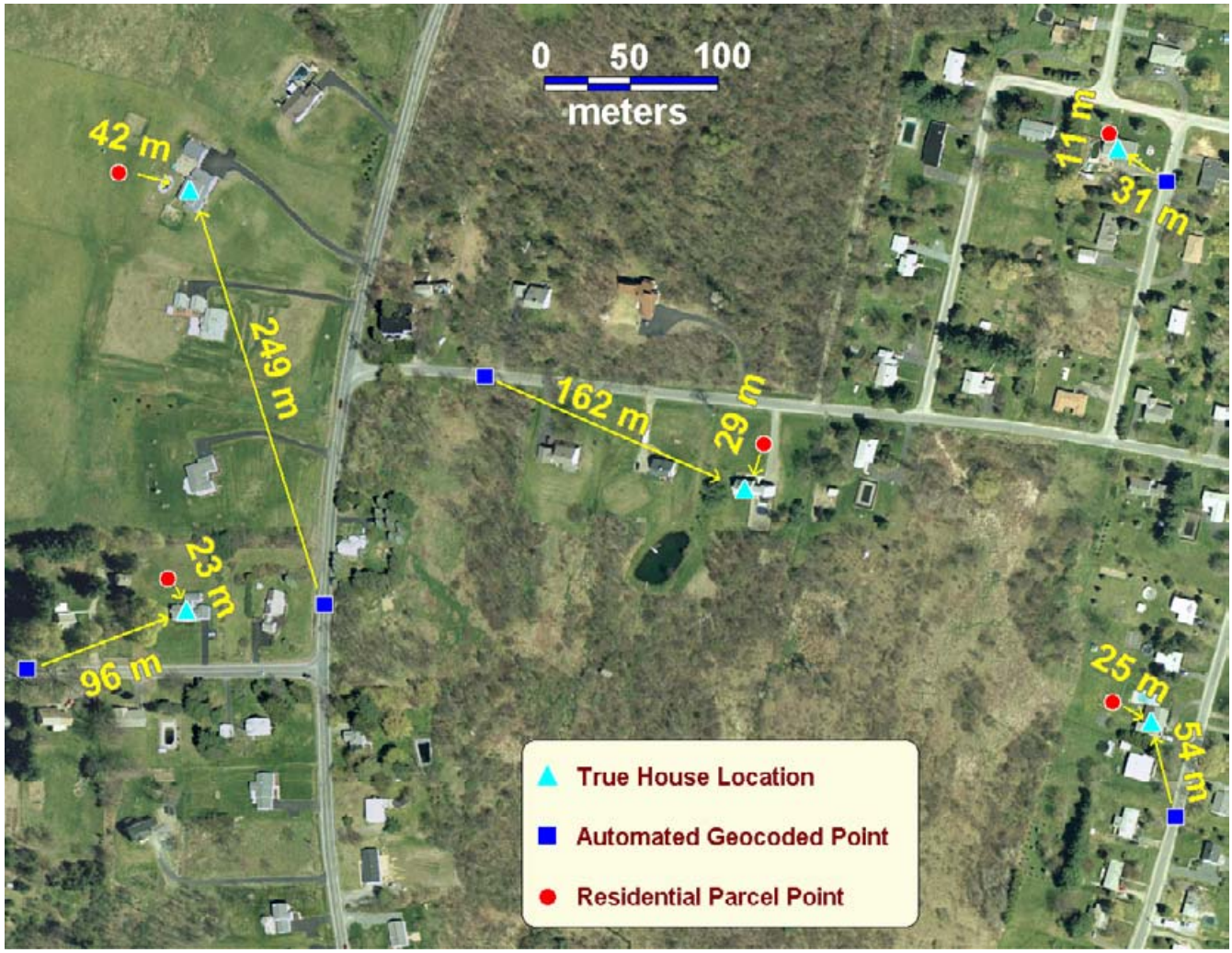


**Figure 3**
**Measuring positional error using two methods.** Distances are measured from the true locations (triangles) to the automated geocoded points (squares) and to the residential parcel points (circles) to determine the positional error.

**Table 2: TIGER based positional error. Positional error is calculated by measuring the distance between address locations determined by automated geocoding methods using enhanced TIGER files and the true location of the houses. RMSE = Root Mean Square Error (radial). N = 1000 for each density class.**

| Percentiles | Urban (m) | Suburban (m) | Rural (m) |
|---|---|---|---|
| 50% | 38 | 78 | 201 |
| 75% | 62 | 158 | 498 |
| 90% | 96 | 306 | 1,544 |
| 95% | 152 | 421 | 2,872 |
| 99% | 379 | 1,219 | 5,706 |
| Max | 1,088 | 2,584 | 18,742 |
| Mean | 58 | 143 | 614 |
| RMSE | 102 | 259 | 1,578 |

## Results

We found substantial differences in positional error between the automated geocoded points and the corresponding true locations. In rural areas, 95 percent of the addresses geocoded to within 2,872 m of their true location, while in suburban areas the same percent of addresses geocoded to within 421 m. Urban areas showed the least error, where 95 percent of the addresses geocoded to within 152 m of their true location. The mean error for rural areas was 614 m, 143 m for suburban areas and 58 m in the urban area. Table 2 summarizes the error in automated geocoding using TIGER based files by percentile and density class. We also provide the root mean square error (RMSE). The RMSE provides a measure of the variation of this error following the National Standard for Reporting Spatial Data Accuracy [30]. The cumulative density distribution plot in Figure 4 can be used to estimate the percent error at any distance. Addresses having errors in excess of 5 km were examined more closely. It appeared that many of these large errors are due to inaccurate ZIP code boundary information in the software's reference files. This resulted in addresses being placed in an adjacent ZIP code several kilometers from their true location.

We found the optimal combination of the street offset and corner inset for the entire sample to be 15 m and 50 m respectively. This combination of values, however, only reduced the overall mean positional error from 272 to 265 m. Optimal values were actually determined for the rural, suburban and urban areas separately, but provided little additional benefit from using an average setting for all density areas. Using unique values for each area provided an additional reduction in the mean error of 2.1 m in rural areas, 0.1 m in the suburban areas, and 0.7 m in urban areas.

The use of property parcel coordinates significantly reduced the positional error. In rural areas, 95 percent of the parcel points were within 195 m of the true location. In suburban areas, this distance was 39 m and in urban areas was 21 m. The mean positional error of the parcel points for rural areas was reduced to 55 m compared to 614 m found in automated geocoding. The mean parcel error for suburban areas declined to 15 m from 143 m, while in urban areas the mean parcel error was reduced to 10 m from 58 m. Table 3 summarizes the parcel based positional error by percentile and density class. The scatter diagram in Figure 5 illustrates the dramatic differences between positional error using parcel points and traditional methods relying on enhanced TIGER files. As expected, the greatest improvements were seen in the rural areas.

A visual inspection of the rose diagram showed that the directions of the error were well dispersed. The Rayleigh test confirmed that the angles of the errors were uniform for both the automated geocoded points and the parcel points.

## Discussion

This project used address data typical of that which are geocoded for health studies. We calculated error only for the addresses which had an exact match on house number, street name and ZIP code to the reference files. If we considered the addresses that matched on less stringent criteria, both match rates and positional error would have increased. Yu showed that small improvements in achieving higher match rates by relaxing the matching criteria results in large decreases in positional accuracy [17]. Researchers often sacrifice positional error in order to reduce subject loss from lower match rates when resources are limited for accurately geocoding study subjects.

Several factors explain the positional error in the geocoded locations. The original TIGER files have a horizontal positional accuracy of ± 167 feet (51 m) [31]. The geoco-

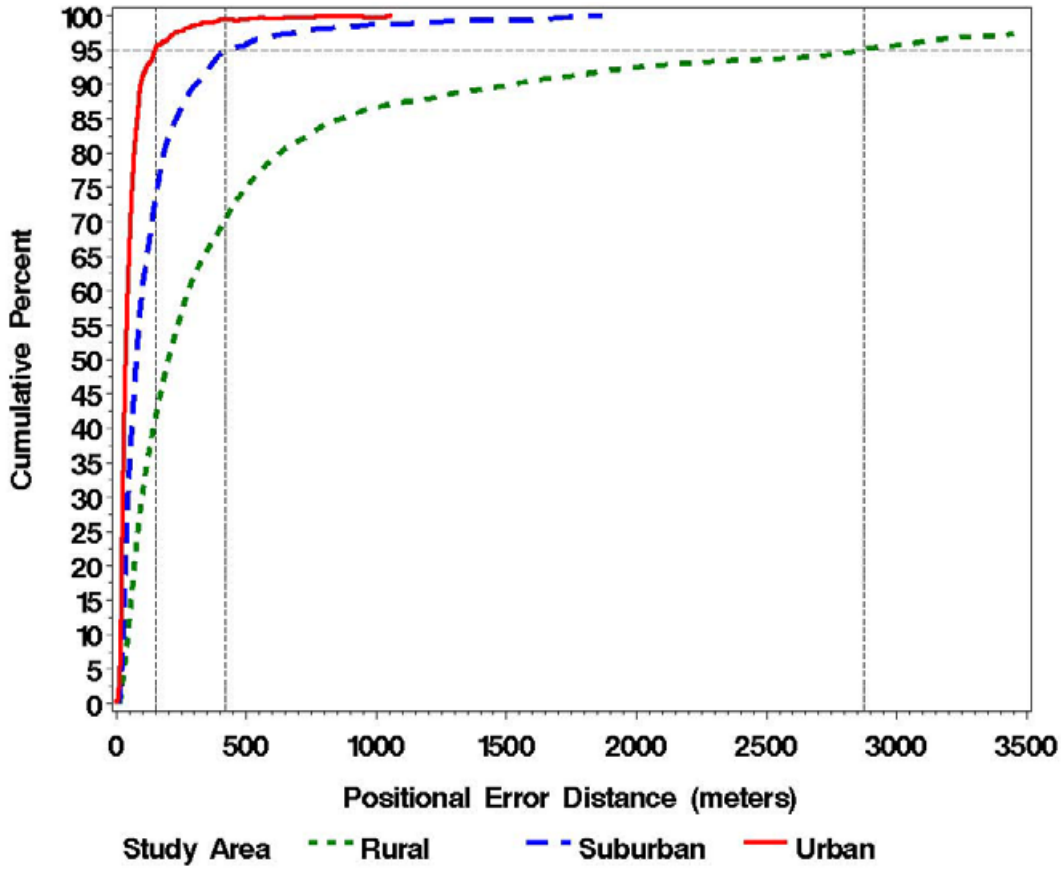


**Figure 4**
**TIGER based positional error cumulative density distribution.** This plot can be used to estimate the percent error at any distance for the three density classifications. The vertical dashed lines show the error distance at the 95$^{th}$ percentile.

**Table 3: Parcel based positional error. Positional error is calculated by measuring the distance between property parcel locations and the true location of the houses. RMSE = Root Mean Square Error (radial). N = 1000 for each density class.**

| Percentiles | Urban (m) | Suburban (m) | Rural (m) |
|---|---|---|---|
| 50% | 8 | 8 | 15 |
| 75% | 11 | 14 | 43 |
| 90% | 15 | 22 | 113 |
| 95% | 21 | 39 | 195 |
| 99% | 62 | 177 | 582 |
| Max | 299 | 921 | 3,567 |
| Mean | 10 | 15 | 55 |
| RMSE | 18 | 46 | 211 |

ding engine used in this project incorporates enhanced versions of these files. We are unaware of the improvement in geometric accuracy of these street centerlines over original TIGER files. Although it is difficult to measure, we feel that positional accuracy of the enhanced files represents a significant source of positional error in the geocoded addresses. Further research is needed to quantify this contribution to the error.

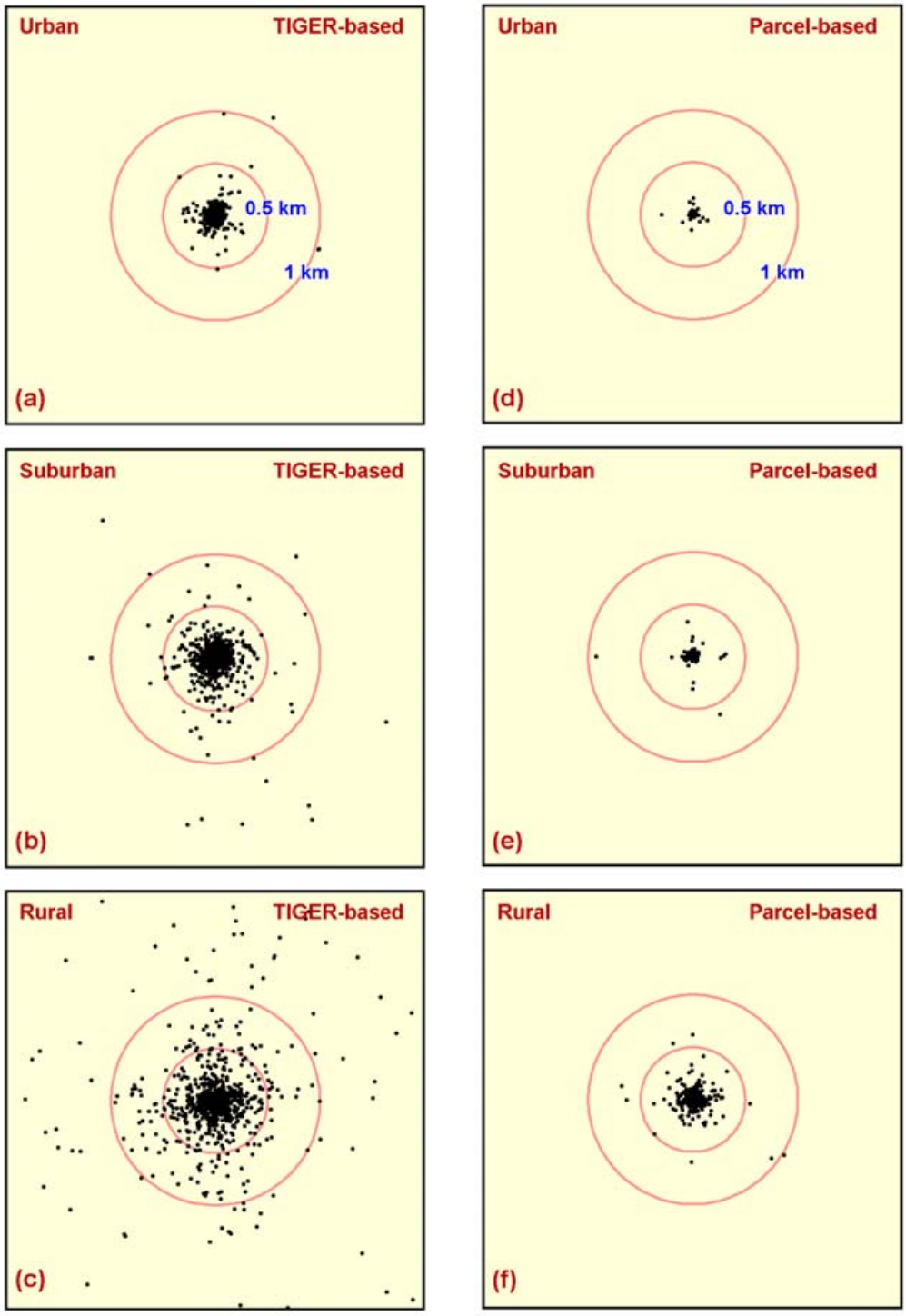


**Figure 5**
**Scatter diagram of positional error by geocoding method and population density.** Each point represents one residential address and there are 1000 addresses for each plot. The true location of each house is set to the center of the circles. The positional errors by density class of TIGER based geocoded addresses as measured from true location are shown in a,b,c; The positional error by density class of parcel-based points as measured from true location are shown in d,e,f.

A more dominant source of error originates in the interpolation algorithms used to determine an address along a street centerline. Address ranges can be incorrect or reversed in the reference files, which causes houses to be geocoded to either the wrong side or wrong end of the street. Larger positional error was observed in rural areas. Generally, rural areas consist of longer streets with fewer intersections. The software must interpolate where to place an address based on the street numbers assigned to the ends of each street segment. As the street segments increase in length, the interpolation error will also increase. In a study of vehicle accident locations, Levine et al. reached a similar conclusion that geocoding error is a function of street segment length and urban areas typically contained shorter segments [11]. Since the software often assumes uniform intervals between street numbers along a street segment, interpolation errors increase when homes are not evenly spaced along a street. Parcels tend to be larger and less consistent in size in less densely populated areas. The median parcel size in our random sample was found to be 472 $m^2$ in the urban areas, 1214 $m^2$ in the suburban areas and 3035 $m^2$ in the rural areas. The variation in parcel size showed a similar trend. In those properties classified urban, the standard deviation was 445 $m^2$, for suburban was 5024 $m^2$, while in rural areas increased to 56,046 $m^2$. Finally, there is a greater variation in the distance houses are located from the street centerlines in rural areas. In the urban settings, a common problem was row type housing or condominiums. Reference files space the addresses uniformly along the street when in fact the addresses are clustered together.

A spatial non-stationary process is evident if statistical parameters such as the mean and variance change with location. Non-stationarity of the positional error may have some important implications in environmental health studies. For example, in urban areas where the error is small we may notice an association between an environmental risk factor and a particular health outcome. In other areas having greater error, associations may be more difficult to detect. In addition, some types of environmental exposures, such as exposures to air pollution from traffic or exposures to agricultural pesticides, are associated with population density. If the level of error varies as the level of exposure changes, the study parameters which estimate the relationship between a risk factor and a health outcome may also be impacted. Global statistical methods perform poorly at uncovering important associations in which the statistical parameters vary locally due to non-stationarity. [32]. This has lead to an increase in the use of local spatial statistical methods for detecting clustering and localized associations between health outcomes and risk factors.

Though no systematic directional bias in our random sample of addresses was found, we did not determine if systematic error may be present in localized areas. Addresses on a particular street, in close proximity to each other, or within the same ZIP code may all have error of similar direction and distance. For example, the geocoding software may place all the addresses in a local area at some distance from the true street location if a street is misnamed, has incorrect address ranges, or if a ZIP code is incorrect in the street reference file. Burra et al. [10] demonstrated that very localized geocoding errors in which less than one percent of mortality cases are placed in the wrong census area can lead to differences in up to 75 percent of comparative mortality figures in Hamilton, Ontario, census tracts. Once they had more accurately geocoded the cases they found that approximately 80 percent of the difference in number of cases occurred in only 4 of the 88 census tracts studied. In addition, the size and shape of the clusters that were detected using the local Moran I statistic changed when the geocoding errors were corrected. This was a result of errors being concentrated in localized areas. Further work is needed to measure the strength of the spatial autocorrelation of the geocoding errors by distance and direction.

We attempted to reduce the positional error by optimizing the offset and inset default values in the software. Changing these values contributed very little to reducing overall error. Previous work by Ratcliffe [22] also did not yield significant reduction in the positional error by altering the offset and inset distances.

The use of property parcel points provides one solution for reducing error when the level of error in traditional geocoding methods is not acceptable. The parcel data clearly contains more accurate locations for the individual houses compared to TIGER based files. Parcel centroids are rarely at the exact location of the house. In urban areas the centroid will more closely represent the location of the actual house because of smaller parcels and more uniform spacing of homes. In rural areas this becomes less likely. The use of parcel data may also help to improve match rates since the parcel data is updated on a yearly basis for tax purposes, while commonly used street centerline files are often updated less frequently.

Though the use of parcel points provides greater positional accuracy, the parcel addresses are often not standardized. Residential and commercial addresses are collected by thousands of local governments across the country. This can lead to a lack of standardization in the way addresses are stored in the data files. The challenge is to standardize the millions of New York addresses and add a ZIP code to each property parcel record. Commercially available software programs are available which can

be used to help standardize the parcel addresses. Once standardized, linkage to health outcome data could be achieved more efficiently and with the same effort as using currently available TIGER based files.

We considered using data from local county emergency E911 systems to improve geocoding accuracy. However, we found that each county in New York State developed their own E911 system for providing route directions to emergency responders. These systems range from simple text based to more elaborate systems using GIS. The files used in E911 systems come from a variety of sources. Some counties rely exclusively on TIGER based files, some use real property assessment data, while others use files from telephone or electric utility companies. The county E911 data is often considered either confidential or proprietary depending on the source. For example, E911 systems often contain the addresses of unlisted telephone numbers. The advantages of using New York State real property data are that the format is more consistent across the state and is available through freedom of information requests. In addition, most of the counties and municipalities report the data directly to NYSORPS. This minimizes the number of requests needed in developing a statewide reference file.

There are some limitations in our study. Since the TIGER files are often derived from data provided by state and local governments, the geometric accuracy and address range completeness may differ in other areas. For this reason, it is difficult to predict if the magnitude of the geocoding error resulting from positional inaccuracy and interpolation error would be similar in other areas of the country. However, we would expect that interpolation issues contributing to positional error will remain the dominant source and correlate highly with population density in most areas. This is due to such issues as longer street segments and houses being spaced further apart in less densely populated areas. In addition to population density, there may be other predictors of positional error such as population growth or sociodemographic variables. Further research is needed in this regard. This study assumes our true locations to be error free. We recognize there is some positional error in the true locations assigned. However, this error is quite small compared to the error caused by the automated geocoding process and should not have a major impact on our results.

We only provide results from one geocoding package. We are uncertain of how the results would change if other products were used on the same set of address data. Most products we are aware of rely on the use of TIGER or enhanced TIGER files. As the geometric accuracy and completeness of the street centerline files improve, we would expect positional error to decrease. However, because houses are often not spaced evenly along streets, there will continue to be greater error using linear interpolation techniques compared to using parcel points to locate addresses.

## Conclusions

It is important that researchers determine if the level of error caused by a chosen method of geocoding may affect the results of their project. In the past, researchers appeared to pay little attention to understanding positional error from geocoding. Foote and Huebner report that only recently has more attention been devoted to problems introduced by error, inaccuracy, and imprecision in spatial data and how this can "make or break" a GIS project [33]. The location derived from the geocoding process is often used as input to other operations such as assignment of exposure or socioeconomic class. These assignments are often based on models which also have inherent error. When multiple operations are strung together, errors are often compounded making it difficult to evaluate the accuracy of the final result [34,35]. Burra et al. [10] suggest that small geocoding errors, when combined with other types of error in the data, may be amplified into large errors in the final results. Though researchers may be aware that error propagates through the various analyses, they are unable to estimate the accuracy of the final results without first recording the errors of intermediate operations such as geocoding.

Krieger et al. [16] recommends "that all public health projects involving geocoding evaluate and report on methods to verify the accuracy of their geocoding methodology". If the error caused by traditional methods is not acceptable, one consideration is the use of property data to geocode health data.

We are currently conducting further analyses to determine the implications positional error has on the misclassification of individuals with respect to exposure. Copeland et al. [36] provides examples of how to measure the underlying true value of a study's odds ratio or relative risk if the sensitivity and specificity of a classification procedure can be measured. We also need to examine whether the errors are random and bias study results towards the null, or whether there are systematic errors which could lead to erroneous positive results.

## Authors' contributions

This project was a joint collaboration between authors MRC and TOT. Both contributed to all phases of design concepts, data acquisition, processing, fieldwork, and analysis. Both authors prepared and approved the final manuscript and figures.

## Acknowledgements

We thank Chris Pantea and Valerie Haley for their assistance with statistical analysis, Pat Steen and Frank Schoonbeck for their assistance in fieldwork, and Jim Bowers and Deepa Varadarajulu for their assistance in determining photo corrected true locations for residential addresses. We thank Syni-An Hwang, Steve Forand, Francis Boscoe, and Gwen Babcock for providing editorial comments on this manuscript.

## References

1. Geschwind SA, Stolwijk JAJ, Bracken M, Fitzgerald E, Stark A, Olsen C, Melius J: **Risk of congenital malformations associated with proximity to hazardous waste sites.** *Am J Epidemiol* 1992, **135:**1197-1207.
2. English P, Neutra R, Scalf R, Sullivan M, Waller L, Zhu L: **Examining associations between childhood asthma and traffic flow using a geographic information system.** *Environ Health Perspect* 1999, **107:**761-767.
3. Lewis-Michl EL, Melius JM, Kallenbach LR, Ju CL, Talbot TO, Orr MF, Lauridsen PE: **Breast cancer risk and residence near industry or traffic in Nassau and Suffolk Counties, Long Island, New York.** *Arch Environ Health* 1996, **51:**255-265.
4. Reynolds P, Von Behren JV, Gunier RB, Goldberg DE, Hertz A, Smith DF: **Childhood cancer incidence rates and hazardous air pollutants in California: An exploratory analysis.** *Environ Health Perspect* 2003, **111:**663-668.
5. Kitto ME, Kunz CO, Green JG: **Development and distribution of radon risk maps in New York State.** *J Radioanal Nucl Chem* 2001, **249:**153-157.
6. Rushton G, Lolonis P: **Exploratory spatial analysis of birth defect rates in an urban population.** *Stat Med* 1996, **15:**717-726.
7. Talbot TO, Kulldorff M, Forand SP, Haley VB: **Evaluation of spatial filters to create smoothed maps of health data.** *Stat Med* 2000, **19:**2399-2408.
8. Anselin L: **Local indicators of spatial association - LISA.** *Geogr Anal* 1995, **27:**93-115.
9. Kulldorff M, Nagarwalla N: **Spatial disease clusters: detection and inference.** *Stat Med* 1995, **14:**799-810.
10. Burra T, Jerrett M, Burnett RT, Anderson M: **Conceptual and practical issues in the detection of local disease clusters: a study of mortality in Hamilton, Ontario.** *Can Geogr* 2002, **46:**160-171.
11. Levine N, Kim KE: **The location of motor vehicle crashes in Honolulu: a methodology for geocoding intersections.** *Comput Environ Urban* 1998, **22:**557-576.
12. Dearwent SM, Jacobs RR, Halbert JB: **Locational uncertainty in georeferencing public health datasets.** *J Expo Anal Environ Epidemiol* 2001, **11:**329-334.
13. Fulcomer MC, Bastardi MM, Raza H, Duffy M, Dufficy E, Sass MM: **Assessing the accuracy of geocoding using address data from birth certificates: New Jersey, 1989 to 1996.** *http://www.atsdr.cdc.gov/gis/conference98/proceedings/proceedings.html (accessed 2003). In proceedings of the Third National Geographic Information Systems in Public Health Conference, San Diego, August 18-20 1998, 547-560.*.
14. Krieger N, Chen JT, Waterman PD, Soobader MJ, Subramanian SV, Carson R: **Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: Does the choice of area-based measure and geographic level matter? The Public Health Disparities Geocoding Project.** *Am J Epidemiol* 2002, **156:**471-482.
15. Krieger N, Waterman P, Chen JT, Soobader MJ, Subramanian SV, Carson R: **ZIP code caveat: Bias due to spatiotemporal mismatches between ZIP codes and US census-defined geographic areas. The Public Health Disparities Geocoding Project.** *Am J Public Health* 2002, **92:**1100-1102.
16. Krieger N, Waterman P, Lemieux K, Zierler S, Hogan JW: **On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research.** *Am J Public Health* 2001, **91:**1114-1116.
17. Yu Lixin: **Development and evaluation of a famework for assessing the efficiency and accuracy of street address geocoding strategies.** *PhD thesis. University at Albany, State University of New York - Rockefeller College of Public Affairs and Policy*; 1996:1-164.
18. Nie J, Bonner MR, Vito D, Willett NH, Freudenheim JL: **Validation of TIGER (Topologically Integrated Geographic Encoding and Referencing System) to geocode addresses for epidemiologic research.** *Am J Epidemiol* 2001, **Suppl 179:**647.
19. Howe HL: **Geocoding NY State Cancer Registry.** *Am J Public Health* 1986, **76:**1459-1460.
20. Gregorio DI, Cromley E, Mrozinski R, Walsh SJ: **Subject loss in spatial analysis of breast cancer.** *Health Place* 1999, **5:**173-177.
21. **Using Geographic Information Systems Technology in the Collection, Analysis, and Presentation of Cancer Registry Data: A Handbook of Basic Practices.** Edited by: *Wiggins. Springfield (IL), North American Association of Central Cancer Registries,*; 2002.
22. Ratcliffe JH: **On the accuracy of TIGER-type geocoded address data in relation to cadastral and census areal units.** *Int J Geogr Inf Sci* 2001, **15:**473-485.
23. US Bureau of the Census: *TIGER/Line 1997. CD-TGR97-01. Washington, DC, US Department of Commerce*; 1998.
24. Bonner MR, Han D, Nie J, Rogerson P, Vena JE, Freudenheim JL: **Positional accuracy of geocoded addresses in epidemiologic research.** *Epidemiology* 2003, **14:**408-412.
25. New York State Office of Real Property Services: **1999 Real Property Data.** *NYS Office of Real Property Services*; 2000.
26. US Bureau of the Census: **1990 Census of population and housing summary tape file 1A.** *Prepared by Bureau of the Census, Washington, DC*; 1991.
27. MapInfo Corporation: **MapMarker Plus - Version 6.0.** *Troy, NY*; 2000.
28. New York State Department of State: **New York State 2000 Digitally Enhanced OrthoImagery.** *NYS Department of State*; 1999.
29. Mardia KV, Jupp PE: *Directional Statistics Chichester, John Wiley & Sons Ltd*; 2000.
30. Federal Geographic Control Subcommittee: **Geospatial Positioning Accuracy Standards. FGDC-STD-007.** *Reston, Virginia, Federal Geographic Data Committee*; 1998.
31. US Bureau of the Census: **TIGER/Line Files 1997 Technical Documentation.** *Washington, DC, US Department of Commerce*; 1997.
32. Fotheringham AS, Brunsdon C, Charlton M: *Quantitative Geography. Perspectives on Spatial Data Analysis London, SAGE Publications Ltd*; 2000.
33. Foote KE, Huebner DJ: **Error, Accuracy, and Precision.** *http://www.colorado.edu/geography/gcraft/notes/error/error.html (accessed 2003). The Geographer's Craft Project, Dept. of Geography, University of Colorado at Boulder*; 2000.
34. Griffith DA, Haining RP, Arbia G: **Chapter 2: Uncertainty and error propagation in map analyses involving arithmetic and overlay operations: inventory and prospects.** *Spatial Accuracy Assessment; Land Information Uncertainity in Natural Resources. Michigan, Sleeping Bear Press, Inc.*; 1999:11-25.
35. Heuvelink GBM: **Chapter 14: Propagation of error in spatial modelling with GIS.** *Geographical Information Systems. Volume1*. 2ndth edition. Edited by: *Longley PA, Goodchild MF, Maguire DJ and Rhind DW. New York, John Wiley & Sons, Inc.*; 1999:207-217.
36. Copeland KT, Checkoway H, McMichael AJ, Holbrook RH: **Bias due to misclassification in the estimation of relative risk.** *Am J Epidemiol* 1977, **105:**488-495.