Paper presented at the GIS Research UK (GISRUK) 2006 Annual Conference, University of Nottingham, 5-7 April, p. 285-291

How segregated are name origins? A new method of measuring ethnic residential segregation

Pablo Mateos, Richard Webber, and Paul Longley

Centre for Advanced Spatial Analysis
Department of Geography
University College London
1-19 Torrington Place, London,
WC1E 7HB, United Kingdom
www.casa.ucl.ac.uk
p.mateos@ucl.ac.uk
richardwebber@blueyonder.co.uk
plongley@geog.ucl.ac.uk

KEYWORDS: ethnicity, segregation indices, ethnic ontology, names, London

1. Introduction

Two major events in 2005 have reopened a long-standing debate about the residential segregation of ethnic minorities in European cities; the London bombings of July 7th, 2005 and the urban riots in France in November 2005. These events triggered a heated public debate in which diverse issues were all linked to an apparent failure of European society to assimilate immigrant communities (Leppard, 2005, Economist, 2005). A perceived manifestation of such failure is the residential segregation of ethnic minorities, whether through constraint or choice, to the frustration of ethnic minority integration policies.

Geographers have made some interesting contributions to this debate (Simpson, 2005, Observer, 2005), that begin with the difficult task of ascertaining the degree to which neighbourhoods and cities are segregated, through the well established five dimensions of residential segregation (evenness, exposure, concentration, centralisation, and clustering: (Massey and Denton, 1988). Related to this is the task of measuring whether such segregation is increasing or decreasing, and the apportionment of any such change to factors such as immigration, out-migration, and natural growth (Stillwell and Duke-Williams, 2005, Simpson, 2004). These two general issues, determining segregation levels and change, relate to classic problems in geographical research: the former with the MAUP and spatial autocorrelation (Fotheringham et al., 2000), and the latter with the measurement of population structure and dynamics. However, the majority of the research in this field in most countries depends solely upon census of population. This imposes major limitations because censuses are usually conducted only every ten years, and typically utilise only a few self-assigned ethnic categories, the number, definition and reporting of which may be highly inconsistent through time (Kertzer and Arel, 2002). The subjective essence of self-perceived ethnic identity has been highly contested, and that has made race and ethnicity to be considered the most politically loaded question in the US Census (Skerry, 2000). Taken together, poor temporal granularity, inconsistent or incoherent ethnic categories and reporting methods, and the often coarse spatial resolution at which Census data are made available (through units of highly variable size), make the task of determining residential segregation change using censuses an unreliable and error prone process that bears too little correspondence with the rapidity of change in contemporary cities.

This paper presents a new methodology of measuring ethnic segregation in cities at fine temporal and spatial resolutions. It uses data collected and made available, subject to safeguards, at the level of the individual and develops a large number of independently assigned ethnic categories based on the probable ethnic origins of names. Our basic hypothesis is that the ethnic classification of family

names and first names provides a new and more reliable means of measuring residential segregation. It is argued that this methodology offers improved monitoring of residential segregation processes at the neighbourhood scale (based upon unit postcodes of an average of 30 people), develops a more detailed and meaningful classification of people's origins categories (over 150 categories based on name origins), offers improved updating (annually through electoral or patient registers), and better accommodates changing perceptions of self-identity (through independent assignment of ethnicity according to name).

2. Methodology

Name analysis techniques have been independently used and validated to ascribe population ethnic origins in the fields of Epidemiology and Genetics (Lauderdale and Kestenbaum, 2000, Lasker, 1985) in studies since the 1950s (U.S. Census Bureau, 1953, Buechley, 1961, Shaw, 1960). Drawing on this research legacy a dataset has been generated coding each individual elector in the UK Electoral Roll[1] with their most likely ethnic origin based on their personal and family name and a purpose built classification of 280,000 family names (surnames) and 110,000 personal names (given names). The classification of ethnic origins by name provides an ontology of ethnicity that is made up of *Cultural, Ethnic and Linguistic* categories (hereinafter 'CEL'), of which 150 types are defined. This paper will not cover how this classification and coding was derived (see Webber and Mateos, forthcoming), but rather focuses upon the use of the CEL classification to measure residential segregation at very fine scales, and comparison of the geographies of ethnicity in Greater London suggested by CEL rather than the UK Census.

The first dataset used in this analysis is the 'CEL-coded' Electoral Roll for Greater London (hereinafter the *CEL dataset*), which is the most ethnically diverse region of the country. The 5.2 million electors were coded by 150 CEL types, and were assigned to the 131,721 unit postcodes of the Capital. The 150 CEL types were then also further aggregated into 18 meaningful CEL groups that are represented throughout London. In order to facilitate comparison between the CEL dataset and the Census, the left part of Table 1 shows the 16 Census 2001 ethnic categories and their corresponding 18 CEL types. Since the ontologies of ethnicity underlying both datasets are different (the CEL one based on onomastic origin and the Census on self-reported identity), the categories do not conceptually match and the CEL groups are furthered aggregated to 14 effective groups.

The second dataset used, is the Census 2001 Key Statistics KS06 table (Ethnic Group) for the 33 London Boroughs at Output Area (OA) level (comprising 7,158,904 Census respondents and 24,100 OAs). The CEL dataset and the Census dataset were linked together using a Postcode to Output Area lookup table from the All Fields Postcode Directory (AFPD November 2005 version: (ONS, 2005). The AFPD directory was also used to aggregate both postcode units and OAs up to higher level geographies (ordered in increasing size; Lower Super Output Areas -*LSOA*-, Wards, and London Borough). All these geographies were mapped through a GIS using *OS CodePoint* boundaries for the postcode units and the Census administrative geographies for the OAs and their higher level administrative aggregations.

The first step of the analysis entailed calculating correlation coefficients between the CEL dataset and the Census ethnicity responses at different levels of geography (OA, Lower SOA, Ward, and London Borough), in order to establish the validity of using the CEL dataset in the future and its most appropriated geographical scale. Since the two datasets do no use the same denominator (the CEL file only includes adults while the Census enumerates all of the population), the comparison was

[1] The file used is more precisely known as the 'Consumer Dynamics' file and has been developed by Experian UK Ltd, using the Electoral Roll complemented with a variety of commercial data sources to correct undercount and people not entitled to vote. Experian kindly provided access to the 2004 version of this file to Richard Webber for research purposes. It covers over 46 million adult individuals resident in the UK.

performed using the proportion of people in each ethnic group for each geographical unit, upon which a correlation matrix was calculated (Robinson, 1998). A summary of the results is offered in Table 1.

		Correlation at Geographical Levels			
Census 2001 Ethnic Group	**CEL Group Aggregation**	**OA**	**LSOA**	**Ward**	**Borough**
White British	British + Jewish	**0.83**	**0.86**	**0.89**	**0.91**
White Irish	Irish	0.01	0.05	0.16	0.34
Other White	W. & E. European + Hispanic + Greek & G. Cypriot	**0.75**	**0.85**	**0.90**	**0.94**
White and Black Caribbean	Black Caribbean	0.26	0.55	**0.81**	**0.92**
White and Black African	Black African + Somali	**0.27**	**0.47**	**0.58**	**0.67**
White and Asian	*Not Assigned*				
Other Mixed	*Not Assigned*				
Indian	Hindu + Sikh	**0.96**	**0.98**	**0.99**	**1.00**
Pakistani	Pakistani	0.06	0.09	0.12	0.06
Bangladeshi	Bangladeshi	**0.94**	**0.98**	**0.99**	**1.00**
Other Asian	Sri Lankan	-0.01	-0.04	-0.07	-0.17
Black Caribbean	Black Caribbean	0.47	**0.75**	**0.94**	**0.99**
Black African	Black African	**0.78**	**0.87**	**0.89**	**0.93**
Other Black	Avg. of Black Carib + African	**0.40**	**0.64**	**0.80**	**0.87**
Chinese	Chinese	**0.64**	**0.78**	**0.89**	**0.90**
Other ethnic group	Other Muslim + Japanese	0.34	0.43	0.56	**0.76**
Total Population (Ethnicity Question)	**Total Adults** (Persons in the CEL file)	**0.63**	**0.73**	**0.90**	**0.99**
	Avg. Persons / Geog. Unit	285	1,443	10,931	208,011
	Nr. of Geographical Units	24,100	4,758	628	33

Table 1 Summary of correlations between the CEL and Census datasets (>0.6 in bold)
(OA= Output Area, LSOA= Lower Super Output Area)

In order to assess the scale effect in the levels of residential segregation measured with each of the datasets, the second phase of the analysis involved the calculation of a set of well-established residential segregation indices for both datasets at each of the different levels of geography described above. Because of the computer intensive process of dealing with very small geographical units, the segregation indices at postcode unit level were not calculated for this analysis. A software application called *Segregation Analyser* was used to compute the residential segregation indices, a tool developed by Apparicio *et al* (2005) significantly simplifying this task. It can be downloaded from:

http://www.inrs-ucs.uquebec.ca/inc/Groupes/LASER/Segregation.zip (last accessed 05/12/2005)

Drawing upon (Massey and Denton, 1988) and other sources, we calculated the following 6 indices of evenness, exposure, concentration, and clustering (no index of centralisation was used because of the multiplicity of historic town centres in London):

- *Evenness*
  - IS(adj) Segregation Index (adjusted for contiguous boundaries)
  - D Index of Dissimilarity (Multi-group)

- *Exposure*
  - xPx Isolation Index
  - R Relative Diversity Index
- *Concentration*
  - RCO Relative Concentration Index
- *Clustering*
  - ACL Absolute Clustering Index

For a review of these indices equations and their theoretical justification see (Massey and Denton, 1988) and Wong (2004), and for their implementation in *Segregation Analyser* see Apparicio *et al* (2005).

3. Results and Discussion

Table 1 presents a summary of the correlation matrices between each of the ethnic categories in the CEL and the Census datasets at the different levels of geography for which the correlations were calculated (OA, LSOA, Ward and Borough). The number of geographical units at each level and their population size are indicated at the bottom of the table. Most of the ethnic categories present a high degree of correlation, that generally increases with area size, although there are some groups for which anomalies occur. These are the White-Irish, Pakistani, 'Mixed' and 'Other' categories, and to a lesser extent the Black-Caribbean group. The main reasons for this divergence are, on the one hand, the inherent vagueness of some Census categories (Mixed, Other) together with their lack of exact correspondence to the CEL categories (indicated by significant correlation with other categories in the full correlation matrices), and on the other hand, some problems detected in the distinction of Irish, Pakistani and Caribbean names, that are due to historic differences or methodological biases in the allocations that will be corrected in subsequent analysis. Nevertheless, the correlation coefficient for most of the categories is significant, especially at the LSOA- Lower Super Output Area level (1,443 persons in average) and coarser geographies. Some groups perform extraordinarily well across all scales (White British, Indian, Bangladeshi, Chinese, and to a lesser extent Black African) probably indicating the robustness of their Census ethnic categories as well as a strong linkage between current self-perception of ethnic identity and name origins for those groups. The correlation between the geographic distribution of the total population in both datasets is significant at the LSOA level and above, meaning that the fact that the CEL dataset does not cover children and people not entitled to vote (Non-EU and Non-Commonwealth citizens) makes its population estimates too sensitive to small number variation at Output Area level.

Figures 1, 2, and 3 show a summary of five of the six residential segregation indices calculated for each of the four geographical scales of study. As expected, Figure 2 shows that the level of segregation increases as the size of the geographical unit is reduced (Wong, 2004), although the strength of this scale effect varies by ethnic group (for example, in Figure 1a, the White & Black African group is much more segregated at OA than at LSOA level, and the same applies for Black Caribbeans in Figure 1b). As regards to differential experiences of segregation, Figure 1a shows the Bangladeshi group, Mixed (White & Black African), and Other Black, as the most segregated Census ethnic groups in London at OA level, and in general the Non-White groups present a higher rate of segregation in both the Census and CEL datasets. Moreover, the advantage of the much finer CEL categories is apparent in Figure 1b where the broad Census categories are broken down into subgroups (CEL) revealing the differential patterns of residential segregation. For example, the Greek group's Index of Segregation at OA level is nearly double (0.55) that of Western Europeans (0.3). In general the CEL dataset produces a more segregated pattern than the Census one for the same areal units, because of its much finer ethnic group categories and thus a more robust representation of true segregation patterns.

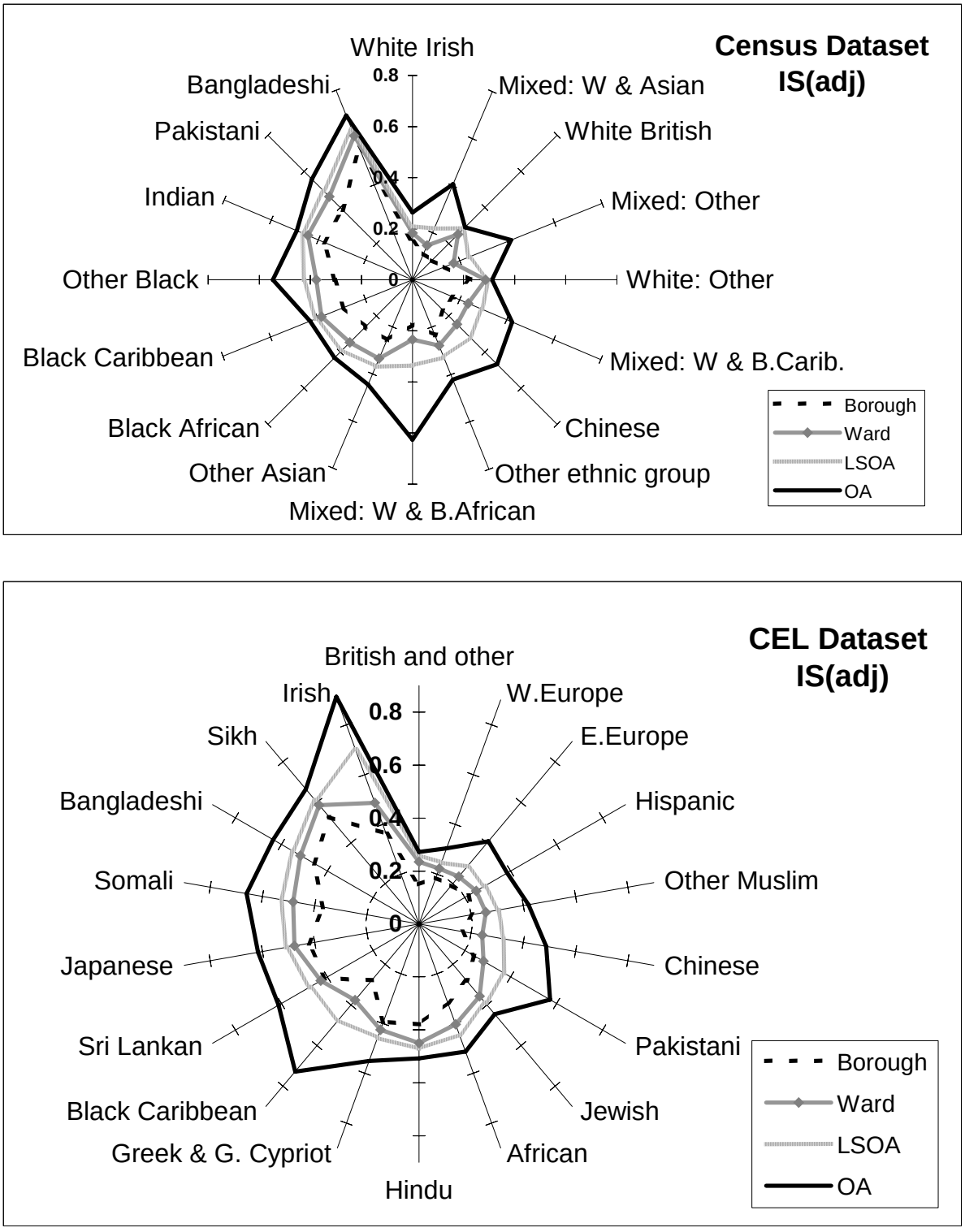

Figure 1 – Segregation Index IS(adj) calculated for the Census (above) and CEL (below) datasets at four different geographical scales. The ethnic categories are ordered by their average IS(adj) index value, showing increasing segregation in a clockwise direction from '12:00 o'clock'.

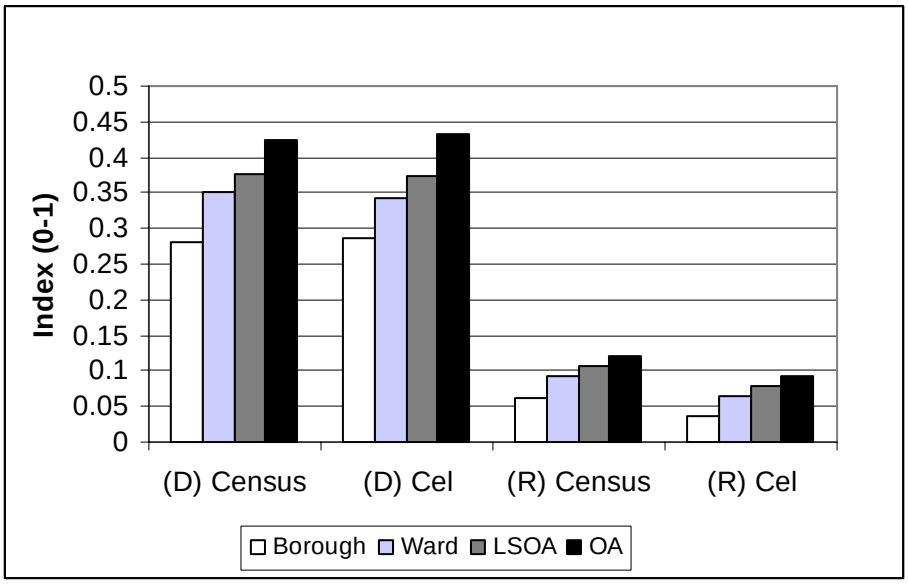

Figure 2 – Scale effect in the dimensions of Evenness - Index of Dissimilarity (D) - and Exposure – Index of Relative Diversity (R) – at various geographical levels for both datasets and all the ethnic groups (multi-group indices of segregation)

Finally, Figure 3 presents a comparison of indices of exposure (isolation) and clustering (ACL) between both datasets, showing the power of the CEL classification in identifying clusters of very detailed groups such us Sikh, Jewish, Greek, Japanese, or Somali, categories that in most cases escape the Census ethnicity classification.

4. Conclusion

This paper has presented an innovative data source to measure ethnic residential segregation at a very fine spatio-temporal resolution and using an alternative ontology of ethnic categories based on the ethnic origins of names. It has illustrated how it overcomes some of the limitations of solely using Census derived data. The analysis performed here provides a preliminary validation of this technique through its comparison with the Census responses in Greater London at various geographical scales. Further research is required to establish the broad and detailed patterns of ethnic minority segregation that this methodology might help us to reveal.

5. Acknowledgements

This research has been partially funded by DTI/ESRC and Camden Primary Care Trust under Knowledge Transfer Partnership Nr. 00037. Special thanks to *Experian UK Ltd* for providing access to their 'Consumer Dynamics' file, and to Philippe Apparicio and his team at INRS in Canada for providing free access to the *Segregation Analyser* tool through their website. The following datasets used in this research are subject to Crown Copyright restrictions; ONS Census data, ONS All Fields Postcode Directory, OS Census boundary files, OS CodePoint

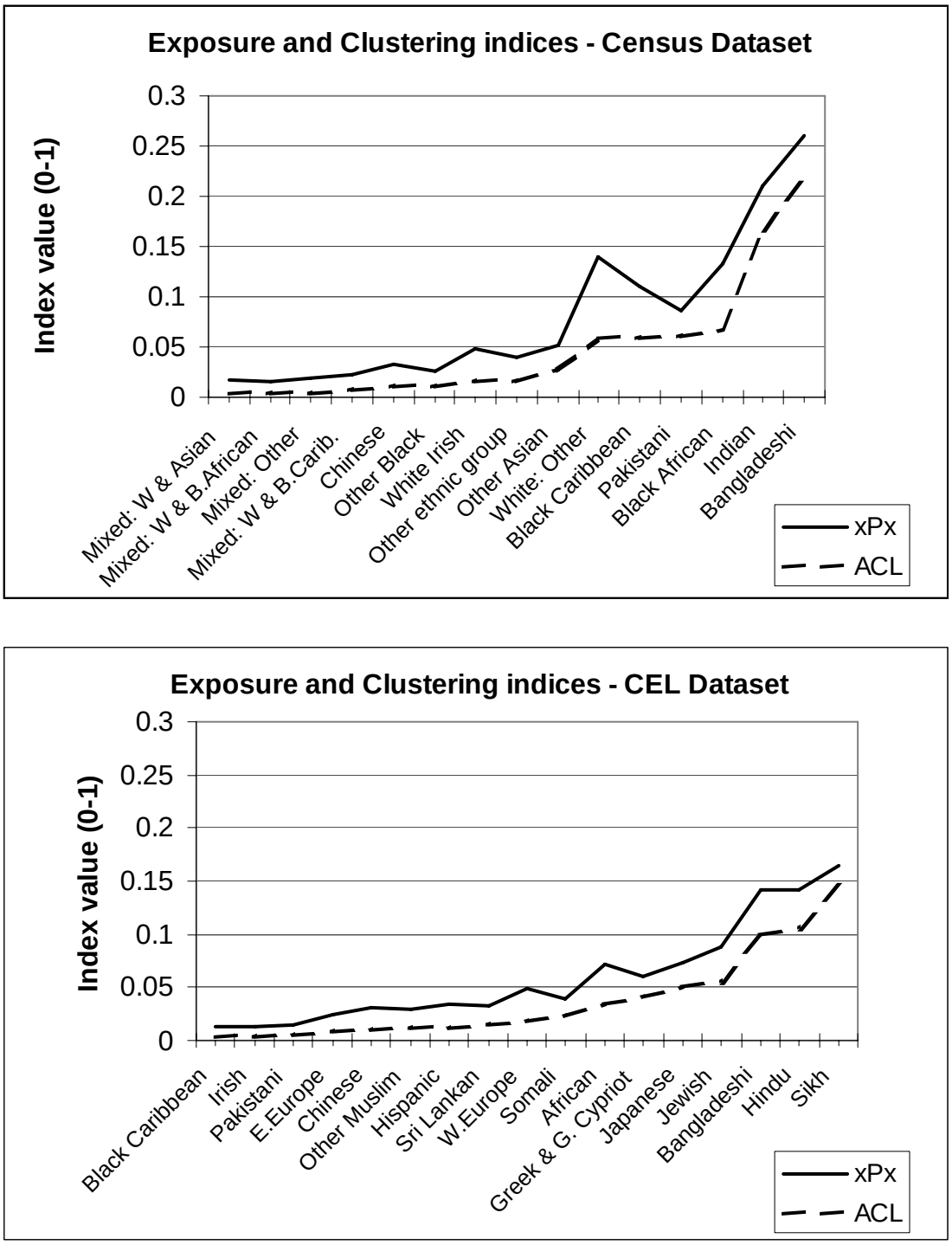

Figure 3 – Exposure - xPx Isolation Index – and Clustering – ACL Absolute Clustering Index -calculated for the Census (above) and CEL (below) datasets at OA geographical level only. The ethnic categories are ordered by their ACL index value from least to most segregated (left to right).

6. References

APPARICIO, P., PETKEVITCH, V. & CHARRON, M. (2005) Une application C#.Net pour le calcul des indices de ségrégation résidentielle. INRS Urbanisation Culture et Societé *Document de recherche (2005-02)* Available at http://www.inrs-ucs.uquebec.ca/pdf/inedit2005_02.pdf Accessed 10/12/2005

BUECHLEY, R. W. (1961) A Reproducible Method of Counting Persons of Spanish Surname. *Journal of the American Statistical Association,* 56**,** 88-97.

THE ECONOMIST (2005) One man's ghetto. *The Economist* 24th September, 16

FOTHERINGHAM, S. A., BRUNSDON, C. & CHARLTON, M. (2000) *Quantitative Geography,* London, Sage.

KERTZER, D. I. & AREL, D. (2002) *Census and Identity. The Politics of Race, Ethnicity, and Language in National Censuses,* Cambridge, Cambridge University Press.

LASKER, G. W. (1985) *Surnames and genetic structure,* Cambridge, Cambridge University Press.

LAUDERDALE, D. & KESTENBAUM, B. (2000) Asian American ethnic identification by surname. *Population Research and Policy Review,* 19**,** 283-300.

LEPPARD, D. (2005) Race chief warns of ghetto crisis. *The Sunday Times* September 18,

MASSEY, D. & DENTON, N. A. (1988) The dimensions of residential segregation. *Social Forces,* 67**,** 281–315.

THE OBSERVER (2005) Why Trevor is wrong about race ghettos. *The Observer* Sunday September 25,

ONS (2005) All Fields Postcode Directory User Guide. London Available at http://www.statistics.gov.uk/geography/downloads/AFPDUserGuide.pdf Accessed 23/11/2005

ROBINSON, G. M. (1998) *Methods and Techniques in Human Geography,* Chichester, John Wiley and Sons.

SHAW, R. F. (1960) An index of consanguinity based on the use of the surname in Spanish-speaking countries. *Journal of Heredity,* 51**,** 221-230.

SIMPSON, L. (2004) Statistics of racial segregation: measures, evidence and policy. *Urban Studies,* 41**,** 661.

SIMPSON, L. (2005) Measuring residential segregation. *Census: present and future.* Leicester.

SKERRY, P. (2000) *Counting on the Census? Race, Group Identity, and the Evasion of Politics,* Washington, Brookings Institution Press.

STILLWELL, J. & DUKE-WILLIAMS, O. (2005) Ethnic population distribution, immigration and internal migration in Britain. What evidence of linkage at the district scale. IN BSPS (Ed.) *British Society for Population Studies Annual Conference.* University of Kent at Canterbury.

U.S. CENSUS BUREAU OF THE CENSUS (1953) *Persons of Spanish Surname.,* Washington D.C., U.S. Government Printing Office.

WEBBER, R. & MATEOS, P. (forthcoming) Using Names to Classify People and Neighbourhoods by Their Cultural, Ethnic and Linguistic Origins. CASA *Working Paper* London Available at http://www.casa.ucl.ac.uk/publications/full_list.htm Accessed 10/12/2005

WONG, D. W. S. (2004) Comparing Traditional and Spatial Segregation Measures: A Spatial Scale Perspective. *Urban Geography,* 25**,** 66-82.

7. Biography

Pablo Mateos is a PhD student at CASA, University College London, as well as research associate at Camden PCT (NHS). He gained a Business Studies BA (1994), a Geography BA (2001), and an MSc in GIS (2004). His research interests are the analysis of socioeconomic and ethnic inequalities through applications of GIS and Geodemographics in Population, Social and Urban Geography.

Richard Webber is Visiting Professor at University College London. He has worked for thirty years developing neighbourhood classifications, the most famous geodemographics systems he created being ACORN and MOSAIC. He has published widely in the academic literature, is a Fellow of the Institute of Direct Marketing and is an editorial board member of the Journal of Interactive Marketing.

Paul Longley is Professor of Geographic Information Science at University College London. His publications include ten books, and over 100 refereed journal articles or contributions to edited collections. He is Editor of Computers, Environment and Urban Systems and reviews editor of Environment and Planning B.