Towards a Seamless World Names Database

Leak A*[1], Longley P†[2] and Adnan M‡[2]

[1]Department of Security and Crime Science, UCL
[2]Department of Geography, UCL

November 7, 2014

Summary

This paper sets out to address limitations in a global database of personal names (worldnames.publicprofiler.org: WND). A synthesis of 26 publicly available electoral register and telephone directory datasets. However, it has proven difficult to source further data that are representative of some other countries resident populations. Thus, this study seeks to evaluate the potential for proxy registers based on geotagged Twitter data and further, to devise a method for evaluating the datas' quality. The paper concludes with a discussion of the problems arising where there are no registers or directories against which population registers or directories might be compared.

KEYWORDS: GIS, Social Media Analysis, Surnames, Twitter.

1 Introduction

This paper sets out to address limitations in a global database of personal names (worldnames.publicprofiler.org: WND) which is representative of approximately 2 billion of the Earth's population. A synthesis of electoral roll and telephone directory data from 26 countries, the WND provides a valuable resource in the analysis of populations (Longley et al., 2011; Mateos et al., 2007). However, it has proven difficult to source further data that are representative of many countries' resident populations, limiting the ability to perform truly global analyses. Thus, this paper seeks to assess the use of proxy population registers based on geotagged Twitter data. To this end, a framework for the creation of Twitter based population registers will be demonstrated and subsequently a method for their verification. The paper will conclude with a proposal as to how such registers may be verified in the absence of any reference data.

*a.leak.11@ucl.ac.uk
†p.longley@ucl.ac.uk
‡m.adnan@ucl.ac.uk

2 Methods

This section describes the methods employed in the construction of a Twitter based population registers comparable in structure to that of the UK Enhanced Electoral Roll (id, forename, surname and address). The method is applied to three countries, Poland, Spain and the United Kingdom such that three of the largest naming regions are represented. Each register is subsequently validated against a reference dataset drawn from the existing WND.

2.1 Data

2.1.1 Twitter

The Twitter dataset comprised 1.4 billion geotagged tweets harvested using the Twitter sample stream API during the period December 2012 through January 2014. It is important to note that whilst the sample stream is commonly cited as being just 1% of all tweets, this figure is in reference to the total number of tweets which may be returned. As such, where only the geographically referenced content are specified (circa 1% of all tweets) the majority are returned (Morstatter et al., 2013).

2.2 Base population registers and geography

The three baseline registers used were the 2011 UK Enhanced Electoral Roll (EER), the 2004 Spanish telephone directory and the 2005 Polish telephone directory. These data account for 53 million, 10.4 million and 8.2 million individuals for the UK, Spain and Poland accounting for 82%, 22% and 21% of the countries' populations respectively. In each case the data are referenced to GADM levels 0,1 and 2. Whilst the GADM boundaries lack the precision of many national administrative datasets, they have a consistent data structure.

2.2.1 Estimation of residential location

Prior to the extraction of users' personal names, the users believed to be resident within the study area were identified. For this exercise, the location information recorded in the users' tweets was used in preference to their declared locations. The assumption being made that the users are resident in the administrative areas in which they tweet most frequently. The process is as follows:

1. All unique users to tweet within the country boundary are identified.
2. All tweets, regardless of location, by the previously identified users are extracted from the primary dataset.
3. Extracted tweets are spatially joined to GADM administrative geography.

4. User are ascribed a location where they have 5 or more tweets and greater than 50% of all their tweets within a single spatial unit.
5. Only those users who are assigned a location within the country are included in the final register.
6. Users who do not meet the inclusion criteria are omitted.

The accuracy of the residential location ascription algorithm, assessed in the UK at GADM level 2, was 84% based on a stratified sample of 1073 users.

2.2.2 Extraction of personal names

Having identified those users believed to be resident within the target area, the next phase was the extraction of the users' probable given and family names. Rather than the distinct given and family name attributes found in conventional population registers, the full name is provided as a single string. This string was split using heuristics based on an enhanced version of a name extractor developed by Muhammad Adnan of UCL. The process is as follows:

1. All non-alpha characters are removed from the user's screen name.
2. Screen names are split into multiple tokens based on white spaces.
3. Tokens are compared against a list of titles and surname prefixes.
4. Surname prefixes are affixed to the family name segment.
5. Cleaned names are recorded against the user's id.

2.2.3 Population scaling

Lastly, whilst not critical in terms of population composition, it is necessary to scale the Twitter population for inclusion in the WND. To achieve this, the correct distribution of the countries populations was ascertained based on the best available data and in turn, this data was used to scale up or down the Twitter population such that it accurately represented the resident population distribution.

2.3 Assessment of population registers

Whilst the registers created are accurate in terms of their structure, little is known as to their representative capability. Subsequently, a testing framework is proposed which examines three key properties of the proxy registers; the most common names, population distribution and similarity in composition of surnames.

2.3.1 Location quotient

The comparison between observed and expected Twitter population size is performed using Equation 1, the Location Quotient.

$$LQ = \frac{p_i/p}{P_i/P} \tag{1}$$

- $LQ > 1$ indicates proportionally **more** Twitter users than expected.
- $LQ < 1$ indicates proportionally **fewer** Twitter users than expected.
- p_i Local Twitter population count.
- p Local Electoral population count.
- P_i Total Twitter population count.
- P Total Electoral population count.

2.3.2 Similarity of composition

The similarity in name composition is measured using Equation 2 the Morisita-Horn index of overlap (Horn, 1966). The index, developed in ecology, is recognised for ability to deal with populations of different sizes and diversities (Wolda, 1981). In this analysis, comparison is made between the resident and Twitter derived population for each administrative unit at each spatial resolution. Where an area had 100 or few individuals in either register, the area was declared null and omitted.

$$C_H = \frac{2\sum_{i=1}^{S} x_i y_i}{(\frac{\sum_{i=1}^{S} x_i^2}{X^2} + \frac{\sum_{i=1}^{S} y_i^2}{Y^2})XY} \tag{2}$$

- C_H 1 indicates equal composition of surnames.
- C_H 0 indicates no overlap in name composition.
- S is the number of unique surnames shared between the two populations.
- x_i and y_i are the number of individuals sharing a specific surname in region X and Y.
- X and Y are the number of unique surnames in regions X and Y respectively.

3 Results

UK				Spain				Poland			
Surname	Res.	Tw.	Dif.	Surname	Res.	Tw.	Dif.	Surname	Res.	Tw.	Dif.
Smith	1	1	0	Garcia	1	1	0	Kowal	1	11	-10
Jones	2	2	0	Fernandez	2	6	-4	Nowak	2	2	0
Williams	3	3	0	Gonzalez	3	5	-2	Wojniak	3	9	-6
Brown	4	5	-1	Rodriguez	4	4	0	Kowalczyk	4	9	-5
Taylor	5	4	+1	Lopez	5	2	+3	Wozniak	5	15	-10
Davies	6	6	0	Martinez	6	7	-1	Mazur	6	12	-6
Wilson	7	7	0	Perez	7	8	-1	Kaczmarek	7	11	-4
Evans	8	8	0	Martin	8	10	-2	Krawczyk	1	4	+4
Thomas	9	9	0	Gomez	9	9	0	Zajac	1	-	-
Johnson	10	11	-1	Ruiz	10	11	-1	Krol	1	5	+5

Table 1: Comparison of surname ranks for the 10 most common family names between the resident and Twitter population registers for the UK, Spain and Poland.

UK		Morisita-Horn			Location Quotient			
GADM Level	Valid Locations	Min	Mean	Max	Min	1st Qu.	2nd Qu.	Max
0 n=1	416,819	-	0.98	-	-	-	-	-
1 n=4	413,461	0.91	0.97	0.99	0.94	-	-	1.08
2 n=192	346,900	0.10	0.67	0.97	0.45	0.77	1.08	8.77

Spain		Morisita-Horn			Location Quotient			
GADM Level	Valid Locations	Min	Mean	Max	Min	1st Qu.	2nd Qu.	Max
0 n=1	261,131	-	0.91	-	-	-	-	-
1 n=18	248,190	0.68	0.86	0.94	0.36	0.82	1.22	2.14
2 n=51	241,706	0.57	0.82	0.94	0.33	0.68	1.32	0.28

Poland		Morisita-Horn			Location Quotient			
GADM Level	Valid Locations	Min	Mean	Max	Min	1st Qu.	2nd Qu.	Max
0 n=1	13,178	-	0.37	-	-	-	-	-
1 n=16	12,721	0.00	0.04	0.16	0.38	0.49	1.38	2.75

Table 2: Results of the Location Quotient and Morisita-Horn Similarity Analysis for the UK, Spain and Poland.

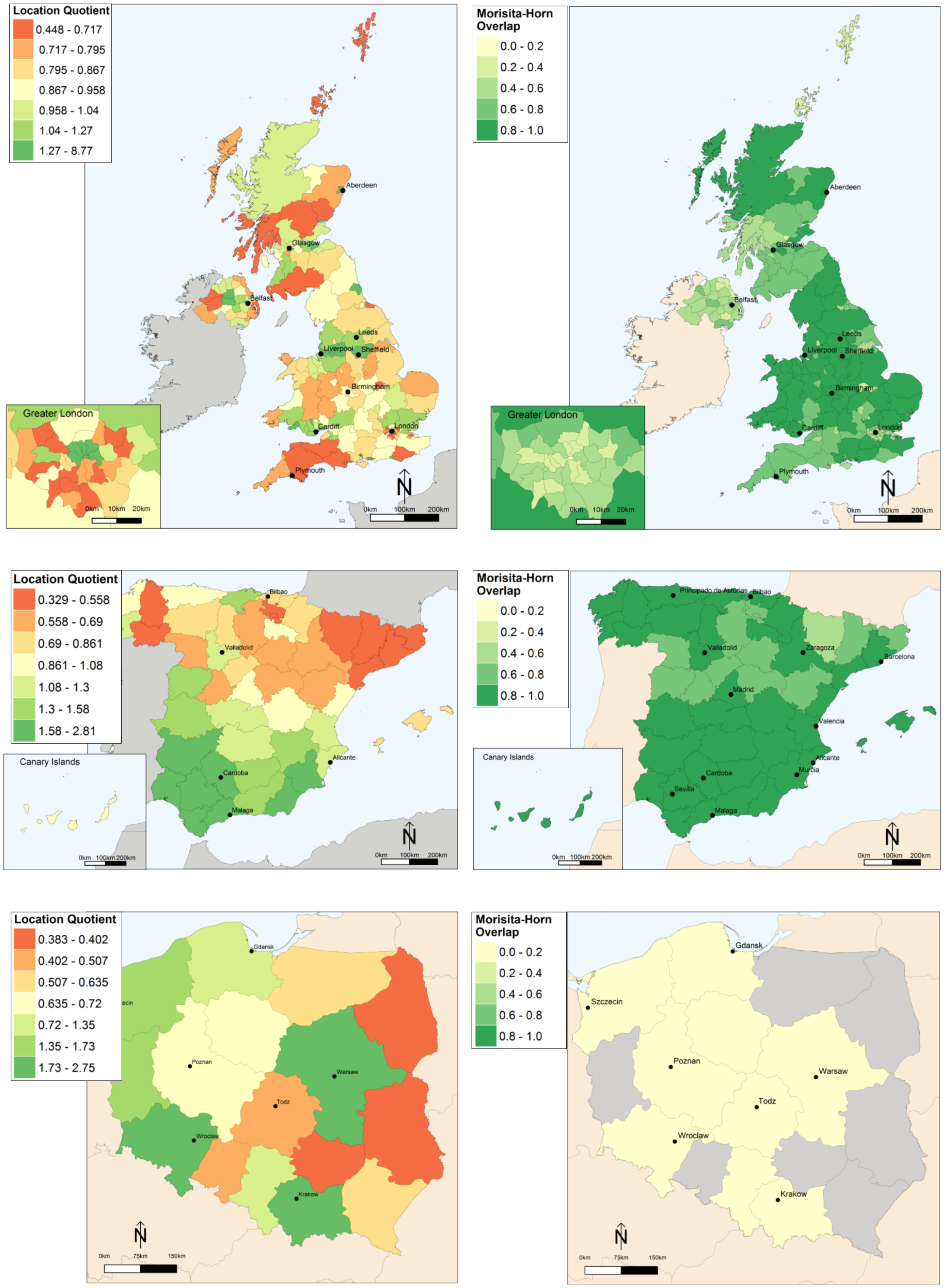

Figure 1: Location Quotient (left) and Morisita-Horn Similarity (right) maps for the UK, Spain and Poland at GADM levels 2, 2 and 1 respectively

4 Discussion

The results presented in this report provide a number of early indicators as to the potential utility of social media based population registers. The performance in both the UK and Spain was impressive at both national and regional scales. However, the results for Poland highlighted several potential limitations of the approach. Unlike its counterparts, Poland exhibited poor surname composition similarity at all spatial resolutions. Table 1 provides an early indicator as to the success of the register creation methodology. However, whilst strong rank similarity may be observed in the UK and Spain, this is not reflected for Poland. The disappointing performance in Poland is further observed in Table 2 which provides a breakdown of the Morisita-Horn and Location Quotient results. Of particular significance is the low national surname similarity observed in Poland. A further observation was the high Location Quotient value observed in the City of London, UK, which is likely a consequence of the highly transient population. The geographic distribution of Location Quotient and Morisita-Horn analysis are illustrated in Figure 1.

A key concern to arise from the preliminary analysis was as to how a country could have been deemed suitable for a Twitter based population register in the absence of baseline data. That is to say that whilst the UK and Spain may have been well represented by a Twitter based register, Poland most certainly was not; a fact that may not have been evident without some form of reference. Thus, there is a requirement for a series of diagnostic tools which may be applied to the Twitter derived registers in the absence of a reference dataset. To this end, Twitter population registers are to be created for all countries presently in the WND. In turn, a series of diagnostic measures will be performed on each of the proxy register to determine if, and to what extent, they are indicative of the registers representative capacity versus their WND counterparts. These measures will include the proportion of identified users to the true population, surname diversity and population structure as determined by the conformity to power laws. These diagnostic tools may also assist in the selection of an appropriate spatial resolution for each new country to be included in the WND.

5 Conclusions

Overall, the study has demonstrated how population registers may be created using geotagged Twitter data and how such registers may be validated. The preliminary analysis provided strong positive evidence to support the methodology though it was unfortunate how poorly Poland performed throughout the analysis. That being said, the results of the Polish analysis have allowed for validation of the register creation method in a Slavic country and helped determine how a country's suitability may be determined in the absence of baseline data. Finally, the work has taken the first steps in creating arguably the most complete record of human population ever created; a resource of significant value in academia.

6 Acknowledgements

This work was completed as part of the DSTL National PhD scheme (12/13NatPHD 61) and the EPSRC research Grant The Uncertainty of Identity: Linking Spatiotemporal Information in the Real and Virtual Worlds (EP/J005266/1).

7 Biography

Alistair Leak is a Ph.D. student in the department of Security and Crime Science at University College London. His research interests include the analysis of names and the application of geographical information systems to 'Big' and open data.

Paul Longley is Professor of Geographic Information Science at University College London. His publications include 14 books and more than 125 refereed journal articles and book chapters. He is a former co-editor of the journal Environment and Planning B and a member of four other editorial boards. He has held ten externally-funded visiting appointments and given over 150 conference presentations and external seminars.

Muhammad Adnan is a Senior Research Associate at Consumer Data Research Centre, University College London. His research interests are in data mining, social media analysis, and visualisation of large spatio-temporal databases.

References

Horn, H. S. (1966). Measurement of "overlap" in comparative ecological studies. *American naturalist*, pages 419–424.

Longley, P. A., Cheshire, J. A., and Mateos, P. (2011). Creating a regional geography of Britain through the spatial analysis of surnames. *Geoforum*, 42(4):506–516.

Mateos, P., Webber, R., and Longley, P. (2007). The cultural, ethnic and linguistic classification of populations and neighbourhoods using personal names.

Morstatter, F., Pfeffer, J., Liu, H., and Carley, K. M. (2013). Is the sample good enough? comparing data from twitters streaming api with twitters firehose. *Proceedings of ICWSM.*

Wolda, H. (1981). Similarity indices, sample size and diversity. *Oecologia*, 50(3):296–302.