Using People's Names to Classify Ethnicity

Pablo Mateos

Lecturer in Human Geography

Department of Geography

University College London, UK

ESRC Research Methods Festival

1st July 2008

p.mateos@ucl.ac.uk

www.casa.ucl.ac.uk/pablo

www.spatial-literacy.org

Contents

1. Context and justification
2. Names and ethnicity
3. Methodology: Onomap classification
4. Validation
5. Applications
6. Conclusions

Migration, Ethnicity & Religion

- Growing debate in Europe on issues of:
  - Migration policy
  - Ethnic relations
  - Religion & the State, specially Islam
  - National identity
- 2001-2005 Shift from multicultural to assimilationist policies
- Segregation vs. integration debate
- Fear of ‘the other’ / ‘Identity crisis’

London bomber video aired on TV

The definition of ethnicity

- ***Ethnic groups*** *are those human groups that entertain a subjective belief in their common descent because of similarities of physical type or of customs or both, or because of memories of colonization and migration, regardless of blood ties*. (Max Weber, 1922)

- A multi-dimensional concept that encompasses different aspects of identity (Bulmer, 1996):

• Kinship	• Shared territory
• Religion	• Nationality
• Language	• Physical appearance
• Culture	

- Group's affinity is defined in opposition to other groups perceived as 'different' and with whom contact is required (Eriksen, 2002)

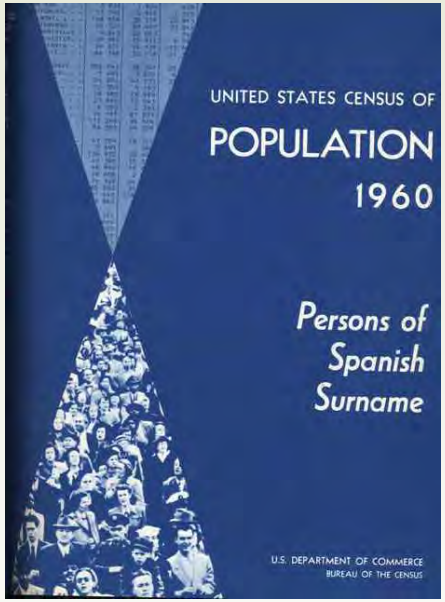
Problems with official ethnicity classifications

- Lack of sufficient granularity
  - Ethnic groups
  - Geographic dissagregation (in combination with other variables)
- Low frequency of update
- Lack of routine ethnicity data collection, or poor quality and comparability

  (surveys and admin. data sources)
- Need for complementary methodologies to study ethnic inequalities
  - (e.g. residential segregation measurement)

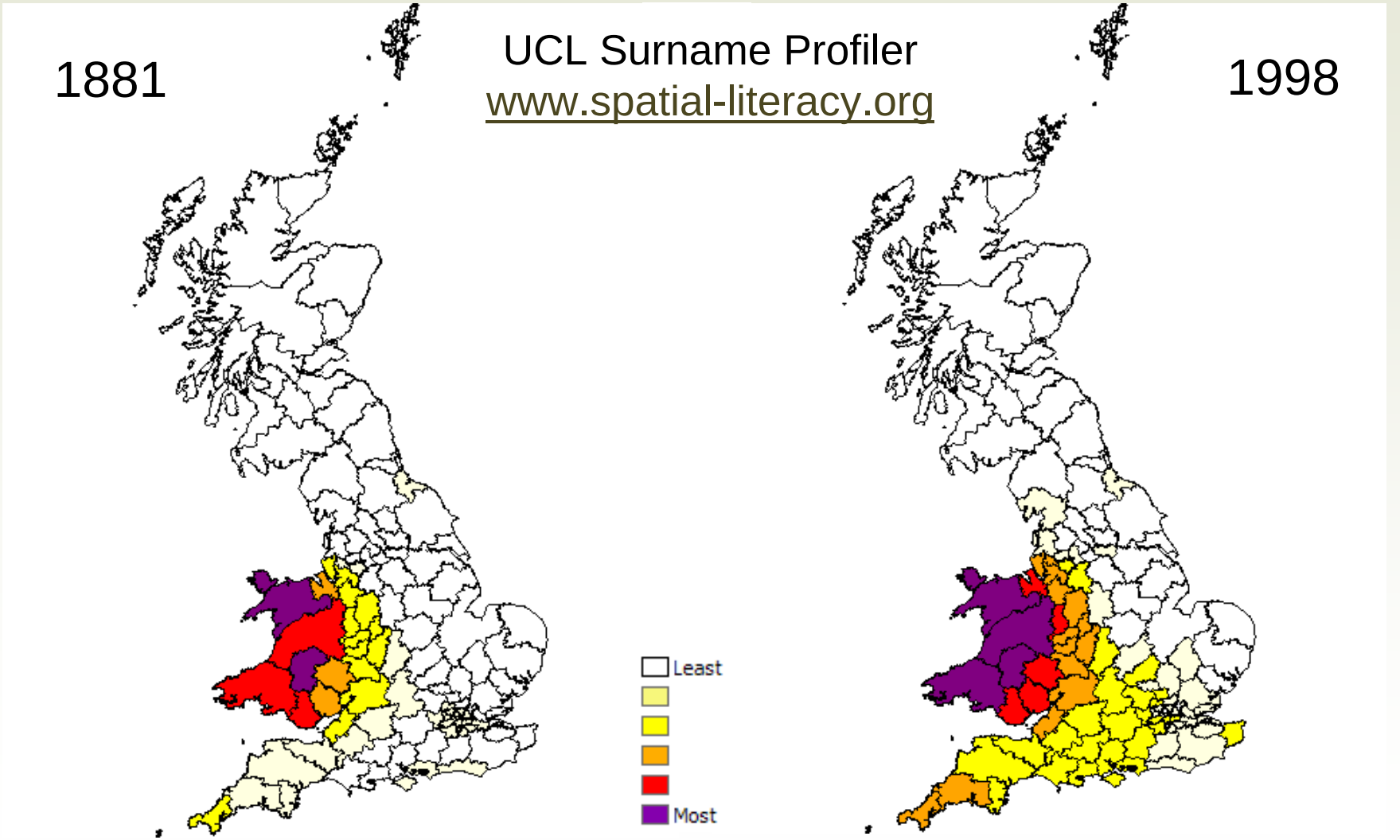
Names in Kreuzberg, Berlin

Research on names and identity



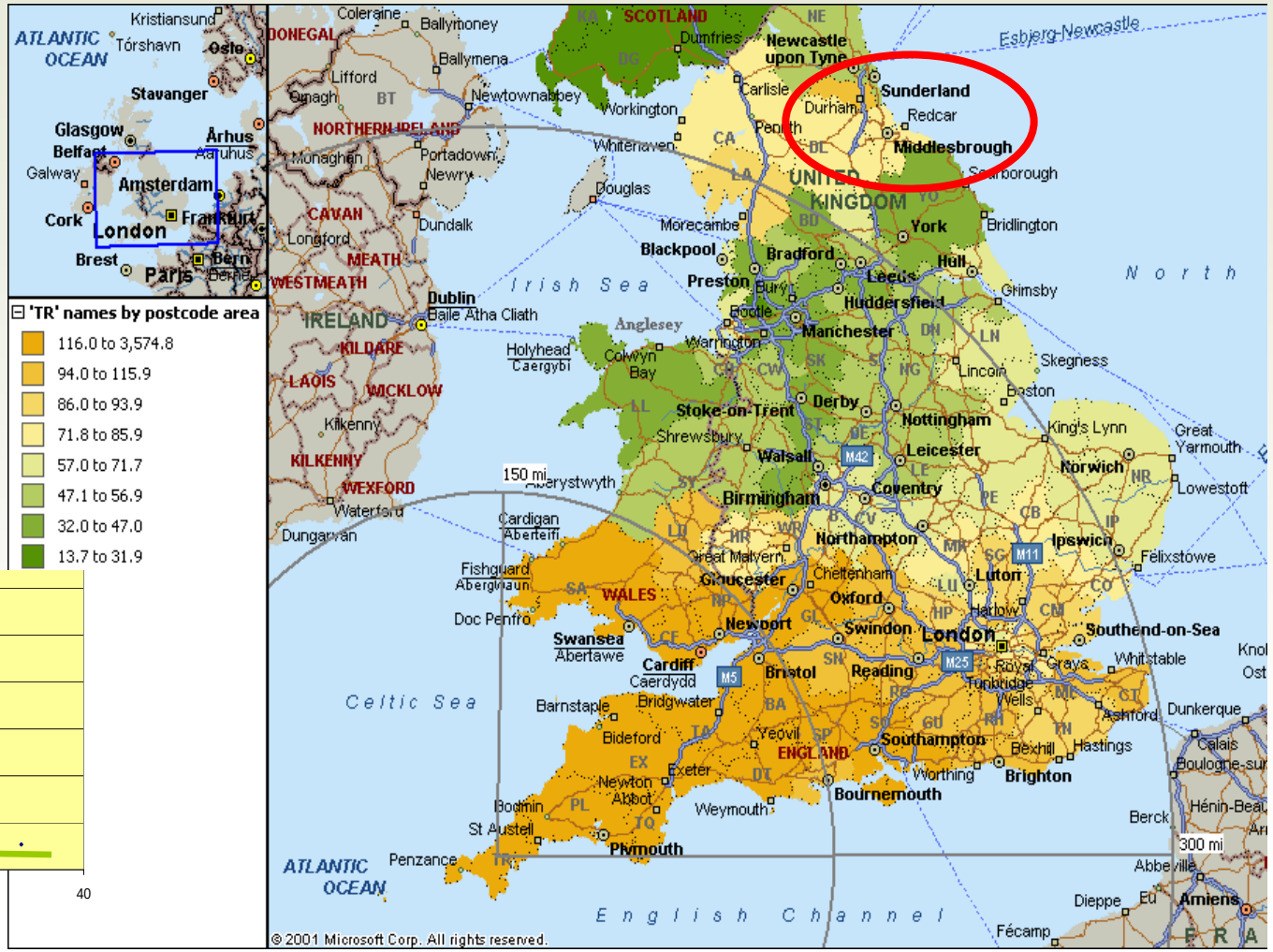
- Demography and epidemiology; subdivision of populations by ethnicity
  - US Census Hispanic names list (Passel and Word since 1950s)
  - Asian surnames in US (Lauderdale, 2004)
  - South Asian names in UK (Nam Pechan & SANGRA)
- Genetics; Population structure and geography, endogamy and gene mutations
- Economics; Name discrimination in labour, housing, and credit markets
- Geography and Sociology; cultural transmission, migration and spatial diffusion

Welsh surnames 1881-1998

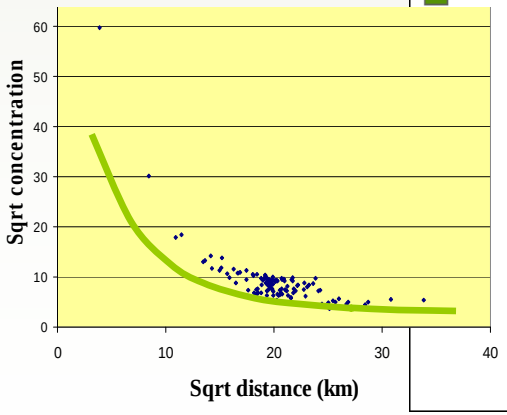


‘Cornish’ names relative frequency 1998

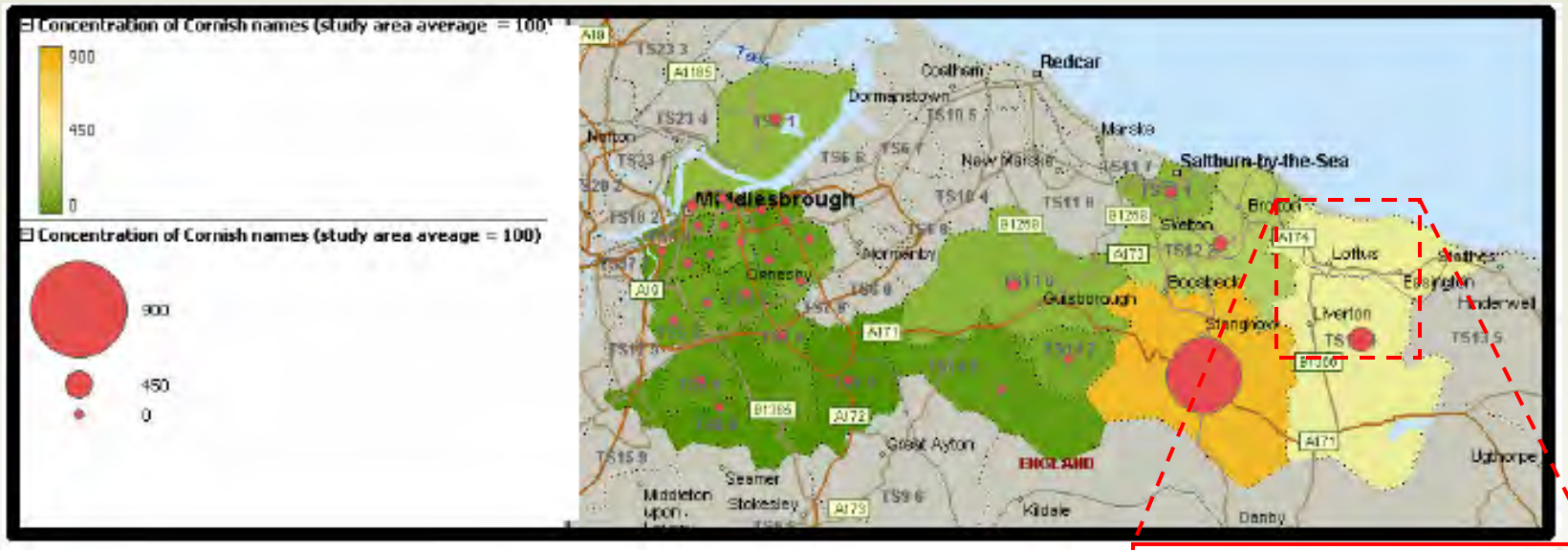
(Webber, 2005)

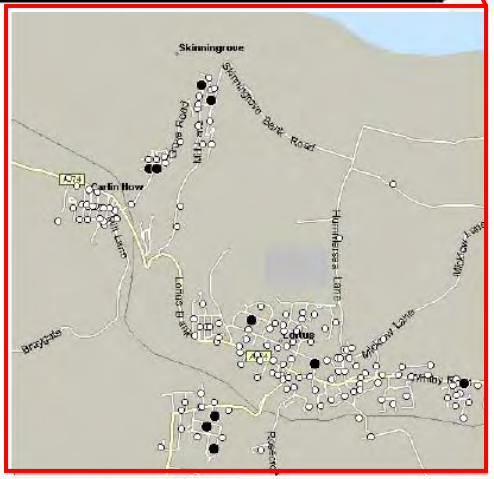


Distance Decay



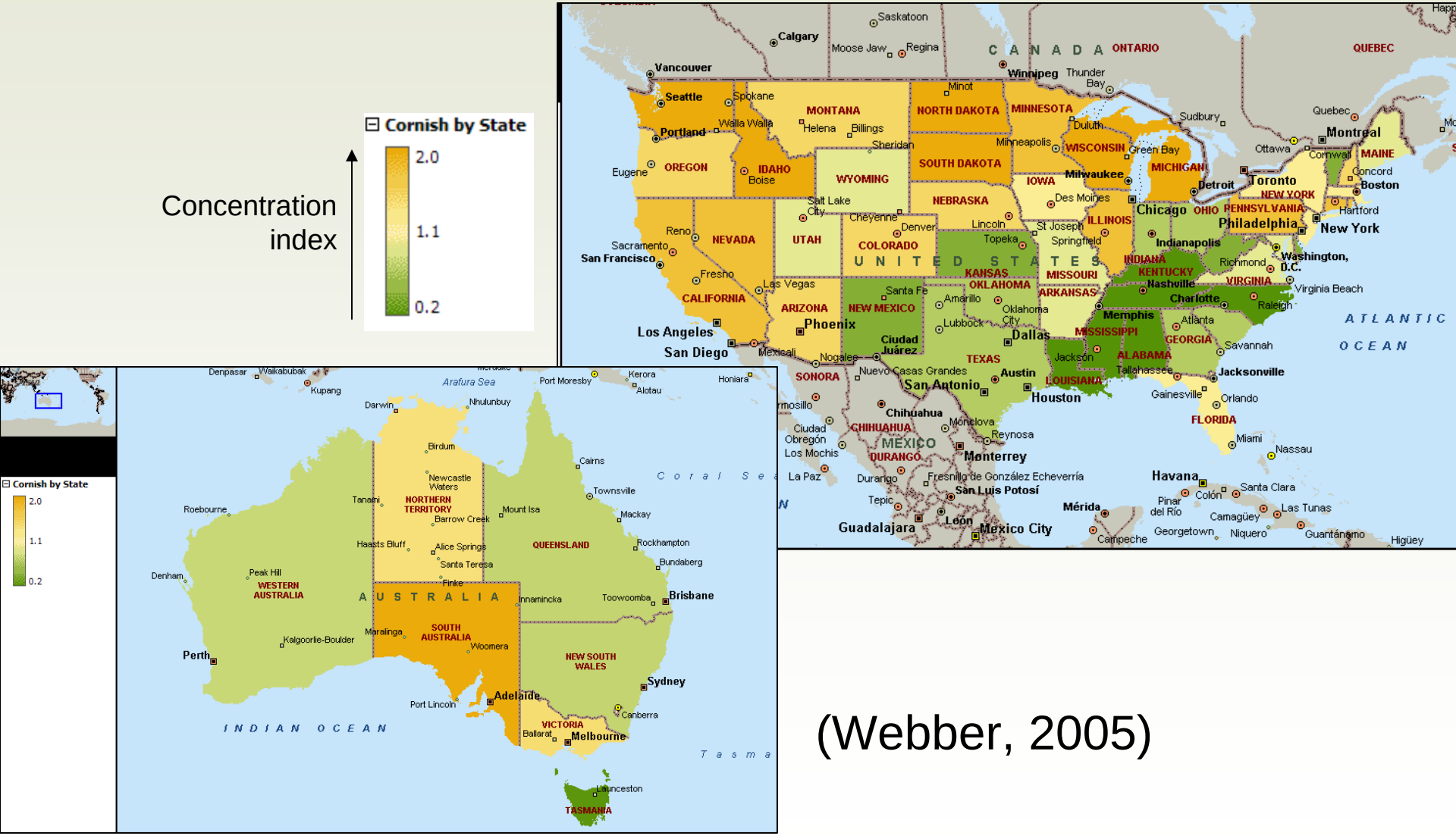
‘Cornish’ names in Middlesbrough





(Longley et al, 2007)

‘Cornish’ names & Anglosaxon diaspora



(Webber, 2005)

Decoding ethnicity from names

- Names can potentially provide information about:

Aspect	Etimology/ Onomastics	Space-time Distribution
Surname & Forename	**Language**	**Geographic Origin**
	Religion	**Migration flows**
Forename	**Gender**	***Age…***

- Review paper of name-based classifications of ethnicity; Mateos (2007) *Population, Space and Place*
- Primarily public health applications
- Main groups: Hispanic, South Asian, Chinese, and Muslim

Name-based ethnicity classifications

- 13 studies analysed in review paper

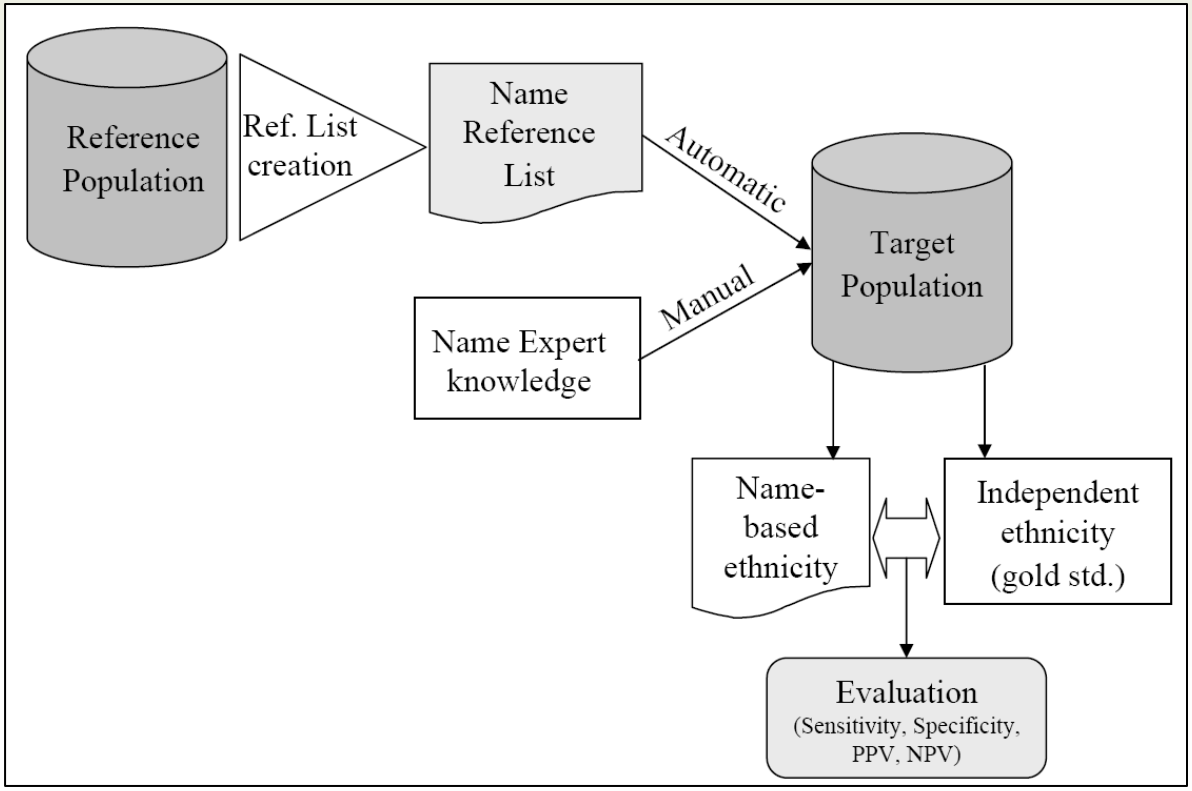
Paper reference	Geographical area of study: country and (region)	Ethnic minorities (EM) classified	Name to ethnicity assignment: Method: (Automatic Manual)	Name to ethnicity assignment: Name components: (Surname Forename Middle name)
Choi *et al.* (1993)	Canada (Ontario)	Chinese	A	S
Coldman *et al.* (1988)	Canada (British Columbia)	Chinese	A	F, S, M
Lauderdale and Kestenbaum (2000)	US (national)	Chinese, Japanese, Filipino, Korean, Indian and Vietnamese	A	S
Razum *et al.* (2001)	Germany (Rhineland Palatinate & Saarland)	Turkish	A	F, S
Word and Perkins (1996)	US (national)	Hispanic	A	S
Harding *et al.* (1999)	UK (Bradford & Coventry)	South Asian and Hindu, Muslim and Sikh	A	F, S
Cummins *et al.* (1999)	UK (Thames, Trent, W.Midlands & Yorkshire)	South Asian	A	F, S
Nanchahal *et al.* (2001)	UK (London, W. Midlands, Glasgow)	South Asian	A	F, S, M
Sheth *et al.* (1999)	Canada (national)	South Asian and Chinese	A/M	S
Martineau and White (1998)	UK (Newcastle; four general practices)	Bangladeshi, Pakistani, Indian Muslims, nonSouth Asian Muslims, Sikh, Hindu, White, Other	M	F, S and Gender
Bouwhuis and Moll (2003)	Netherlands (Rotterdam; one hospital)	Turkish, Moroccan, Surinamese	M	F, S
Nicoll *et al.* (1986)	UK (selected areas)	South Asian	M	F, S
Harland *et al.* (1997)	UK (Newcastle)	Chinese	M	F, S

Method of name to ethnicity assignment: A, Automatic; M, Manual. Name components used in the classification: S, Surname; F, Forename; M, Middle Name.

Mateos (2007)

Name to ethnicity assignment method

- Process flow to classify names by ethnicity and evaluate the method's accuracy

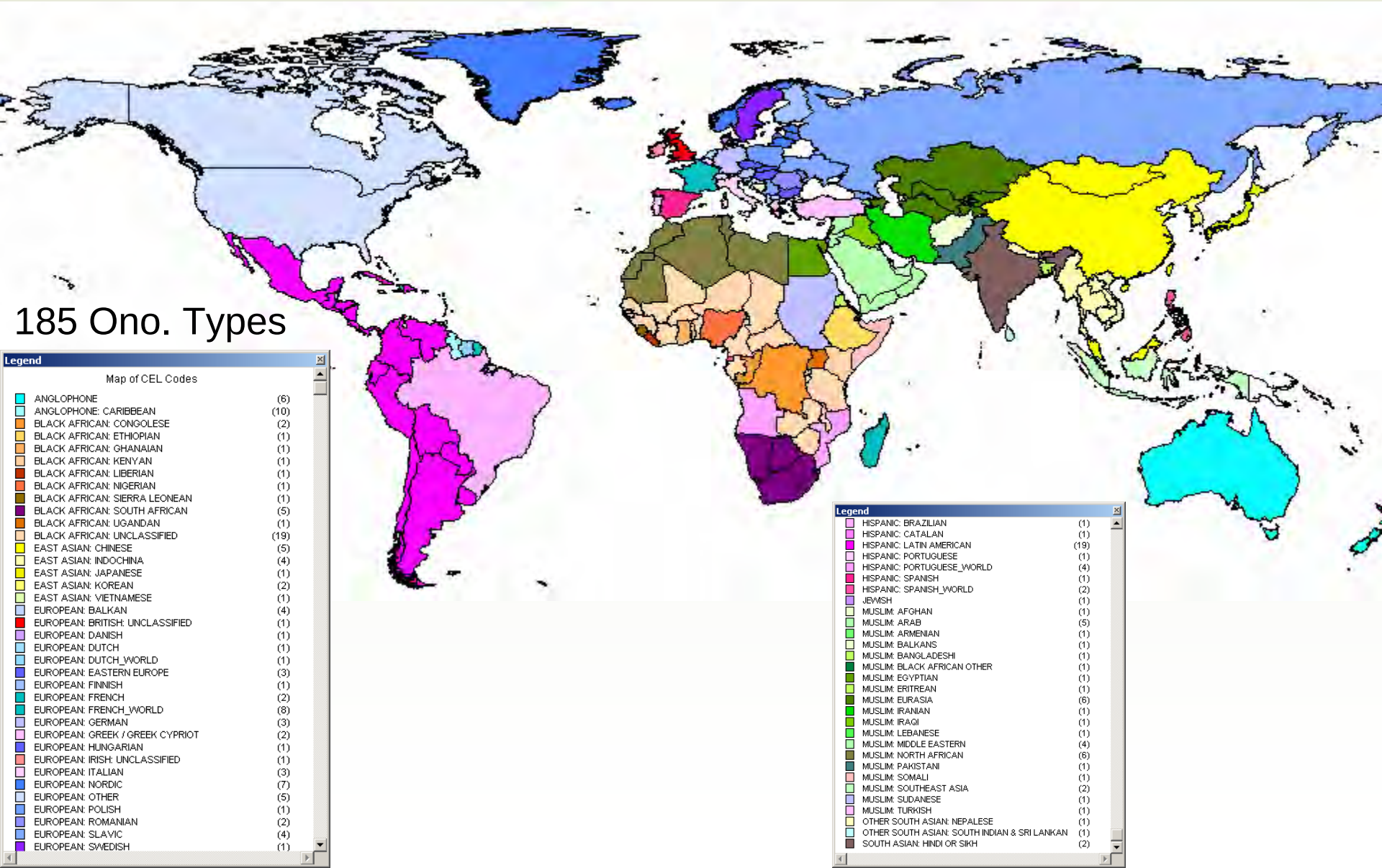


Mateos (2007)

Creating a name-based ethnicity classification

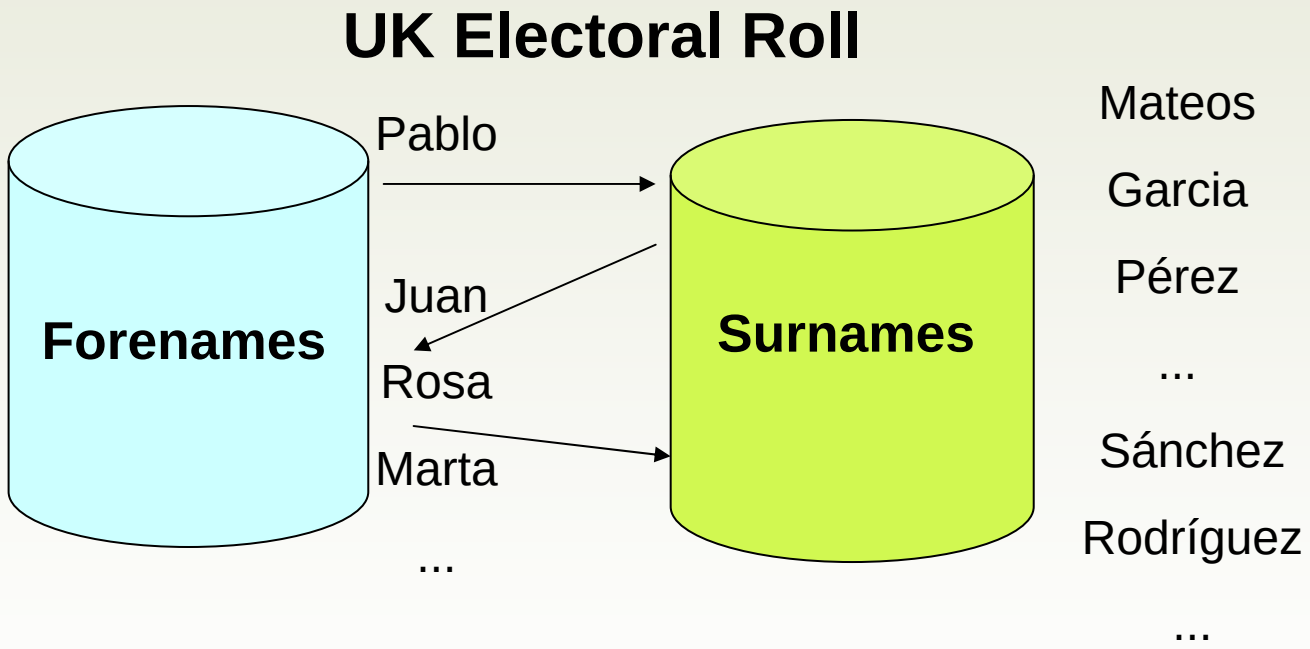
- Objective:
  - To create a classification of forenames and surnames by fine ethnic groups, covering the whole of the population in 28 countries
  - EU-21, North Am., AU & NZ, Japan, India, & Argentina
- Data sources:
  - UK Electoral Register 2001-2006 (46 million adults)
  - Telephone directories for 27 countries (300 million subscribers)
  - Covering a total population of 1 billion people
  - Individual level data (full name and address) for:
    - 10.8 million unique surnames
    - 6.5 million unique forenames
- Analysis:
  - Over 1 million names coded into “Onomap classification”
  - 185 Onomap Types, aggregated into 66 Subgroups and 15 Groups

World map of Onomap categories



Onomap classification

- **Forename-Surname clustering (based on Hanks and Tucker, 2000)**

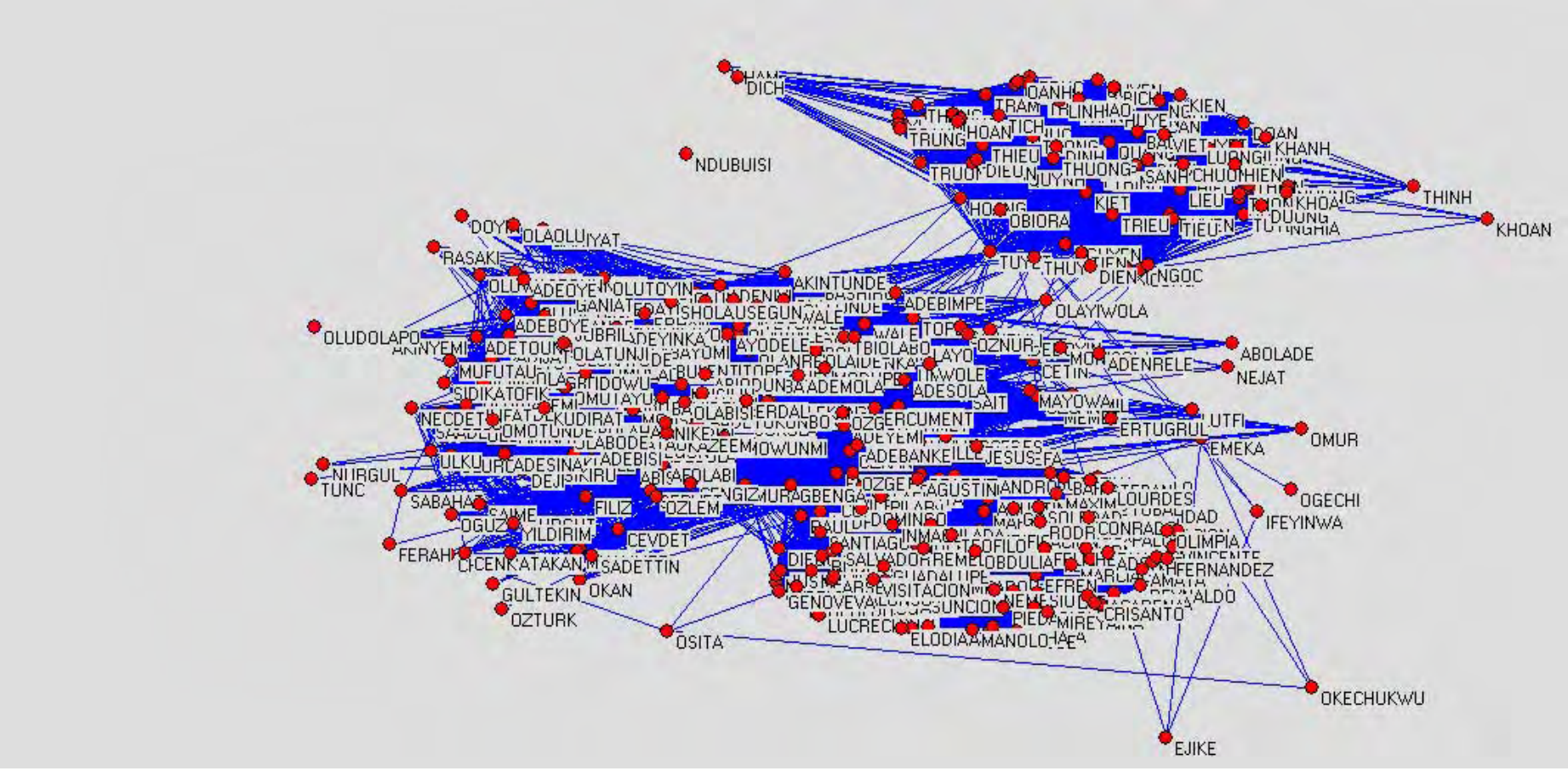


- Several iterations until self-contained cluster is exhausted
- Cluster assigned a cultural, ethnic & linguistic Onomap type
- Probability of ethnicity assigned to each name

Mateos et al (2007) CASA Working Paper 116

‘Surname distance’ between forenames

A sample of 401 forenames



Sociogram created in Pajek, selecting over 8,000 forename-surname pairs

Social networks parallel

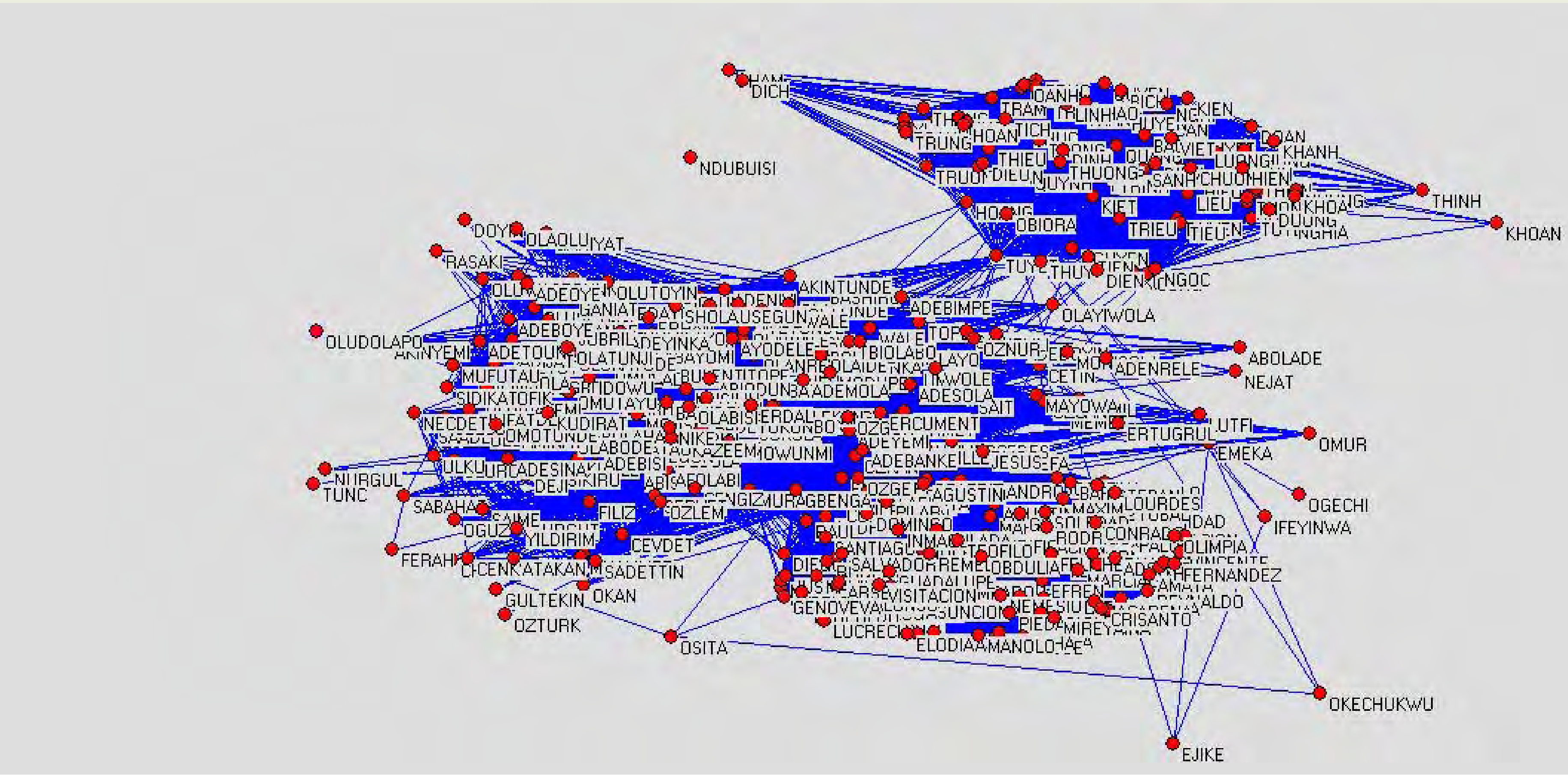
- Granovetter (1973) The Strength of Weak Ties
- Weak ties play an essential role in the diffusion of information and innovation
- Cliques of highly related names are separated by bridges or weak links (sparse links)
- Key: find those bridges to remove ‘the strength of the weak ties’

Social network analysis measures

- **Betweenness centrality**
  - First proposed by Freeman (Freeman, 1977)
  - 'The betweenness centrality of a vertex is defined as the number of shortest paths between pairs of other vertices that run through *i*' (Girvan and Newman, 2002: 7822)
- **Bi-components**
  - A bi-component is a component of minimum size k that does not contain a vertex whose deletion would increase the number of components in the network (a cut-vertex) (De Nooy et al, 2005: 141)

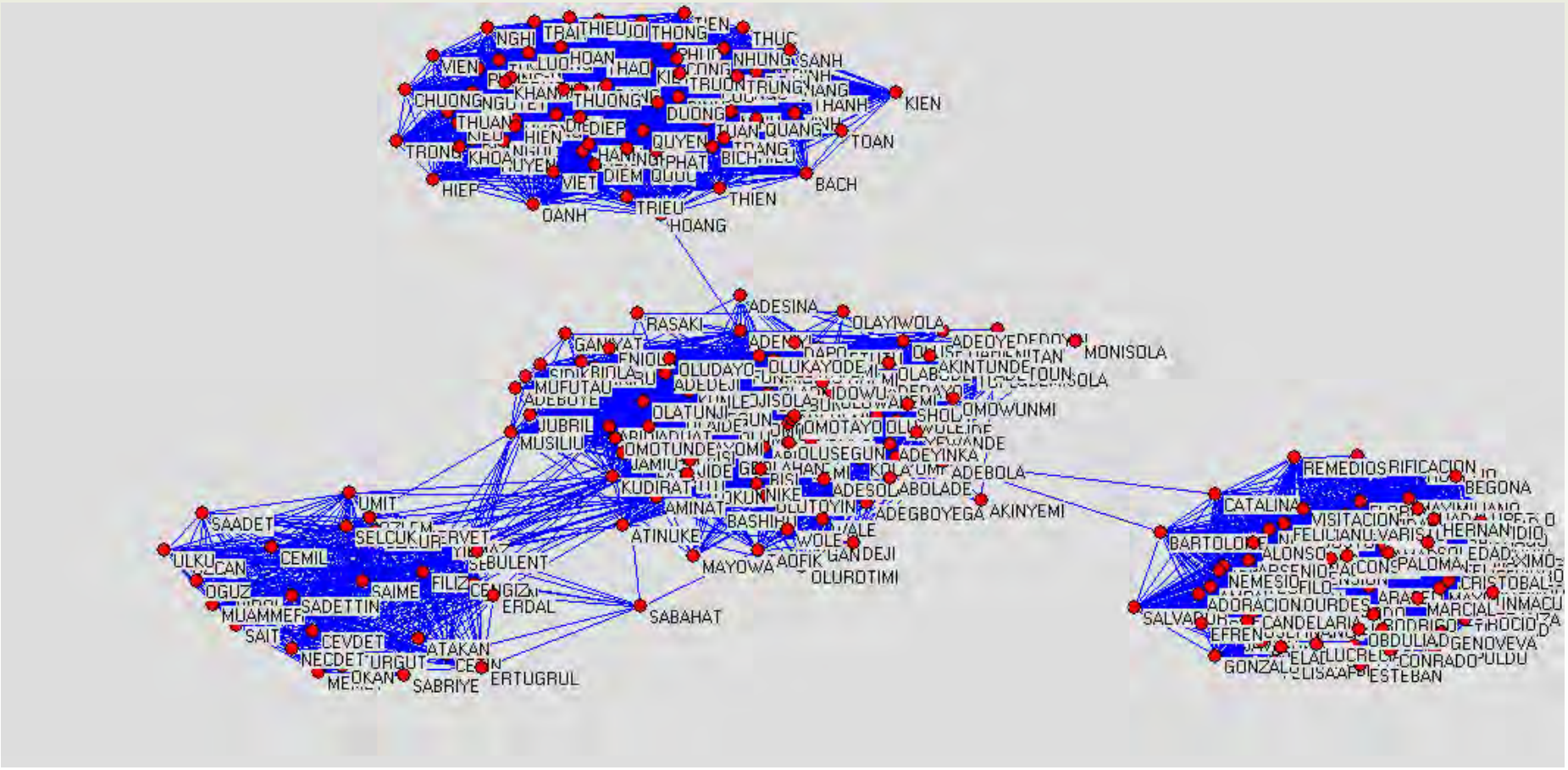
‘Surname distance’ between forenames

A sample of 401 forenames

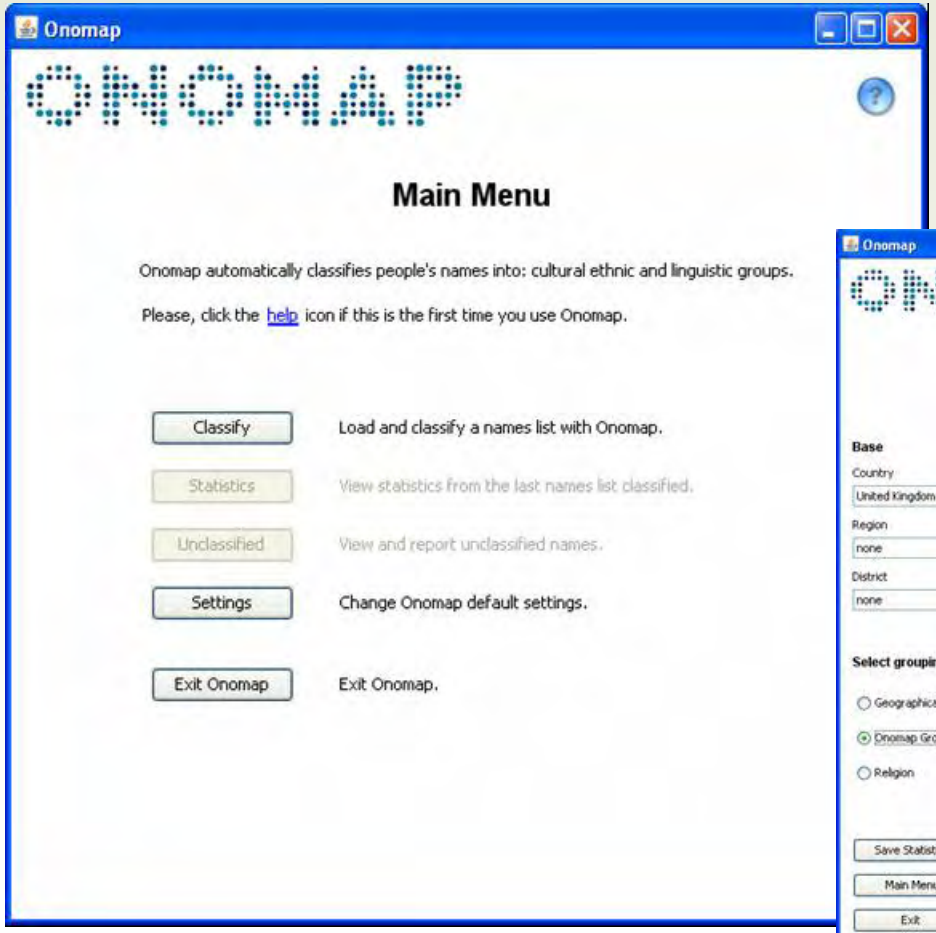


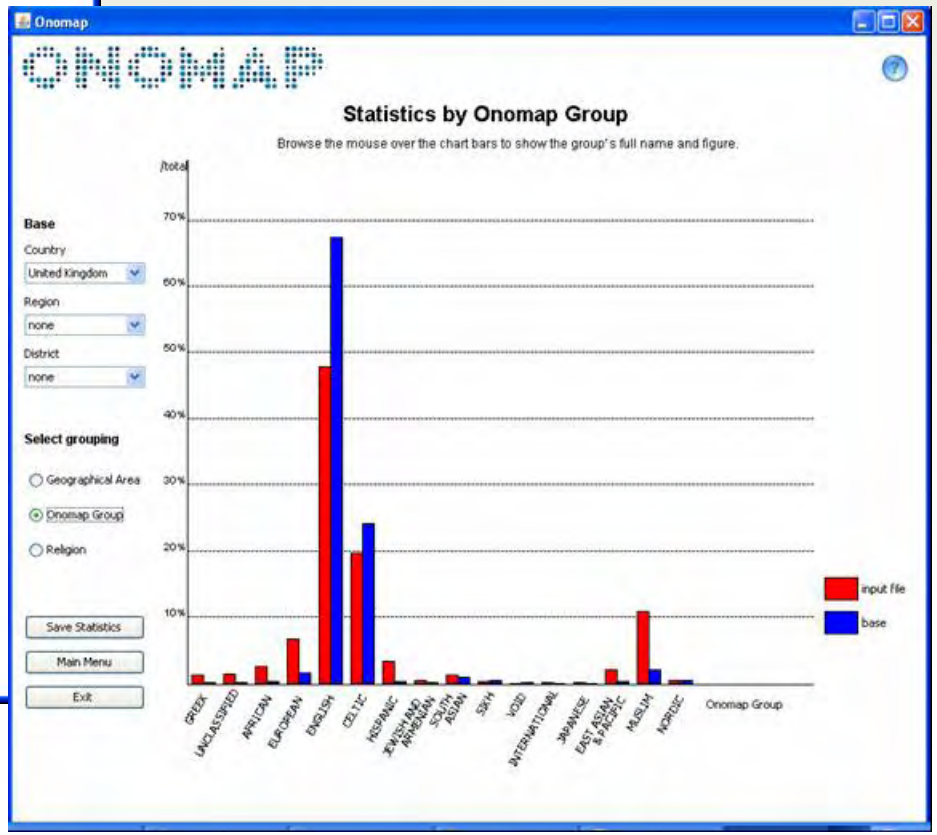
‘Surname distance’ between forenames (II) Clustered cliques

Resulting clusters from sample of 401 forenames (328)



Onomap software





Correlations Onomap vs Census (GB)

Pearson Correlation Coefficient between:

2001 Census and 2004 GB Electoral Roll classified by CEL Types

	Geographical Unit of Comparison			
Ethnic Group	***OA***	***LSOA***	***WARD***	***LA***
A) White - British	**0.88**	**0.93**	**0.93**	**0.95**
B) White - Irish	0.32	0.37	0.42	0.46
C) White - Any other White background	0.74	**0.85**	**0.88**	**0.93**
H) Asian or Asian British - Indian	**0.92**	**0.95**	**0.96**	**0.98**
J) Asian or Asian British - Pakistani	**0.90**	**0.93**	**0.93**	**0.91**
K) Asian or Asian British - Bangladeshi	**0.91**	**0.93**	**0.95**	**0.98**
L) Asian or Asian British - Any other Asian background	-0.06	0.11	0.24	0.62
M) Black or Black British - Caribbean	0.32	**0.77**	**0.91**	**0.98**
N) Black or Black British - African	**0.83**	**0.95**	**0.97**	**0.99**
R) Other Ethnic Groups - Chinese	0.65	**0.79**	**0.84**	**0.97**
S) Other Ethnic Groups - Any other ethnic group	0.38	0.66	**0.77**	**0.88**
Number of Units valid for analysis	218,037	40,883	10,072	408

*GB = England, Wales and Scotland. Values over 0.75 are highlighted in **bold***

OA = Output Area, LSOA = Super Output Area, LA = Local Authorities

Evaluation at the individual level - HES-

		Actual Ethnicity from HES data										
	Predicted by CEL	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Total**
0	White	**150,574**	7,971	4,468	2,535	595	68	160	488	17,383	73,920	258,162
1	Black - Caribbean	92	**226**	21	32	3				69	197	640
2	Black - African	857	283	**5,996**	698	53	14	41	23	1,695	4,716	14,376
3	Black - Other											0
4	Indian	1,066	96	562	125	**2,184**	85	171	30	1,679	3,503	9,501
5	Pakistani	856	60	1,736	306	690	**861**	2,390	17	2,507	4,625	14,048
6	Bangladeshi	284	30	373	122	687	194	**6,086**	5	1,174	3,777	12,732
7	Chinese	227	39	72	21	11	2	7	**1,473**	531	1,088	3,471
8	Any other ethnic group	3,811	111	990	228	202	112	280	358	**5,858**	5,747	17,697
9	*Unclassified*	3,364	328	1,706	322	164	32	107	47	2,199	**4,079**	12,348
	Total	161,131	9,144	15,924	4,389	4,589	1,368	9,242	2,441	33,095	101,652	**342,975**

	1991 Census Categories	***Sensitivity***	***Specificity***	***PPV***	***NPV***
0	White	**0.93 - 0.98**	0.58 - 0.62	**0.82 - 0.90**	**0.82 - 0.89**
1	Black - Caribbean	0.02 - 0.03	**1.00 - 1.00**	0.51 - 0.62	**0.96 - 0.96**
2	Black - African	0.38 - 0.45	**0.98 - 0.99**	0.62 - **0.76**	**0.96 - 0.96**
3	Black - Other	n/a	n/a	n/a	n/a
4	Indian	0.48 - 0.52	**0.98 - 0.99**	0.36 - 0.50	**0.99 - 0.99**
5	Pakistani	0.63 - **0.70**	**0.96 - 0.97**	0.09 - 0.12	**1.00 - 1.00**
6	Bangladeshi	0.66 - 0.69	**0.99 - 0.99**	0.68 - **0.79**	**0.99 - 0.98**
7	Chinese	0.60 - **0.73**	**1.00 - 1.00**	0.62 - **0.80**	**1.00 - 1.00**
8	Any other ethnic group	0.18	**0.97**	0.49	**0.88**
9	Not Given	n/a	n/a	n/a	n/a

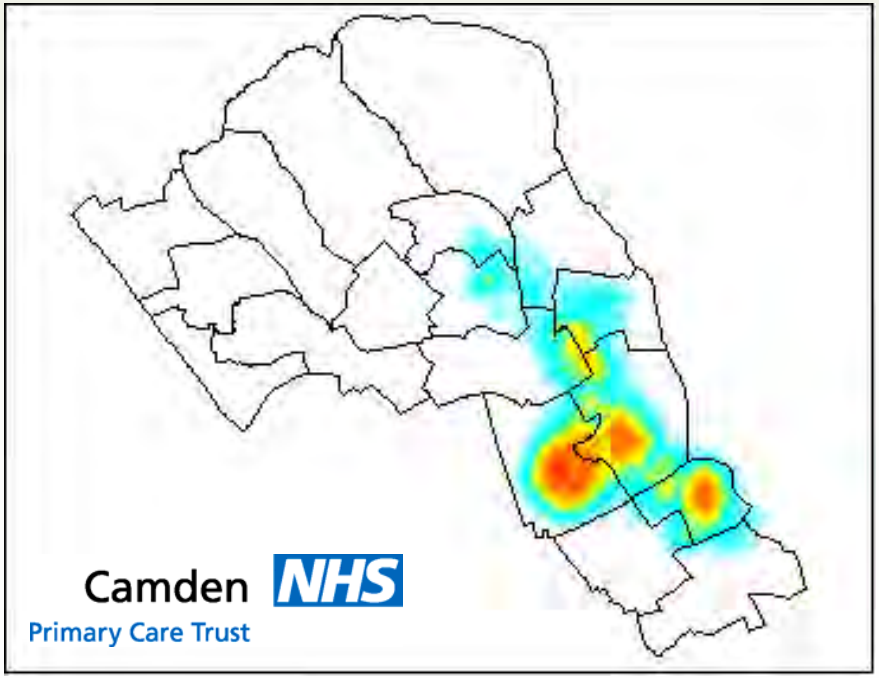
Applications

- Cancer study (5 m. patients) – LSHTM
- Public Health
  - PCTs (Camden, Islington, Southwark)
  - University of Edinburgh & GROS (Onomap evaluation)
  - University of Essex
- Political party representation
  - ANU, Australia; Princeton Univ.
- Residential Segregation
  - Univ. Paris 8

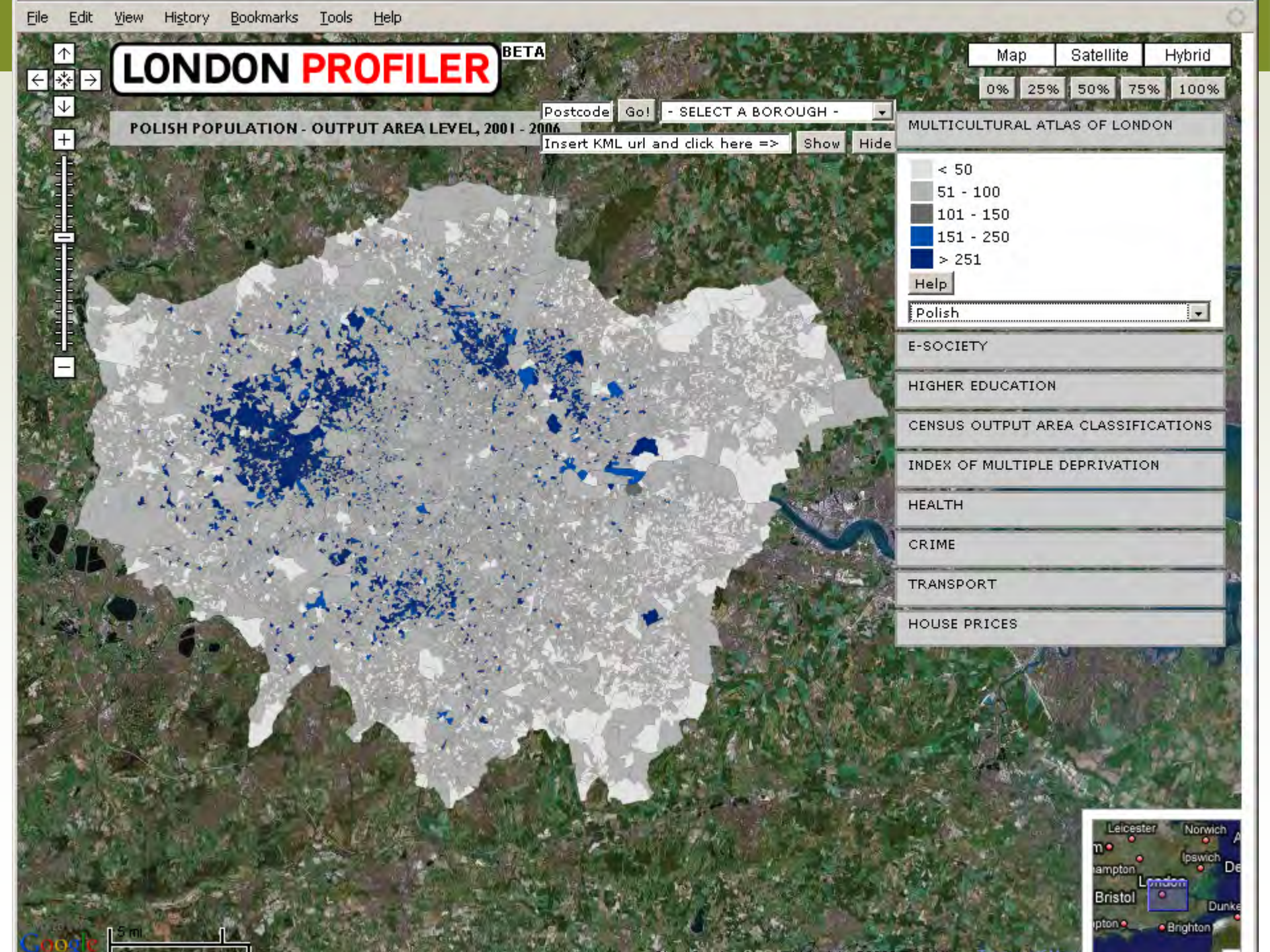
Applications in Public Health

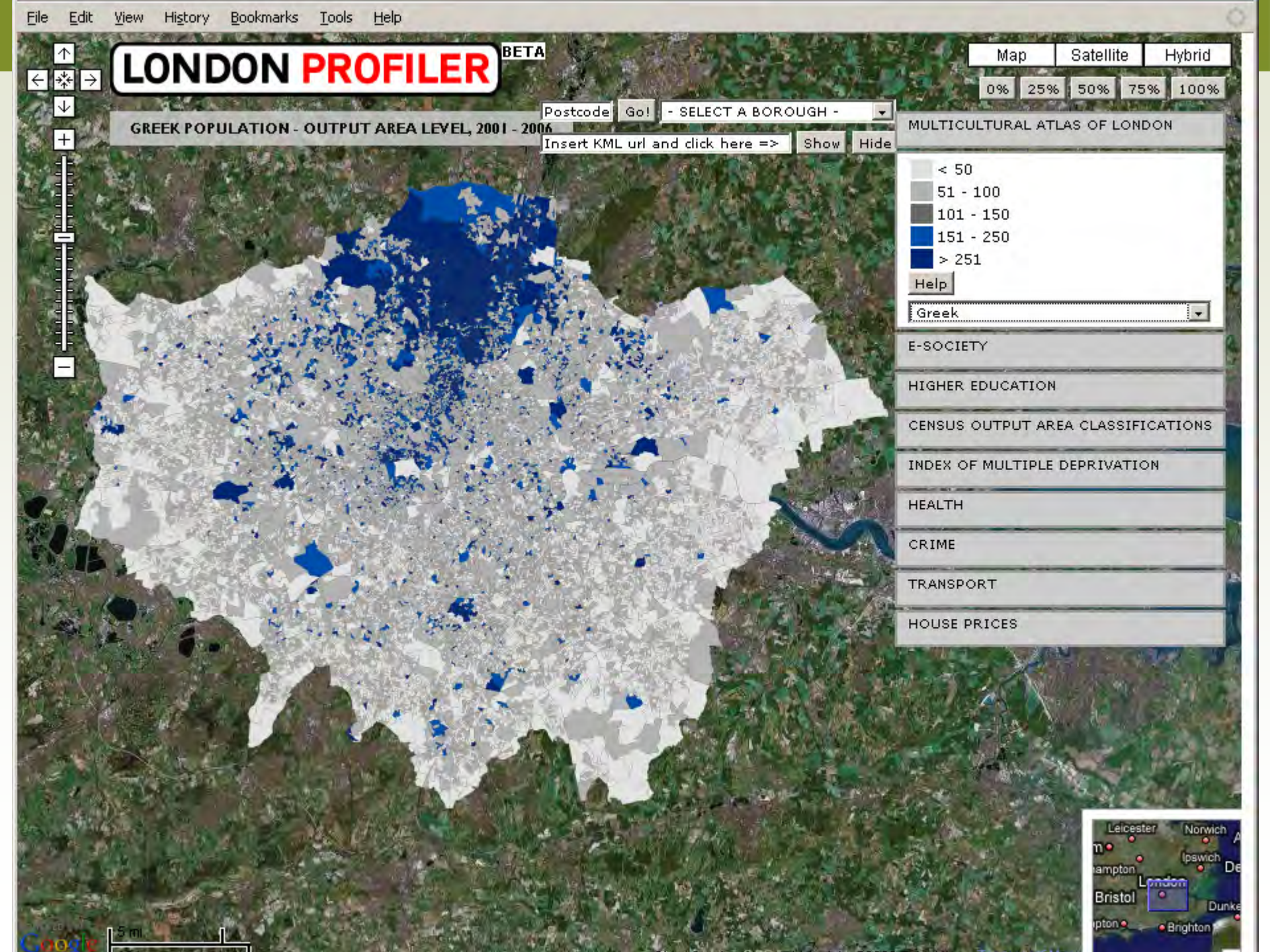
- Reducing the number of non-responders to breast screening (Camden PCT, London)

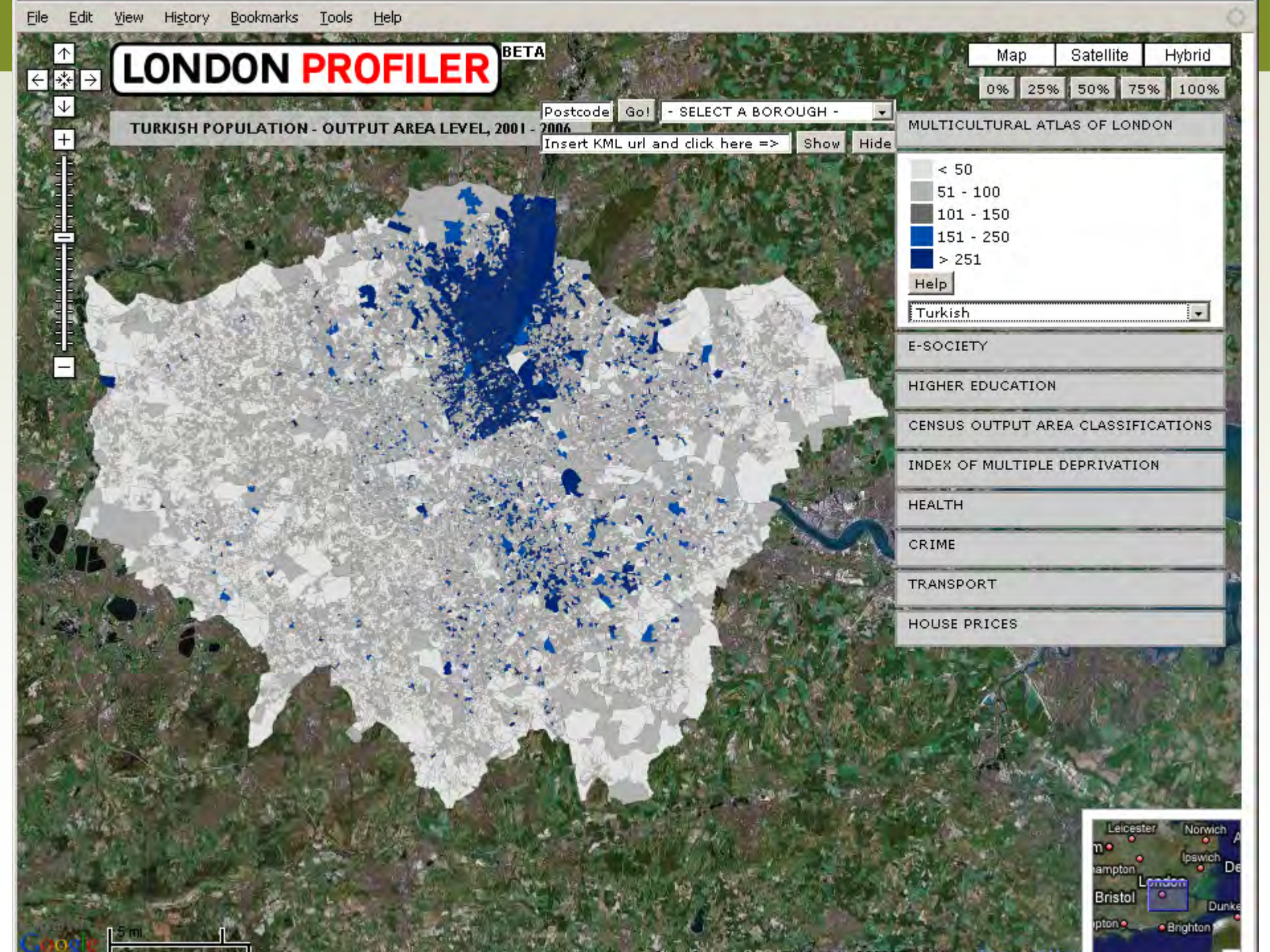
Concentration of non-screened women with Bangladeshi names

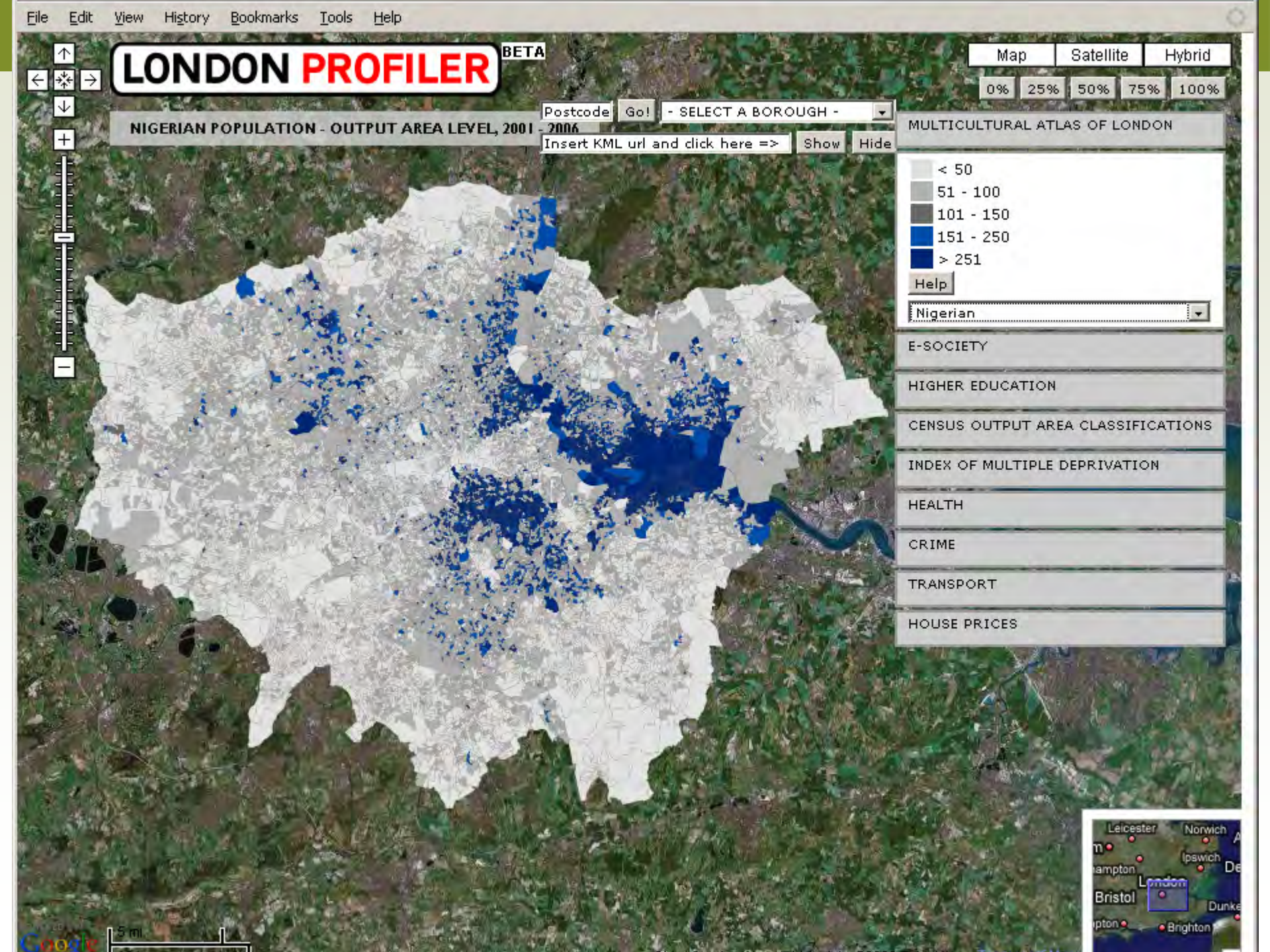


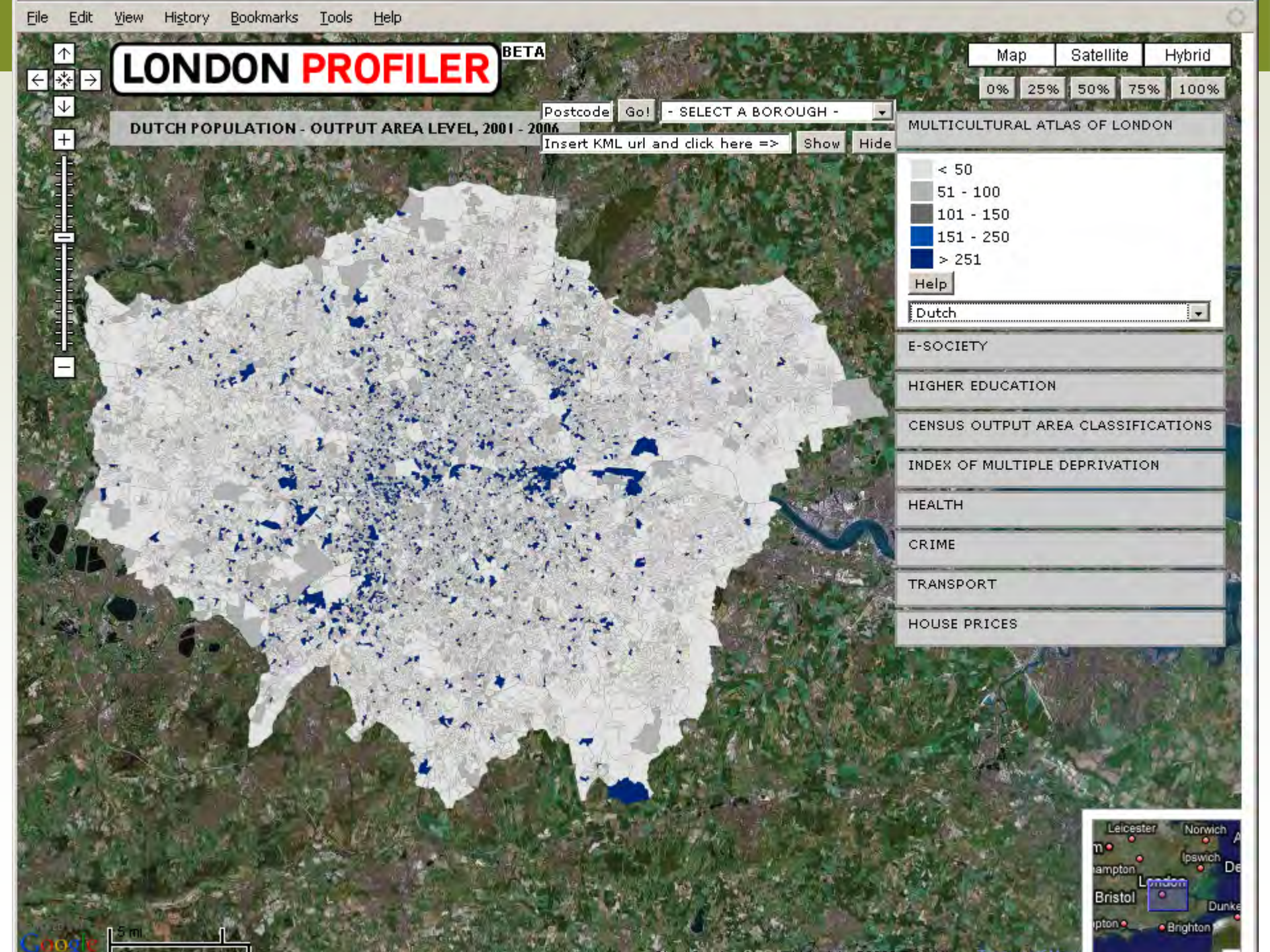
Jones and Mateos (2005)

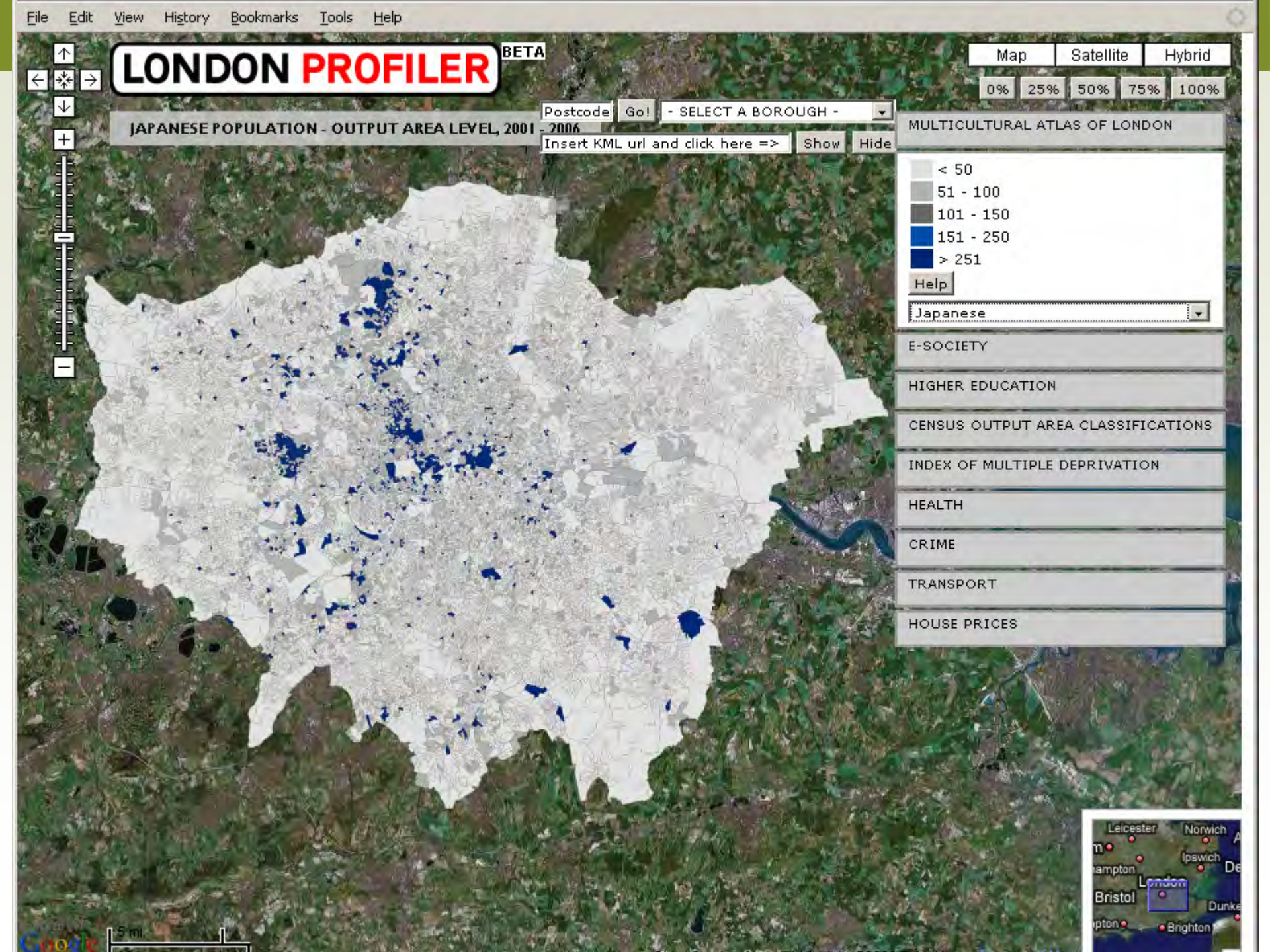

File Edit View History Bookmarks Tools Help
LONDON PROFILER
BETA
Map
Satellite
Hybrid
0%
25%
50%
75%
100%
Postcode
Go!
- SELECT A BOROUGH -
POLISH POPULATION - OUTPUT AREA LEVEL, 2001 - 2006
Insert KML url and click here =>
Show
Hide
MULTICULTURAL ATLAS OF LONDON
< 50
51 - 100
101 - 150
151 - 250
> 251
Help
Polish
E-SOCIETY
HIGHER EDUCATION
CENSUS OUTPUT AREA CLASSIFICATIONS
INDEX OF MULTIPLE DEPRIVATION
HEALTH
CRIME
TRANSPORT
HOUSE PRICES
Leicester
Norwich
Ipswich
London
Bristol
Brighton
5 mi


File Edit View History Bookmarks Tools Help
LONDON PROFILER
BETA
Map
Satellite
Hybrid
0%
25%
50%
75%
100%
Postcode
Go!
- SELECT A BOROUGH -
GREEK POPULATION - OUTPUT AREA LEVEL, 2001 - 2006
Insert KML url and click here =>
Show
Hide
MULTICULTURAL ATLAS OF LONDON
< 50
51 - 100
101 - 150
151 - 250
> 251
Help
Greek
E-SOCIETY
HIGHER EDUCATION
CENSUS OUTPUT AREA CLASSIFICATIONS
INDEX OF MULTIPLE DEPRIVATION
HEALTH
CRIME
TRANSPORT
HOUSE PRICES
Leicester
Norwich
Ipswich
London
Bristol
Brighton
5 mi


File Edit View History Bookmarks Tools Help
LONDON PROFILER
BETA
Map
Satellite
Hybrid
0%
25%
50%
75%
100%
Postcode
Go!
- SELECT A BOROUGH -
TURKISH POPULATION - OUTPUT AREA LEVEL, 2001 - 2006
Insert KML url and click here =>
Show
Hide
MULTICULTURAL ATLAS OF LONDON
< 50
51 - 100
101 - 150
151 - 250
> 251
Help
Turkish
E-SOCIETY
HIGHER EDUCATION
CENSUS OUTPUT AREA CLASSIFICATIONS
INDEX OF MULTIPLE DEPRIVATION
HEALTH
CRIME
TRANSPORT
HOUSE PRICES
Leicester
Norwich
Ipswich
London
Bristol
Brighton
5 mi

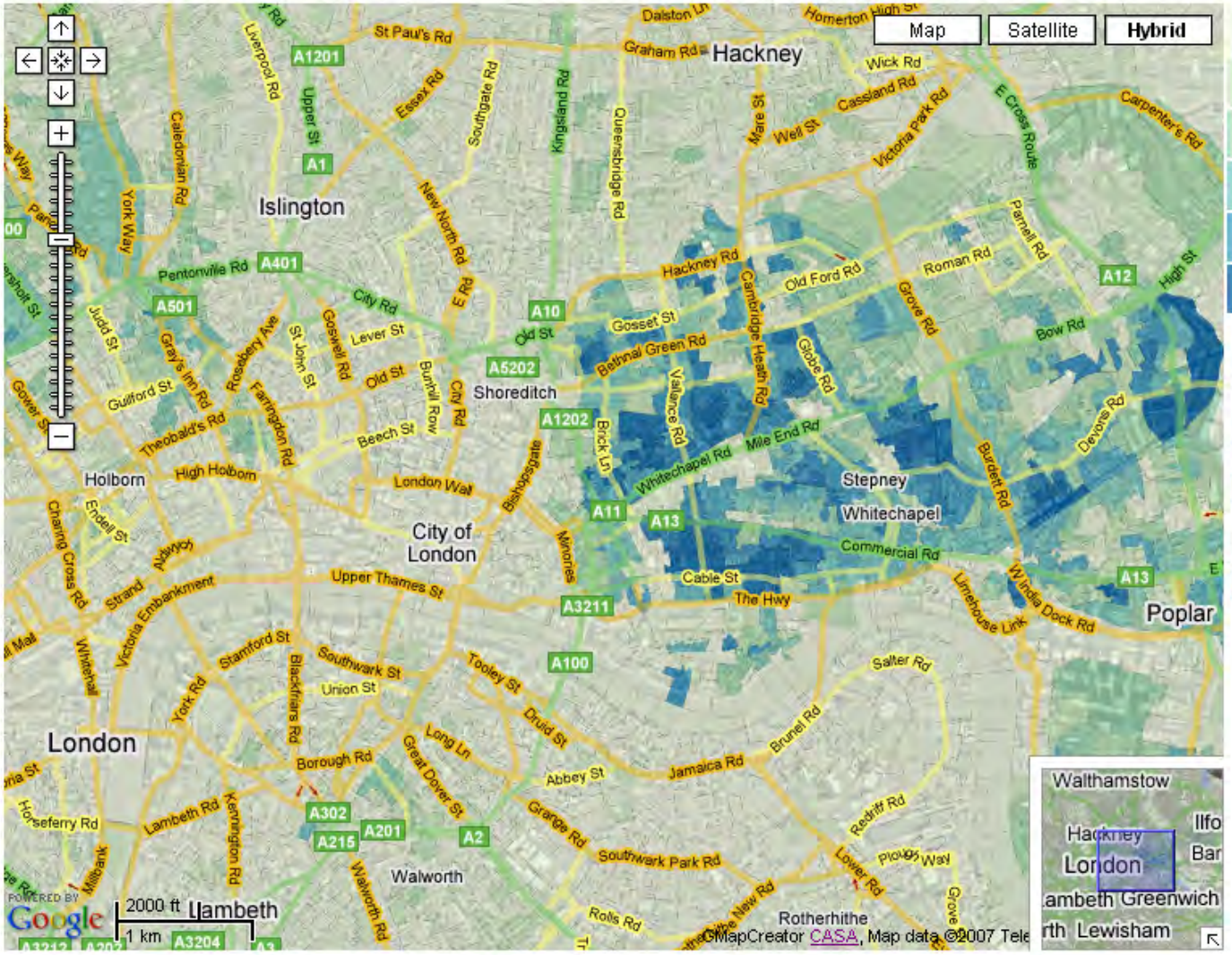

File Edit View History Bookmarks Tools Help
LONDON PROFILER
BETA
Map
Satellite
Hybrid
0%
25%
50%
75%
100%
Postcode
Go!
- SELECT A BOROUGH -
NIGERIAN POPULATION - OUTPUT AREA LEVEL, 2001 - 2006
Insert KML url and click here =>
Show
Hide
MULTICULTURAL ATLAS OF LONDON
< 50
51 - 100
101 - 150
151 - 250
> 251
Help
Nigerian
E-SOCIETY
HIGHER EDUCATION
CENSUS OUTPUT AREA CLASSIFICATIONS
INDEX OF MULTIPLE DEPRIVATION
HEALTH
CRIME
TRANSPORT
HOUSE PRICES
Leicester
Norwich
Ipswich
London
Bristol
Brighton
5 mi


File Edit View History Bookmarks Tools Help
LONDON PROFILER
BETA
Map
Satellite
Hybrid
0%
25%
50%
75%
100%
Postcode
Go!
- SELECT A BOROUGH -
DUTCH POPULATION - OUTPUT AREA LEVEL, 2001 - 2006
Insert KML url and click here =>
Show
Hide
MULTICULTURAL ATLAS OF LONDON
< 50
51 - 100
101 - 150
151 - 250
> 251
Help
Dutch
E-SOCIETY
HIGHER EDUCATION
CENSUS OUTPUT AREA CLASSIFICATIONS
INDEX OF MULTIPLE DEPRIVATION
HEALTH
CRIME
TRANSPORT
HOUSE PRICES
Leicester
Norwich
Ipswich
London
Bristol
Brighton
5 mi


File Edit View History Bookmarks Tools Help
LONDON PROFILER
BETA
Map
Satellite
Hybrid
0%
25%
50%
75%
100%
Postcode
Go!
- SELECT A BOROUGH -
JAPANESE POPULATION - OUTPUT AREA LEVEL, 2001 - 2006
Insert KML url and click here =>
Show
Hide
MULTICULTURAL ATLAS OF LONDON
< 50
51 - 100
101 - 150
151 - 250
> 251
Help
Japanese
E-SOCIETY
HIGHER EDUCATION
CENSUS OUTPUT AREA CLASSIFICATIONS
INDEX OF MULTIPLE DEPRIVATION
HEALTH
CRIME
TRANSPORT
HOUSE PRICES
5 mi
Leicester
Norwich
Ipswich
London
Bristol
Brighton

file:///F:/Pablo/PhD/MulticulturalLondon/Multicultural_London.html

Bangladeshi population in London - Output Area level, 2004



0 - 5%
5% - 10%
10% - 20%
20% - 30%
30% - 50%
> 50%

Cultural and ethnic group:

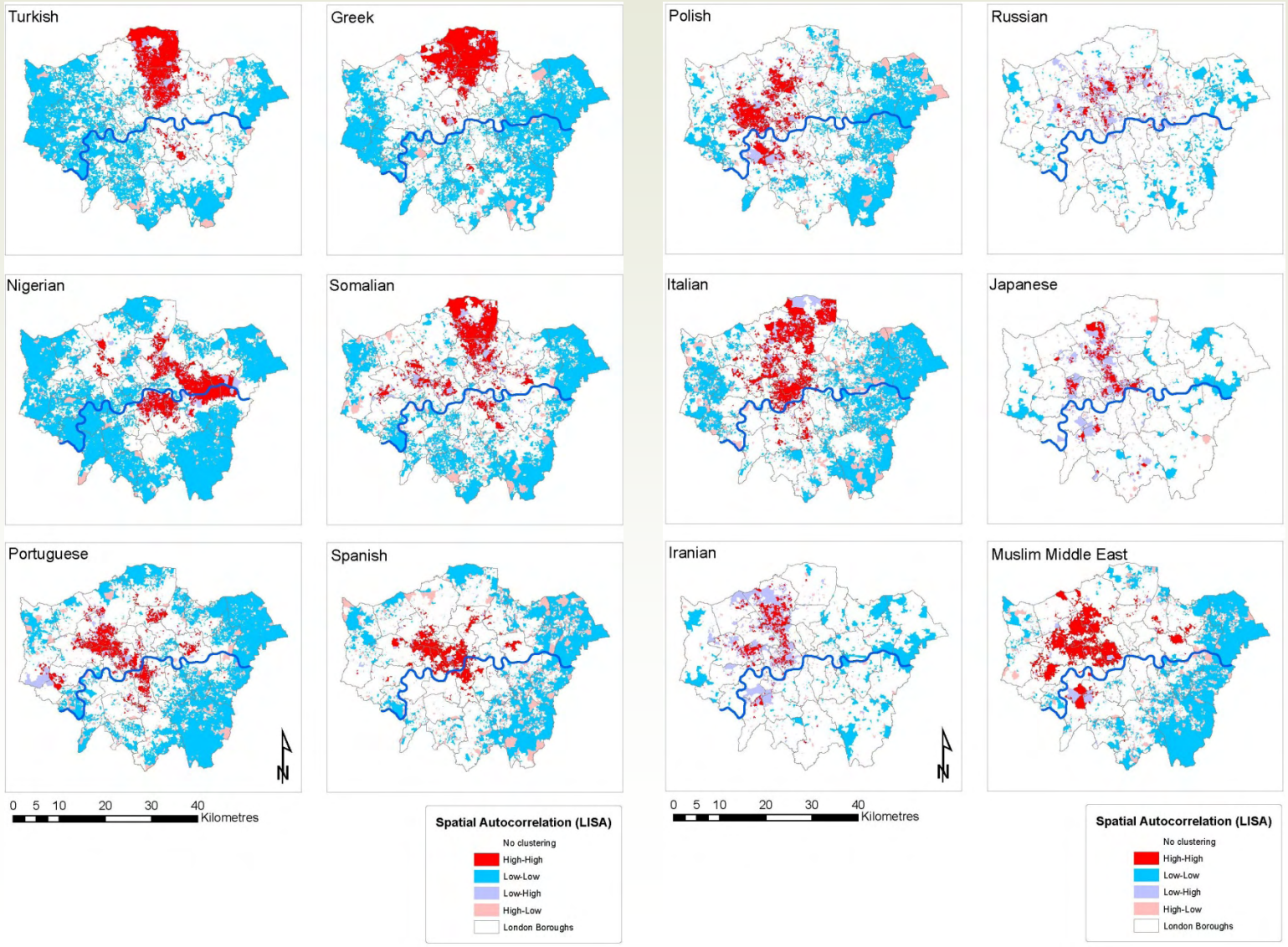
Bangladeshi

Hide thematic map

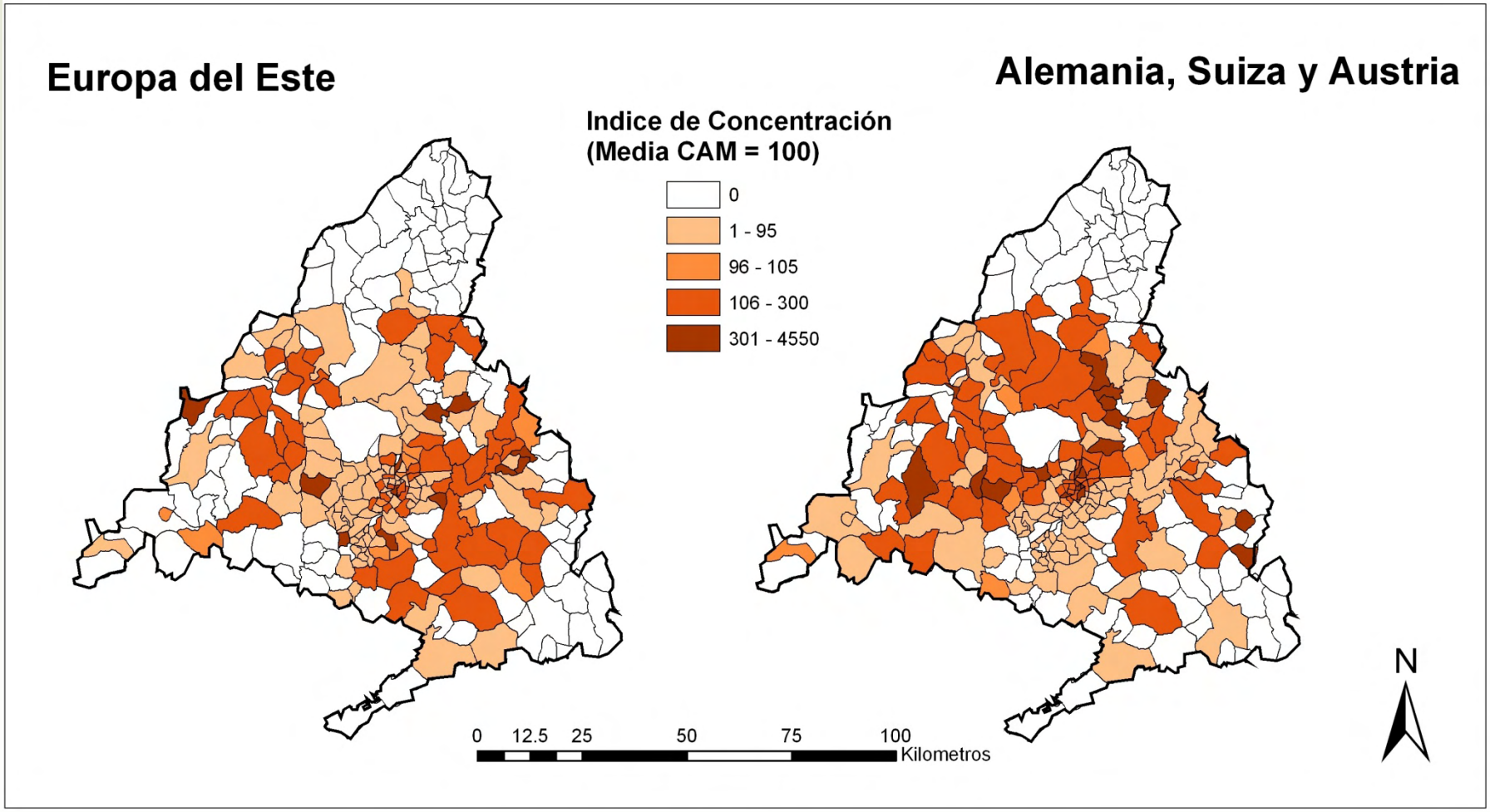
Set transparency:

25% 50% 75% 100%

Spatial Autocorrelation (LISA) by Onomap type



Residential segregation in Comunidad de Madrid



$$IC_i = \frac{\%CP}{\%PT} x100$$

%CP = % del *grupo Onomap i* en la población del código postal
%PT = % del *grupo Onomap i* en la población total.

Issues of aggregation and geographical scale

- Most and least segregated Onomap units; Index of Dissimilarity (ID) at Ward level (avg. size 10,000 people)

Top 20

Onomap Subgroup	Total Pop.	% Pop	ID
CONGOLESE	598	0.01%	0.69
UGANDAN	812	0.02%	0.65
KOREAN	1,139	0.02%	0.62
ETHIOPIAN	918	0.02%	0.61
ERITREAN	1,053	0.02%	0.59
ARMENIAN	2,436	0.05%	0.59
SIERRA LEONEAN	3,854	0.08%	0.58
SIKH	83,968	1.68%	0.58
MUSLIM STANS	1,155	0.02%	0.56
ALBANIAN	1,908	0.04%	0.54
BALTIC	1,061	0.02%	0.54
ROMANIAN	1,085	0.02%	0.54
MUSLIM	2,335	0.05%	0.54
BANGLADESHI	72,829	1.45%	0.53
UKRANIAN	1,629	0.03%	0.53
VIETNAMESE	8,415	0.17%	0.53
BLACK SOUTH AFRICAN	2,161	0.04%	0.51
SRI LANKAN	39,269	0.78%	0.50
JAPANESE	3,469	0.07%	0.50
MALAYSIAN	891	0.02%	0.50

Bottom 20

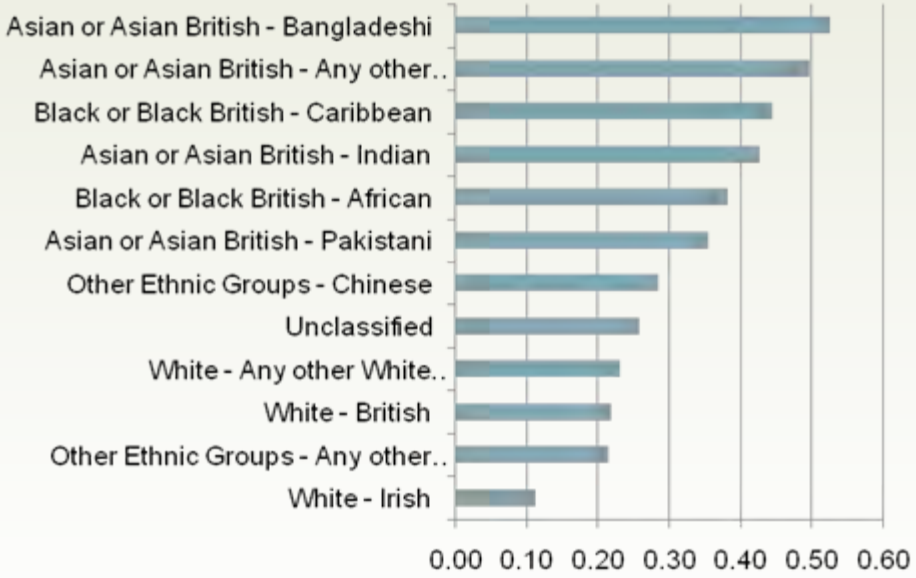
Onomap Subgroup	Total Pop.	% Pop	ID
WELSH	222,429	4.44%	0.09
SCOTTISH	323,847	6.47%	0.10
IRISH	414,038	8.27%	0.11
ENGLISH	2,876,980	57.47%	0.18
NORWEGIAN	24,927	0.50%	0.23
POLISH	33,270	0.66%	0.23
PORTUGUESE	44,780	0.89%	0.24
SPANISH	44,679	0.89%	0.24
FRENCH	40,264	0.80%	0.24
ITALIAN	71,967	1.44%	0.25
UNKNOWN NAME	101,261	2.02%	0.26
PAKISTANI KASHMIR	32,061	0.64%	0.27
MUSLIM MIDDLE EAST	48,114	0.96%	0.27
HONG KONGESE	35,609	0.71%	0.29
EUROPEAN OTHER	9,091	0.18%	0.29
GERMAN	33,264	0.66%	0.30
INDIA NORTH	31,888	0.64%	0.31
MUSLIM SOUTH ASIAN	11,380	0.23%	0.33
INTERNATIONAL	6,214	0.12%	0.34
BALKAN	9,035	0.18%	0.35

- Mateos, *et al* (forthcoming) *Journal of Ethnic and Migration Studies*

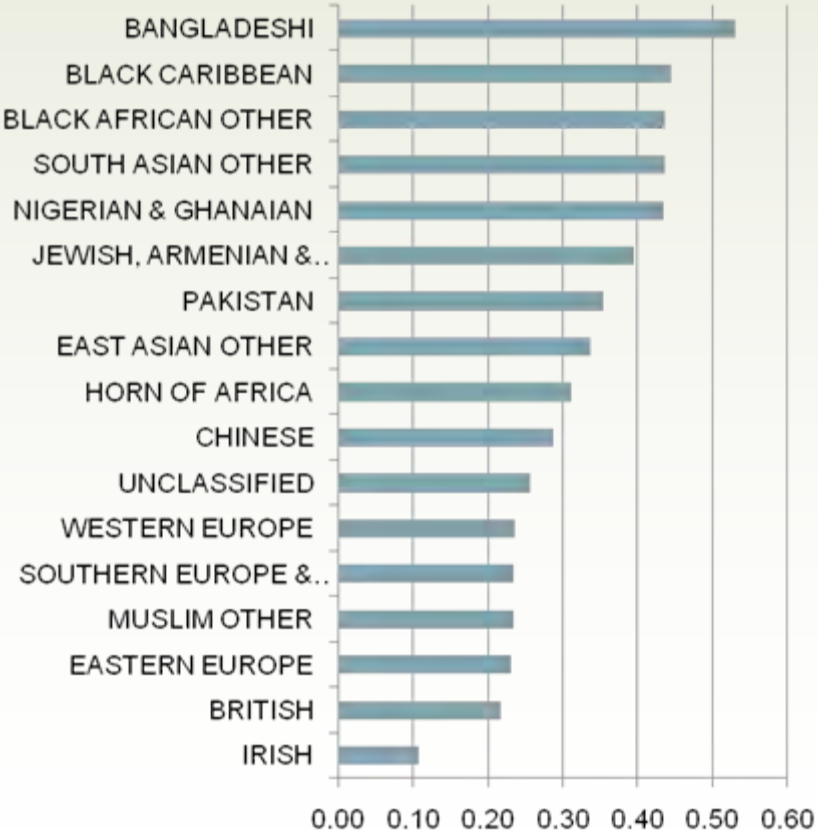
Ethnic groups aggregation effects

Index of Dissimilartiy for each of two proposed groupings of Onomap units at Ward level

A- 2001 Census equivalent



B- Alternative grouping



- Mateos, *et al* (forthcoming) *Journal of Ethnic and Migration Studies*

Overview of the methodology

- Advantages
  - Facilitates ethnicity analysis using finer spatial, temporal, and nominal granularity
  - Cost-efficient alternative when ethnicity data is missing/ low quality
  - Ethnicity categories can be re-aggregated in different ways
  - Probability scores; tailor classification to specific applications

- Disadvantages
  - Only reflects patrilineal heritage (problem of mixed ethnicity)
  - Different histories of surname adoption, naming conventions & name change rules in each language and country
  - Name normalisation decisions are required
  - Publicly available registers of names have biases
  - Not appropriate for reporting ethnicity at individual level
  - Ethical considerations and privacy issues

Thank you for listening!

Pablo Mateos
p.mateos@ucl.ac.uk

www.casa.ucl.ac.uk/pablo