Centre for Advanced Spatial Analysis (CASA)
Department of Geography

UCL

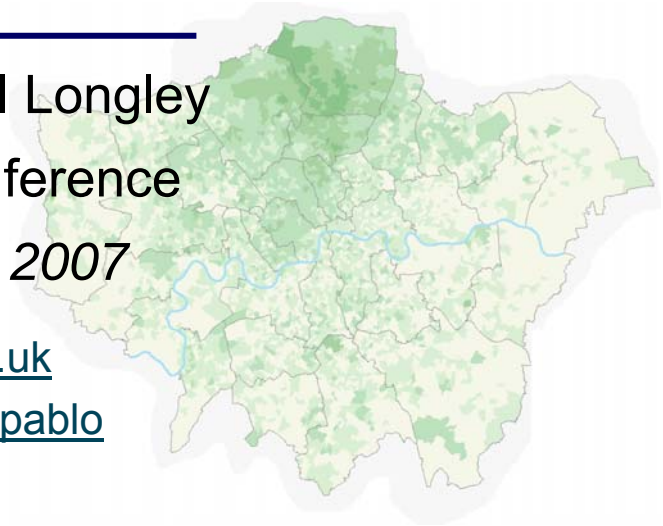
# A Name-based Ethnicity Classification to Subdivide Populations in Groups of Common Origin

Pablo Mateos & Paul Longley

DNA Sampling Conference

*Paris, 16th March 2007*

p.mateos@ucl.ac.uk
www.casa.ucl.ac.uk/pablo

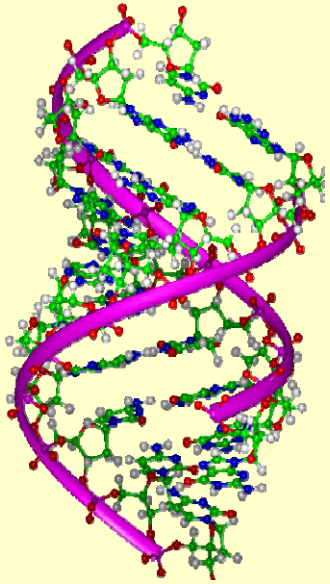


casa

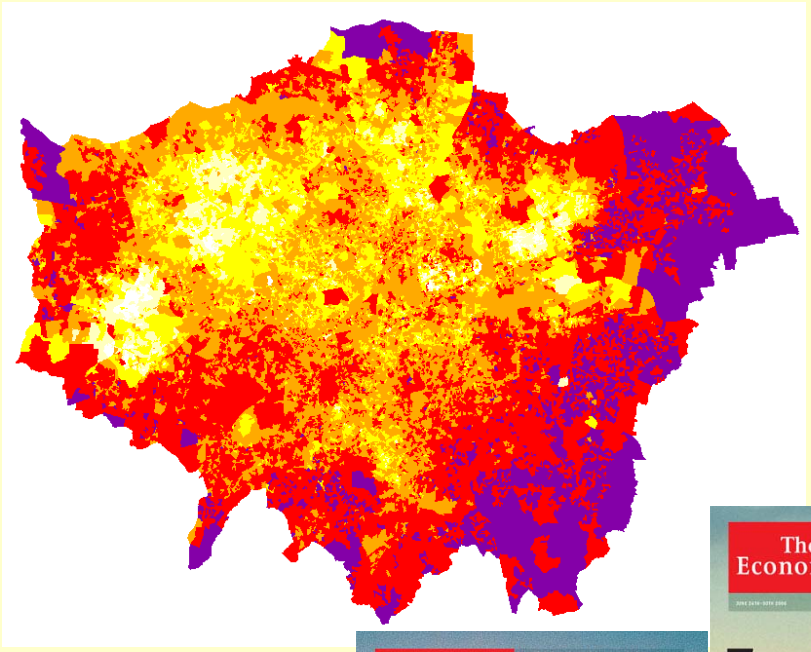
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Camden NHS Primary Care Trust

# Word of warning: Geographers and DNA?



??



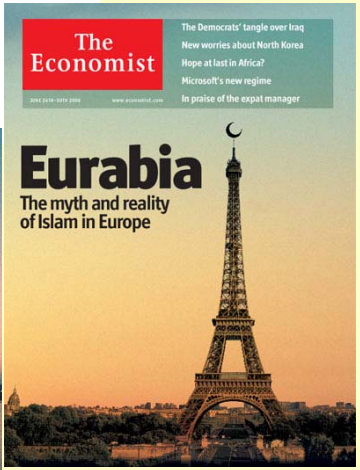
The Sunday Times

The Sunday Times September 18, 2005

**Race chief warns of ghetto crisis**

DAVID LEPPARD

BRITAIN'S race relations chief is to warn that the country is "sleepwalking" into New Orleans-style racial segregation, with Muslim and black ghettos dividing cities.

# Contents

1. Context: Surnames in Genetics, Epidemiology and Geography
2. Developing a Name-based Ethnicity Classification
3. Validation of the Method
4. Some Applications
5. Conclusion
6. Recommendations

# 1 – Context: Surnames in Genetics, Epidemiology, Geography

# 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)
- Socially constructed and very difficult to measure
- A multi-dimensional concept that encompasses different aspects of identity:

| | | |
|---|---|---|
| • Kinship | • Shared territory | • Culture |
| • Religion | • Nationality | • Others |
| • Language | • Physical appearance | |

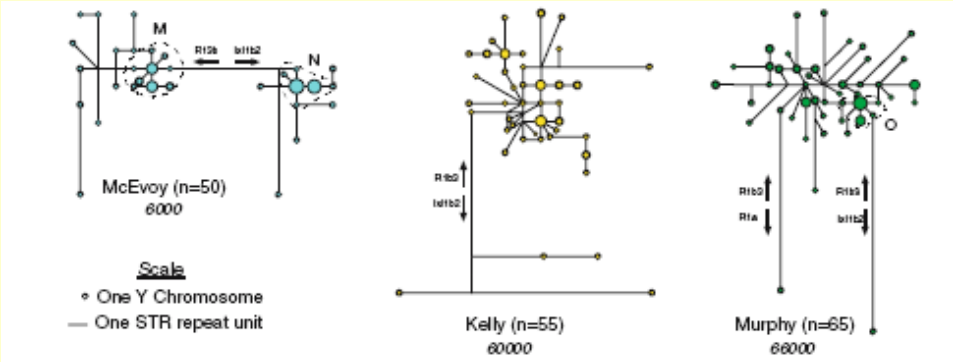
# Research on the geography of ethnic inequalities

- The problem:
  - Ethnicity data is not routinely collected & updated
  - Ethnic categories and reporting methods are not consistent over time or between data sources
  - Country of birth or Nationality commonly used as a proxy
- The result:
  - Current data sources are unable to monitor the rapidly changing nature of contemporary societies
- Our approach:
  - Use a name-based method to ascribe ethnicity at the individual level using population registers

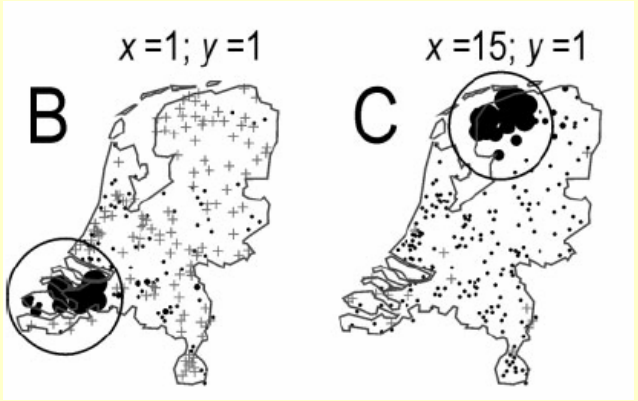
# Surname Analysis in Genetic Research

## Key Features: *Lasker (1985), Colantonio et al (2003)*

- Individual Surname Frequency -> Patrilineal lineages & Isonymous marriages
- Regional Surname Frequency -> Population structure
- Differences across Space and Time:
  - Geography: Sub-national surname structures
  - Temporal time-frame: Evolution in the last 200-700 years




*McEvoy & Bradley (2006)*




*Manni et al (2005)*

# Names & Ethnicity in Epidemiology

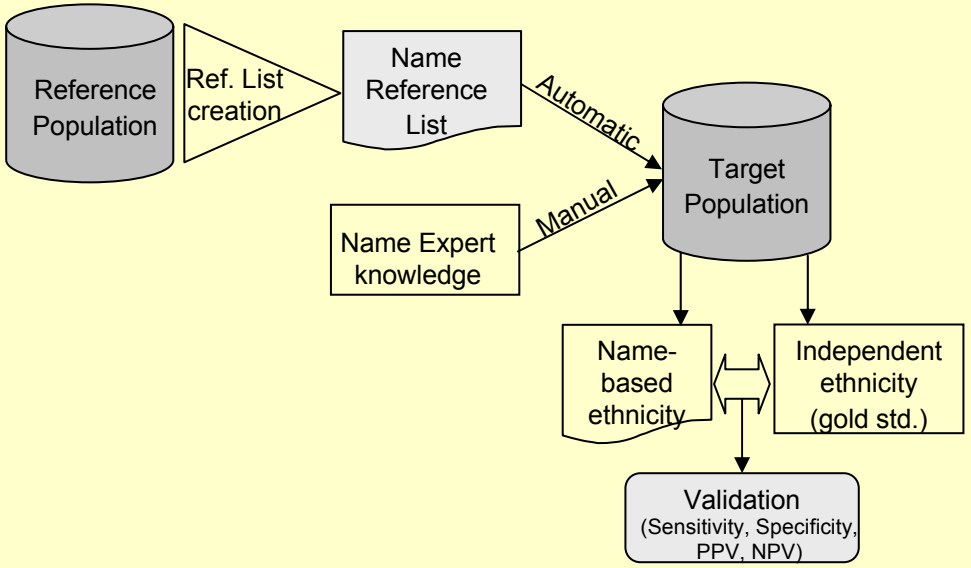
| *Aspect* | Etimology/ Onomastics | Space-time Distribution |
|---|---|---|
| Surname & Firstname | **Language** | **Geographic Origin** |
| | **Religion** | **Migration flows** |
| Firstname | **Gender** | **Age** |

- Have proved useful to subdivide populations into ethnic groups in epidemiology since the 1950s
- Largest ethnic minorities in main immigration countries
  - Hispanic, South Asian, Chinese, East Asian, Turkish, & others
  - US, Canada, Australia, UK, Germany, Netherlands
- For a review see:

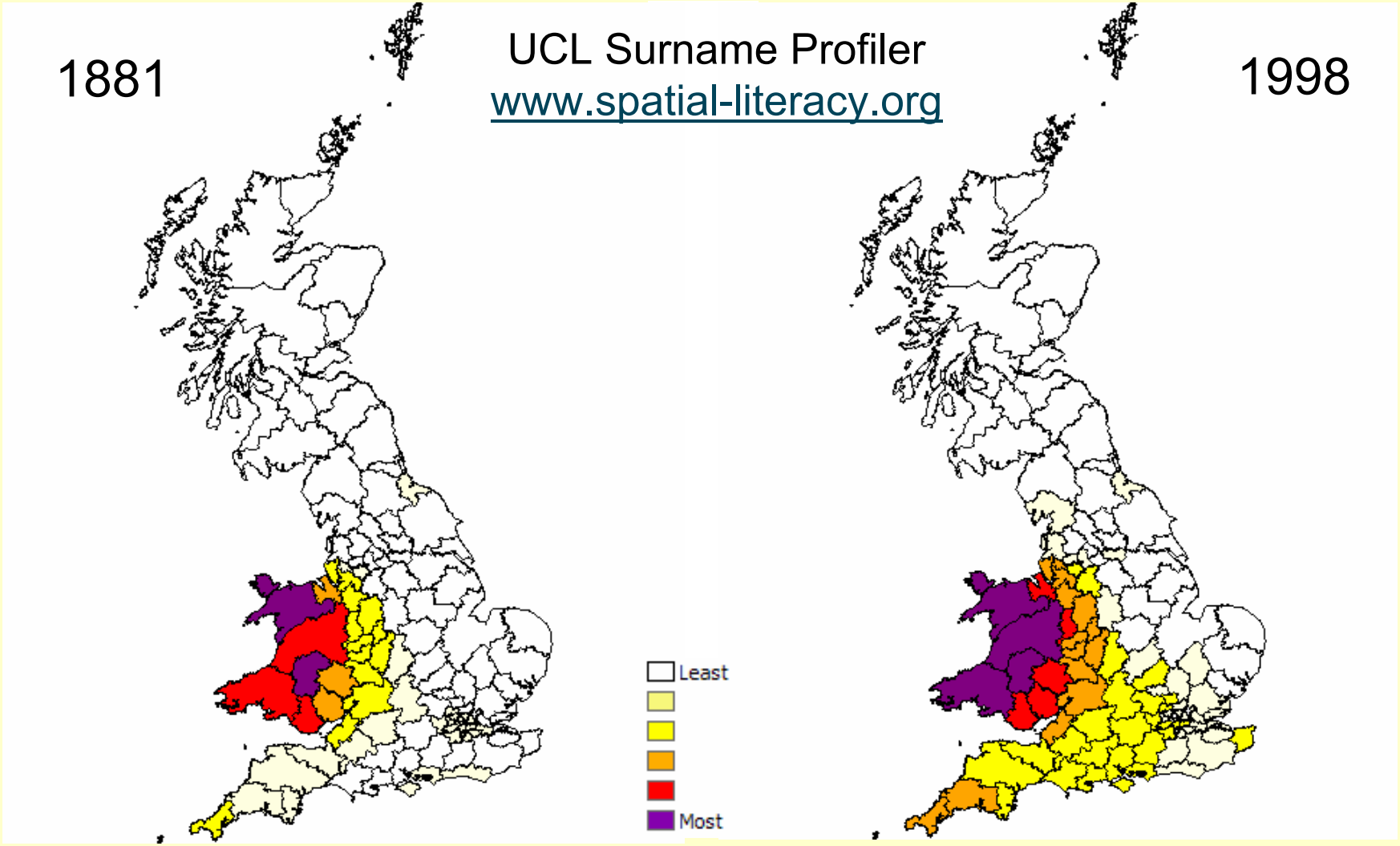
Mateos (2007) A review of name-based ethnicity classification methods and their potential in population studies, *Population Space and Place* ,13 (*in press*)

# Names & Ethnicity in Epidemiology (II)

- Manual Vs Automated methods
- Single ethnic group Vs Multi-group classifications (but only up to 5 groups)
- Name dictionaries: from 1,000 to 25,000 surname types
- Common processes to build a name classification:



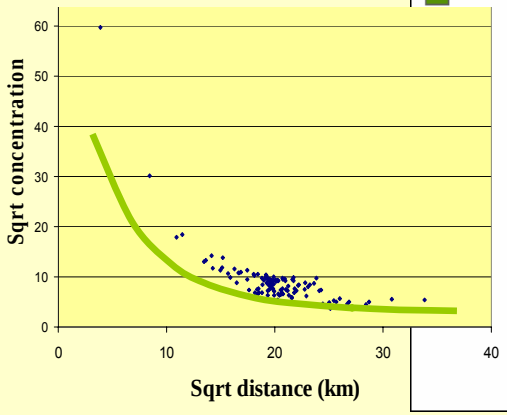
# Welsh Surnames 1881-1998



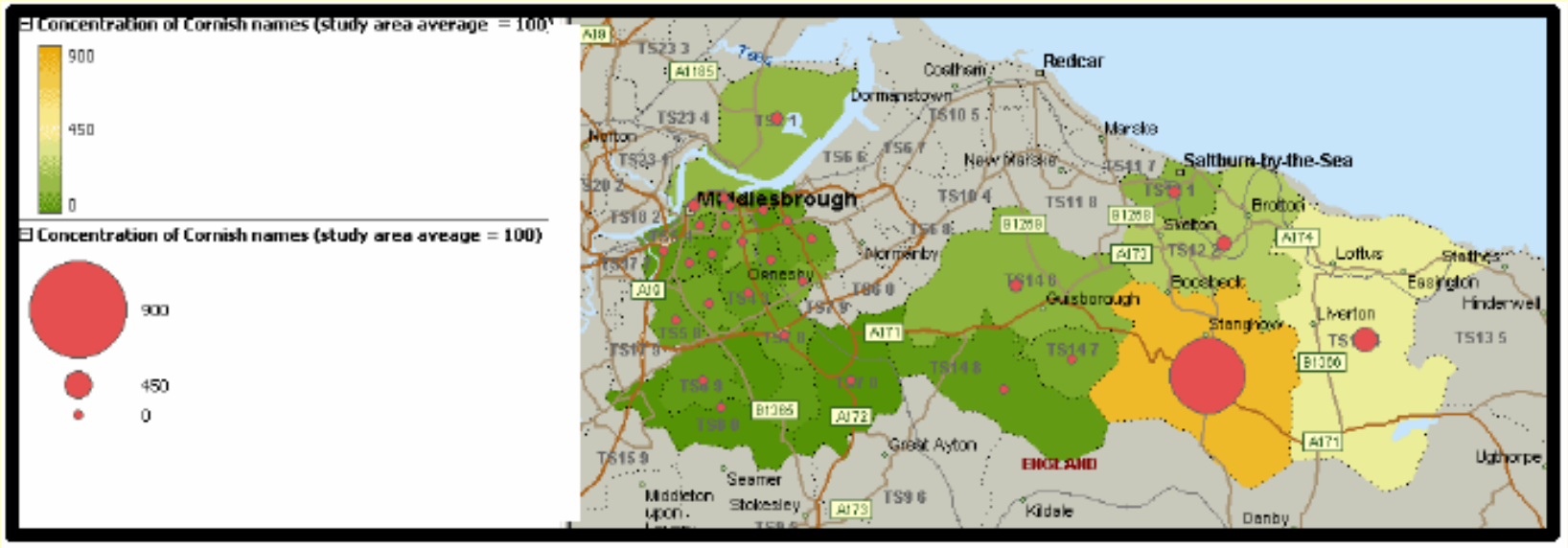
# ‘Cornish’ names relative frequency 1998

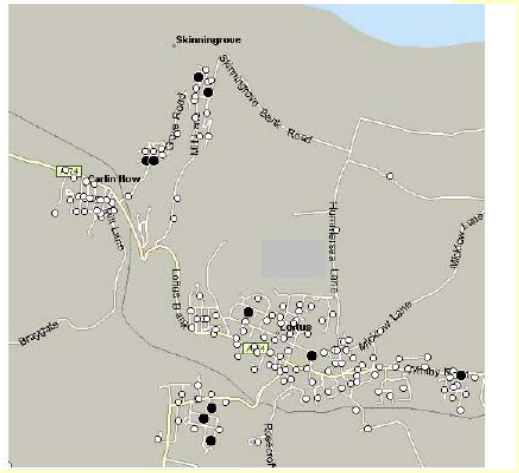
(Webber, 2005)


Distance Decay



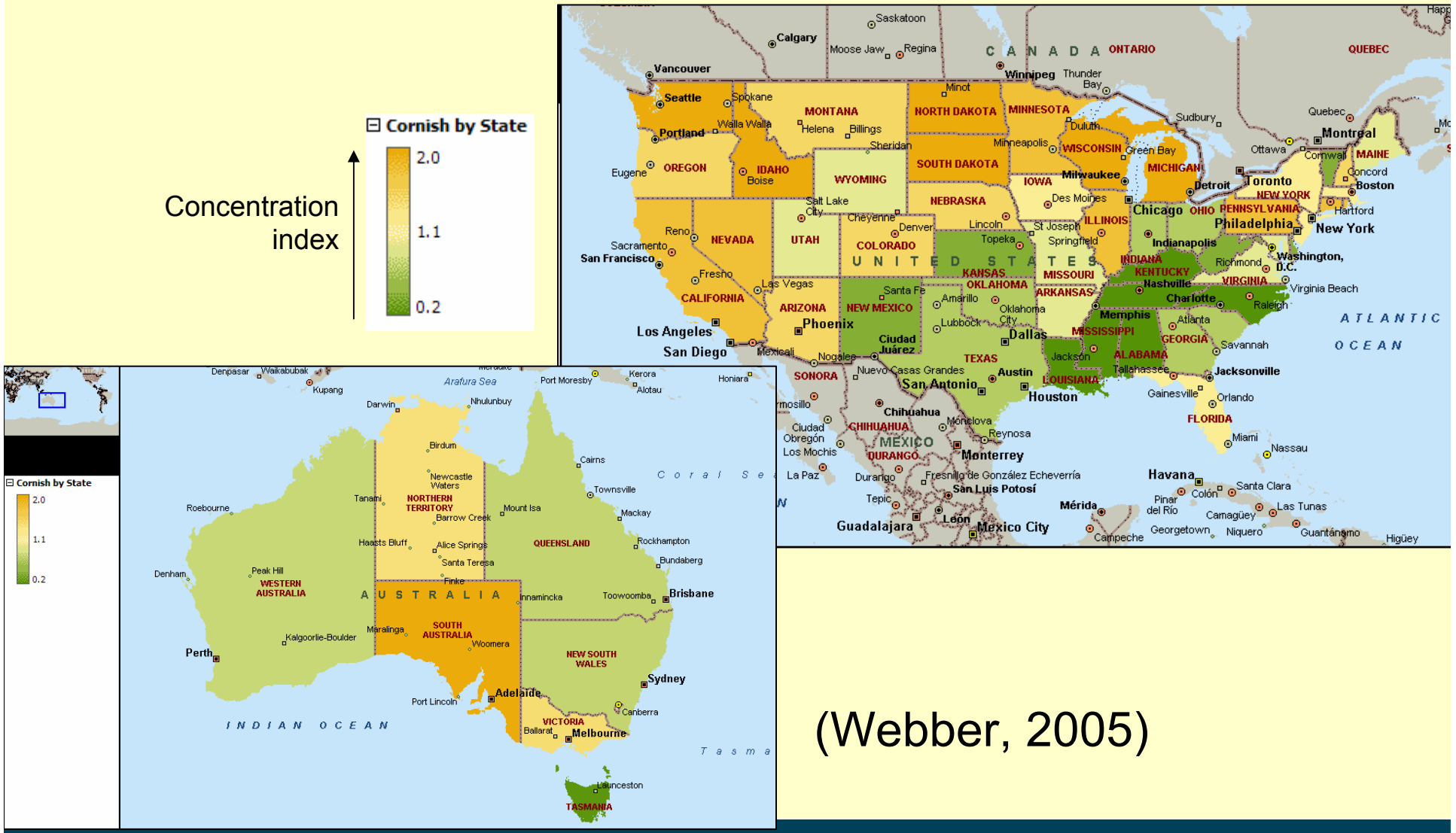
# ‘Cornish’ names in Middlesbrough




(Longley et al, 2007)

# ‘Cornish’ names & Anglosaxon diaspora




(Webber, 2005)

# 2 – Developing a name-based ethnicity classification

# Cultural Ethnic Linguistic (CEL) classification

- Objective: To create a 'universal' classification of Forenames and Surnames by fine ethnic groups
- Data source: UK Electoral Roll & Consumer Dynamics file (Experian)
  - 46,300,000 adults
  - Full name and Postcode Unit
- Analysis:
  - 250,000 Surname Types & 120,000 Forename Types
  - A taxonomy of Cultural Ethnic and Linguistic groups
    - +150 CEL Types aggregated into 15 CEL Groups
  - All names coded by CEL Type

# Cultural Ethnic Linguistic (CEL) taxonomy

- CEL taxonomy: follows an onomastic criteria to ascribe names to groups of common Cultural Ethnic and Linguistic origin
- Initially based on Hanks (2003) *Dictionary of American Family Names*, Oxford Univ. Press
- Developed empirically using UK name distribution data
- Written up in a working paper:

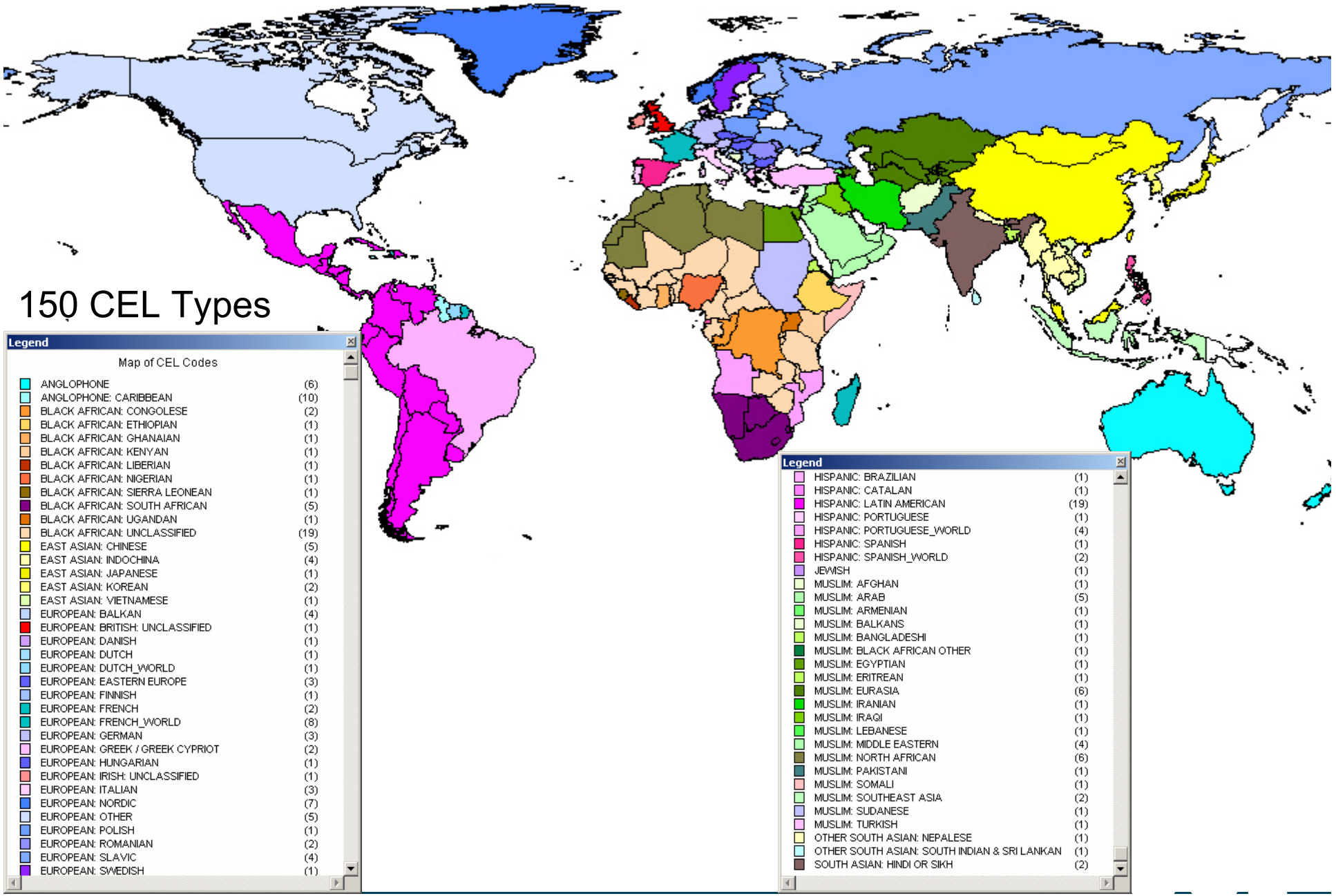
  Mateos, Webber and Longley (2007) CASA working paper 116
  www.casa.ucl.ac.uk/publications

# CEL Taxonomy & alternative groupings

| CEL Type Code | CEL Group | CEL Type | People GB 2004 | Geographical Area | Religion | Major Language | 1991 Census Ethnic Group | 2001 Census Ethnic Group |
|---|---|---|---|---|---|---|---|---|
| EU839 | EUROPEAN | BULGARIA | 109 | EASTERN EUROPE | CHRISTIAN: RUSSIAN ORTHODOX | Bulgarian | 0- White | C) White - Any other White background |
| EU840 | EUROPEAN | ROMANIA | 744 | EASTERN EUROPE | CHRISTIAN: RUSSIAN ORTHODOX | Romanian | 0- White | C) White - Any other White background |
| EU841 | EUROPEAN | ROMANIA BANAT | 29 | EASTERN EUROPE | CHRISTIAN: RUSSIAN ORTHODOX | Romanian | 0- White | C) White - Any other White background |
| EU842 | EUROPEAN | ROMANIA DOBREGA | 28 | EASTERN EUROPE | CHRISTIAN: RUSSIAN ORTHODOX | Romanian | 0- White | C) White - Any other White background |
| EU843 | EUROPEAN | ROMANIA MANAMURESCRIANA | 331 | EASTERN EUROPE | CHRISTIAN: RUSSIAN ORTHODOX | Romanian | 0- White | C) White - Any other White background |
| EU844 | EUROPEAN | ROMANIA MOLDOVA | 200 | EASTERN EUROPE | CHRISTIAN: RUSSIAN ORTHODOX | Romanian | 0- White | C) White - Any other White background |
| EU845 | EUROPEAN | ROMANIA MUNTENIA | 364 | EASTERN EUROPE | CHRISTIAN: RUSSIAN ORTHODOX | Romanian | 0- White | C) White - Any other White background |
| EU846 | EUROPEAN | ROMANIA TRANSILVANIA | 835 | EASTERN EUROPE | CHRISTIAN: RUSSIAN ORTHODOX | Romanian | 0- White | C) White - Any other White background |
| EU947 | EUROPEAN | RUSSIA | 11,118 | EASTERN EUROPE | CHRISTIAN: RUSSIAN ORTHODOX | Russian | 0- White | C) White - Any other White background |
| EU948 | EUROPEAN | BELARUS | 27 | EASTERN EUROPE | CHRISTIAN: RUSSIAN ORTHODOX | Belarusan | 0- White | C) White - Any other White background |
| EU949 | EUROPEAN | UKRAINE | 3,948 | EASTERN EUROPE | CHRISTIAN: RUSSIAN ORTHODOX | Ukrainian | 0- White | C) White - Any other White background |
| EU950 | EUROPEAN | AZERBAIJAN | 12 | CENTRAL ASIA | MUSLIM | Azerbaijani, North | 0- White | C) White - Any other White background |
| EU951 | EUROPEAN | GEORGIA | 185 | CENTRAL ASIA | CHRISTIAN: RUSSIAN ORTHODOX | Georgian | 0- White | C) White - Any other White background |
| GR110 | GREEK ORTHODOX | GREECE | 29,134 | SOUTHERN EUROPE | CHRISTIAN: GREEK ORTHODOX | Greek | 0- White | C) White - Any other White background |
| GR211 | GREEK ORTHODOX | GREEK CYPRUS | 79,304 | SOUTHERN EUROPE | CHRISTIAN: GREEK ORTHODOX | Greek | 0- White | C) White - Any other White background |
| EA221 | EAST ASIAN | SINGAPORE | 583 | EAST ASIA | BHUDDIST | Chinese, Min Nan | 7- Chinese | R) Other Ethnic Groups - Chinese |
| EA225 | EAST ASIAN | TIBET | 13 | EAST ASIA | BHUDDIST | Tibetan | 8- Any other ethnic group | S) Other Ethnic Groups - Any other ethnic group |
| EA316 | EAST ASIAN | INDONESIA | 116 | EAST ASIA | MUSLIM | Javanese | 8- Any other ethnic group | S) Other Ethnic Groups - Any other ethnic group |
| EA317 | EAST ASIAN | MALAYSIA | 2,092 | EAST ASIA | MUSLIM | Malay | 8- Any other ethnic group | S) Other Ethnic Groups - Any other ethnic group |
| EA319 | EAST ASIAN | MYANMAR | 1,601 | EAST ASIA | BHUDDIST | Burmese | 8- Any other ethnic group | S) Other Ethnic Groups - Any other ethnic group |
| EA323 | EAST ASIAN | SOUTH KOREA | 2,315 | EAST ASIA | BHUDDIST | Korean | 8- Any other ethnic group | S) Other Ethnic Groups - Any other ethnic group |

UCL

# World map of CEL types



# Cultural Ethnic Linguistic (CEL) classification

- CEL Groups and UK Electoral Roll frequency

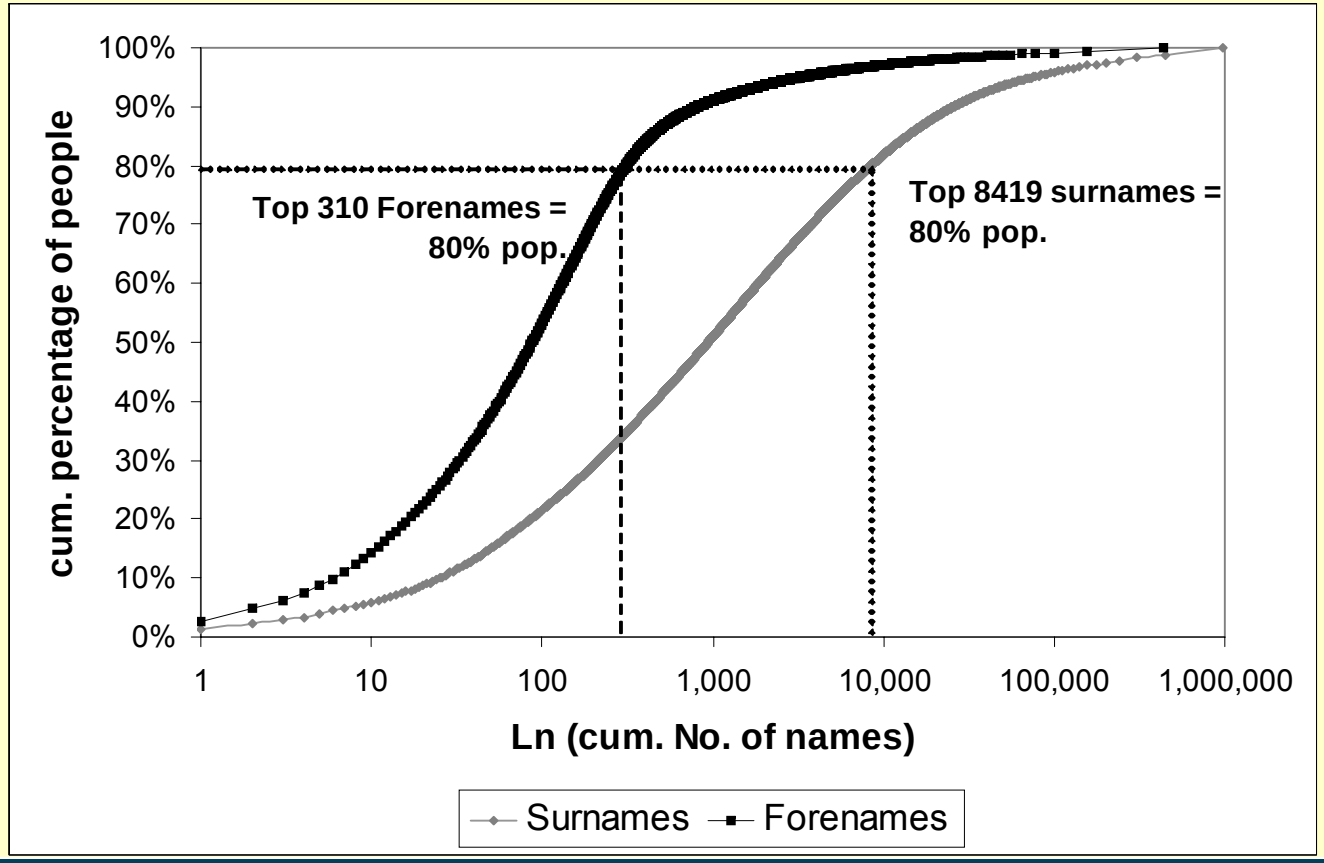
| *CEL Group* | *People* | *%* |
|---|---|---|
| ENGLISH | 29,455,761 | 67.6% |
| CELTIC | 10,485,126 | 24.1% |
| MUSLIM | 987,422 | 2.3% |
| EUROPEAN | 735,105 | 1.7% |
| SOUTH ASIAN | 475,834 | 1.1% |
| SIKH | 275,939 | 0.6% |
| NORDIC | 222,859 | 0.5% |
| HISPANIC | 186,381 | 0.4% |
| EAST ASIAN | 159,668 | 0.4% |
| AFRICAN | 149,076 | 0.3% |
| GREEK | 102,646 | 0.2% |
| JEWISH AND ARMENIAN | 80,650 | 0.2% |
| JAPANESE | 5,829 | 0.0% |
| | | |
| INTERNATIONAL | 35,763 | 0.1% |
| VOID | 210,803 | 0.5% |
| UNCLASSIFIED | 20,942 | 0.0% |
| | | |
| TOTAL | 43,589,804 | 100.0% |
| Total valid CELs | 43,322,296 | 99.4% |
| Total non-valid CELs | 267,508 | 0.6% |

# Asymmetry of Name Distributions

Rank-size power law: Zipf and Lerch distributions

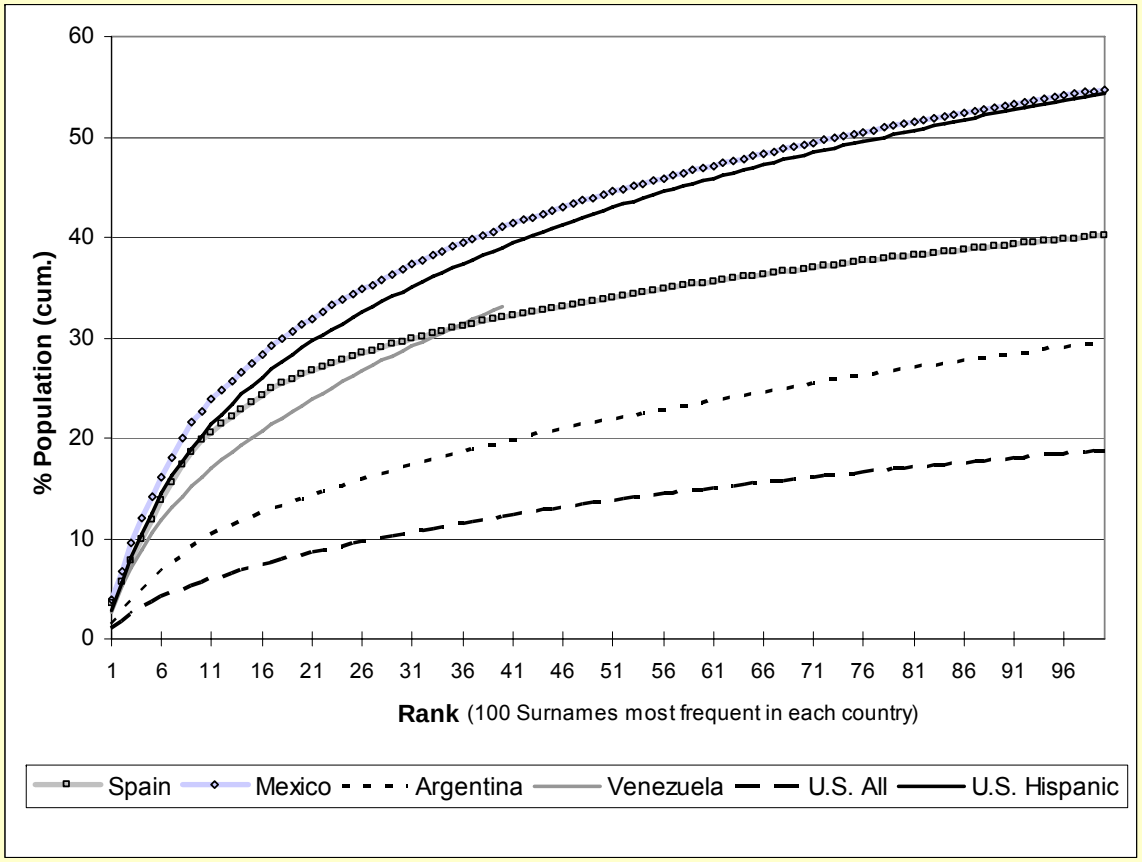
size ~ (rank)$^{\alpha}$ (Zorning & Altmann, 1995)

Cumulative frequency distribution of names in GB Electoral Roll



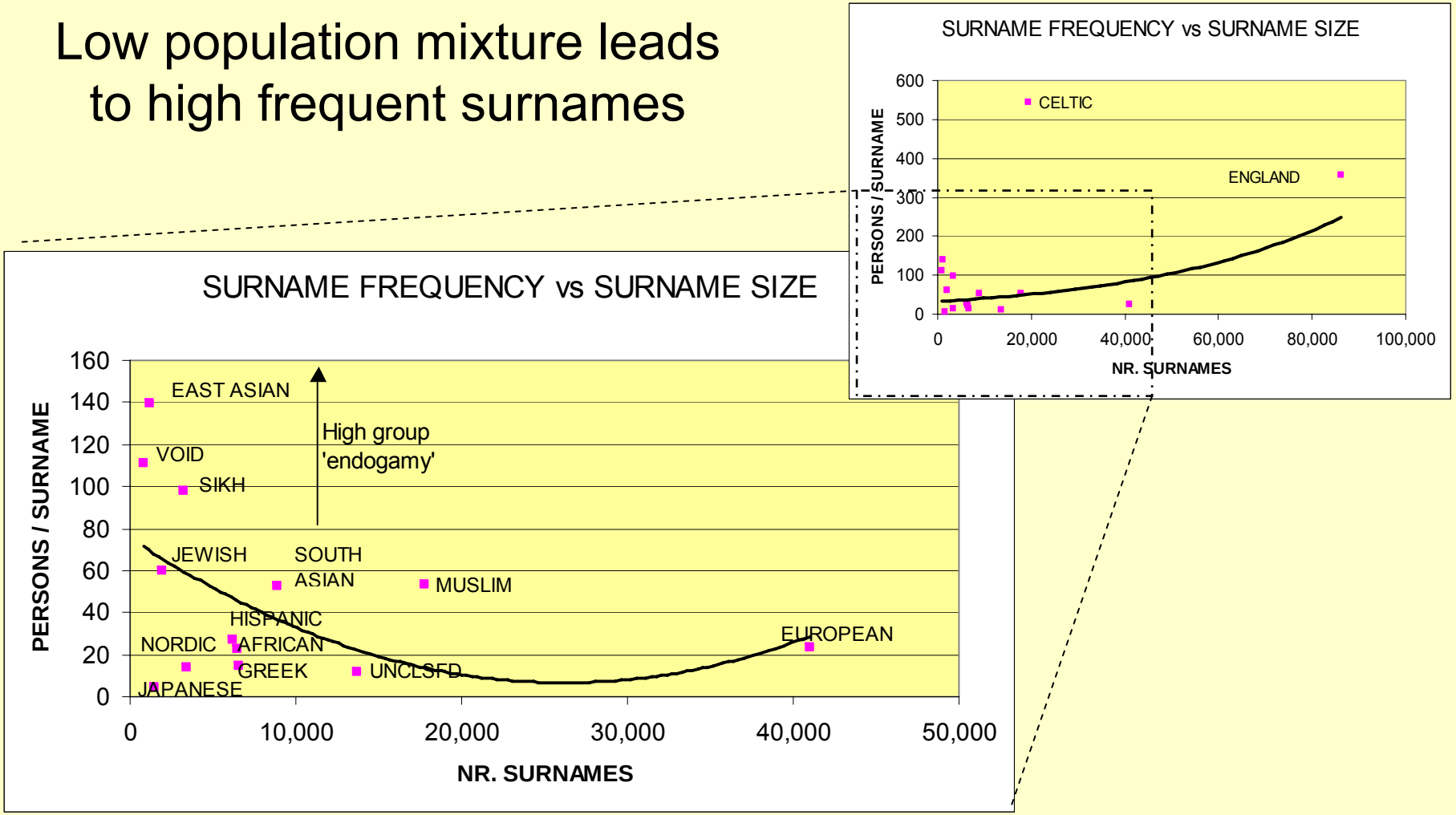
# Top Surname Sizes and Population Mix

International comparison of the top 100 surnames cumulative frequency in 6 Spanish-speaking populations

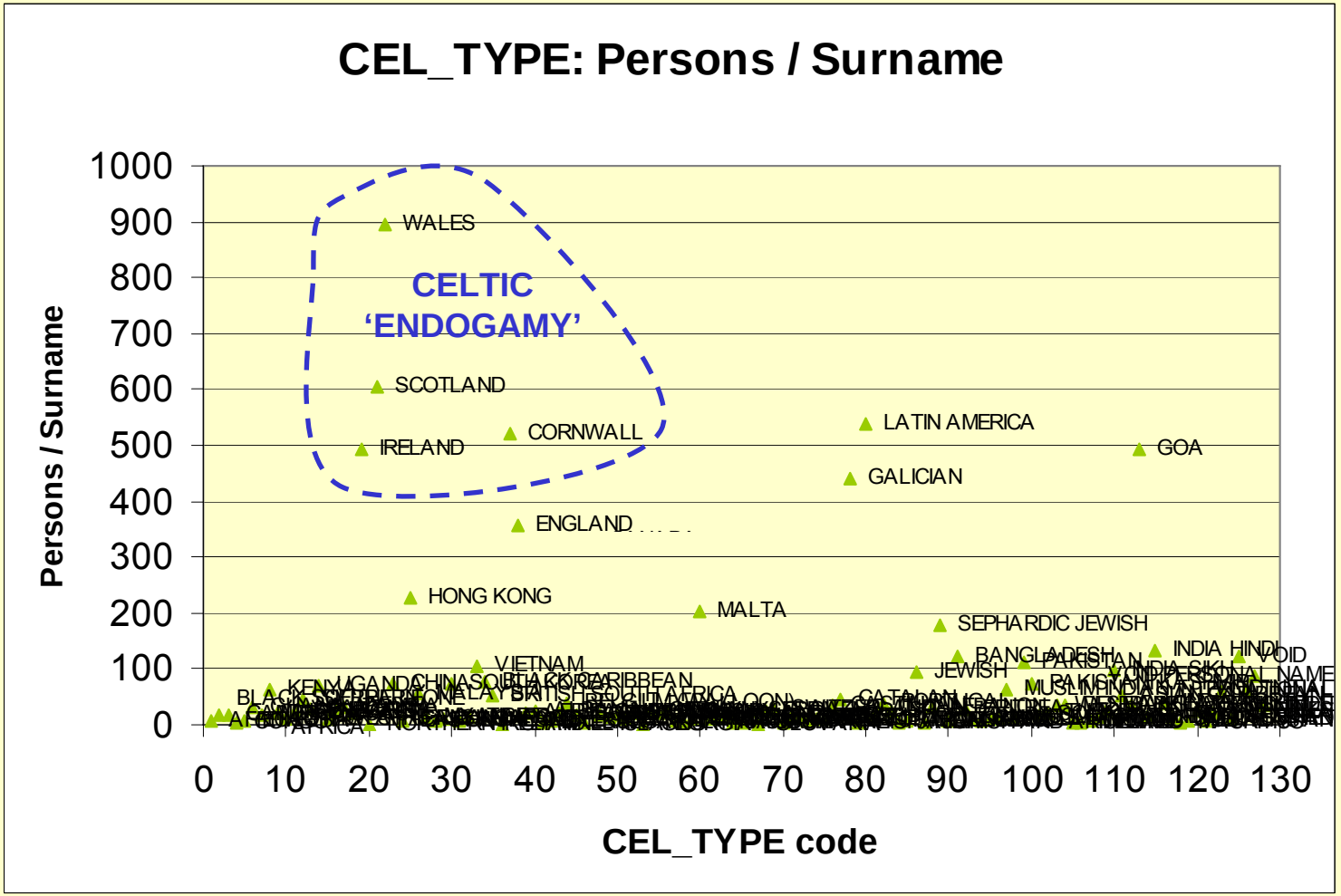


# Surname Frequency and Size: CEL Groups

Low population mixture leads to high frequent surnames



# Surname Frequency and Size: CEL Types

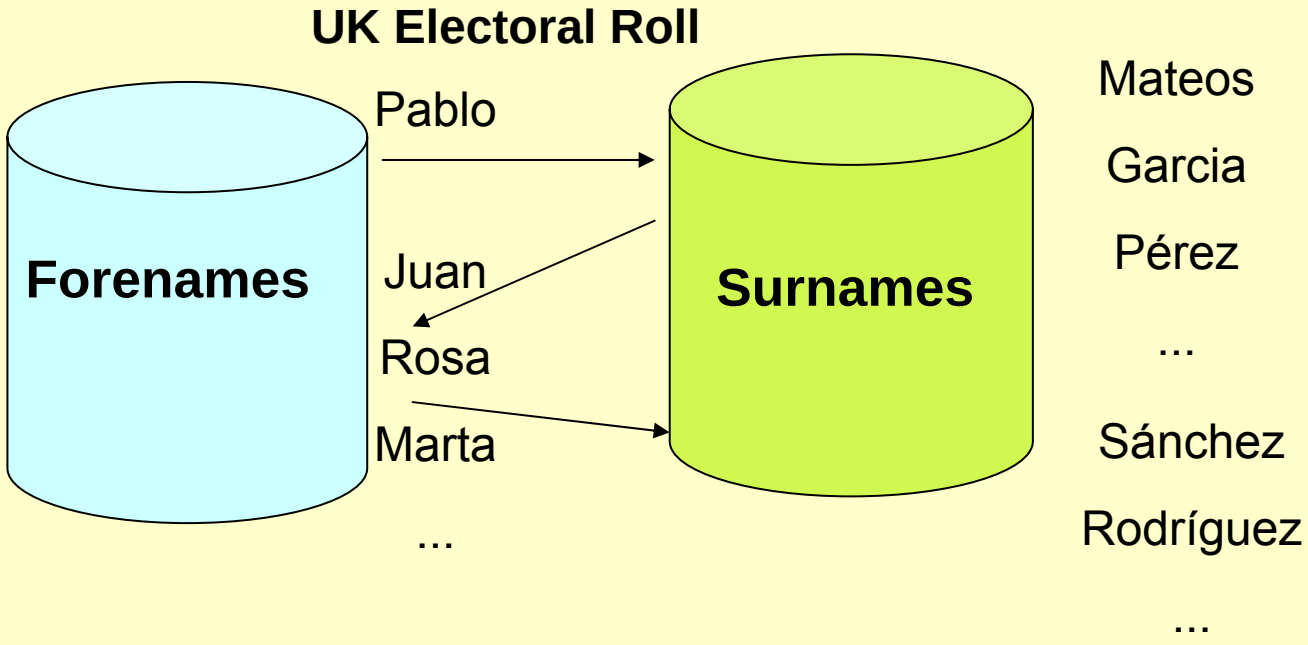


# Main methods used to classify names

1. Analysis of Forename-Surname cross-occurrences
2. Census and Geodemographic area data
3. Geographical distribution & clustering
4. Text mining
5. Birthplaces & names
6. Lists of names by country
7. ‘Googling’ individual names

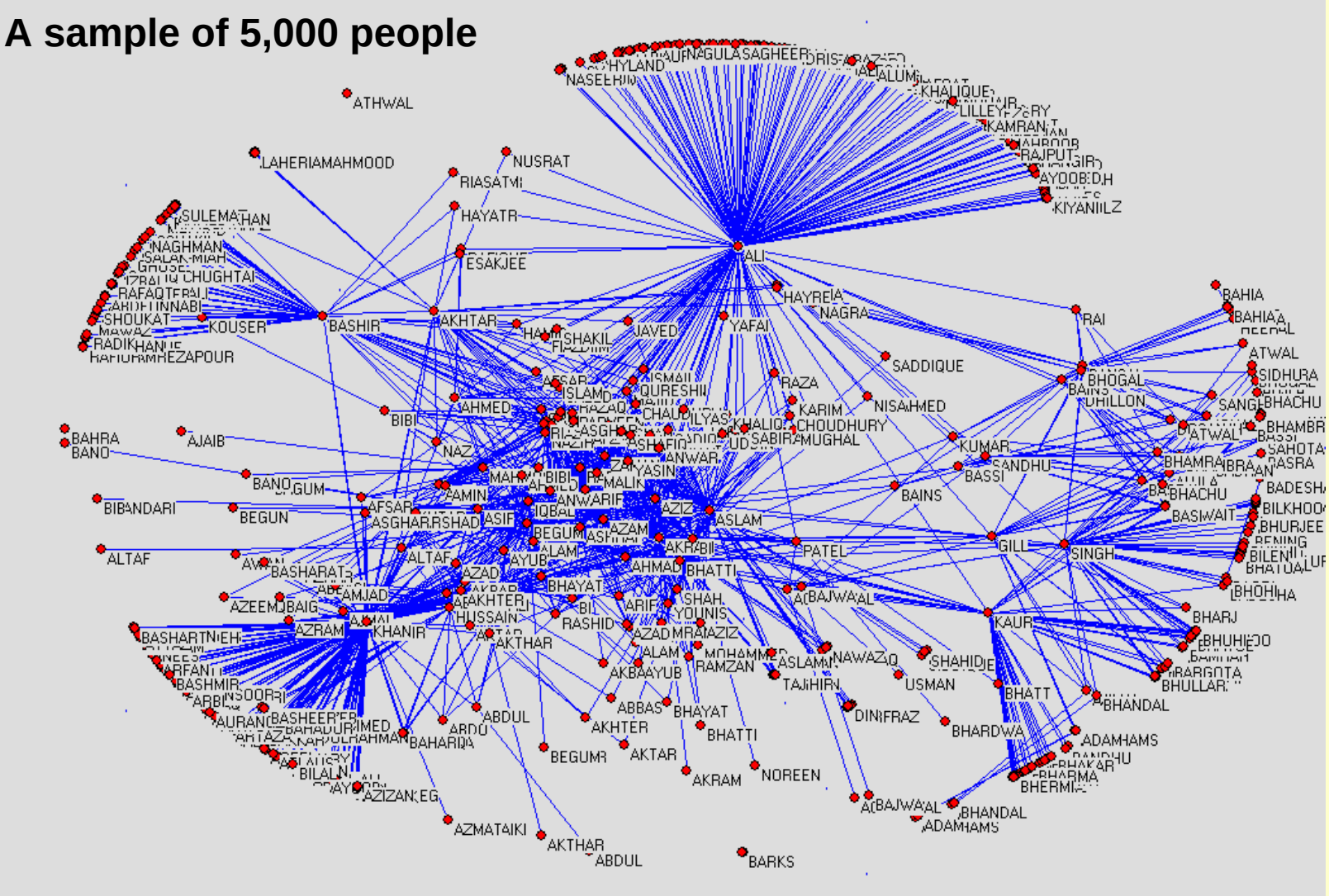
# Analysis of Forename-Surname cross-occurrences

- A technique also called ‘Triage’ proposed by Tucker (2003) to find self-contained clusters of Forenames and Surnames




- Several iterations are run until self-contained cluster is exhausted. The whole cluster is then CEL coded

# ‘Forename distance’ between Surnames



# Main methods used to classify names

1. Analysis of Forename-Surname cross-occurrences
2. Census and Geodemographic area data
3. Geographical distribution & clustering
4. Text mining
5. Birthplaces & names
6. Lists of names by country
7. ‘Googling’ individual names

(See CASA Working Paper 116 for full methodology)

# Assigning Name Scores

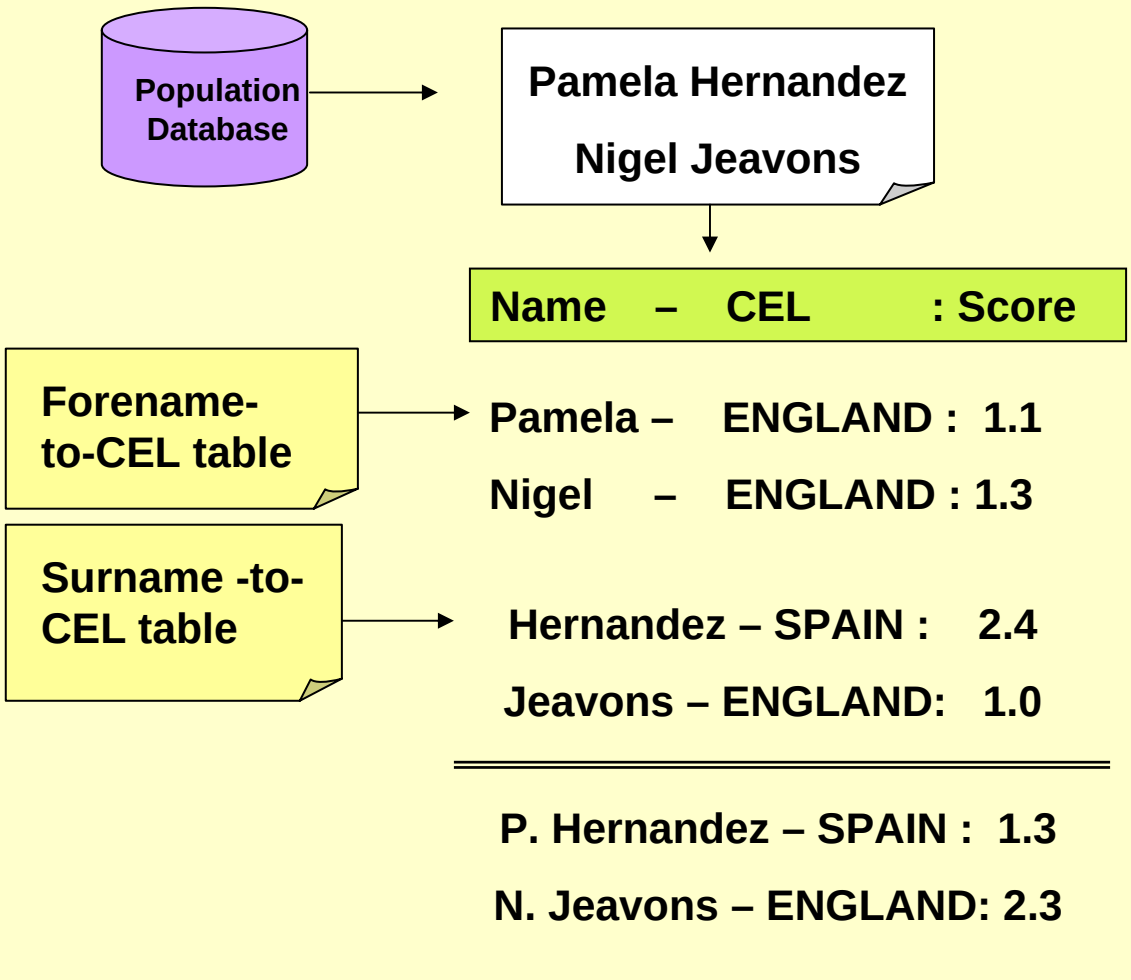
- Score: Measures the strength of association between a name and its CEL type
- A ratio of Forename-Surname CELs distributions

*Forename: **Lorcan** Frequency- Total: 90; British or Jewish: **83**, Others: 7*

| | English | Welsh | Scottish | **Irish** | Jewish |
|---|---|---|---|---|---|
| a) Lorcan SCEL Types | 30% | 8% | 18% | 41% | 2.4% |
| b) GB Average SCEL Types | 69.4% | 11.4% | 10.3% | 6.9% | 2% |
| c) Ratio c=a/b | 0.43 | 0.74 | 1.75 | **5.94** | 1.20 |

*(SCEL = Surname CEL)*

# Coding Name Databases using Scores




- 85% of cases the 2 CELs in a name are the same.
- >99% hit rate using both when we might get only 90% hit rate using just one.

User can choose which score threshold suits their purpose, dropping weakest CEL assignments

# Names from Myanmar in the UK

| Name | Score | Name | Score |
|---|---|---|---|
| SAN.....NAIDU | 0.00 | SEIN.....NGWE | 8.51 |
| TUN.....WILLIAMS | 0.00 | AUNG.....HLAING | 8.52 |
| JOSEPHINE.....TOE | 0.01 | MAUNG.....HLAING | 8.53 |
| LEONA.....MOE | 0.01 | AYE.....NGWE | 8.60 |
| MAUNG.....SAW | 0.01 | LWIN.....HLAING | 8.78 |
| WIN.....GILL | 0.01 | AUNG.....THANT | 8.80 |
| MARLIS.....ZIN | 0.02 | KYAW.....THANT | 9.04 |
| FREDERICK.....WIN | 0.03 | KYI.....HLAING | 9.04 |
| GEOFFREY.....LATT | 0.03 | WIN.....HTUT | 10.31 |
| KHIN.....MURPHY | 0.03 | KHIN.....HTUT | 11.19 |
| PAULA.....THEIN | 0.03 | THAN.....HTUT | 11.20 |
| SHAN.....SHWE | 0.03 | AUNG.....HTUT | 11.28 |
| SOE.....JOWES | 0.03 | MAUNG.....HTUT | 11.29 |
| ANGELA.....THEIN | 0.04 | ZAW.....HTUT | 11.69 |

# 3 – Validation

# Issues with Names Analysis

- Only reflects patrilineal heritage
- Different history of surname adoption, naming conventions & surname change
- Name normalisation is required
- Family/Household Autocorrelation
- Limited names lists, due to temporal & regional differences in name distribution
- Lack of consistency in self-conceived identity

(Senior & Bhopal, 1994; Martineau 1998, Word & Perkins, 1996; Jobling 2001)

# Correlations CEL 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

Camden NHS Primary Care Trust

- Hospital Episode Statistics (HES) in Camden & Islington 1998-2006
- CEL classification of names compared with self-reported ethnicity
- 343,068 patients matched to a unique ethnic code (1991 Census categories)
- Problem of bad quality HES data
  - Inconsistent data (Aspinall and Jacobson, 2004)
  - Incomplete data. Only 75% of records contain ethnicity (London Health Observatory, 2005).

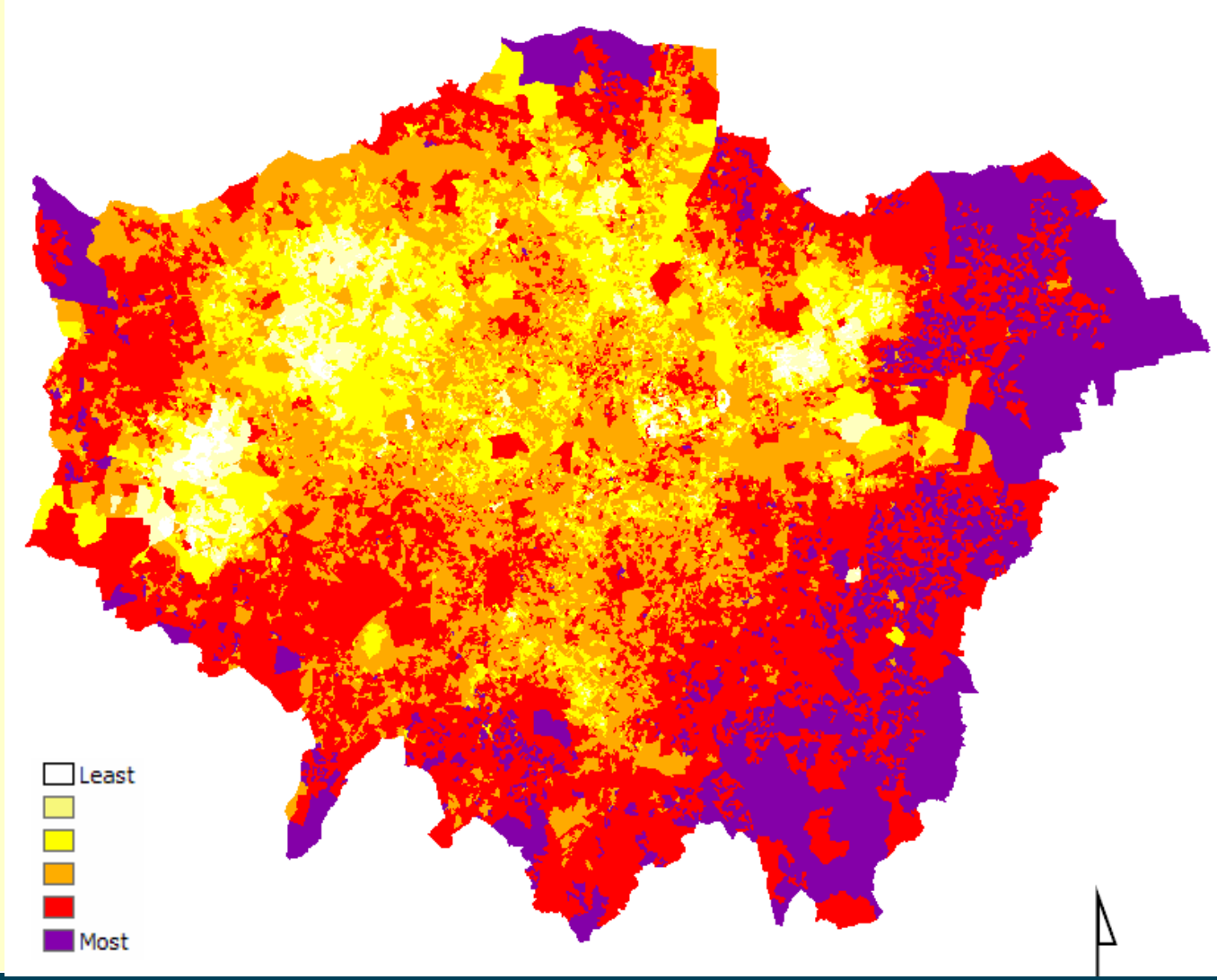
# HES Evaluation Results

| | | 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** |

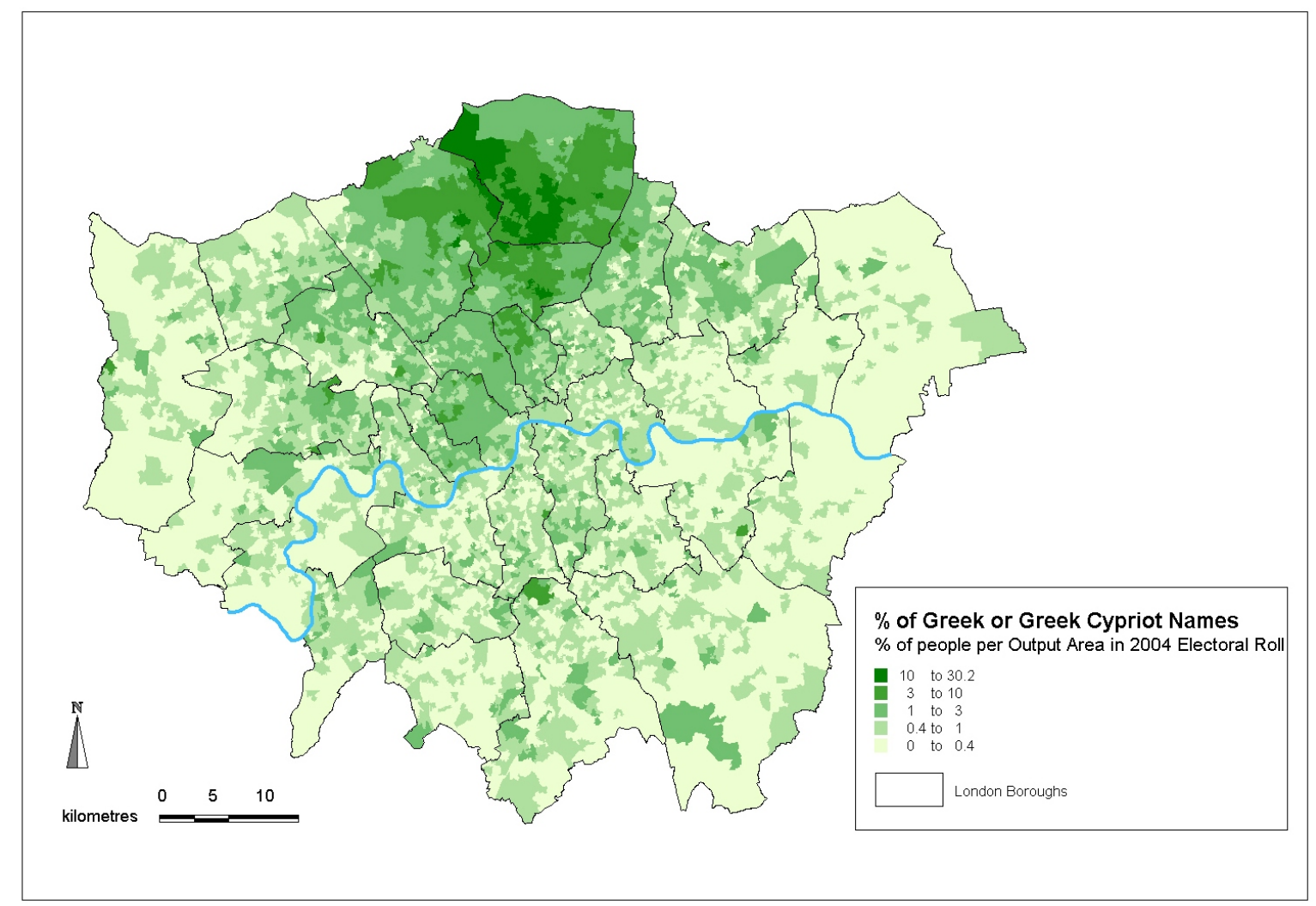
| | *1991 Census Categories* | *Sensitivity* | *Specificity* | *PPV* | *NPV* |
|---|---|---|---|---|---|
| 0 | White | **0.93 - 0.98** | 0.58 - 0.62 | **0.82 - 0.90** | **0.82 - 0.89** |
| 1 | Black - Caribbean | 0.02 - 0.03 | **1.00 - 1.00** | 0.51 - 0.62 | **0.96 - 0.96** |
| 2 | Black - African | 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 |

# 4 – Applications

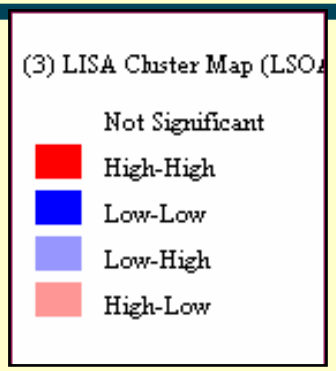
# English Names in London



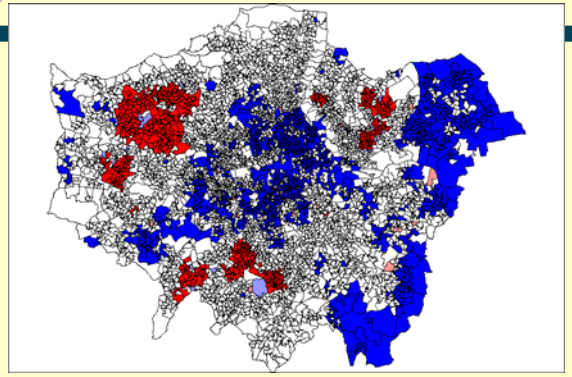
# Greek Names in London



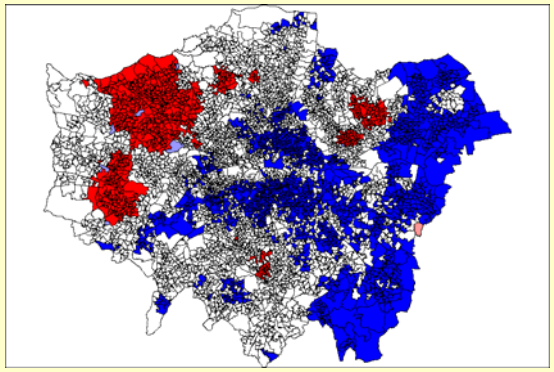
# CEL Clusters in London by LSOA



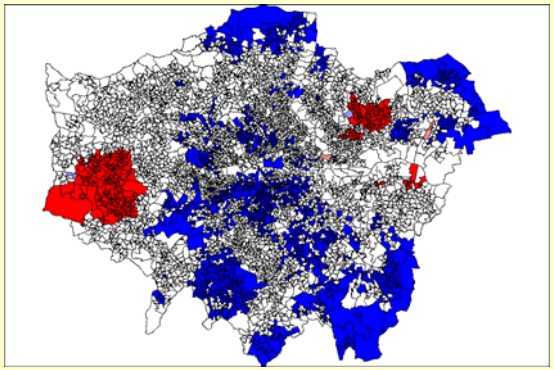

Local Indicators of Spatial Association (LISA) (Anselin, 1995) using GeoDA



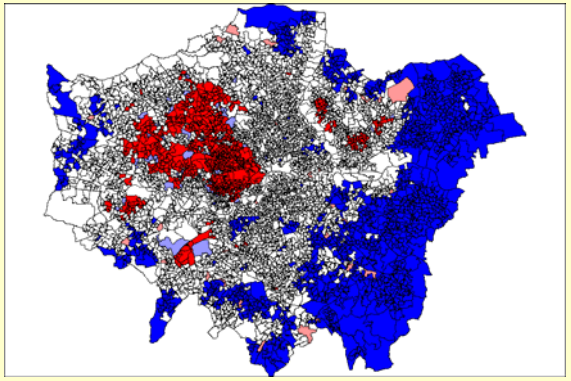
Somali



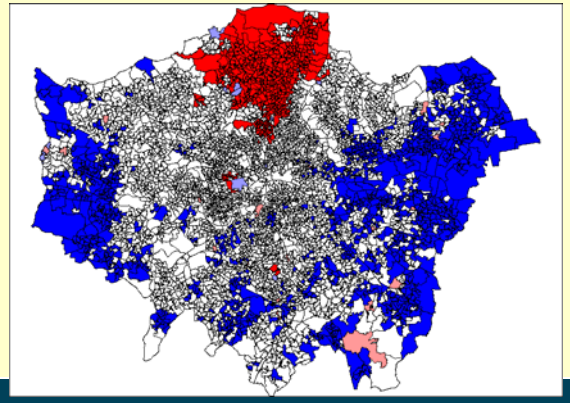
Hindu



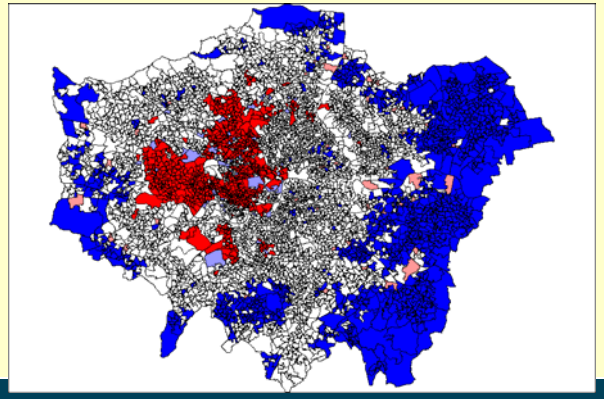
Sikh



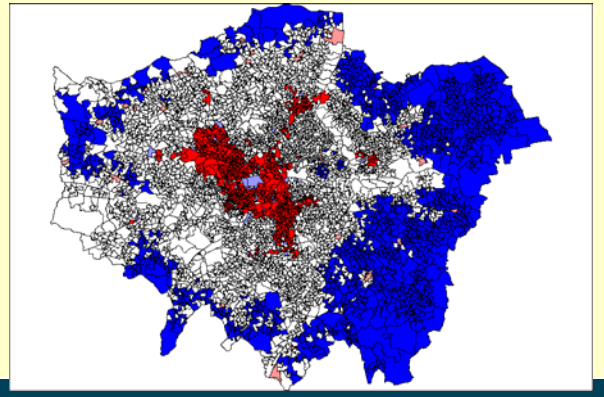
Other Muslim



Greek & G. Cypriot



Eastern Europe



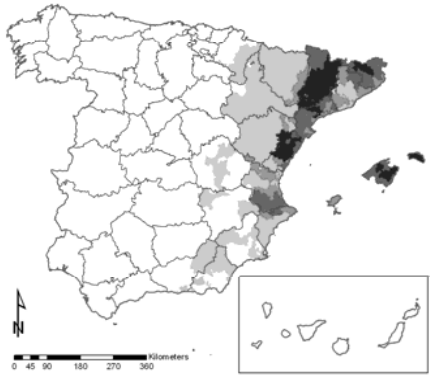
Hispanic

UCL

# Spain & Population Settlement History

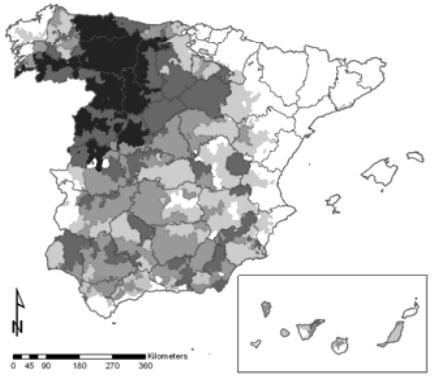
Basque

Catalan & Valencian

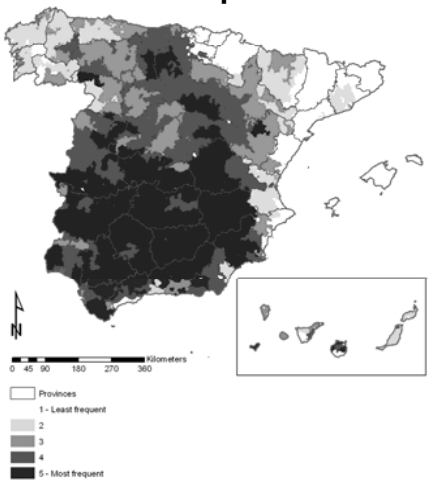


Galician

Castilian



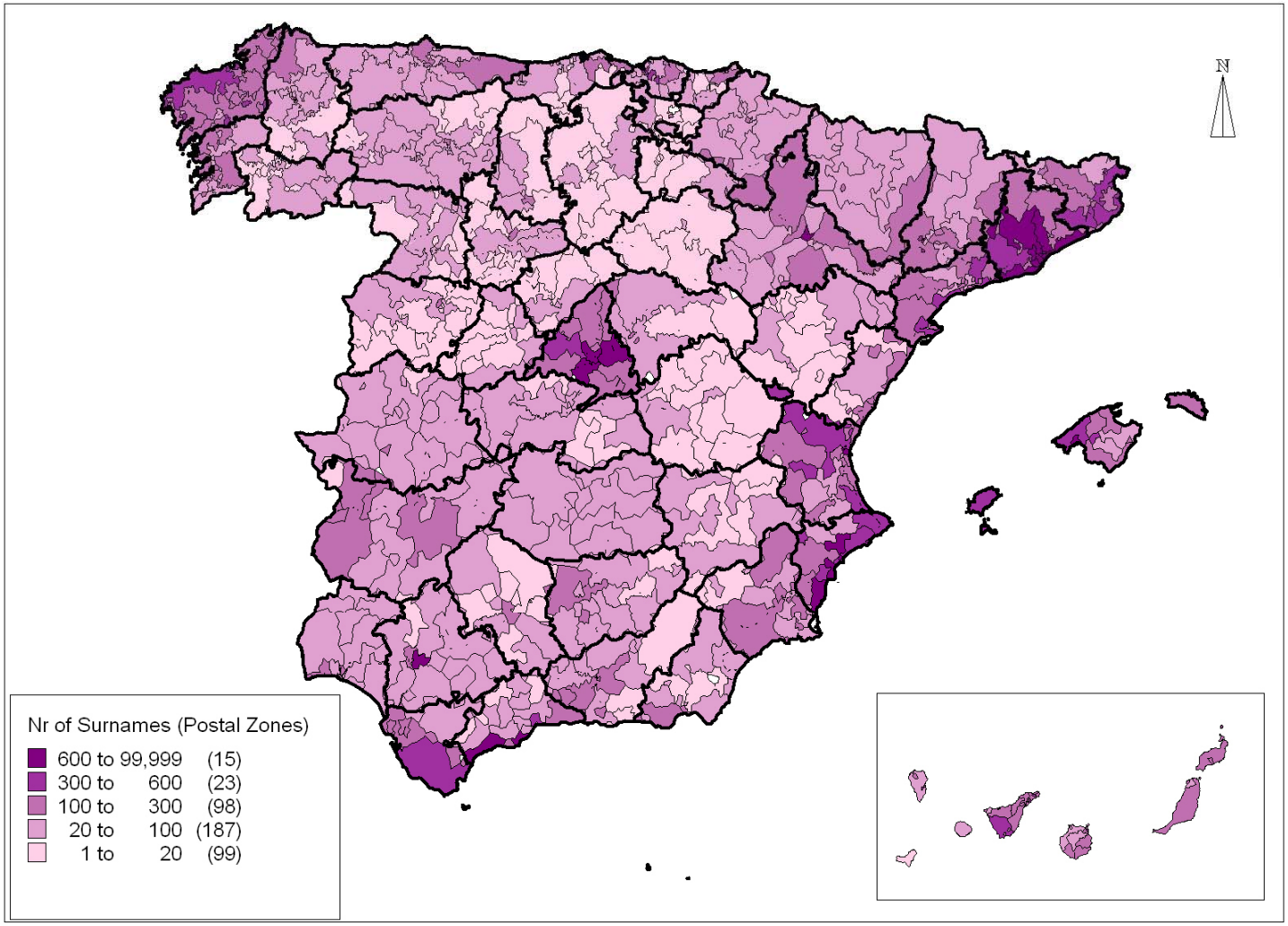

Other Spanish



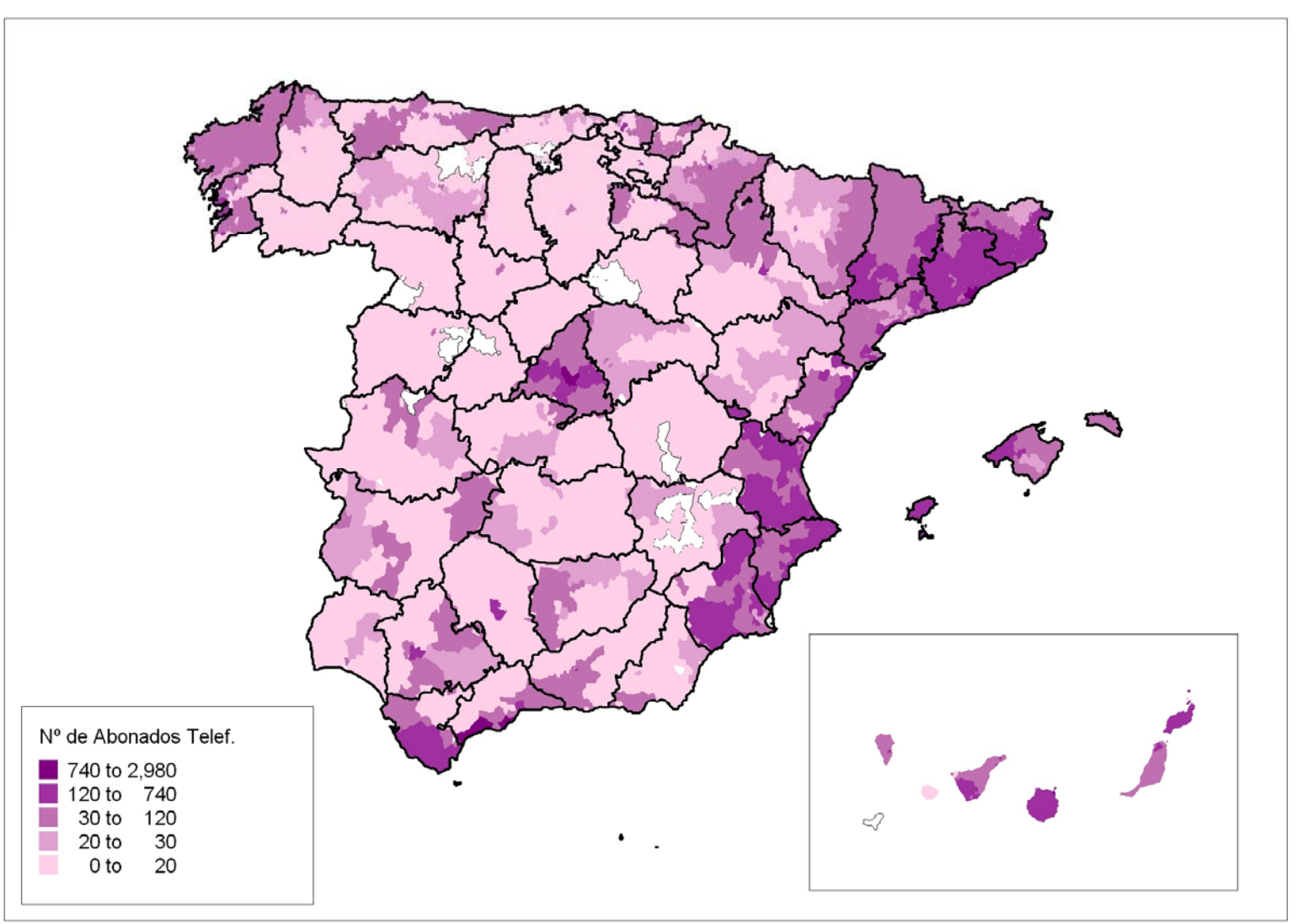

Surname frequencies, grouped by CEL (2004 telephone directory)

Tucker & Mateos (2007) *Names* (in press)

# Italian names in Spain (absolute frequency)



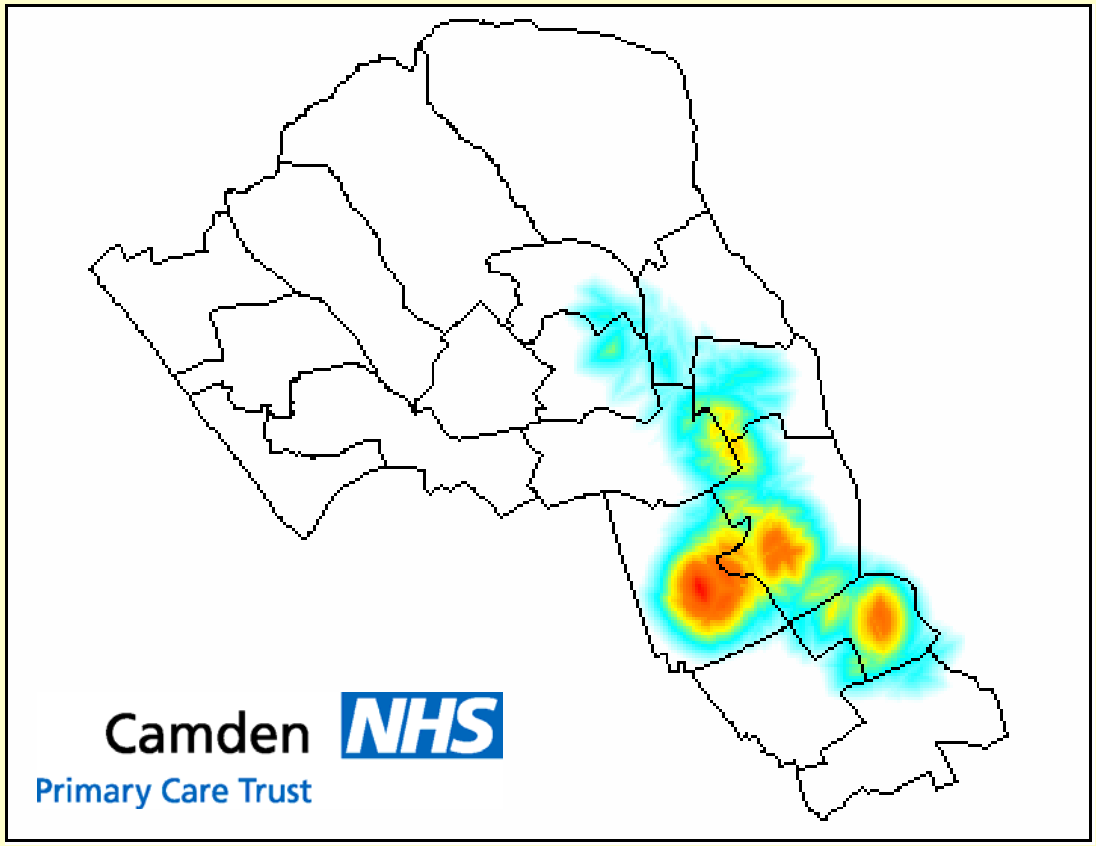
# Muslim Names in Spain (absolute frequency)



# Public health campaigns

Non-responders in a list of Breast Screening calls:

*Women with Bangladeshi names*



# Accident and Emergency profiling

Increase in patients with Polish names in A&E since May 2004 at Princess Royal Hospital,Telford.

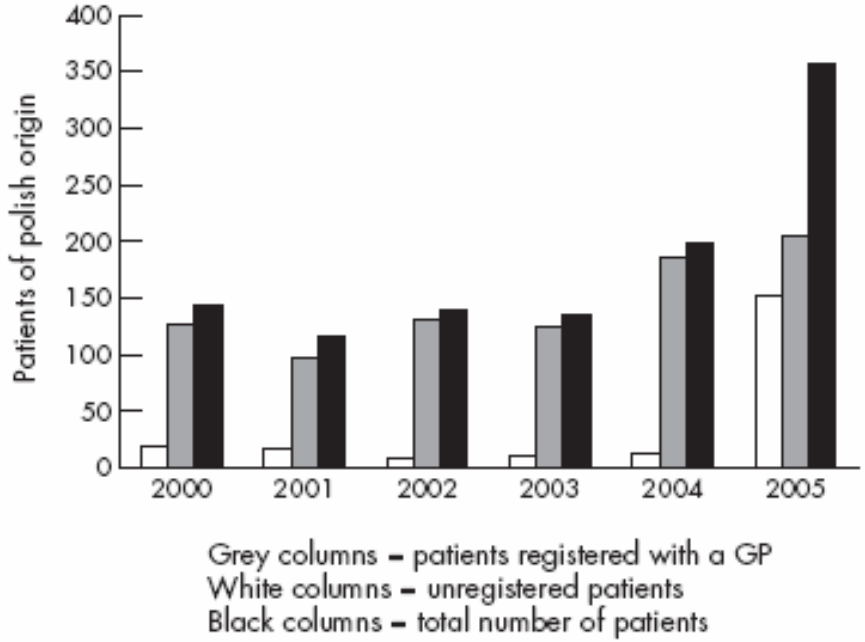



**Figure 1** Annual attendances at the emergency department by patients of Polish origin.

Leaman AM, Rysdale E, Webber R. 2006. Use of the emergency department by Polish migrant workers. *Emerg Med J* 23(12): 918-919

# Applications in Population Genetics

- Reconstruct recent migrations in highly mixed populations (e.g. USA, Argentina):

  Barrai et al (2000) did not find any surname structure across the USA, based on regional frequencies. But what about grouping surnames by ethnic group?

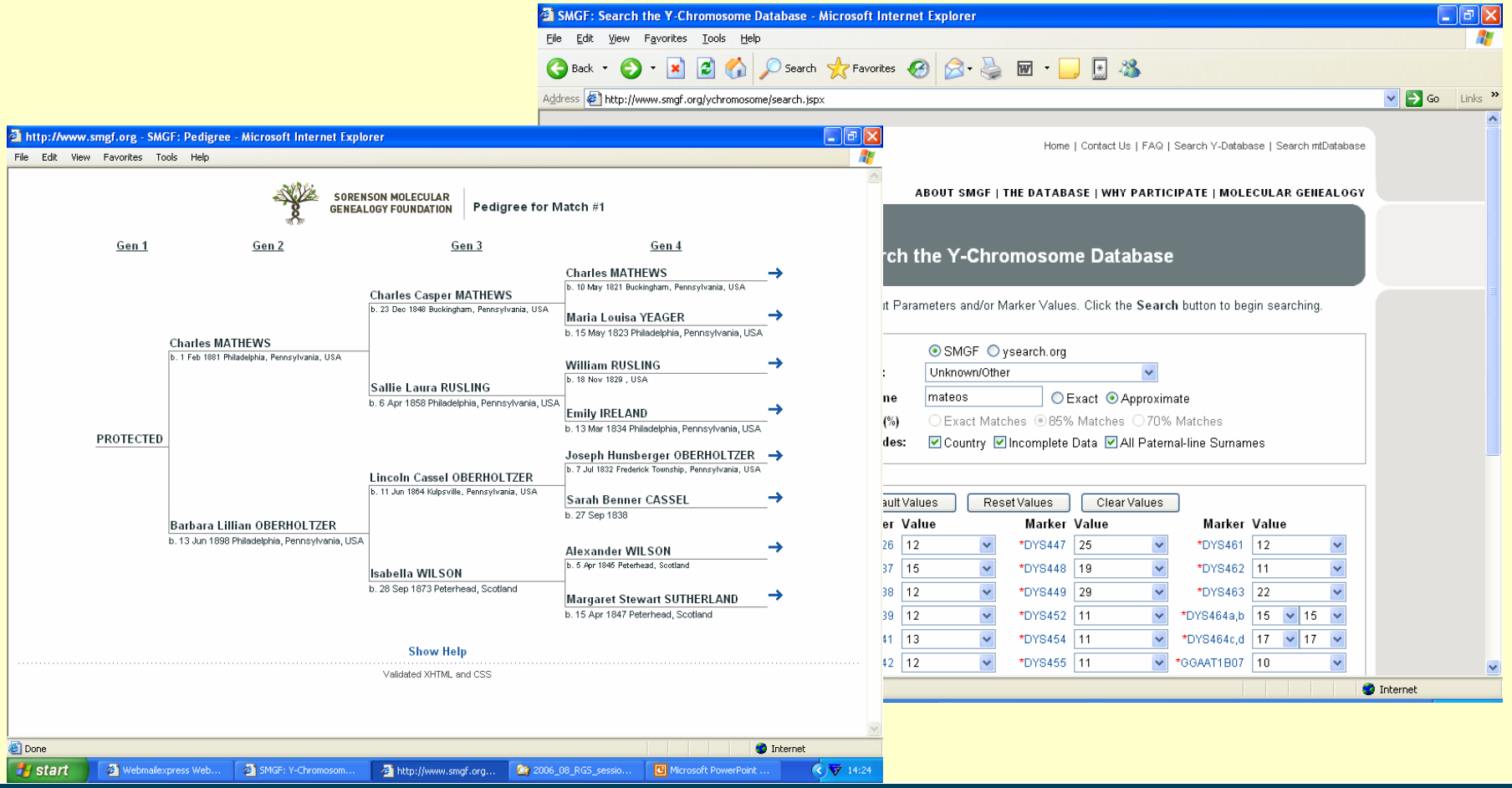
  Barrai et al (2000) Isonymy Structure of USA Population, *Am. J. Phys. Anthrop. 114, 109*

- Selection of individuals from a particular ethnic group using a name classification and a telephone directory / electoral register as a sampling strategy

  Extensive literature: Hispanic, Vietnamese, Korean, Cambodian, Chinese, South Asian, Japanese, Irish, Jewish, Iranian and Lebanese names, in USA, Canada, U.K. and Australia.

# Applications in Population Genetics

- DNA Banks - data mining of the ethnic composition of names from pedigrees



# 5 – Conclusions

# Conclusions: Review of CEL methodology

- Advantages
  - Finer spatial, temporal, and nominal scales
  - Cost-efficient method that can be applied to Population & Patient Registers, Telephone Directories, etc.
  - CELs can be re-aggregated in different ways and score thresholds adapted to each specific application
- Challenges
  - Enhance the classification to consider other context data (e.g. country of birth and postcode)
  - Internationalisation; different CEL allocation for a name in different countries
  - Ethical considerations and privacy issues

# 6 – Recommendations

# Recommendations: DNA sampling

- Methodological:
  - Use forename data in surnames studies to cluster names by 'forename-surname distance'
  - Develop your own classification of relevant 'ethnic groups' and select a threshold score, suited to your specific study purpose
- Potential applications:
  - Cost-effective large scale sampling of populations by ethnic group, both in the region of origin or of destination of migrations
  - Find relationships between individuals in existing DNA databases, grouping names that are 'closer' to each other.
  - Multiple applications on genetic diseases research to identify ethnic minorities in the countries of migration

# Acknowledgements


- Professor Richard Webber, UCL (initial study)
- Experian Ltd, UK Data Archive (datasets providers)
- Economic and Social Research Council and Camden Primary Care Trust (sponsors)
- Franz Manni for his invitation to participate in the conference

# Thank you!
# Any Questions?

**www.casa.ucl.ac.uk/pablo**

**p.mateos@ucl.ac.uk**

UCL