Volume 17, Number 3–4
March–April 2006

NSW Public Health Bulletin

NSW HEALTH

ISSN 1034 7674
State Health Publication PH 060106

AUTOMATED GEOCODING OF ROUTINELY COLLECTED HEALTH DATA IN NEW SOUTH WALES

Richard Summerhayes, Paul Holder and John Beard
Northern Rivers University Department of Rural Health

Geoff Morgan
Northern Rivers University Department of Rural Health and North Coast Area Health Service

Peter Christen
Department of Computer Sciences
Australian National University

Alan Willmore and Tim Churches
Centre for Epidemiology and Research
NSW Department of Health

Almost every record of an individual's contact with the NSW health system contains some form of spatial information, whether a street address or postcode. This information can be used to assign a geocode to the record, which in turn can be used to examine the spatial distributions of disease and health service utilisation. This article describes a study that compares two different methods of geocoding addresses from routinely collected administrative health data to enable small area analysis.

WHAT IS GEOCODING?

Geocoding is the process of allocating geographical coordinates (such as latitude and longitude) to an address, thus defining the position of the address on the Earth's surface. The geocode itself can be used in the analysis or, alternatively, the geocoded record can be assigned to a spatial unit, such as a census collection district, that is smaller than other spatial units generally available (for example, postcodes), and then analysed. Spatial analysis of routinely collected health data in Australia has generally been limited to larger spatial units such as local government areas or area health service boundaries.[1,2]

continued on page 34

CONTENTS

While spatial analysis at this level can be useful, many interesting spatial features that occur at a smaller geographical level can be lost in the aggregation of data to larger units.

GEOCODING PACKAGES

FEBRL (Freely extensible bio-medical record linkage)

NSW Health and the Australian National University have recently developed FEBRL 'freely extensible bio-medical record linkage'.[3] Before geocoding, FEBRL firstly 'cleans' the data by transforming the original text address into a standardised format that corrects for missing, erroneous or abbreviated data (for example, 'st' is transformed to 'street', 'pde' is transformed to 'parade'). The data are then 'parsed' by separating the address into individual standardised elements (for example, the element for 'street' or 'parade' is 'wayfare type'). FEBRL can then match the cleaned address data to the Australian Geocoded National Address File (G-NAF)[4], using a probabilistic algorithm, and allocate a geocode. The G-NAF contains 12.6 million unique geocoded addresses derived from a variety of national and state-based datasets. The geocode is provided for the centre (or centroid) of a property parcel. The G-NAF is updated every quarter.

MapMarker

MapMarker is a commercial geocoding package developed by MapInfo Australia[5]. MapMarker cleans and parses the address data and then links the cleaned address to an Address/Co-ordinate Dictionary, using fuzzy logic and Soundex indexing. Soundex is an algorithm for phonetic name encoding which indexes names by their English pronunciation to overcome minor differences in spelling. Secondly, the Address/Co-ordinate Dictionary, which is MapInfo's StreetWorks Australia database, is based on street centerlines and town/postcode centroids. Each street is broken into linear segments with a coordinate pair at both endpoints. Linear interpolation between the endpoints provides geocode coordinate values for a given address. Figure 1 illustrates the allocation of geocodes using the MapMarker street centre line interpolation method compared with the allocation of geocodes using the G-NAF land parcel centroid method.

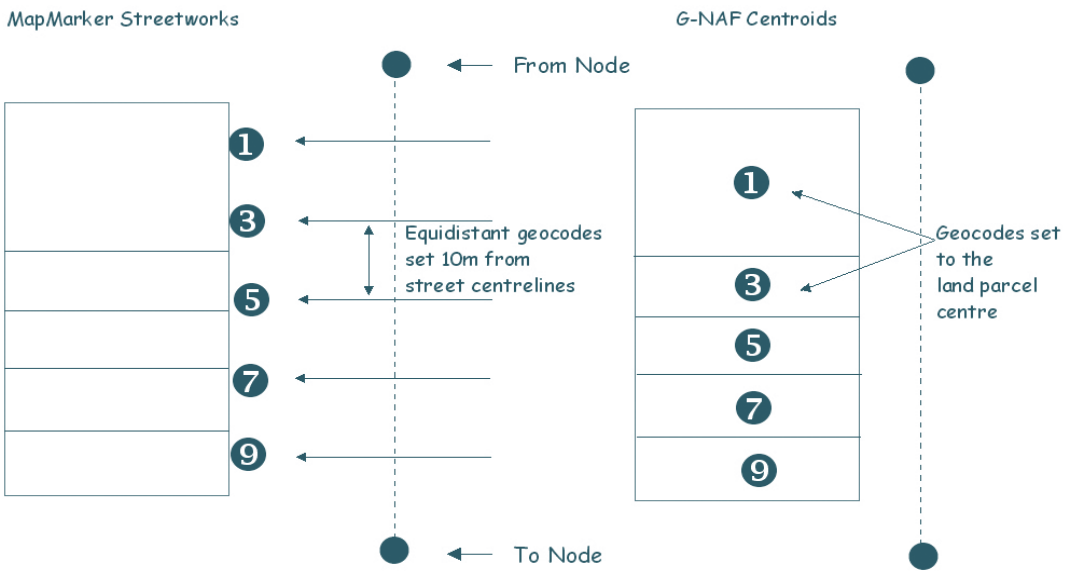
APPLYING GEOCODING

There is growing recognition of the value of geocoding administrative datasets[6] to enhance their usefulness for service planning and resource allocation.[7,8] Geocoded health events can also be used to investigate the possible effect of environmental hazards (often using proximity of residential address to a source as a proxy for exposure) such as: congenital malformation and proximity to a hazardous waste sites[9]; childhood asthma and proximity to roads[10]; and childhood leukaemia and exposure to electromagnetic fields. [11] Geocoded health data is also being used to investigate the epidemiology of specific diseases including adverse birth outcomes and childhood leukaemia.[12]

While a small number of outcomes can be geocoded individually, geocoding large numbers of records in routinely collected health databases requires well-defined geocoding procedures. Australia has lagged behind internationally in

FIGURE 1

EXAMPLE OF MAPMARKER INTERPOLATED HOUSE NUMBER WITH AN OFFSET OF 10 METRES FROM STREETLINE CENTRE, COMPARED TO A G-NAF LAND PARCEL CENTROID FOR THE SAME ADDRESSES.



developing standards for defining a geocoded address.[13] Limitations for accurate geocoding include: errors in the address/street reference data; imprecision in the algorithms used by geocoding software[14]; and errors in the original addresses such as post box addresses or property names.[15] Lack of precision in addresses, especially in rural areas, can bias urban and rural comparisons and limit the opportunity for small area or point source analysis.[16] Studies generally find that 60–80 per cent of addresses can be assigned a geocode; however, few studies have examined the accuracy of these geocoded addresses.[17,18]

This study compares the performance of FEBRL and MapMarker in the geocoding of addresses using routinely collected administrative health data from the NSW Central Cancer Registry.

METHODS

The authors obtained ethics approval for use of NSW Central Cancer Registry data, including residential addresses, on 888 cases of childhood leukaemia in children aged up to 14 years and diagnosed in NSW between 1990 and 2002. Using both FEBRL (version 0.3 and G-NAF version May 2005, MapInfo 2005 postcode boundaries) and MapMarker (version 7.0, MapInfo 2003 postcode boundaries), we geocoded these cases and compared the match status for each case.

Both products assign a geocode 'match status' category that indicates the precision of the assigned geocode, summarized in Table 1. FEBRL has several non-geocoded categories for addresses that have multiple possible streets or localities. MapMarker provides a geocoded result for all addresses, at least to the postcode centroid. We developed a hierarchical protocol to identify the most accurately geocoded address between the two products, where an exact address match in FEBRL and MapMarker was considered the most accurate and a 'many' match in FEBRL and a postcode centroid in MapMarker the least accurate. The least accurately geocoded cases (addresses that were not allocated a geocode or were allocated only a locality centroid in FEBRL and only a postcode centroid in MapMarker) were then considered for clerical review.

Clerical review is an essential part of all automated geocoding procedures for verifying and improving the accuracy of geocoding. The options for clerical review are largely dictated by the resources available and the degree of spatial precision required. We used the following approach for clerical review:

1. The addresses of the most imprecise cases were checked for misspelling using the Internet site Whereis[19], a street address mapping website based on UBD digital street map data. Potentially misspelt addresses were amended and resubmitted to FEBRL for geocoding.
2. If FEBRL did not return an improved match status we used Whereis to get an approximate location for the address. This Whereis location was often a local name for a locality or wayfare, or an address not yet included in G-NAF. We then used GIS-Epi, a mapping tool developed by the NSW Department of Health, to try to find this locality or wayfare using StreetPro 8.5 (which displays street names below 1:20,000 scale).
   a. Where the local wayfare address coincided with a G-NAF address or vice versa (for example, a Les Darcy Drive Maitland address in the G-NAF coincides with an address on New England Highway Maitland), we assigned the G-NAF geocode.

TABLE 1

DEFINITIONS OF MATCHING STATUS IN FEBRL AND MAPMARKER GEOCODING SOFTWARE

FEBRL	
Exact Address	Matched to wayfare number, name, type and locality—unique latitude and longitude
Exact Street	Matched to wayfare name, type and locality—latitude and longitude based on the street centroid
Average Address	Multiple addresses found close enough in space to produce an average match (eg units in an apartment block)—unique latitude and longitude based on average address
Exact Locality	No matching on the street level—latitude and longitude of the locality (postcode) centroid
Many Addresses[a]	Multiple addresses found but not close enough in space to produce an average match—no geocode
Many Streets[a]	Match on street name but no other item and the streets appears in many localities—no geocode
Many Localities[a]	Match only on locality but there may be many localities with same name—no geocode
MapMarker[b]	
Exact street address match (S5)	Single close match—the record has been geocoded to an interpolated house number offset from the street
Street centroid match (S4)	Single close match—the record has been geocoded to the street centroid
Locality centroid (Z1)	No street address match—the record has been geocoded to the suburb or postcode centroid

a. This match status does not produce a geocoded result in FEBRL (beta version)

b. MapMarker (v7.0) further classifies S5 and S4 by subcodes: H (House number match), P (Street prefix match), N (Street name match), T (Street type match), S (Street suffix match), C (Town name match), Z (Postcode match)

b. If GIS-Epi found the wayfare, we identified the address, a nearby address, or wayfare centroid to estimate the geocode.
c. If GIS-Epi found the locality, we assigned the geocode to the estimated locality centroid.

RESULTS

The geocoding results for MapMarker and FEBRL are summarised in Table 2. FEBRL assigned a geocode for an 'exact' or 'average' address match to 719 (81 per cent) cases. Street centroid interpolation within a locality was given to 73 (8.2 per cent) cases as an exact street match. Where a locality (postcode or suburb) could be identified but no unique address matched, the locality's centroid was allocated to a further 58 (6.5 per cent) cases. The 38 (4.3 per cent) remaining cases that could not be matched by FEBRL were not assigned a geocode. MapMarker assigned a street level interpolated geocode to 766 (86.3 per cent) cases. Street centroids were allocated to 72 cases (8.1 per cent). The remaining 50 (5.6 per cent) cases were assigned the postcode centroid.

The major reasons for addresses not being geocoded in FEBRL or assigned only a postcode in MapMarker were due to errors in the original health record, such as incomplete addresses (eight records), post boxes (five records) or lot/property names (16 records).

The hierarchical protocol developed to identify the most accurately geocoded addresses between the two products is summarized in Table 3. This protocol can be used to select the most accurate geocode assigned between the two products depending on the positional accuracy required by a study. It also helps people to understand the level of spatial imprecision and the limitations this may have in any spatial analysis.

TABLE 2

GEOCODING MATCH STATUS FOR FEBRL AND MAPMARKER FOR 888 CASES OF CHILDHOOD LEUKAEMIA DIAGNOSED IN NEW SOUTH WALES FROM 1990 TO 2002

Geocoding method	MapMarker match status							
FEBRL match status	Interpolated Exact Address (S5)[a]		Street Centroid (S4)[a]		Postcode Centroid (Z1)[a]		Total	
	n	Row %	n	Row %	n	Row %	n	Row %
Exact Address	659	94.7	19	2.7	18	2.6	696	100.0
Average Address	23	100.0	0	0	0	0	23	100.0
Exact Street	41	56.2	29	39.7	3	4.1	73	100.0
Exact Locality	20	34.5	18	31.0	20	34.5	58	100.0
Many Addresses[b]	17	85.0	2	10.0	1	5.0	20	100.0
Many Streets[b]	1	16.7	3	50.0	2	33.3	6	100.0
Many Localities[b]	5	41.7	1	8.3	6	50.0	12	100.0
Total	**766**	**86.3**	**72**	**8.1**	**50**	**5.6**	**888**	**100.0**

a Codes used in MapMarker (v7.0)

b No geocode produced in FEBRL

TABLE 3

HIERARCHICAL APPROACH TO REFINING THE GEOCODING OUTPUT USING FEBRL AND MAPMARKER SOFTWARE

FEBRL output	Approach
Exact Address	Use FEBRL output.
Average Address	Use FEBRL output.
Exact Street Address	Use MapMarker result of Exact Address (S5); otherwise use FEBRL output.
Exact Locality	Use MapMarker output for cases with a MapMarker result of Exact Address (S5) or Street Centroid (S4). Depending on accuracy required, the remaining cases coded to postcode centroid (Z1) are either excluded depending on the required spatial resolution for the study or undergo clerical review.
Many Addresses, Many Streets or Many Localities	Use MapMarker output for cases with a MapMarker result of Exact Address (S5) or Street Centroid (S4). Depending on accuracy required, the remaining cases coded to postcode centroid (Z1) are either excluded, depending on the required spatial resolution for the study, or undergo clerical review.

The hierarchical protocol allocated a FEBRL geocode to 771 cases (87 per cent), a MapMarker exact address or street centroid geocode to a further 108 cases (12 per cent), leaving 29 cases (3 per cent) for clerical review. These 29 cases were considered the most imprecise, having either no FEBRL geocode but a MapMarker postcode centroid (9 cases) or only a postcode/locality centroid provided by both products (20 cases). These cases then underwent clerical review, resulting in a geocode being assigned to all 29.

DISCUSSION

Both products gave above average results for exact address matches (a match result of 70 per cent is often considered acceptable). [6]

For environmental epidemiological studies examining very small areas or point event analysis, where the address of individual cases of a disease can be modeled as the data unit rather than an area[20], the positional accuracy required is more likely to be met with FEBRL than MapMarker because of the use of property centroids in FEBRL compared to street interpolation in MapMarker.

There can be enough difference between the actual position and a street centreline interpolation for an address to be assigned to the wrong spatial unit, especially if the study uses small areas.[14] In a previous Australian study using a street centre-line product to compare property geocodes, an estimated 5-7.5 per cent of addresses were misclassified to another census collection district. [21] The size and shape of a study's spatial unit can influence the estimated disease rate (referred to as the Modifiable Area Unit Problem) and assignment of area level covariate/exposure data. [22–24] The use of property parcel centroids could reduce this misclassification. Verifying the actual positional accuracy of the geocode with ground proofing of the address was beyond the scope of this study.

Our evaluation of two geocoding products suggests both are acceptable for use with large data sets, providing geocoding cheaply and quickly, with each having trade-offs. FEBRL accurately assigned an exact address geocode to 81 per cent of subjects while MapMarker geocoded slightly more (86 per cent).

Our study was limited to 888 subjects with a particular diagnosis, but there is no reason to suspect that these results would not be applicable to other outcomes. Indeed, the authors recently geocoded over one million records for the Midwives Data Collection using a protocol similar to the one described above with similar results.

The G-NAF is the most complete, up-to-date and accurate coverage of Australian addresses available and is updated quarterly. When the FEBRL probabilistic mapping algorithm is complete, it will be one of the most sophisticated freely available geocoding software products in Australia.

ACKNOWLEDGEMENT

The authors acknowledge the support of the Australian Research Council Linkage Grant LP0348628, the North Coast Area Health Service, the NSW Department of Health and the Commonwealth Department of Health and Ageing.

REFERENCES

1. NSW Department of Health. The health of the people of NSW. In *Report of the Chief Health Officer, NSW Health. 2005*, NSW Health: Sydney. Available from: www.health.nsw.gov.au/public-health/chorep/.
2. Glover J, Harris KR, Tennant S. *A social health atlas of Australia*. 2nd edition. Adelaide: Openbook Publishers, 1999.
3. Christen P, Churches T, Hegland M. A parallel open source data linkage system. In: *Proceedings of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining, May 2004*. Sydney: Springer Lecture Notes on Artificial Intelligence (3056), 2004. At: www.springer.com.
4. Paull DL. *A geocoded national address file for Australia: The G-NAF what, why, who and when?* 2003, Griffith ACT Australia: PSMA Australia. At: www.psma.com.au/about-g-naf, accessed 20 April, 2006.
5. MapInfo Corporation. MapMarker® Australia—Overview. (homepage on the Internet) 2005, MapInfo Corporation. At: http://extranet.mapinfo.com/products/Overview.cfm?productid=152&productcategoryid=1,2,3, accessed 20 April 2006.
6. Christen P, Churches T, Willmore A. *A probabilistic geocoding system based on a National Address File.* Cairns: Australasian Data Mining Conference, 2004. At: http://datamining.anu.edu.au, accessed 20 April 2006.
7. Rural Retention Program – GPARIA Category Maps. At: www.health.gov.au/internet/wcms/publishing.nsf/Content/Rural+General+Practice+Programs-1.2004, accessed 20 April 2006.
8. *Pharmacy Access/Rural Retention Program- Remote Index of Australia (PhARIA)*. At: www.gisca.adelaide.edu.au/projects/pharia.html, accessed 20 April 2006.
9. Geschwind SA, Stolwijk JA, Bracken M, Fitzgerald E, Stark A, Olsen C, et al. Risk of congenital malformations associated with proximity to hazardous waste sites. *Am J Epidemiol* 1992; 135(11): 1197-1207.
10. English P, Neutra R, Scalf R, Sullivan M, Waller M, Zhu L.Examining associations between childhood asthma and traffic flow using a geographic information system. *Environ Health Perspect* 1999; 107(9): 761–7.
11. Washburn E, Orza MJ, Berlin JA, Nicholson WJ, Todd AC, Frumkin H, et al. Residential proximity to electricity transmission and distribution equipment and risk of childhood leukemia, childhood lymphoma, and childhood nervous system tumors: systematic review, evaluation, and meta-analysis. *Cancer Causes & Control*, 1994; 5(4): 299–309.
12. Northern Rivers University Department of Rural Health (homepage on the Internet), North Coast Area Health Service: Lismore Australia. (Updated 2005) At: www.nrudrh.edu.au/publish/research/environmentalresearch.php, accessed 20 April 2006.

13. Street Address Working Group. *Issues affecting construction of a geocoded address file*. Intergovernmental Committee on Surveying and Mapping. 1999. At: http://icsm.gov.au/icsm/street/geocode_construction.html, accessed 20 April 2006.
14. Cayo MR, Talbot TO. Positional error in automated geocoding of residential addresses. *Int J Health Geogr* 2003. 2(1): 10.
15. Hurley SE, Saunders TM, Nivas R, Hertz A, Reynolds P. Post Office Box Addresses: A challenge for Geographic Information Systems-based studies. *Epidemiology* 2003; 14(4): 386–91.
16. Skelly C. Disease surveillance in rural communities is compromised by address geocoding uncertainty: a case study of campylobacteriosis. *Aust J Rural Health* 2002. 10: 87–93.
17. 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(7): 1114-6.
18. Bonner MR, Han D. Positional accuracy of geocoded addresses in epidemiologic research. *Epidemiology* 2003; 14(4): 408–12.
19. Whereis® OnLine. (homepage on the Internet) 2005, Sensis Pty Ltd: Australia. At: www.whereis.com/, accessed 20 April 2006.
20. Elliott P, Martuzzi M, Shaddick G. Spatial statistical methods in environmental epidemiology: a critique (Review). *Stat Methods Med Res* 1995; 4(2): 137–59.
21. Ratcliffe JH. On the accuracy of TIGER-type geocoded address data in relation to cadastral and census units. *Int J Geogr Inf Sci 2001*: 15(5): 473–85.
22. Krieger N, Waterman P, Chen JT, Soobaderm, MJ et al. Zip code caveat: Bias due to spatio-temporal mismatches between zip codes and US census-defined geographica areas—the public health disparities geocoding project. *Am J Pub Hlth* 2002; 92(7): 1100–02.
23. Lim ST. Controlling for a proximate determinant of breast cancer - a creative use of geographic information systems (GIS) for analyzing data with sparse background information. In *Proceedings of National Conference on Health Statistics, 1999*. Washington DC, 1999.
24. Hyndman JC, Holman CD, Hockey RL, Donovan RJ, Corti B, Rivera J. Misclassification of social disadvantage based on geographical areas: comparison of postcode and collector's district analyses. *Int J Epidemiol* 1995; 24(1): 165–76.