A geodemographic analysis of the ethnicity and identity of Twitter users in Greater London

Muhammad Adnan, Guy Lansley, Paul A Longley

University College London, Department of Geography, Gower Street, London, WC1E 6BT.
Tel: +44 (0)20 7679 0510 Fax: +44 (0)20 7679 0565
Email m.adnan@ucl.ac.uk, g.lansley@ucl.ac.uk, plongley@geog.ucl.ac.uk

KEYWORDS: Twitter, Geodemographics, GIS, Identity

Abstract

The availability of social media data for research is continuously increasing. This paper presents a preliminary comparison between residential geographies of different ethnic groups as inferred from an enhanced version of the Electoral Register and the corresponding mobile geographies as recorded by Twitter in the day and night time.

1. Introduction

Use of social media continues to increase day by day, with implications for the creation of 'big' data – Twitter alone was forecast to have created 1.8 zettabytes of data in 2011 (equivalent to the information storage capacity of 150bn Apple iPads). Users of the likes of Twitter, Facebook, Flickr, LinkedIn, Bebo, and Orkut are frequently mobile users of the developing range of smartphones and tablet devices. The availability of such data has profound implications for the geodemographic analysis of human settlement structure. Hitherto, small area measures of neighbourhood conditions have been based only upon the night-time socioeconomic characteristics of residential areas and selected physical characteristics of their built environments (Martin et al 2012). Although useful for guiding resource allocation decisions for many private and public goods (Longley 2005), these static and essentially cross sectional views provide only limited insights into the functioning of settlement systems and the temporal heterogeneity that characterises their component parts. The advent of new data sources derived from social media also has profound implications for our understanding of behaviour in virtual as well as observable space, and interactions between the two. Taken together, the prospect of developing composite cyber-geodemographic measures offers the prospect of better understanding the dynamic as well as the static organisation of human settlements, and has potential in investigations of a range of spatial equity issues first broached a generation ago (Pahl 1970).

This paper provides an initial empirical exploration of these issues using data from the Twitter social-networking and micro-blogging service. Twitter was launched in 2006 and within six years had accrued more than 140 million active users, who typically send 340 million tweets every day. The site is one of the ten most visited on the Internet (Twitter, 2012). The Twitter API (http://dev.twitter.com) allows download of a sample of the live Twitter data, and it is possible to focus download activity upon the small subset of users that disclose their geographic locations. Exploiting such a huge data source for research can potentially yield insights into the residential and travel geographies of individuals in different geographical areas at different times of day: in short, a selective but numerically large representation of activity patterns across geographically extensive areas.

Previous research at UCL has identified the cultural, ethnic and linguistic characteristics that can be ascribed to individuals on the basis of forename and surname conventions and pairings (Mateos et al 2011). In terms of the residential geography of conventional geodemographics, this makes it possible to pinpoint small area geographies of groups that are not specifically identified in population censuses, and at any point in time for which names data are available (see Figure 1). The same principles can be extended towards classification of Twitter data for mobile users, at different times of the day and night. This paper undertakes a preliminary investigation of the observed differences between the residential geographies of different ethnic groups as recorded in an enhanced version of the GB Electoral Register, and those obtained from a subset of Twitter users.

2. Residential geography of different ethnic groups

The night-time geography of Polish residence in London, shown in Figure 1, uses the ‘Onomap’ methodology of Mateos et al (2011) and can be extended to any societal group that has distinctive naming practices. The basis to this method entails associating distinctive forename and surname pairs to their corresponding ethnic groups.

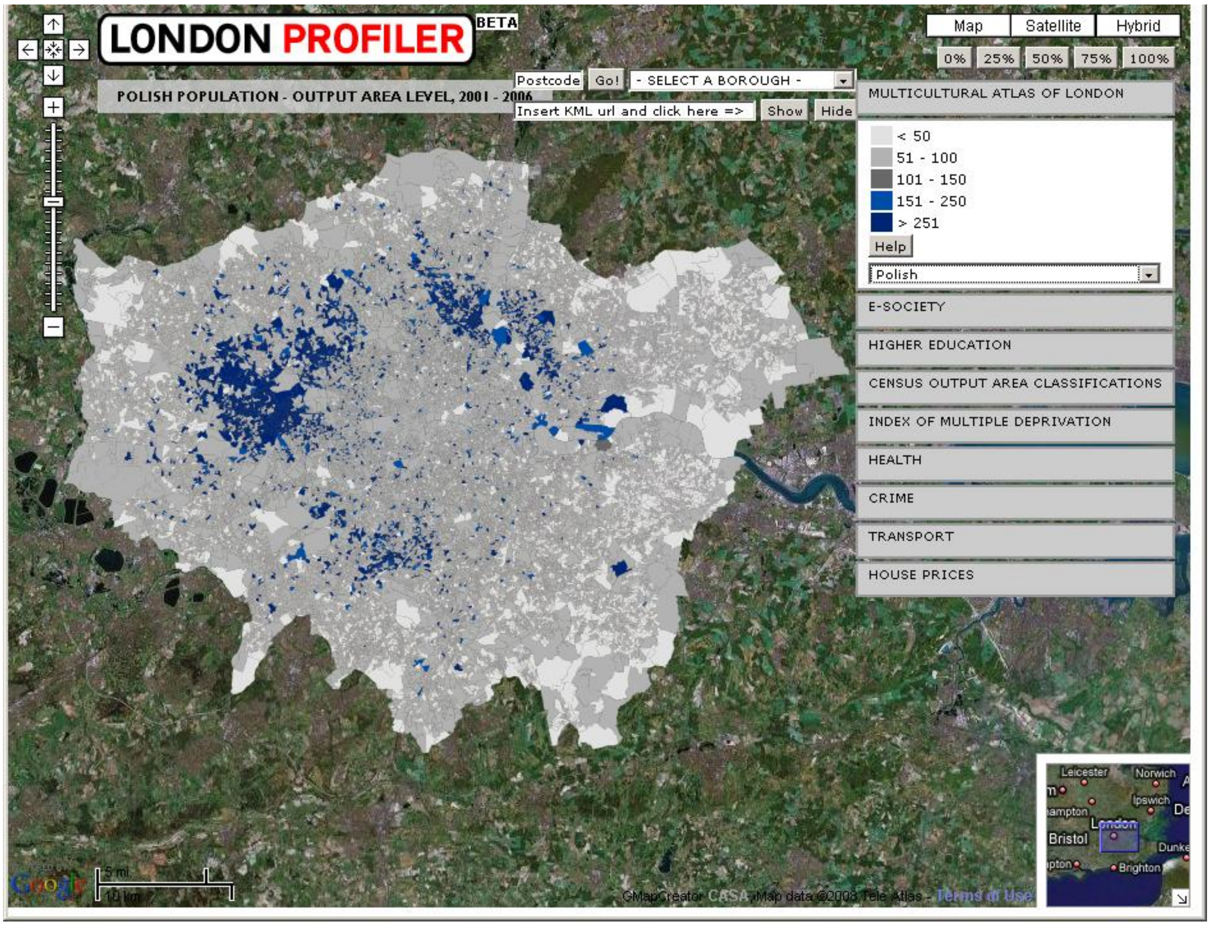

Figure 1: Distribution of Polish residents London based upon classification of names (source: www.londonprofiler.org)

3. The geography of Twitter users

The twitter API allows developers to download geo-tagged and non-geo-tagged tweets. A non-geo-tagged tweet does not have an associated "latitude" and "longitude" information with the tweet message. Non-geo-tagged tweets could be sent from mobile devices or desktop computers. Geo-tagged tweets, in contrast, are only sent from mobile devices, and are a small fraction of the total number of tweets sent. Geo-tagged tweets have an associated "latitude" and "longitude" information with the tweet message. The Twitter API was used to download approximately 1 million geo-tagged Tweets for London during September and October 2012. The data downloaded from the API included the "User Name", "Date and time of the Tweet", "Latitude of the Tweet", and "Longitude of the Tweet". Figure 2 shows all of the latitude/longitude coordinates plotted for Greater London, illustrating a higher density of Tweets per unit area in the central areas of the city than in its suburbs. It is also clear from the map, however, that the medium has London wide coverage.

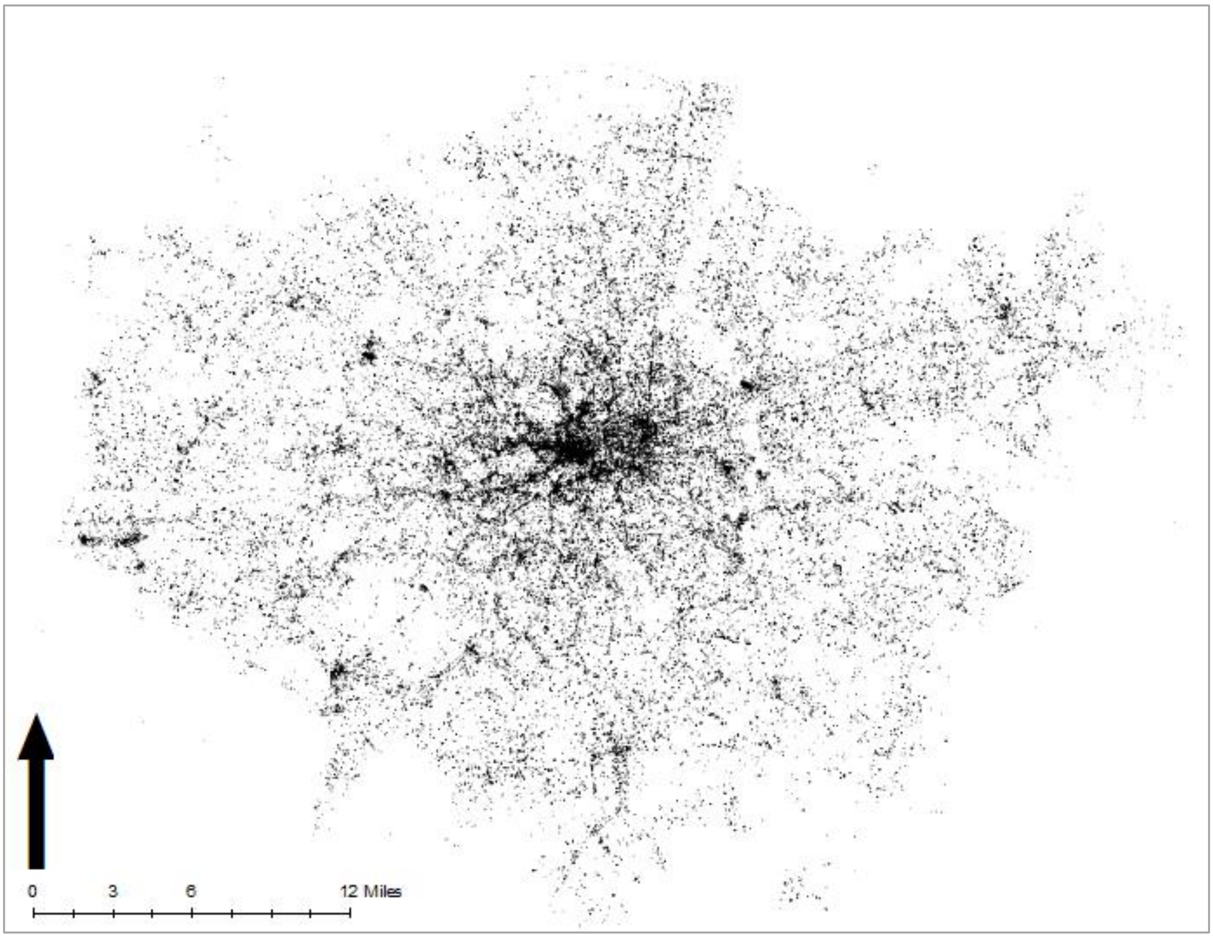

Figure 2: Locations of geo-tagged tweets data downloaded during September and October, 2013

Twitter requires a user to enter their name or other identifying data in the 'User Name' field when they create their user profiles. In many cases, tokens other than given and family names are entered, as in 'MysticMIND', 'What is Love', etc. Our text analytic work suggests that approximately 60% of Tweets are identified by recognisable forename-surname pairs. The 'User Name' field was divided into separate 'forename' and 'surname' fields for these users, as illustrated in Table 1.

User name	Forename	Surname
Kevin Hodge	Kevin	Hodge
Jose De Dranco	Jose	De Franco

Carolina Thomas, Dr.	Caolina	Thomas
Prof. Martha Del Val	Martha	Del Val
Dame Alexia Singleton	Alexia	Singleton

Table 1: ‘User Name’ divided into separate ‘forename’ and ‘surname’ fields

In the next step, Onomap (Mateos et al 2011) was used to assign forename and surname pairs to predicted ethnic groups, and the results mapped for different ethnic groups across Greater London. These maps can also be decomposed into different times of day, as shown in Figure 4: on the basis of the activity shown in this figure we set the arbitrary boundary between day and night at 07.00 and 19.00, although this assumption will be re-evaluated for different classes of user in future research. The following figures 4 and 5 show the “distribution of different ethnic groups in the day time” and “distribution of different ethnic groups in the night time”. Figure 3 shows the legend of Figures.

Figure 3: Map Legend

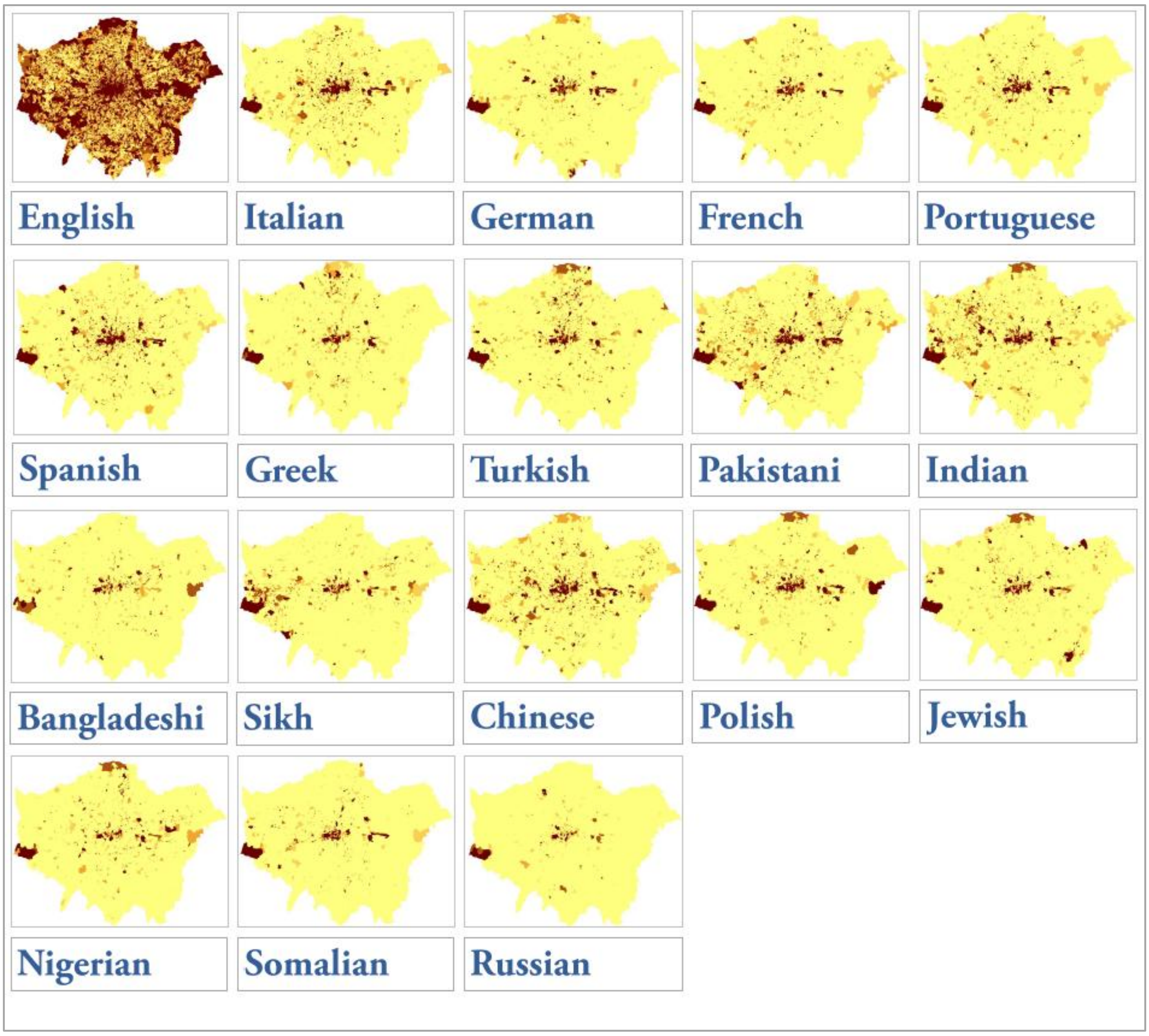

Figure 4: Distribution of different ethnic groups in the day time

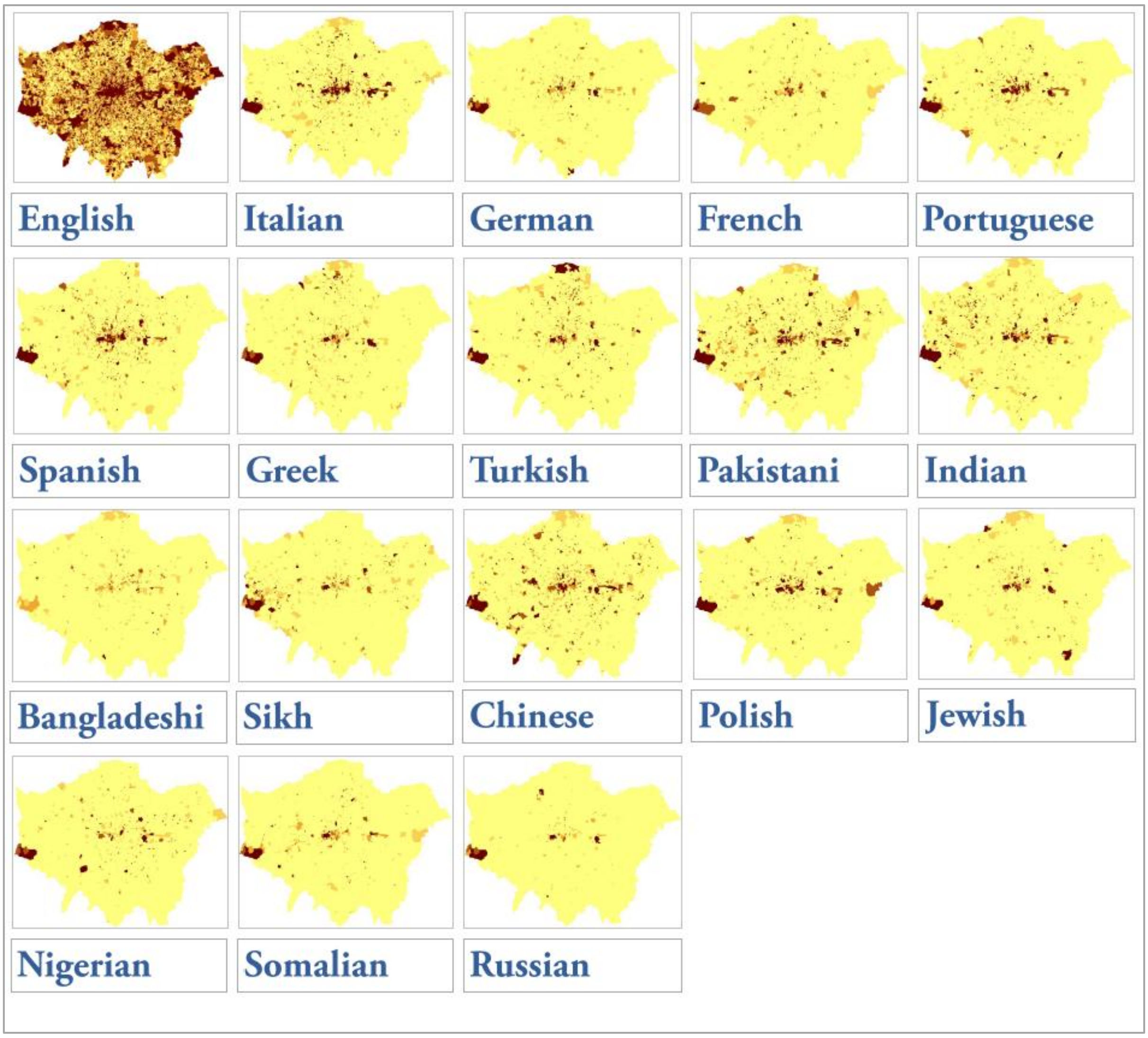

Figure 5: Distribution of different ethnic groups in the night time

3. Discussion

The empirical results reported in this paper are an initial foray into a very promising area of research. For example, we are presently investigating the observed differences between the conventional night time geographies of different ethnic groups and their day time activity patters, as an investigation of likely work and leisure activity patterns at different times of the week. We are also investigating the merits of assigning observations to different age group categories based upon changing fashions in naming. Finally, it must be remembered that the sources and operation of bias in our (very large) dataset are unknown: by no means everybody Tweets, and there is no a priori reason to assume that Tweeters who disclose their locations are representative of those that do not. A related theme to our work is therefore establishing the basis to generalisation using our Twitter data. It is also the case that language, whether as declared in Twitter registration or as used in communication, is a valuable indicator of ethnicity and linguistic group membership. Our own predilection is that the Onomap classification can provide a more useful and broad based assignment of cultural, ethnic and linguistic group, but this position does require further investigation.

4. Conclusion

This paper has presented a preliminary comparison between residential geographies of different groups as inferred from an enhanced version of the Electoral Register and the corresponding mobile geographies as recorded by Twitter in the day and night time.

Future research work will extend this analysis to more extensive geographical areas, such as the entire United Kingdom. We also envisage extending the analysis to other social media datasets (Flicker, Four Square) to investigate location specific social and linguistic characteristics of social media usage.

5. References

Longley P A 2005. A renaissance of geodemographics for public service delivery. **Progress in Human Geography**, 29: 57-63

Martin D, Cockings S, Harfoot, A 2012. Development of a geographical framework for Census workplace data. **Journal of the Royal Statistical Society: Series A (Statistics in Society)** (In Press).

Mateos P, Longley P A, O'Sullivan D 2011. Ethnicity and population structure in personal naming networks. **PLoS ONE (Public Library of Science)**, **6(9)** e22943, 1-12

Pahl R 1970. **Whose City?** Penguin, Harmondsworth

Twitter (2012). "What is Twitter ?". From https://business.twitter.com/basics/what-is-twitter/.

6. Biographies

Muhammad Adnan is a Postdoctoral Research Associate at the Department of Geography, University College London. His research focus is on open geodemographics, data mining, algorithm optimisation, and visualisation of large spatio-temporal databases.

Guy Lansley is a teaching fellow at the Department of Geography at University College London, and currently convening modules on population geography. His research interests are primarily in geodemograpics and socio-spatial analysis.

*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 co-editor of the journal **Environment and Planning B** and a member of five other editorial boards. He has held ten externally-funded visiting appointments and given over 150 conference presentations and external seminars.*