INT. J. GEOGRAPHICAL INFORMATION SCIENCE, 2001, VOL. 15, NO. 5, 473–485

Research Article

On the accuracy of TIGER-type geocoded address data in relation to cadastral and census areal units

JERRY H. RATCLIFFE

School of Policing Studies, Charles Sturt University, NSW Police College,
McDermott Drive, Goulburn NSW 2580, Australia; e-mail:
jerry.ratcliffe@bigfoot.com

(*Received 20 January 2000; accepted 16 January 2001*)

Abstract. In many applications of Geographical Information Systems (GIS) a common task is the conversion of addresses into grid coordinates. In many countries this is usually accomplished using address range TIGER-type files in conjunction with geocoding packages within a GIS. Improvements in GIS functionality and the storage capacity of large databases mean that the spatial investigation of data at the individual address level is now commonly performed. This process relies on the accuracy of the geocoding mechanism and this paper examines this accuracy in relation to cadastral records and census tracts. Results from a study of over 20 000 addresses in Sydney, Australia, using a TIGER-type geocoding process suggest that 5–7.5% (depending on geocoding method) of addresses may be misallocated to census tracts, and more than 50% may be given coordinates within the land parcel of a different property.

1. Introduction

There are now a large number of socio-economic applications that require a researcher to manipulate address-based data in a Geographical Information System (GIS). As the capabilities of databases and GIS improve, a greater level of resolution is possible and the individual address is becoming a standard level for spatial investigation. Such address-based data may refer to a wide range of records such as insurance policy holders, medical records or vehicle ownership, and application areas include market research, the provision of public utilities and the work of the emergency services. To utilise fully the spatial nature of an address it is usually necessary to create a point associated with the address record. Most GIS have the capability to geocode - the process of associating an address record with a point on a map— and to perform spatial queries on the result. Such spatial queries may only require the original point data and a number of spatial investigative algorithms have been designed for point pattern analysis (Bailey and Gatrell 1995), though other techniques may require a point-in-polygon operation to compare the number of point records with attribute data associated with an areal unit.

Early work by Gatrell and colleagues (Gatrell 1989, Gatrell *et al.* 1991) examined the relationship between points attributed to UK postcodes at a 100 m resolution,

International Journal of Geographical Information Science
ISSN 1365-8816 print/ISSN 1362-3087 online © 2001 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/13658810110047221

with enumeration districts, the smallest UK census tract. Martin and Higgs (1997) reviewed a number of spatial data sets in the UK, comparing council tax registers with individually geocoded locations, and variations in the number of properties in an enumeration district by various means. These studies concentrated on the extraction of maximum accuracy from postcode information (variously the zip code or postal code depending on country), used regularly in the UK as a means to protect the confidentiality of individuals in a study. Much of the research was conducted at a time when automated individual address mapping was in its infancy, expensive and generally beyond the scope of many researchers unable to access more descriptive address records. Recent advances in the last decade have raised expectations in location referencing, and reduced costs to the level that the centreline data necessary for geocoding a city such as Canberra, Australia (population 330 000) costs less than US$500. Once access to confidential address-based data has been negotiated (often the hardest part of the process) it is now possible to use as a matter of routine the individual property as the unit of spatial investigation. Within the specific field of law enforcement, the imperfections in the geocoding of address data for policing purposes have been recognised (Harries 1999, PFCML 2000), though not articulated or quantified clearly. The level of imprecision in geocoding is important to law enforcement so that a better understanding of the limitations of the data can be appreciated prior to further spatial analysis, analysis that might be spurious given the spatial limitations of the geocoded data.

This study aims to build on the previous work in the area, by concerning itself with the accuracy 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 from the Australian national census. We start with an outline of the geocoding process before examining the study area in Sydney, Australia, in §3 and §4.

2. Geocoding address records

Point-in-polygon operations require both points and polygons. Point data for urban locations are commonly created from geocoded address records. Current geocoding tools in the USA and Australia are derived from US TIGER files, collections of street line segments that hold street names, and the range of house numbers on each side of the street as attribute data. The TIGER composition was developed from the DIME structure used originally to map US census data and provides a US-wide address matching standard (Cooke 1998).

An example is shown in figure 1 where a polyline for Smith Street has a From node, To node and attribute data. The attribute data for the line indicates that odd house numbers are on the left side of the street and range from 1 to 17, and the even numbers found on the right side of the street range from 2 to 18. This is not the only method of geocoding addresses and some commercial organisations have created address registers that contain individual coordinates for every house and address. One of the most widely used in the UK is the ADDRESS-POINT data set that contains *x* and *y* coordinates for millions of British addresses. Derived from Ordnance Survey mapping (www.ordsvy.gov.uk[1]) with a resolution of 0.1 m the use

[1] Information relating specifically to ADDRESS-POINT is available at www.ordsvy.gov.uk/productpages/addresspoint/index.htm (accessed January 2001).

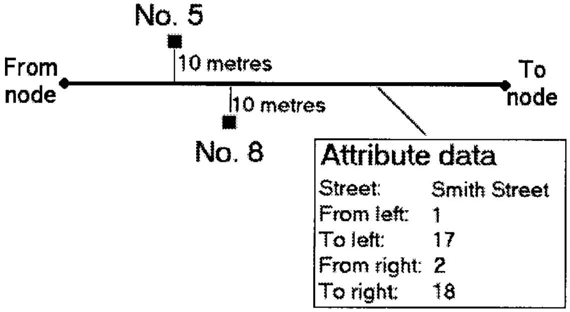

Figure 1. Example geocode line segment with an offset of 10 m (inset not shown).

of individual address coordinate systems has the potential to provide excellent geocoding accuracy though some limitations have been noted in relation to correlation between ADDRESS-POINT georeferencing and other data sets (Martin and Higgs 1997). Outside the UK and the few other areas where individual address co-ordinate data sets exist however, the use of TIGER-type line segment geocoders is still the norm.

Most vector GIS have built-in geocoding tools that use derivatives of TIGER files to interpolate a suitable *x* and *y* coordinate for an address. The target address is used to identify a suitable line segment with an appropriate number range and then a location along the line is interpolated between the From and To node based on the number of houses along the line. In the two most popular desktop GIS programs (ArcView and MapInfo) it is now possible to select an offset that relocates the derived point a number of metres perpendicular to the line segment. ArcView and the most recent version of MapInfo (at the time of writing version 6.0) both allow the user to select an offset distance, though the new version of MapInfo was only made available at the time of final draft of this paper, and all previous versions had a hardcoded offset of 10 m. This paper is therefore written to assist the vast majority of MapInfo users who are using previous versions of MapInfo (versions 5.5 and earlier), though some consideration to future use of version 6.0 and beyond is considered. This paper will also be of benefit to all geocoders seeking clarification of the importance of the offset in geocoding. Version 6.0 of MapInfo also permits the selection of an inset, a distance from the From and To node that is ignored before geocoding begins. The relative import of this feature is discussed later.

Given that the line segment should ideally follow the street centreline this helps to locate the geocoded address closer to the more likely house position and away from the road. This is shown in figure 1 where number 5 on the left and number 8 on the right side of the street are offset by 10 m, the offset in most MapInfo versions.

There are a number of potential problems with this type of geocoding technique and a number of different sources of error (Harries 1999). Some of these are outlined below.

Out-of-date street directories. The geocoding base files might be out of date, and this can mean that new addresses or even new housing estates of dozens of streets are not known. While some vendors endeavour to keep their street directories as current as possible the problem is worst in urban fringe areas where development and change are most rapid.

Abbreviations or misspelling. Street names can be misspelled or abbreviated. For example Smith Street can also be shown as Smith St., Smythe Street, or Smithe Street. Many geocoding engines may only be able to recognise some of the variations. Often an alias type of table to replace misspellings and abbreviations must be manually created. Rigid database entry tied to street naming standards can limit the impact of this type of error. The Australian standard (www.auslig.gov.au) has been in use for about 5 years and has been developed to establish a consistent set of rules for street naming, while similar standards exist in other countries.

Local name variations. In some areas the street directory name of a road may not reflect the local name of the road and this can mean that the target address cannot be matched to the geocode address file.

Address duplication. In many cities roads are named after historical figures or places and can be duplicated in nearby suburbs. For example, 40 occurrences of Smith Street can be found in the Sydney street directory. Without the geocoding process specifying a suburb or a limiting boundary polygon it is possible for a geocoding engine to find multiple matches of common addresses.

Non-existent addresses. Simple typographical errors can easily turn 30 Smith Street into 300 Smith Street, an address that may not exist.

Line simplification. The sections of line segment may not reflect accurately the geography of the road or housing layout. This will reduce the geocoding accuracy and can affect streets in hilly areas that can be more winding or new streets in residential areas that are built on sweeping bends.

Noise in the address file. Unnecessary additional data can complicate an address field. Some householders may choose to give their house a name or a target address field may start with a company name and this can be difficult for a geocoding engine to interpret accurately. For example some geocoders would struggle with: ‘Dunroamin, 24 Smith Street’. Similarly a business address can often start with a company name.

Geocoding non-address locations. A number of organisations such as the emergency services might want to record the location of incidents that do not occur in buildings. Some geocoding engines can use special characters to recognise intersections but, for example, an ambulance supervisor might want to record a road traffic accident that happened ‘50 metres West of the junction with Brown Street’ or ‘outside number 12 Smith Street’. These locations would be rejected by most geocoders, or they would geocode to the exact intersection or house.

Geocoding imprecision. The geocoded point might be some distance from the actual address, and it is the extent of this particular problem that the next sections will examine.

Ambiguous or vague addresses. Ambiguity of addresses in the target file (such as simply ‘Smith Street’) can make geocoding impossible. Some geocoders can be configured to geocode ambiguous addresses to either the nearest existing address or to the centroid of the line segment. Neither is an ideal solution as a number of vague addresses can create a cluster of points at an arbitrary location.

3. Source data

This study uses three sources of data. The geocoded points have been determined from centreline data for Sydney in StreetWorks version 5.0, available from MapInfo Australia. This was used in conjunction with the standard geocoding engine found in MapInfo version 5.5—the most current version at time of analysis. The geocoded locations have been compared to the cadastral base for the study area. Cadastre records are a catalogue of interests in land parcels. Usually these are now retained as digital maps that contain descriptions of land parcels as well as unique identifiers that can be used to identify who has ownership rights to the land as well as other legal interests. An enhanced digital cadastral file has been made available for this study by the Land Information Centre, Bathurst, NSW, Australia. The Land Information Centre (LIC) is the government mapping organisation for the state of New South Wales. Individual polygons in the cadastre are identified by a unique Lot/Deposited Plan (Lot/DP) number. There is no indication from this unique identifier as to the conventional address of the property or even the name of the street. During 2000 the LIC have been attempting to correlate disparate data sources from a number of government bodies to try and identify a street number and road name for each land parcel in the cadastre. Road names exist as an integral part of the Digital Cadastre Database (DCDB) for New South Wales and are encoded into the road polygons but due to technical constraints the names are not included in the other parcel tags. This is still under development and the choice of study area was dictated by the availability of data from a semi-completed area. The Eastern Suburbs of Sydney, New South Wales, Australia has attracted interest from a number of organisations since this area was the victim of a severe hail storm in 1998 that caused a considerable amount of property damage. Figure 2 shows this region.

The cadastre for this area has had house numbers attached to polygons by the LIC, where these can be determined, though no street names. Where a house number was available the complete street name was attached manually by the author. This

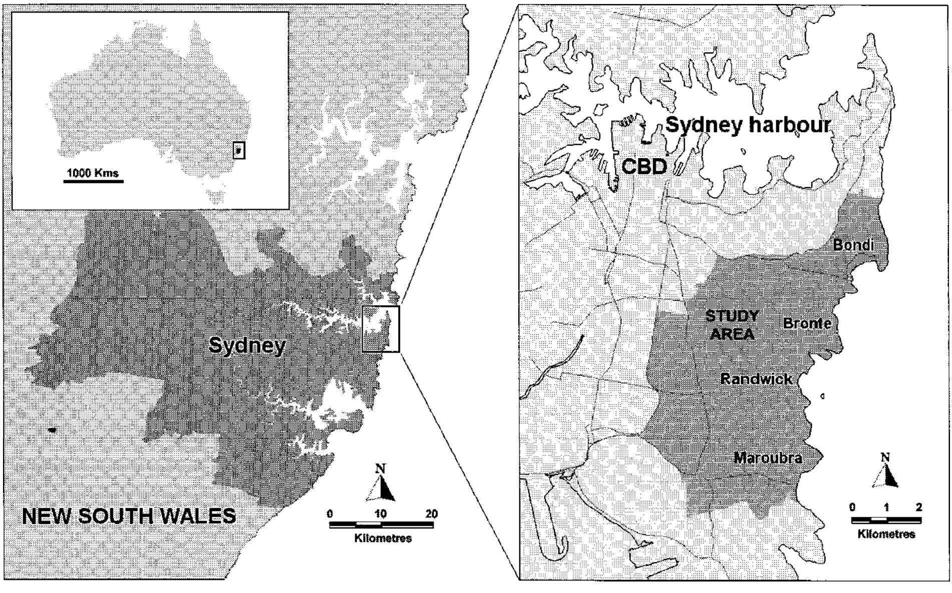

Figure 2. The study area in the Eastern suburbs of Sydney.

involved identifying the nearest street for the property and confirming that the house range for the street matched the house number, and that they were on the correct side of the street, running in the correct direction (see figure 1). Where any ambiguity existed the property was not included in the final study set. The majority of ambiguous locations were on street corners where a house number could have been in either street. These accounted for less than 2% of the original data set, resulting in over 20 000 addresses included in the following analysis.

The third source of data was the polygon boundaries of the census collection districts for the study area. Collection districts (CD) are the smallest geographical area defined in the Australian Standard Geographical Classification (ASGC). In urban areas, such as the Eastern Suburbs of Sydney, there is an average of 225 dwellings in each of these census tracts. The boundary files were obtained from CData96, a production of the Australian Bureau of Statistics that includes census data along with boundary files in MapInfo format.

4. Analysis

A list of 21 890 addresses was obtained by cross-comparison of the StreetWorks file and the cadastral data set. This ensured that both files knew an address. The purpose of the study was not to correct every error in each data set, but to extract enough addresses common to both files that a meaningful analysis could be completed. There are over 21 890 addresses in the study area shown in figure 2, though a small proportion did not have house numbers shown in the cadastre file, or their house number could have referred to more than one street at an intersection. This lack of completeness in the data set is understandable given that this address enhanced cadastral data set is still under development. Version 5.5 of MapInfo was used for the following analysis (this version has a hardcoded offset of 10 m).

4.1. *Geocoded points and cadastre polygons*

The first part of the study performed a standard point-in-polygon operation to examine the match between geocoded points and their corresponding cadastral polygons. Of the 21 980 addresses 7774 (35%) geocoded points were located within a polygon from the cadastral set. Unfortunately the majority of these were located in the wrong polygon. Only 2165 (10% of the 21 980) were correctly located in the concordant cadastral polygon, while 5609 (26%) were located in the polygon for another address. If we consider these as a proportion of only those points that were located within a polygon, then this equates to an error level of 72%. This figure does not include those geocoded points that were referenced within another cadastral property polygon that was not part of the 21 980 study set.

These figures would seem to raise cause for concern for researchers and mappers interested in micro-level mapping applications of address related records. The TIGER system was not designed originally along a cadastral basis but was created to provide a relatively accurate large scale mapping capability. It is therefore understandable that to some degree it does not hold up to a micro-level examination. The low number of records that were located correctly (10%) did however suggest that further work was necessary to better understand the extent of the geocoding error. Visual inspection of a number of misplaced points showed that a high number were near the correct polygon, often located in a neighbouring land parcel, though this was not always the case. The next part of the study endeavoured to get a more empirical measure of the error level.

4.2. *Geocoded points and cadastre centroids*

The geocoded point for each address was extracted as before, as was a centroid for each known address in the cadastre. The computed centroid for each property in the cadastre does not show the location of the actual house. One of the advantages of using the Eastern Suburbs of Sydney is that this is a more established and older part of the city, and a desirable area in which to live, and the land plots tend to be compact and small. The mean area of the land parcels used in this study is only 434 square metres. Examination of aerial photography of the study region with cadastre polygons used as an overlay, indicated that the majority of properties dominate the whole land parcel with little room for a garden or, surprisingly for such an affluent area in Sydney, a swimming pool. Due to the dominance of the domestic unit within the land parcel the centroid was located within the building perimeter the vast majority of times. The use of the centroid as an indication of the location of an individual property is used in this study as the centroid (or similar central point identifier) is likely to form the basis of any future individual address referencing system (Hickson 2000) though a firm commitment to a national standard has yet to be confirmed. It is therefore of interest to measure the distances between the geocoded points and the cadastral centroids.

The two sets of location coordinates from the geocode set and the cadastre set can be compared by computing the distance between paired points, and the statistics of the distribution of these distances can be examined. The mean separation is 47 m, with a standard deviation of 187 m. A histogram of the distances revealed that there are a number of outliers with distances of over one kilometre. These outliers are suspected errors in either the StreetWorks file, the enhanced cadastral data set, or the geocoding engine. This was confirmed by plotting the points and comparing back to the original cadastre files. The actual cause of the error is not possible to determine without the resources to ground truth hundreds of addresses, and when plotted on a map the erroneous addresses were scattered across the study area without any apparent pattern. Therefore, as in other studies that have examined distances between estimated points and actual locations (Gatrell 1989), a 'trimmed mean' that does not include the smallest and largest 5% of values can be used as a more accurate reflection of the data. The mean distance between the geocoded points and their corresponding cadastre centroid in the reduced set of 19 791 concordant points is 31 m.

4.3. *Offset adjustment*

The 'trimmed mean' of 31 m (standard deviation 14.7 m) from the previous section would appear to be a fairly acceptable degree of error, given that a cadastre centroid is probably rarely more than about 10 m from the boundary of the land parcel. It may be that adjusting the offset used by the geocoding engine (see figure 1) could reduce this measure of separation. An adjustment in the offset may improve the separation between linked points, but only if the cadastre centroid is found in the area of a perpendicular line extrapolated from the geocoding line segment through the geocoded point at the 10 m offset.

Figure 3 shows a line segment that has geocoded a point A, while the centroid for the cadastral land parcel is shown as point B. The solid line shows an offset of 10 m and the continuance of the perpendicular line through point A. The dashed line shows the degree of error (θ) between the perpendicular extrapolation and the centroid location. This angle can be measured in the data set by extracting the

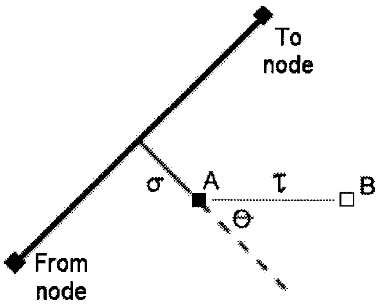

Figure 3. Line segment from a standard TIGER-type geocode file with an offset to geocoded location A. The centroid for the cadastre polygon is shown as B and the angle θ indicates the amount of variation from the perpendicular.

relevant line segment from the StreetWorks file and comparing this angle to the line A–B. Ideally we would hope to see low angles for θ indicating that an adjustment in the offset distance may help improve geocoding accuracy.

To test this, the two line angles were extracted from the various data sets. All analyses took place within a single coordinate system (Australian Map Grid 84 Zone 56) and this allowed for the extraction of metric coordinates, and therefore the use of plane geometry. The line segment nodes were used to calculate the perpendicular offset angle (σ in figure 3) and the geocoded point and the cadastral point were used to calculate the separation line A–B (τ in figure 3). This operation was performed on a subset of the whole data set comprising of 10 000 addresses.

If an adjustment of the offset distance in the geocoding engine would improve the separation distance then we would expect to see angles closer to zero indicating that the actual location of the cadastre centroid lay roughly along the perpendicular line extending from the line segment. In figure 4 the angles relative to the line segment have been shown in a quadrant radar diagram, rounded to whole degrees, along with the number of records for each angle. The length of each line from the quadrant's hub indicates the number of records that are angled in that direction. Because the direction of the line segment has not been taken into consideration, only a quadrant of the radar diagram is shown with all angles of θ taken to be acute. An exact correlation with the perpendicular line is shown at the zenith. Close to the horizontal indicates a greater angle of θ, as seen in figure 3. The results from this analysis are shown in figure 4.

If most of the cadastral centroids had been found roughly in line with the perpendicular street segment then we would have expected to see a large number of records appearing close to the vertical in figure 4. However as can be seen, the majority of the addresses are found to have a displacement angle greater than 45°. This would suggest that varying the amount of offset from the geocode file line segment is not alone likely to have the desired affect on the accuracy of the geocoded points. Because this study has only measured the displacement angle (θ) and not the direction of the error, it may be that increasing or decreasing the offset may serve in some cases to reduce the displacement angle but actually increase the Euclidean distance between the geocoded point and the cadastre centroid.

As an adjuvancy to geocoders, repetitive testing of the study area data sets prior to final draft established that minor improvements in the accuracy of this study data could be achieved by increasing the offset to 25 m. It should be cautioned that this

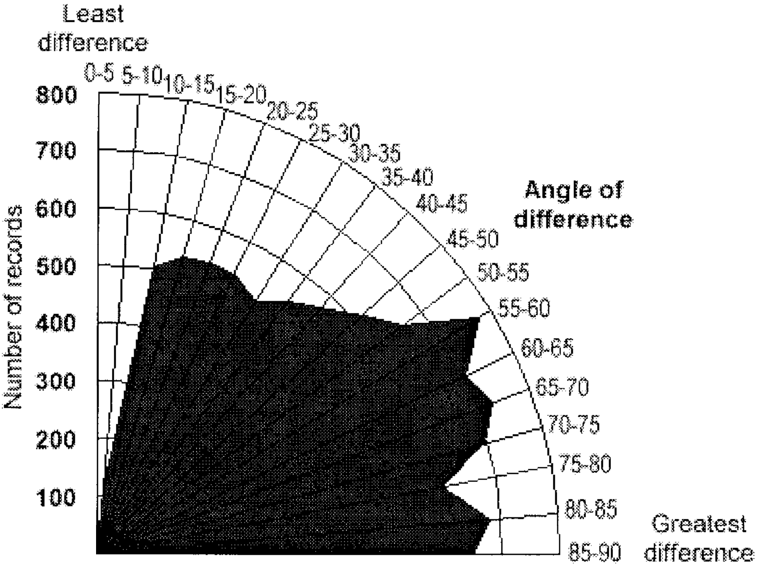

Figure 4. Quadrant radar diagram indicating the degree of difference between a line perpendicular to a geocode street segment and the line between the geocoded point and the concordant cadastral centroid.

figure works well for the Eastern Suburbs of Sydney with correspondingly low property sizes, while other areas may need a different adjustment. It is clearly indicative however that 10 m, the previous MapInfo hardcoded value, is insufficient even for areas like the Eastern Suburbs where the centroid is quite close to the road.

4.4. *Point-in-polygon operations*

If a research project is designed to examine the relationship of a point variable with census data, then the important characteristics of the two data sets are not their simple proximity but whether cases are allocated to the correct areal unit (Gatrell 1989). This paper consequently turns its attention to the spatial relationship between census collection districts (CDs) and the geocoded location.

There are many types of areal unit used for socio-economic research, and the aggregation of points within other types of polygons can also be used as a cartographic technique to simplify the display of large numbers of points. Neither the research application nor the cartographic requirement necessarily require CDs, though these have been used in this study as they are a common choice of thematic aggregator, giving a good indication of population density, and are a popular choice of areal unit for research. They are also readily available to University researchers and outside organisations. It is important to stress that the results of point-in-polygon searches depend on the accuracy with which the boundaries of the polygons were digitised (Gatrell 1989) and in the case of census tracts the user's trust is in the hands of the census bureau. Given the metre resolution of geocoded points in this study it must be acknowledged in advance that some degree of geographical uncertainty will exist about points that lie close to a census tract boundary and that for compression purposes the general CD boundaries that are available publicly have undergone a stage of line simplification. The following section of the paper examines this degree of uncertainty.

Using the CD file available from the Australian Census Bureau, a point-in-polygon operation was performed on both the geocoded points and the cadastre centroids for all 21 890 points in the study area. Comparison of the unique identifiers for each concordant pair show that 1634 (7.5%) geocoded points fell into different CDs than their corresponding cadastral centroid. The use of every address in the study was felt reasonable even though it was known from the earlier analysis that there were outliers that indicated errors in the geocoding process. This is because most GIS users will not have the benefit of both data sets (given that the cadastre set used in this study is more or less unique) and in the absence of any indication to the contrary will take the geocoded location as 'truth'. Most users of geocoding tools, such as police analysts wanting to map crime occurrence for example, will neither have the time nor the skills to question and analyse the distribution of 20 000 geocoded locations, and will proceed with their analysis using the points geocoded automatically by their GIS.

When the misallocated points are plotted as in figure 5, it becomes apparent that there is no particular area of density, but that the points that are in the wrong CD are distributed across the study area. A visual inspection of figure 5(*a*) suggests that many of the geocoded points (black dots) are located on or near a CD boundary (white bordered regions). This would appear to be due to line simplification in the CD boundary file where the generalised boundary of the collection district has not mimicked the urban geography of the road layout. The example shown in figure 5(*b*) is common for many of the misplaced geocoded coordinates across the study area.

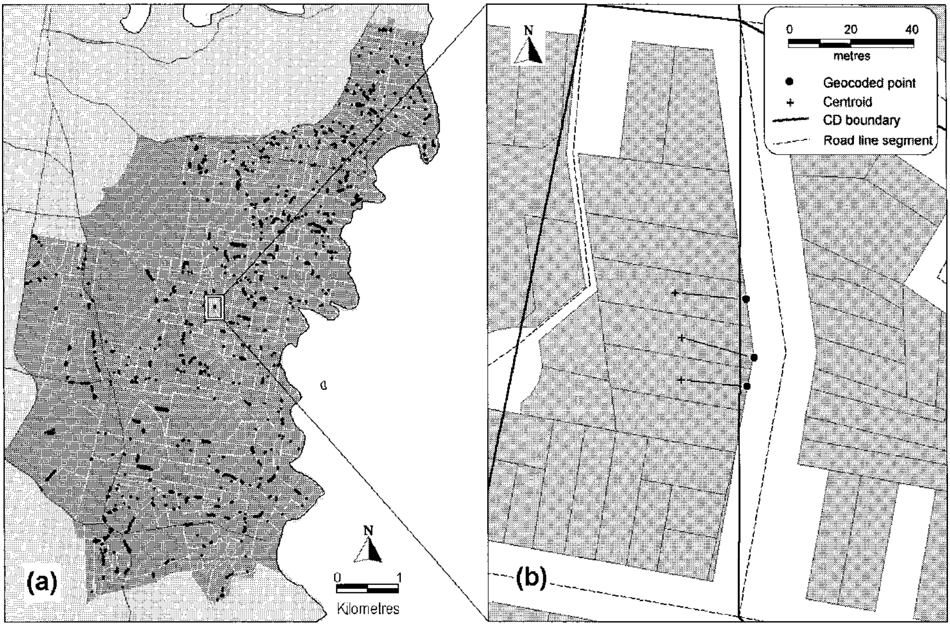

Figure 5. (*a*) Geocoded points (black dots) that are in a different CD (white bordered regions) to their cadastral centroid. (*b*) Displacement from the corresponding centroid CD due to line simplification in the CD boundary file. A small increase in the geocoding offset would be sufficient to place the geocoded points into the correct CD.

A discrepancy of 7.5% does seem high given the importance of collection district level analyses and this figure is likely to be of interest to researchers concerned with address level data and their relationship to census variables. The result of this analysis indicates that this should be another error factor to be considered if the results of a point-in-polygon process indicate only slight statistical significance in micro-level socio-economic analyses. The lack of high variation in variables between adjacent CDs goes some way to alleviating the effects of this discrepancy for counts of cases in polygons. The mean distance between points from the first analysis would suggest that the problem is not a factor if larger areal units are used, though this might reduce the impact and quality of the analysis undertaken. It is clear from figure 5 that an increase in offset would generally improve the point-in-polygon accuracy, and the reader is referred to the end of the previous section that identified some value in the use of an offset of 25 m. This would both improve the point-in-polygon geocoding and bring the final point closer to the cadastral unit centroid. Again the *caveat* should be added that the figure of 25 m worked well for the Eastern Suburbs of Sydney, but may need to be increased for areas with more expansive properties.

The use of an inset is currently only available to MapInfo version 6 users. Given an optimum of a 25 m offset, a number of trials were conducted to identify the impact of the inset. Two types of inset are available: a fixed distance along the segment from the node, or a percentage distance based on the length of the road segment. Tests were conducted on insets of 0 metres to 50 m (5 m increments) and 5% to 40% (5% increments). It is difficult to establish an ideal, as some settings had a greater percentage of points close to the address centroid but also a slightly higher number of points in the wrong census tract. The end choice is likely to be based on the desires of the user, who might not be required to perform many point-in-polygon operations. As an indication to users (with the caveat that these results are based on Eastern Sydney and are indicative of a compact inner-urban suburb) an optimum that achieved the best results overall was the use of a 15 m inset with a 25 m offset. This choice of setting in the geocoding process achieved the results shown in table 1.

5. Summary

An important *caveat* should be in place when reviewing this research. The author did not create the cadastre file and therefore an indication of the level of error in

Table 1. Geocoding characteristics of 23 087 addresses geocoded in the Eastern Suburbs of Sydney with a 25 m offset and a 15 m inset.

Characteristic	Percentage (%)
Points in correct census tract	94.85
Points in correct address plot	46.75
Points not in an address plot	18.62
Points in wrong address plot	34.63
Points within 5 m of centroid	11.69
Points within 10 m of centroid	33.75
Points within 15 m of centroid	49.07
Points within 20 m of centroid	60.11
Points within 50 m of centroid	89.08
Points within 100 m of centroid	96.42

the enhanced file is not available. This will not be possible without a considerable amount of ground-truthing and the author lacks the resources for this. Given this unknown in one of the files there was a need for a conservative study, and when attention was turned to the distances between the centroids and the geocoded locations, it was deemed sensible to trim the highest and lowest 5% from the values. The output from this analysis suggests that in compact, urban areas the accuracy of geocoded points when a 25 m offset and 15 m inset are employed is to 81m in 95% of the cases, 31 m in 75% of the cases, and 16 m 50% of the time. It is not in doubt that the distances involved are small, but the fact over 5% of points are allocated to different CDs from their cadastral centroids is likely to have implications when an attempt is made to relate counts of cases to census variables, or where areal comparisons are made between variables geocoded through two different processes. This figure increases to 7.5% when a 10 m offset is used, as is only available to the majority of current MapInfo users. If accuracy of geocoding is important to a MapInfo user, upgrading to version 6.0 or beyond may be a worthwhile expense. When attempting to relate counts of cases to a small polygon such as a CD no immediate solution to the problem presents itself, other than a conservative interpretation of any subsequent findings. Although the choice of geocoding settings presented here have optimised the point location, it is not a corollary that the same figure would improve geocoding elsewhere and other researchers and geocoders are unlikely to have access to both data sets used in this study. In the reverse situation, where an areal variable is extracted for association with a point record, it is possible to use an areally weighted buffer approach to minimise the impact of misallocated points. A *vicinity*-type approach such as this has been used and demonstrated successfully in the analysis of crime locations, described as points, in relation to deprivation measured areally in enumeration districts (Ratcliffe and McCullagh 1999).

Given the spatial separation found in this study between geocoded coordinates and a variety of geographical units, it would seem prudent to verify the accuracy of any new geocoding processes or products, and if a significant improvement is found, to consider the validity of any important findings that used the older process where any statistical significance is slight. This is a standard scientific principle but it is worth stating again given that it seems unlikely that we have reached the pinnacle of geocoding methodology.

With more than 5% of the geocoded points in this study falling outside the correct cadastral polygon, and more than 50% being coordinated to either the street, or worse the wrong property (rising to 90% with a 10 m offset), the limitations of the current geocoding process become apparent. A number of police services in the UK have been utilising the ADDRESS-POINT file mentioned earlier in this paper and have built the file into a gazetteer (Ratcliffe and McCullagh 1998). A rigid data entry process rejects any location that is not known to the system and this prevents the encoding of erroneous addresses and permits the geocoding at the data point of entry. This type of data entry method requires constant maintenance as new addresses have to be cross-checked and then permitted onto the system, but it does present the possibility of a near 100% geocoding rate. When, if possible, public services have the opportunity to move to this type of system and where accurate geocoding presents real analytical profit, such an adjustment may prove beneficial. For the rest of us it would seem that the present method of geocoding, warts and all, is here to stay for the immediate future.

References

Bailey, T. C., and Gatrell, A. C., 1995, *Interactive Spatial Data Analysis* (London: Longman).

Cooke, D. F., 1998, Topology and TIGER: The Census Bureau's contribution. In *The History of Geographical Information Systems: Perspectives from the Pioneers,* edited by T. W. Foresman (Upper Saddle River, NJ: Prentice Hall), pp. 47–57.

Gatrell, A. C., 1989, On the spatial representation and accuracy of address-based data in the UK. *International Journal of Geographical Information Systems*, **3**, 335–348.

Gatrell, A. C., Dunn, C. E., and Boyle, P. J., 1991, The relative utility of the Central Postcode Directory and Pinpoint Address Code in applications of geographical information systems. *Environment and Planning A*, **23**, 1447–1458.

Harries, K., 1999, *Mapping Crime: Principles and Practice* (Washington: US Department of Justice).

Hickson, J., 2000, *Land Information Centre* (Bathurst, NSW, Australia), personal communication.

Martin, D., and Higgs, G., 1997, Population georeferencing in England and Wales: basic units reconsidered. *Environment and Planning A*, **29**, 333–347.

PFCML, 2000, *Police Federation Crime Mapping Laboratory*, Geocoding in law enforcement: Final report. (Washington DC, US Department of Justice: Office of Community Orientated Policing Services): 17.

Ratcliffe, J. H., and McCullagh, M. J., 1998, Identifying repeat victimisation with GIS. *British Journal of Criminology*, **38**, 651–662.

Ratcliffe, J. H., and McCullagh, M. J., 1999, Burglary, victimisation and social deprivation. *Crime Prevention and Community Safety*, **1**, 37–46.