HAG-GIS: A spatial framework for geocoding historical addresses

Daras K[*1], Feng Z[†1] and Dibben C[‡2]

[1] School of Geography & Geosciences, University of St Andrews

[2] School of Geosciences, University of Edinburgh

November 7, 2014

Summary

The Digitising Scotland (DS) project aims to digitise the 24 million vital events record images (births, marriages and deaths) for all residents in Scotland since 1855 (ie transcribe them into machine encoded text). This will allow research access to information on individuals and their families for those who have ever lived in Scotland between 1855 to the present day. In this paper we present the methodology for geocoding these 24 million historical addresses in Scotland from 1855 to 1974 by introducing the Historical Address Geocoder – GIS (HAG-GIS) spatial framework and its matching algorithms implemented for the needs of the DS project. The matching processes link the historical addresses to the contemporary addresses by exact and fuzzy matching algorithms. Apart from geocoding the historical addresses, we also produce pseudo registration district boundaries using the pilot historical addresses from death event records in 1950 and 1951.

KEYWORDS: Digitising Scotland, HAG-GIS, geocoding, address matching

1. Introduction

In Scotland, the Registration Districts (RDs) were used from 1855 as result of the Registration of Births, Deaths and Marriages (Scotland) Act 1854 and registrars were appointed in each district to maintain the statutory registers (births, marriages and deaths). These records are very detailed and have changed very little in content over time, making them a highly valuable record of change over time and space. Apart from the great importance for research, the detailed nature of the record allows records for the same person to be linked and then link persons into families. Originally the Digital Imaging of the Genealogical Records of Scotland's people (DIGROS) project converted all statutory

[*] kd54@st-andrews.ac.uk
[†] zf2@st-andrews.ac.uk
[‡] Chris.Dibben@ed.ac.uk

vital events records into digital image format (PDF) with a supporting index, where only forenames, surname, year and RD code were indexed. The purpose of DS project is to build on the DIGROS index and extend it through digitisation for research purposes. It has four main objectives: 1) digitise the 24 million vital events record images (births, marriages and deaths) for all residents in Scotland since 1855 (ie transcribe them into machine encoded text), 2) standardise and code occupation descriptions to the Historical International Standard Classification of Occupations (HISCO), 3) standardise and code all deaths to a modified form of the International Classification of Disease 10 (ICD10) and 4) link address information to consistent geographies through time. This will allow research access to information on individuals and their families for those who have ever lived in Scotland between 1855 to the present day. The combination of temporal and spatial historical information will provide a rich demographic database system for the Scottish population of a similar potential depth and breadth as the Scandinavian and Low countries (Edvinsson, 2000; Mandemakers, 2000; Thorvaldsen, 2000).

In this paper we present the methodology for geocoding 24 million historical addresses in Scotland from 1855 to 1974 by introducing the HAG-GIS system and its matching algorithms implemented for the needs of the DS project. The overall matching process links the historical addresses to the contemporary addresses by exact and fuzzy matching algorithms. Apart from geocoding the historical addresses, we produce pseudo registration district boundaries using the pilot historical addresses from death event records in 1950 and 1951. This way we create a historical geographical dataset where new insights into the Scottish geographies of the past are available. A large number of projects (Fitch and Ruggles, 2003; Gregory and Healey, 2007; St-Hilaire et al., 2007) have used similar approaches to assist in developing historical GIS systems where the spatial boundaries are unknown but textual information is available in various public records.

2. HAG-GIS: System framework

The HAG-GIS application has been designed to be used with historical data for Scotland, targeting special data characteristics and providing required tasks of data analysis and normalisation. Its implementation is based on the Python language using the SQLite database for storing the data providing portability to any operational system platform. The geocoding process introduced here is related to efficiently organising historical addresses by eliminating redundancy and ensuring data dependencies. The geocoding process has two distinct phases: a) the automated phase where the HAG-GIS system preforms an exact and a fuzzy matching to link each record to a particular geographical reference point, and b) the manual phase where the clerks check/edit the unmatched addresses using the historical maps from Ordnance Survey (Figure 1). Importantly in the first phase, HAG-GIS exploits the two types of geographical information retained in the records to assign a geocode to the record: firstly the geographically high resolution but ‘dirty’ address information and

secondly the less high resolution but much 'cleaner' Registration District information. The programme cycles between these two types of information to fully geocode the dataset.

During the automated phase, the HAG-GIS system executes the following tasks using a SQLite database to store the data:

a) the cleaning task where the system transforms all the address data into a normal form by lowercasing all the characters of address, stripping possible whitespaces, removing punctuations and removing duplicate addresses from the database,

b) the tokenisation task where each address is split into word tokens and each token could be marked as possible unchanged token or linked to name aliases stored in the database,

c) the address matching task where in order to match historical addresses to contemporary addresses we use exact and fuzzy matching algorithms such as the Measuring Agreement on Set-Valued Items (Passonneau, 2006) and Levenshtein edit-distance respectively,

d) the geocoding task where the system assigns the appropriate reference point for each matched address and uses these reference points to build artificial polygons (Thiessen polygons). In addition we use the K-Dimensional Tree algorithm for nearest neighbours (Maneewongvatana & Mount, 2001) for creating series of point density surfaces for each RD code. Thus, the system can exclude reference points that are assigned to wrong RD codes or distant address locations.

e) Finally the Thiessen polygons are assigned to the pseudo-RDs allowing us to enable a new matching phase through a more relaxed matching criterion and to direct a historical map linked clerical phase.

Even though the manual phase is not yet implemented, its basic framework components are presented for a complete introduction to the HAG-GIS system framework (figure 1). Thus, the pseudo-RDs will be used by the clerks to manually identify and geocode addresses to grid references in combination with the historical maps. The clerks will use the Ordnance Survey six-inch to the mile historical maps in Scotland for the period 1892-1905 for manually edit unmatched historical addresses. If the manual allocation of some addresses to the appropriate grid references is not feasible then those addresses are assigned to the center of the pseudo-RD. The quality of the geocoding process and the pseudo-RDs is secured by calculating the density of matched addresses for each pseudo-RD (index = matched addresses/all addresses). This way, the manual process is concentrated on the pseudo-RDs with low index value.

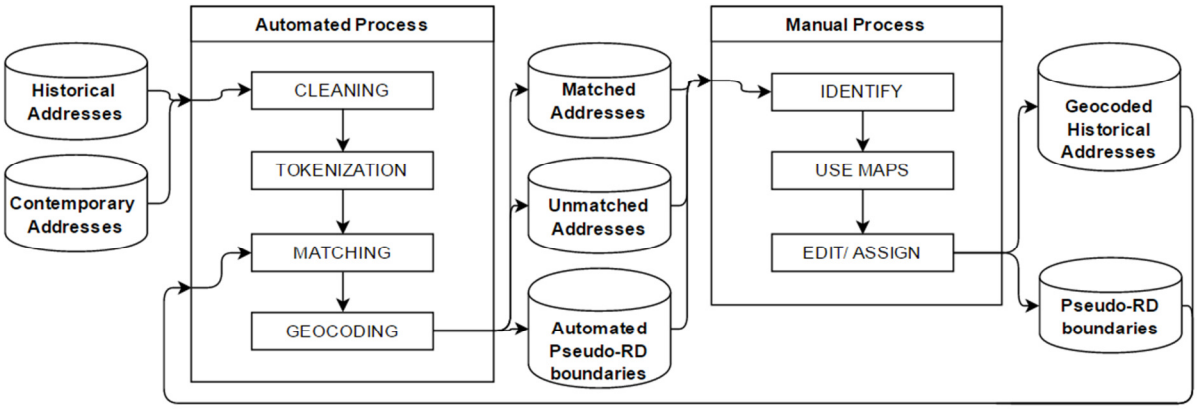

Figure 1 HAG-GIS system framework

3. Pilot study

In this pilot study we utilise 129,694 historical addresses derived from a sample digitisation of the death records for the 1950 and 1951 years in Scotland. Each historical address refers to an individual and provides information about the building number or name of property, the street name, the town name and the registration district (RD) code. In addition, the catalogue of contemporary addresses are provided by the NRS address register, which includes the building number or name of property, the street name, the town name, postcodes and national grid references.

Using the aforementioned historical and contemporary addresses and the HAG-GIS geocoding system, we performed an initial automated address matching process using a 80% match cut-off point in fuzzy matching so that all addresses whose matching scores are above a certain membership will be determined as a match. The initial address matching process was completed in under an hour using a single CPU processor. The HAG-GIS system was able to match 36,412 (42%) of the 87,495 cleaned historical addresses with 25,761 (29%) addresses being exact matched. In Table 1, we can see the initial matched addresses by matching thresholds.

Table 1 The initial matched addresses in the pilot study by matching threshold

Num. of addresses	Match threshold			Un-matched addresses
	100%	≥90%	≥80%	
87495	25761 (29%)	33288 (38%)	36412 (42%)	51083 (58%)

After the end of initial matching process, the system created temporary Thiessen polygons using the 80% matched addresses assigning a weight for each polygon (Figure 2a & 2b). The weights are computed using the nearest neighbours' method (k-dimensional tree) for each RD number. We selected only the 80% matched addresses because the system can overcome the problem of partially

matched addresses with incorrect geo-reference and can provide a clean artificial tessellation for constructing the initial set of pseudo-RD boundaries. Having all the Thiessen polygons weighted for each RD, the system automatically assigned the RD number with the highest weight to the relevant Thiessen polygon. Then, the HAG-GIS system repeated the matching process for each unmatched address focusing on the contemporary addresses that fall within the pseudo-RD's bounding box of the unmatched address with the same RD code. This time we used a less strict threshold (50%) for configuring the system because the safe estimation of the initial pseudo-RD boundaries supported us with the geographical extent of each pseudo-RD and we could narrow down the search results considerably. In later phases of the work we will test the quality of the results by different thresholds in order to establish an optimal value. The HAG-GIS system was able to deliver a final match of 62,602 (72%) of the 87,495 cleaned historical addresses with 24,893 (28%) addresses being unmatched. In Table 2, we can see the final matched addresses by matching thresholds. The final pseudo-RDs have been improved extensively and they are imitating closely the Bartholomew Post Office Plan (1939-40) boundaries (figure 2c & 2d).

Table 2 The final matched addresses in the pilot study by matching threshold

Total addresses		Match threshold						Un-matched addresses
		100%	≥90%	≥80%	≥70%	≥60%	≥50%	
Counts	87495	26067	33865	37671	41364	49869	62602	24893
%	100	30	39	43	47	57	72	28

4. Conclusions

This paper has presented the HAG-GIS spatial framework for geocoding historical addresses in Scotland using the pilot data for two distinct years (1950 and 1951). The preliminary results suggest that the overall geocoding methodology performs well and the matching process is expected to further improve the final results when the pilot data will be replaced with the full data. However, there are possible improvements that can be facilitated to address issues related to the size of urban and rural pseudo-RDs, the precision of geocoder for long or short streets and better handling of various transcribing errors. Further work will involve the effects to the geocoding precision using other matching algorithms such as phonetic match algorithm etc. and the optimisation of essential phases during the automated and manual processes.

5. Acknowledgements

The Digitising Scotland project is funded by ESRC. The support from National Records of Scotland is also gratefully acknowledged.

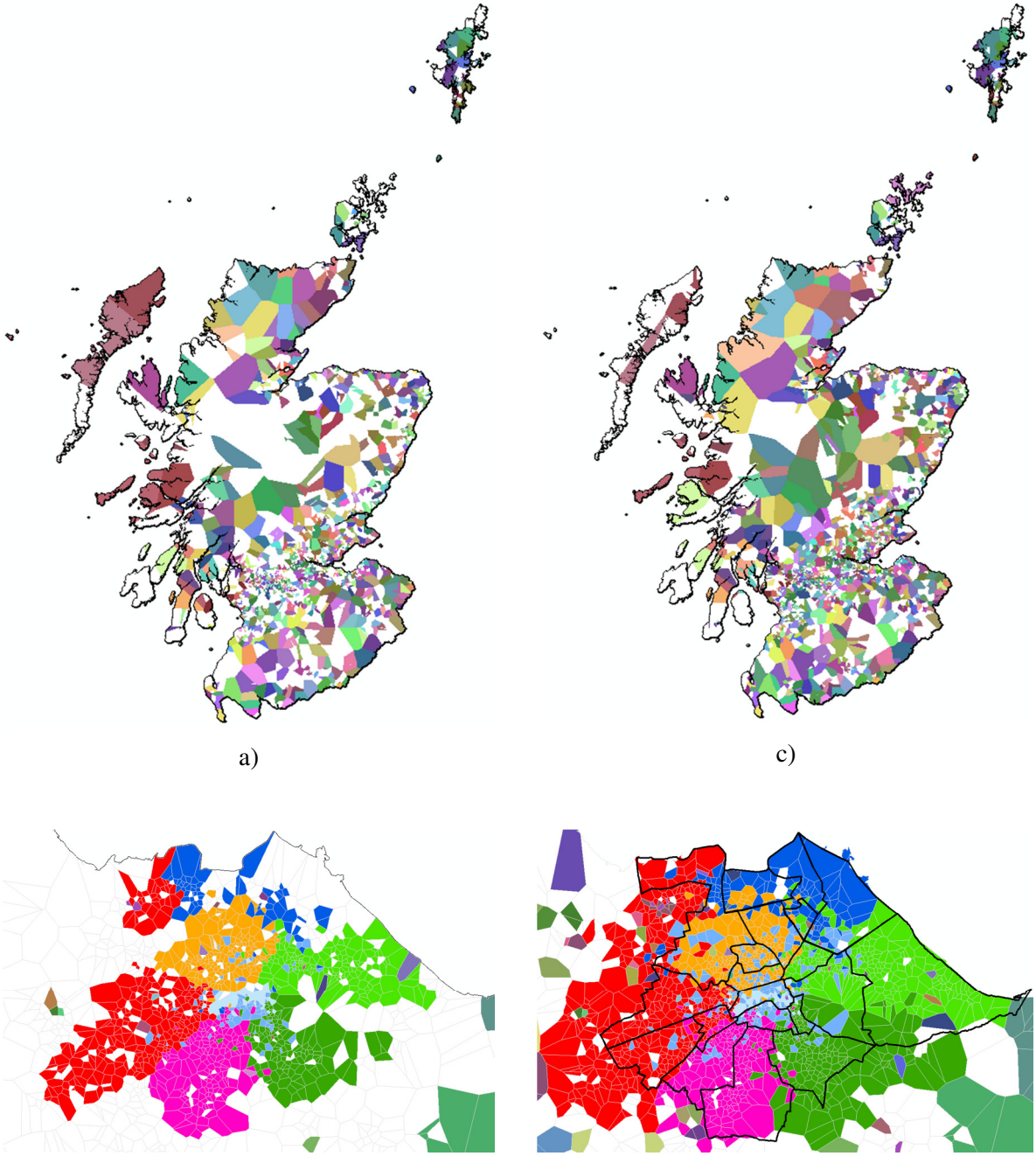

Figure 2 Thiessen polygons created from the HAGGIS system: a) & b) pseudo-RDs of Scotland and Edinburgh at the initial phase, c) pseudo-RDs of Scotland at the final phase & d) Bartholomew Post Office Plan (1939-40) boundaries layered over the pseudo-RDs of Edinburgh at the final.

References

Edvinsson S (2000) The Demographic Data Base at Umeå University—a resource for historical studies. In: Hall PK, McCaa R, and Thorvaldsen G (eds), *Handbook of international historical microdata for population research*, Minneapolis: Minnesota Population Center, pp. 231–248.

Fitch CA and Ruggles S (2003) Building the National Historical Geographic Information System. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 36(1), 41–51.

Gregory IN and Healey RG (2007) Historical GIS: structuring, mapping and analysing geographies of the past. *Progress in Human Geography*, 31(5), 638–653.

Mandemakers K (2000) Historical sample of the Netherlands. In: Hall PK, McCaa R, and Thorvaldsen G (eds), *Handbook of international historical microdata for population research*, Minneapolis: Minnesota Population Center, pp. 149–178.

Maneewongvatana S and David M M (2001) *On the efficiency of nearest neighbor searching with data clustered in lower dimensions,* Springer Berlin Heidelberg, pp. 842-851.

Passonneau R (2006) Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation*.

St-Hilaire M, Moldofsky B, Richard L, et al. (2007) Geocoding and Mapping Historical Census Data: The Geographical Component of the Canadian Century Research Infrastructure. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 40(2), 76–91.

Thorvaldsen G (2000) The Norwegian Historical Data Centre. In: Hall PK, McCaa R, and Thorvaldsen G (eds), *Handbook of international historical microdata for population research*, Minneapolis: Minnesota Population Center, pp. 179–206.