A Spatial Analysis of Psychiatric Patient Record Data

Paul Lewis, Mary O'Brien, Prof. A. Stewart Fotheringham, Martin Charlton,
National Centre for Geocomputation, National University of Ireland Maynooth, Maynooth, Co. Kildare, Ireland.
Paul.lewis@nuim.ie, Mary.obrien@nuim.ie, Stewart.fotheringham@nuim.ie. Mary.obrien@nuim.ie

KEYWORDS: address-matching, geocoding, health patient data

ABSTRACT:

This paper reports on a project which explored whether it was possible to spatially reference data from the National Psychiatric Inpatient Reporting System (NPIRS) database from the Mental Health Division of the Health Research Board in Ireland. NPIRS provides information on patient health (in particular admitting diagnosis) and a spatial reference in the form of address data with address line one omitted for confidentiality reasons. Currently the spatial analysis of this data is limited to county units (34 units in total in Ireland, NUTS 4) providing no useful indication of psychiatric service demand, and so the aim was to increase the granularity and georeference patients to an Electoral Division, the current standard demographic unit in Ireland (totalling 3440 units, NUTS 5). This is not a simple matter given the absence of postcodes and unique addressing in Ireland, and involved the development of a procedure to address-match patient addresses (approx. 20,000 records in 2004) to an Electoral Division through the use of the national address database, GeoDirectory. We report on the address-matching procedure developed and its success - variable over space, the problems encountered with regard to the issue of non-unique addresses and how these were dealt with, and we evaluate the input datasets and the procedure used, providing recommendations to improve the results for future spatial analysis with this dataset.

1. INTRODUCTION

In the exploration of health data it is often useful to include a spatial view of the data – this can be useful *inter alia* in identifying areas of elevated risk, exploring associations between disease incidence and socio-environmental factors, planning efficient and equitable service provision. A requirement of spatial analysis is the use of spatially referenced data, and generally involves integration of a number of spatial datasets, to a common reference system. Typically a post-code or similar spatial reference is used, however in Ireland no such system exists. Thus we wish to develop a procedure to spatially reference a patient dataset (NPIRS) to enable the integration of other useful demographic data, improving the spatial analysis potential of this dataset.

1.1 The Problem

The Health Research Board (HRB) collects patient data containing information on patient health and a spatial reference in the form of address data, with address line 1 removed. Their current capabilities in the spatial analysis of this data are limited to county and national scales, and we wish to add a more detailed spatial reference, ideally one common with other spatial datasets of interest, such as the Census of Population. Hence we wish to reference this dataset to the spatial unit of Electoral Division (ED).

However this is not a simple matter given the characteristics of address geography in Ireland and the dataset available. Importantly
(a) No post-code geography: Ireland is one of the few remaining developed economies without a post-code geography.
(b) Non-unique addresses: This problem is particular to rural areas, where up to 60% of addresses in a county cannot be uniquely identified without local knowledge of family residences.
(c) Omission of address line 1 from the data – This was removed for confidentiality reasons by the Health Research Board, and was expected to be an issue for address-matching in urban areas.

The question therefore is whether, given the data, such geo-referencing is possible? And if so, then what methods would be appropriate to allow it to be undertaken?

2. DATASETS

In the absence of postcodes or similar, the first step involved identifying available datasets required to undertake this address-matching exercise. We needed the patient addresses (to be matched), a national address database (to match against) and a geography (to match to).

2.1 National Psychiatric In-Patient Recording System (NPIRS)

This is an annual dataset of patients admitted to psychiatric hospitals in Ireland, maintained by the Health Research Board (HRB) Mental Health Research Unit. Data were provided for 2000 to 2004, although only data for 2004 is reported here. Variables of interest to this study were *Address Lines 2, 3, 4,* and *County*, which provided location information. *Address Line 1* was omitted for confidentiality reasons.

On exploration of the above data, it was found that data quality was generally poor with

- many null entries.
- non-standard data entry, e.g. varying inclusion of commas, full-stops, abbreviations.
- incomplete address records, e.g. only county recorded.
- inconsistent address data, e.g. line 3 data appearing in line 2, etc.

- no unique patient identifier, and thus nothing to cross-check errors.

Some 87% of data was deemed valid input to address-matching procedure, however less than 40% of this provided full complete address lines.

2.2 GeoDirectory

GeoDirectory is a national address database maintained by An Post, the national mail delivery service. It provides location information in the form of address line data and grid coordinates for each mailing address/letterbox in Ireland, i.e. in excess of 1.7 million records. Within its structure it contains information on a number of geographies at differing scales thus meaning, for this study, that each geography can be keyed with an ED ID number.

GeoDirectory Field	**Geography**
Building	House
Thoroughfare	Street
Locality/ Townland/ Posttown	Area
County	County

Table 1. GeoDirectory geography

As address line one was not provided in the patient data, it was decided to build a gazetteer describing all the unique spatial reference relationships that are available in GeoDirectory and to use this as the search space, thus reducing the search space size to 117,000 entries. Three levels of spatial relationships were identified: Townland (detailed administrative division, 47000 in total), Locality (GeoDirectory defined areas, 55000 in total) and Posttown (postal delivery defined areas, 126 in total), all of which could be linked to an ED.

2.3 Electoral Division (ED) Boundaries

Electoral Divisions are administrative boundaries totaling 3440 polygons and maintained by Ordnance Survey Ireland (OSI), the national mapping agency. They are generally thought of as the smallest spatial unit for demographic analysis in the Republic of Ireland given census data is released at this level, and for this reason they were chosen as the spatial reference unit for the address-matching procedure.

3. Address-Matching Procedure

Ultimately, the purpose of the developed procedure is to geo-code each input address to an OSI ED. However, due to the lack of a postcode system for Irish addresses the geo-referencing requirements were met by developing an address matching program that relied upon textual searching and matching algorithms.

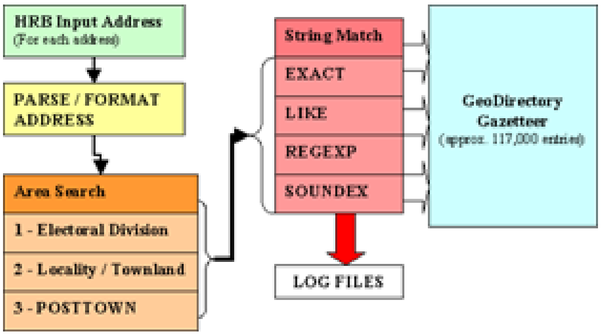

Figure 2. Diagram of the Address-Matching procedure

The program written to geo-code the NPIRS addresses comprises a linear set of operational stages, returning a subset of GeoDirectory records that could possibly match the input address. The stages are:

(1) Firstly issues with basic address structure required a parse and format procedure to be developed. Some of the solutions included setting all entries to upper case, changing Town entries to Urban, removing County and Co references, splitting comma delimited entries, to highlight the more significant ones.

(2) Next, three geographic levels of search were performed. ED, Townland/Locality or Posttown. If an HRB input address was of sufficient spatial detail it could be immediate matched to an ED or, if an immediate match was not possible a Locality, Townland or Posttown search could also determine the appropriate ED. This searching procedure was controlled by county subset searches which also improved operational efficiency.

(3) For each of these geographic searches, four specialised SQL statements were employed to retrieve possible result (address-matched) datasets.

1. EXACT: return a data set where the HRB input address is an exact match for a GeoDirectory address.
   Sample: ‘BRAY’ matches ’BRAY’
2. LIKE: return a data set where the HRB input address is an exact match or subset match with a GeoDirectory address.
   Sample: ‘*HILL*’ matches ’SHILLELAGH’
3. REGEXP: return a data set where the HRB input address characters appear in any order in any GeoDirectory address.
   Sample: ‘HILL’ matches ’KILLEAGH’
4. SOUNDEX: return a data set where the HRB input address has the same phonetic value as any GeoDirectory address.
   Sample: ‘MALLOW’ matches ’ALLOW’

(4) Finally, string comparison procedures were used to determine a correct NPIRS–GeoDirectory match. These procedures generated a numerical similarity score for each GeoDirectory entry that has a possible match to the input address. Based on different evaluation tests a set of thresholds were also defined that acted as validation criteria. Therefore, an HRB address could be successfully matched to a GeoDirectory geo-reference by choosing the highest match score from the resultant data set that also exceeded the extra

threshold criteria. For addresses that did not satisfy these conditions a number of separate log files were created that fully described all matching, scoring and threshold information causing the mismatch. Based on evolutionary program development these mismatch log files helped improve upon previous address matching iterations. They recorded the various steps and decisions taken during the address matching procedure, the types of information recorded include result data sets, address matching scores, matched addresses, unmatched addresses and irrelevant addresses, to name but a few. Two particularly important result files included the matched addresses processing file and the match statistics file. The later contained some statistics calculated on the success rates for the address matching procedure while the former contained a copy of the original HRB address and the matched GeoDirectory equivalent.

4. Results of Address-Matching Procedure

Of the 22,400 HRB input addresses, 15,833 were successfully geo-referenced. This represents a 70.6% match rate. However, given 2838 of the unmatched input referenced just a county, while another 79 had non-Irish data, a relevant match percentage of 81% can be assumed.

Initial analysis of the address matching results concentrated on the production of match success rates per county, which seem to vary significantly.

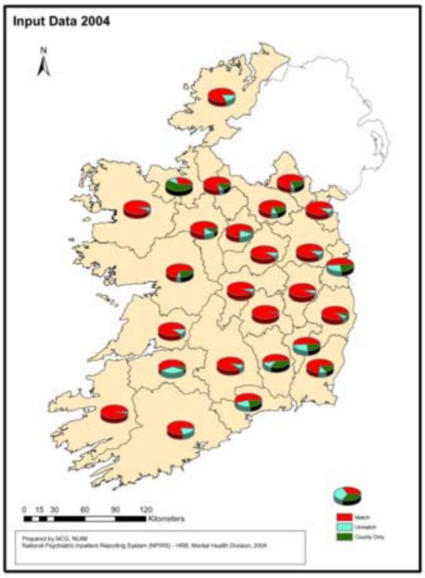
Input Data 2004
N
0 15 30 60 90 120
Kilometers
Match
Unmatch
County Only
Prepared by NCG, NUIM
National Psychiatric Inpatient Reporting System (NPIRS) - HRB, Mental Health Division, 2004

Figure 3. Match success rates per county

From initial exploration of the non-match and multiple-match log files, this variation would appear to be for a number of reasons:

(1) Poor input data:

Not all hospitals provided ‘good’ address data – input data for Sligo (57%), Kilkenny, Dublin, Waterford, Galway all having in excess of 20% of records having blank entries for address lines 2 and 3.

(2) Non-unique addresses:

(i) There are many addresses which can only be distinguished by local knowledge of family residences. Analysis of GeoDirectory showed this to be particularly a rural issue with one third of counties having in excess of 50% addresses being non-unique, the worst cases being Roscommon (63%) and Leitrim (62%).

(ii) Urban EDs can often not be uniquely identified by an address line, for example addresses in EDs named ‘Rathmines East A’, ‘Rathmines East B’, ‘Rathmines West A’, ‘Rathmines West B’, etc. would all simply have the address line ‘Rathmines’ in the dataset. In these cases, where EDs were adjacent, it was decided that boundaries would be dissolved.

(iii) In other cases, an address line may legitimately be found in a number of differing EDs which are not adjacent or even near to each other, see example below. In this case no match was assumed, however it is hoped to explore these types of multiple-matches further and perhaps incorporate some measure of certainty or probability to the address-match results.

Input	Output	
Address Line	*Locality*	*ED*
Castlepollard	Castlepollard	Coolure
	Castlepollard	Milltown
	Castlepollard	Kinnegad
	Castlepollard	Kinturk

Table 4. Example multiple-match

(3) Phonetically similar placenames:

Occasionally similar sounding (see reference to SOUNDEX use in address-matching) names but in different EDs could not be differentiated, e.g. ‘Mallow’ and ‘Allow’, ‘Shannon’ and ‘Mountshannon’, ‘Blackpool’ (ED 17011 in Cork) and ‘Blackpool’ (ED 18079 in Cork). In these cases the counts of patients in each ED were reassigned based on the population.

5. Recommendations

More analysis is needed in order to be able to understand the limitations of the address-matching procedure developed here, and to improve its results. A number of items can however be identified which would aid in this.

(1) Match rates could be improved if the street/thoroughfare element if address line one is included in the patient dataset.

(2) The systems used to record patient information could be improved to record addresses more consistently and include postcodes when they have been introduced.

(3) Further analysis of the multiple-match and no-match log files would provide more insight into the address-matching procedure and help evaluate it. Also the incorporation of these data by assigning probabilities would be useful.

6. Further work

We plan to further analyse the result log files as suggested above, and to report on this. Also a more recent sample of the NPIRS dataset with improved address line information is being examined. This data has been address-matched using both the method described here (with 82% success rate) and using an alternative procedure (with 77% success rate), and we hope to compare the results from both and report on this to provide some evaluation of the address-matching procedure developed here.