Geocoding the NSW Central Cancer Registry database

Implications for cancer control in Sydney, Australia

September 2007

Tracey, E[1], Morgan G[2], Bishop J[1]
[1]Cancer Institute NSW
[2]Northern Rivers University Department of Rural Health - University of Sydney, and North West Area Health Service

Poster 151

BACKGROUND

The NSW Central Cancer Registry records the residential address of all persons with cancer at the time of diagnosis. On the basis of this address, a National Locality Index code (NLI) is applied that determines the Local Government Area of residence at the time of diagnosis of all cases and subsequently, the Area Health Service of Residence.

The NSWCCR is an historical database of all cases of cancer diagnosed in NSW from 1972 to 2004. Currently the entire database is coded to 1991 Local Government boundaries. To update to 2001 boundaries, all addresses on the CCR would need to be remapped to 2001 boundaries so that numerator and population denominator information is compatible.

In addition, the Australian Bureau of Statistics will no longer provide new versions of the NLI coder for population boundaries and is moving to geocoding population information. Geocoding also increases the opportunities for epidemiological analysis that can improve cancer control for NSW.

AIM

1. To allow the CCR database to be easily updated to 2001 Local Government Area boundaries and to new boundaries in the future.
2. To enable cancer rates standardised for age to be determined for smaller more precise geographical areas.
3. To enable socioeconomic status, ARIA and other Census characteristics to be used to explain patterns in cancer incidence, mortality, screening and survival geographically.
4. To enable distance between two points or a point and line, eg., facility and cancer address coordinate to be calculated to provide a better understanding of the influence of distance on outcome.
5. To better understand grouping or clustering of cancer cases by space and time.

METHODS

All addresses of NSW cases recorded on the NSW Central Cancer Registry database diagnosed between 1972 and 2004 were geocoded using two different geocoding methods. These were FEBRL (Freely Extensible Biomedical Record Linkage) and MapMarker. Both methods provide mechanisms that can clean and georeference individual records.

The FEBRL geocoding software is currently under development by NSW Health and employs a probabilistic procedure for cleaning and standardising input records and uses the extensive Geocoded National Address File (G-NAF).

The G-NAF data is the first geocoded national address file that contains discrete address locations derived from a variety of national and state based datasets. MapMarker is a geocoding product from MapInfo that uses a comprehensive database of street centerlines and town and postcode centroids developed from MapInfo's StreetWorks Australia database.

The allocation of a geocode is governed by a set of weights that scores each portion of the address against an Address Dictionary. As a result both methods provide, a longitude and latitude coordinate for every case and a 2001 census boundary code.

The quality of geocoding was checked in two ways. The first was to compare existing boundaries allocated using the National Locality coder for cases diagnosed between 1999 and 2003 with the 2001 boundaries allocated using the longitude and latitude coordinates for a case.

The second was to calculate distance using the Great Distance Circle Calculator a SAS algorithm obtained from the North American Association of Cancer Registries website that calculates the distance in kilometres between any two points.

RESULTS

A longitude and latitude coordinate was obtained for every case, as well as a new 2001 census boundary code. There were, however, approximately 6,000 addresses that did not have a geocode determined and required manual resolution.

The quality of address information on the Registry and levels of confidence associated with different geocoded coordinates is presented in Table 1. Regardless of the methods used to geocode, of the 809,551 cases geocoded:

- 82.7% were geocoded using the address
- 7.8% geocoded to a street
- 8.5% to a locality
- 0.87% not enough information to geocode (Table 1).

Method 2 allocated an address 85% of the time with a further 8% determined using the centroid of the street. Both methods were compared and a consolidated geocode agreed to based on a protocol.

Comparison of the number of cases found in existing NSW 1991 Local Government Area boundaries allocated using the National Locality Coder and new 2001 LGAs based on the geocoding protocol indicates that all cases mapped correctly.

The only case addresses that did not map were those that were designated as indeterminate by both geocoding methods. Investigation of the case addressess for indeterminate geocodes revealed characteristics for the 1144 cases diagnosed between 1999 and 2003 (approx 1%). These were, addresses that were PO boxes, caravan parks, nursing homes and hostels. These addresses will need to be manually coded on the basis of the closest postcode and locality.

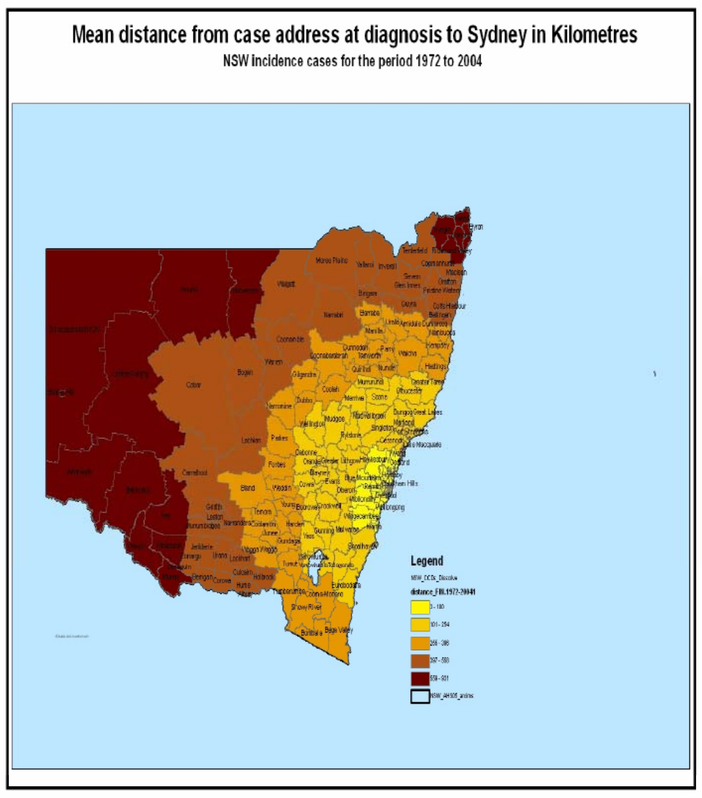
Table 1 NSW Cancer cases 1972 - 2004 Geocoded using two methods

		MapMarker Match Status					
FEBRL Match Status		Exact Address	Exact Street	Locality	Indeterminate geocode	No geocode	Total
	N	411061	83828	54566	36500	196	586151
exact address	%	**50.78**	10.36	6.74	4.51	0.02	**72.41**
average	N	16217	3974	2082	1484	13	23770
address	%	2	0.49	0.26	0.18	0	2.94
	N	19829	31544	13430	10314	79	75196
exact street	%	2.45	3.9	1.66	1.27	0.01	9.29
	N	287	418	140	128	2	975
average street	%	0.04	0.05	0.02	0.02	0	0.12
	N	14341	5531	36144	17607	730	74353
exact locality	%	1.77	0.68	4.47	2.18	0.09	9.19
	N	37	20	22	0	1	80
average locality	%	0	0	0	0	0	0.01
	N	12435	6214	4903	2447	26	26025
many address	%	1.54	0.77	0.61	0.3	0	3.21
	N	526	1948	1726	831	24	5055
many street	%	0.06	0.24	0.21	0.1	0	0.62
	N	4355	1220	7014	2924	87	15600
many locality	%	0.54	0.15	0.87	0.36	0.01	1.93
	N	126	113	1363	560	125	2287
no match	%	0.02	0.01	0.17	0.07	0.02	0.28
	N	479214	134810	121390	72795	1283	809492
Total	%	59.2	16.65	15	8.99	0.16	100

The validity of each geo coordinate allocated to a case was also verified by checking the mean distance of cases within a Local Government Area to the centre of Sydney.

The Great Circle Distance Calculator allocated a distance in kilometres, from the cancer case to the centre of Sydney that was then compared with the known distance for that LGA. As can be seen by the map the mean distance of cases for each LGA from the centre of Sydney is internally consistent (Figure 1).

Figure 1 Mean distance of all cancer cases from the centre of Sydney



DISCUSSION

Geocoding the NSW Central Cancer Registry database has allowed case addresses to be updated to Local Government Boundaries for 2001. In addition, successful comparison of new boundaries to existing 1991 boundaries validates that geocoding of case addresses can be implemented routinely by other Registries. In the future any health or other administrative boundary changes can be implemented with relative ease in NSW.

Geocoded addresses of cancer cases can be spatially joined to population data using ESRI ARCGIS mapping software. The Australian Bureau of statistics are currently finalising small areas of population referred to as mesh blocks. Cases of cancer can then be associated with smaller population grouping than is currently available. Populations will be provided by the ABS in a geocoded format when the 2006 census population is finalised.

For the first time cancer and population data will both be in a geocoded format. Geocoding will improve over time as updating of the GNAF takes place. Urban areas are currently more precisely coded than Rural Area.

Geocoding will also facilitate the wider use of Census data like socioeconomic status and indices of remoteness, as well as enabling a better understanding of spatial patterns of cancers in the population of NSW. Calculating the mean distance for each cancer case to Sydney using the Great Distance Circle Calculator program and then comparing the known distance for each LGA has provided more certainty that the coordinates of latitude and longitude have been correctly allocated.

The ability to calculate and determine a perimeter of cases beyond a known or suspected exposure will lead to a better understanding of cancer patterns. The impact of distance and access to cancer services in Sydney Australia can also now be considered.

CONCLUSION

Geocoding the Central Cancer Registry database ensures that boundary changes in the future in NSW can take place with relative ease. Population and other demographic data can be easily linked to cancer registry data. Other cancer registries can look at implementing geocoding and validating the geocoded address using the methods used in NSW.

A better understanding of access issues can be obtained by considering the distance between cases and treatment facilities.

ACKNOWLEDGEMENTS

Geocoding of the Central Cancer Registry database for the time period 1972 to 2004 was conducted by the Northern Rivers University Department of Rural Health - University of Sydney, and North Coast Area Heath Service as part of a consultancy. The initial research into geocoding methods by the Northern Rivers University Department of Rural Health - University of Sydney, was funded by the Australian Research Council (Grant LP0348628), NSW Health, The Commonwealth Department of Health and Ageing and the North coast Area Health Service.