What's in a name? The frequency and geographic distributions of UK surnames

Are you a Smith, a Featherstonehaugh-Cholomondely-Warnock or a Begum? Your surname can reveal the social history of where you live. **Fiona McElduff**, **Pablo Mateos**, **Angie Wade** and **Mario Cortina Borja** investigate.

In 1875 George Darwin, wrote a paper (Darwin 1875) with the title 'Marriages between First Cousins in England and their effects.'[1] It was published in the *Journal of the Statistical Society of London*, forerunner of the Royal Statistical Society. More relevantly to us here, he used the surnames as his main source of data to measure the probability of marriage between first cousins when both bride and groom surnames coincided. Darwin of course bore a famous surname himself. George was the son of the more famous Charles. Charles Darwin had married his first cousin Emma Wedgwood; George therefore had a particular interest in proving statistically that first-cousin marriages were rather frequent events.

Whether or not he was the first to use surnames to study relationships between different subgroups of the population, many others have followed. And just as understanding of genes and genomes has fed into his father's theory of evolution, it has also fed into studies of the distribution of surnames[2,3]. A well-known relationship exists between surname frequencies, geographic distributions and population genetic structure[4]. Surnames are often passed down through the male line, so they correlate well with Y-chromosomes[5,6] and this is opening up a new era of genetic genealogy[7].

The study of surname frequencies across geographical areas can provide interesting insights into the structure of populations, helping us to disentangle ancestral human movements, historic settlements, distant and local migrations. In addition it enables us to study genetic relationships between populations, marriage within and outside communities, and cultural diffusion and interaction[8]. For example, a recent extensive study of the surname distribution of the total population of eight European countries has concluded that the present surname structure of Western Europe is strictly linked to local languages[9]. However, this study did not include the UK. We have analysed the geographic and frequency distributions of surnames in the UK in 2001 with the aim of showing the potential of this innovative way of looking at cultural diversity, ancestry, migration and the composition of populations.

Surname data

The usual sources for name frequency studies are old-fashioned telephone directories[9,10], such as UCL's WorldNames mapping project (`www.publicprofiler.org/worldnames`). However, phone books have several disadvantages: they represent households not individuals; they usually have a strong male bias; users can "opt-out" of the directory; they tend to have a low coverage of the young and mobile populations who do not use or contract telephone land lines themselves. Some administrative databases such as the National Health Service patient register or the Census of Population both officially cover 100% of the population and would be much better sources of data but, for obvious reasons of personal data privacy, they are not publicly available. For our data we used the UK electoral register, which is a public register containing the names and addresses of all over the age of 16 who are entitled to vote in UK elections. The companies that distribute the electoral register supplement it with additional residents not registered to vote, which they source from commercial surveys and credit-scoring databases: the resulting database is known as the "enhanced electoral register". In 2001 a change in the law allowed voters to "opt-out" from the publicly available version of the electoral register (the level of "opt-out" for 2007 is estimated at 30%[11]). We used an enhanced register of 45.6 million people resident in the UK in October 2001: it is the last version of this data set before opting out was possible and is the best data source on names and locations publicly available.

British surnames can be categorised by their geographical origin using the UCL Surnames Profiler (now rebranded under the National Trust as `http://www.nationaltrustnames.org.uk/`).

Yule's K is defined as

$$K = 10^4 \frac{1}{N^2}\left(\sum_{r=1}^{v} r^2 V_r - N\right),$$

where N is the number of people in the population, v is the highest occurring frequency of surname and V_r $\{r=1, 2, 3,\ldots, v\}$ represents the number of different surnames with frequency r. Larger values of K indicate lower diversity in a population and hence greater uniformity. Independently of the number of individuals, K is proportional to the probability of any two randomly selected persons having the same surname and thus reflects the repetitiveness of surnames in a population. The sequence $\{rV_r/N\}$ provides the proportions of the population with surnames that occur r times; in particular V_1/N is the proportion of people who have surnames appearing uniquely in the population and indicates surname richness. Finally, the ratio of the number of different surnames to the number of people in the population measures the volume of surname diversity.

Methods

Measures of vocabulary richness are used in linguistics[12] to quantify literary style by means of the frequency of the different words in them. These measures can be adapted to analyse the frequency and diversity of surnames in different geographical regions[13].

An established measure of diversity commonly used in linguistics is Yule's K[14] (see box) which varies from 0 (in populations where everyone has a unique surname) to 10 000 (if everyone shares the same surname).

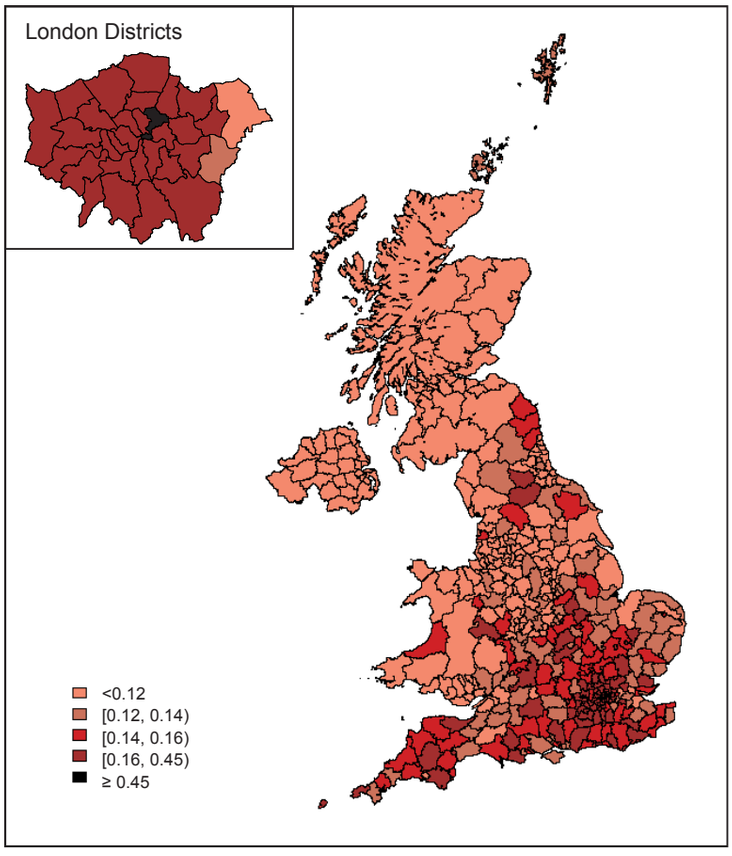

Figure 1. Average number of people in the population per surname

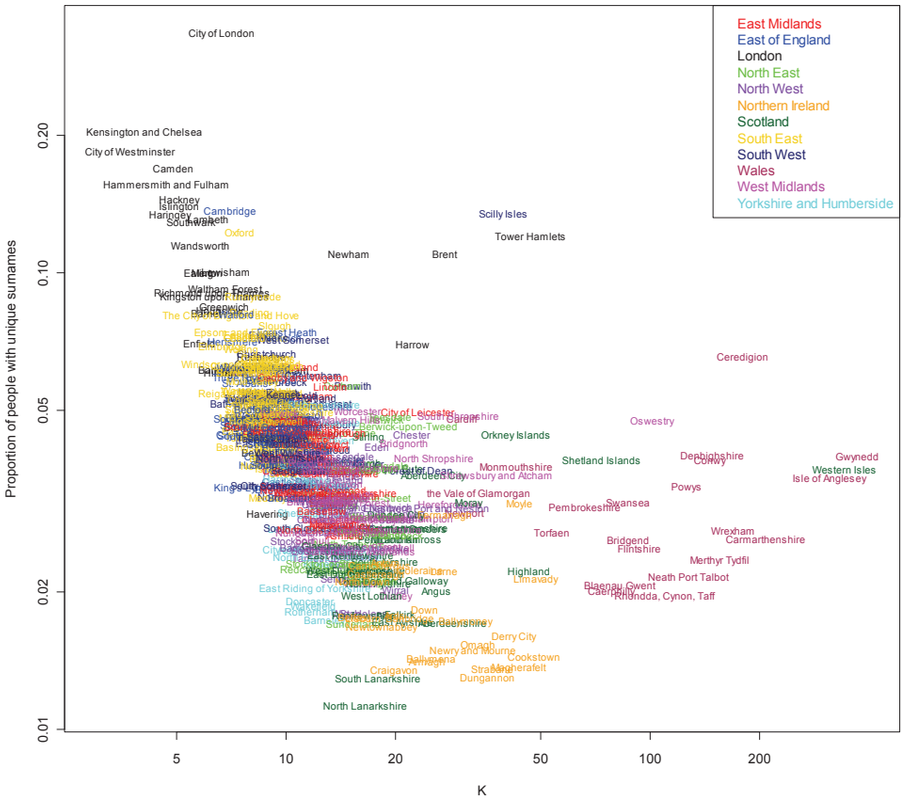

Figure 2. Proportion of unique surnames *versus* Yule's *K*

The geography of surname frequencies

We analysed data from 45 614 126 people on the enhanced electoral register from 434 UK administrative districts, who shared 817 391 different surnames between them. Surprisingly, well over half of those surnames were unique, borne by just one individual on the list; hence, 531 896 surnames represent 65.07% of the total number of surnames, but only 4.41% of the population. The percentage of the population with surnames that occur only twice is 4.93%.

Turning to the other extreme, the most common surnames in the UK are Smith (1.22% of the population), Jones (0.93%) and then Williams (0.64%). A map of the ratio of the number of different surnames to the number of people in each UK district is given in Figure 1. Districts in the south of Britain have a higher proportion of surnames per head of population than those in the north. Put crudely, there are fewer surnames to choose from north of the Scottish border. Gray[15] found that the percentage of names beginning with M—as in MacDonald, McIntyre and the like—in what he called the "crofting counties" of Scotland reached nearly 40%. We found that, in the Western Isles, the equivalent figure was as high as 63.4%. Such names derive from the Gaelic "Mac", meaning "son of".

Figure 2 shows the proportion of unique surnames per head of population (V_1/N) plotted against Yule's K for each of the 434 districts. As expected there is a trend for those districts with a large proportion of unique surnames to have also a greater diversity of surnames (indicated by a low K). However, the extent of correlation between the two measures and the ranges they take varies markedly by districts and, surprisingly enough, also by UK countries and English regions. The correlation is greatest among the English districts. Scotland and Northern Ireland tend to have large K values and Wales clearly has the largest, signifying a higher rate of uniformity of surnames than in the rest of the UK. These three countries tend to have lower proportions of unique surnames than the English districts. There is also a clear pattern among the English regions. Districts in London have much higher proportions of people with unique surnames, and lower values of K, than all other regions, implying a large degree of diversity of surnames. Interestingly, regions in the south of England have higher levels of diversity and unique surnames, compared with those in the north. Oxford and Cambridge are clear outliers with respect to their regions, and Tower Hamlets, Brent and Newham have a large number of unique surnames but much lower diversity than the rest of London.

To illustrate this we compare surname frequencies for three selected districts: the London district of Tower Hamlets, which shows as an outlier on Figure 2, having relatively large K (low diversity) but a large proportion of people with unique surnames; the Isle of Anglesey, in

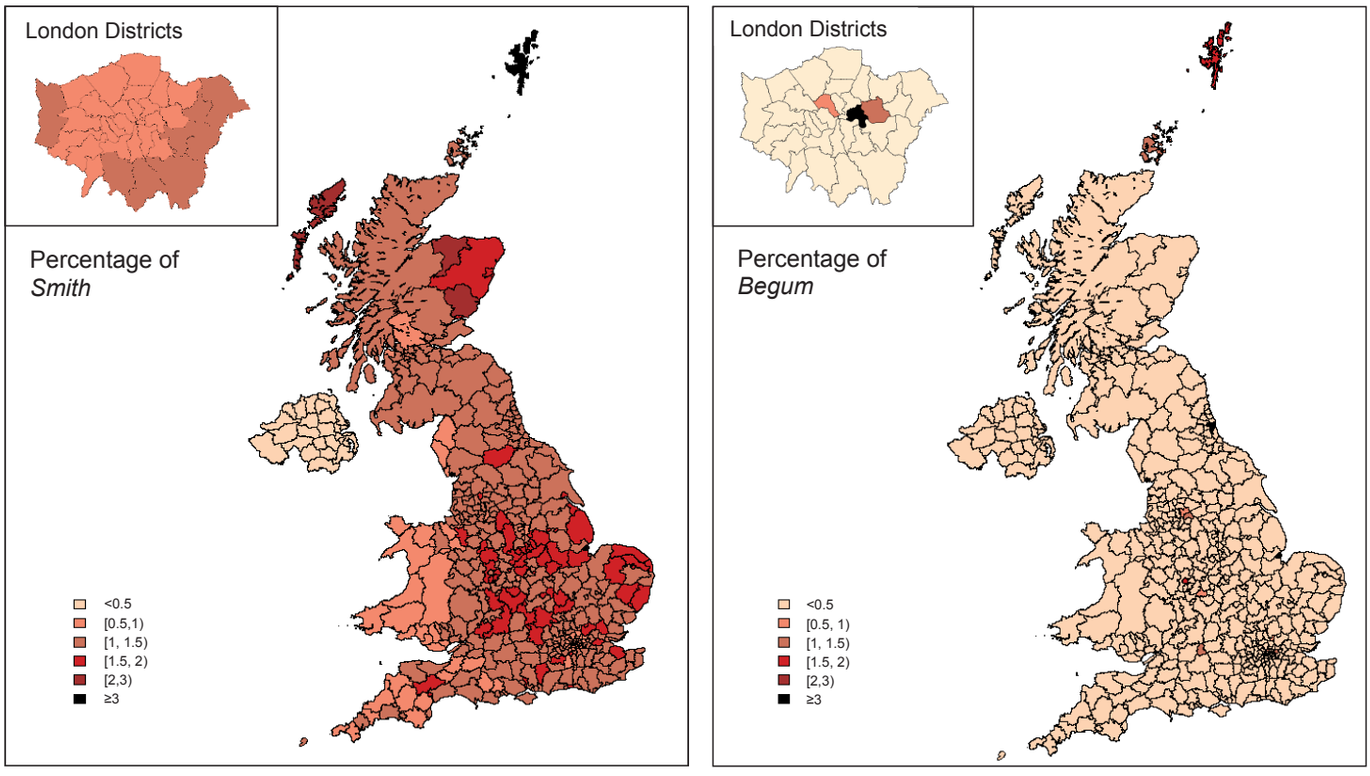

Figure 3. Map percentage of "Smiths" and "Begums" by district

North Wales, which has a very large *K*; Bedford, in the English Midlands, which with a value of *K* of 8.06 and having 5.05% of people with unique surnames is a typical English district in terms of surnames. The Isle of Anglesey had a slightly lower percentage of unique surnames than Bedford and Tower Hamlets, but much lower diversity. Tower Hamlets had the highest proportion of the population with surnames occurring once or twice but also a relatively high percentage of groups of more than 600 with the same surname, suggesting high diversity but also some large clusters of people sharing a surname.

The most common surname in Tower Hamlets is Begum (of Bangladeshi origin, which is actually the honorific in Urdu for Mrs). In Bedford it is Smith, which is also the most common British surname being the most frequent in 308 out of the 434 (71%) districts. By contrast, Begum, placed at rank 104 in the UK with 0.10% of the total population bearing this surname, is the most common name only in the London districts of Tower Hamlets and Camden. In Anglesey the 10 most frequent surnames are of Welsh origin (Jones, Williams, Roberts, Hughes, etc.). Figure 3 shows the geographic distributions of the frequencies of Smith and Begum across districts. The surname Smith has a lower prevalence in districts in Northern Ireland and Wales than in England and Scotland. Its highest percentages are found in Shetland and the Western Isles followed by the Scottish districts of Moray and Angus. Begum, as you would expect, is less prevalent nationally but the districts of Tower Hamlets, Camden, Sunderland, Wolverhampton and, perhaps surprisingly, Shetland have particularly high percentages.

We have calculated the percentage of the population with the top 20 ranked surnames for each country of the UK and present the data in Table 1. Wales has the highest cumulative percentage for these surnames (31.61%) followed by Scotland, Northern Ireland and England (12.40%, 11.51% and 8.30%, respectively). These figures again highlight the much lower diversity of surnames in Wales. Examination of the 20 most common surnames reveals that all of them originate from the UK or Ireland. For example Jones, Williams and Evans are considered to be of Welsh origin, whereas the names Robertson, Thomson and Campbell are considered to be of Scottish derivation. Similarly, Irish surnames (e.g. Kelly, O'Neill and Doherty) and Scottish surnames (e.g. Campbell, Johnston and Thompson) rank highly in the top surnames in Northern Ireland. English origin surnames, however, occur in all four UK countries: Smith (or Smyth) and Brown arise in the top 20 surnames for all countries. Some surnames of Scottish or Welsh origin also appear in the top 20 surnames of England and, more markedly, of Northern Ireland.

Comparisons of the distributions of the top 10 surnames from the 1881 census with our 2001 data (`http://www.nationaltrustnames.org.uk/`) show few changes, with Smith, Jones, Williams, Brown, Taylor, Davies and Wilson appearing in the top 10 in both centuries. This permanency of the top UK surnames across centuries was also found by Tucker[16].

Discussion

In this study we found that London, the South East and the East of England have higher surname diversity than the rest of the UK. Large urban areas attract international and domestic migration and generate a large and ethnically diverse population, with high population mixing, and therefore develop a higher diversity of surnames.

Although all London districts head the top ranks both in terms of number of unique surnames and surname diversity, the London districts of Tower Hamlets, Brent and Newham present unexpectedly low levels of diversity given their large number of unique names. People there seem to have either unique names or very common names with not much in the middle. This finding, which runs counter-intuitively to the overall results otherwise presented here, seems to indicate a new process of sustained immigration combined with local high "ethnic group endogamy" (low intergroup mixing). In other words, the ethnic groups in these districts are not intermarrying. The introduction of a large number of rare surnames from abroad is combined with the local propagation of a few surnames very common among the largest

Table 1. Top 20 most common surnames by country†

Rank	Top surnames by country			
	England	Northern Ireland	Wales	Scotland
1	Smith (1.26)	Wilson (0.75)	Jones (5.75)	Smith (1.28)
2	Jones (0.75)	Campbell (0.75)	Williams (3.72)	Brown (0.94)
3	Taylor (0.59)	Kelly (0.74)	Davies (3.72)	Wilson (0.89)
4	Brown (0.56)	Johnston (0.69)	Evans (2.47)	Robertson (0.78)
5	Williams (0.39)	Moore (0.62)	Thomas (2.43)	Thomson (0.78)
6	Wilson (0.39)	Thompson (0.61)	Roberts (1.53)	Campbell (0.77)
7	Johnson (0.37)	Smyth (0.60)	Lewis (1.53)	Stewart (0.73)
8	Davies (0.34)	Brown (0.59)	Hughes (1.23)	Anderson (0.70)
9	Robinson (0.32)	O'Neill (0.57)	Morgan (1.16)	Scott (0.55)
10	Wright (0.32)	Doherty (0.54)	Griffiths (0.96)	Murray (0.53)
11	Thompson (0.31)	Stewart (0.54)	Edwards (0.93)	MacDonald (0.52)
12	Evans (0.30)	Quinn (0.51)	Smith (0.85)	Reid (0.52)
13	Walker (0.30)	Robinson (0.50)	James (0.82)	Taylor (0.49)
14	White (0.30)	Murphy (0.49)	Rees (0.81)	Clark (0.47)
15	Roberts (0.28)	Graham (0.48)	Jenkins (0.69)	Ross (0.43)
16	Green (0.28)	Martin (0.45)	Owen (0.67)	Young (0.42)
17	Hall (0.28)	McLaughlin (0.45)	Price (0.67)	Mitchell (0.41)
18	Wood (0.27)	Hamilton (0.44)	Phillips (0.65)	Watson (0.41)
19	Jackson (0.27)	Murray (0.43)	Morris (0.63)	Paterson (0.40)
20	Clarke (0.26)	Hughes (0.41)	Richards (0.55)	Morrison (0.40)

†Figures in parentheses are percentages.

minority ethnic groups present in these districts (such as Begum). This is further confirmed in a recent study from the Office for National Statistics[17], which concluded that Brent and Newham are the most ethnically diverse districts in the UK but also that most of the Bangladeshi population in the UK are concentrated just in Tower Hamlets and Newham. This study, based on population projections, seems to corroborate what we have found here based on surnames frequencies.

Regions with a lower level of surname diversity include Northern Ireland, Scotland and, particularly, Wales. Parts of these regions are highly rural, have smaller populations, and present important physical barriers to demographic and cultural interaction. Over time this has generated a lower diversity in their surname pool. This effect is known in population genetics as "isolation by distance"[3].

Such an isolation effect has indeed been found in the island districts, which with smaller, and obviously more detached, populations appear as clear outliers regarding surname diversity. For example, the Western Isles, Orkney, Shetland and Anglesey all present a lower number of unique surnames and therefore less population diversity. An unusual case is the Scilly Isles, which has one of the highest proportions of unique surnames *per capita* in the UK but an average level of diversity, thus indicating relatively large and probably recent levels of immigration of surnames different from those native to the Isles, although Scillonian surnames (e.g. Hicks and Pender) still occupy the largest frequencies. Another interesting outlier is Oswestry, on the English side of the Welsh border, which exhibits a surname structure closer to the Welsh districts' than those of other English border districts.

The prevalence of the most common surnames has not changed between 1881 and 2001. Over time the overall number of different surnames has increased although the rank ordering of the most prevalent has remained relatively static. Important differences in the diversity of surnames can be seen across regions of the UK, where Wales exhibits low diversity and districts in London have a wider variety of surnames. Such differences can be explained by the contemporary and historic differential rates of domestic and international migration as well as a wide variety of degrees of mixing between social and ethnic groups.

Analysis of surname frequencies has the potential to be used as a tool in the study of population structure at different geographical scales in an increasingly globalised world. It can be used to investigate the extent and size of domestic and international migration, regional and local mobility, social and ethnic group mixing and the limits of historical cultural regions. Statistical analysis of surnames can provide better understandings of the geographical and social interaction of people.

References

1 Darwin, G. H. (1875) Marriages between first cousins in England and their effects. *Journal of the Statistical Society of London*, **38**, 153–184.

2 Colantonio, S. E., Lasker, G. W., Kaplan, B. A. and Fuster, V. (2003) Use of surname models in human population biology: a review of recent developments. *Human Biology*, **75**, 785–807.

3 Lasker, G. W. (1985) *Surnames and Genetic Structure*. Cambridge: Cambridge University Press.

4 Piazza, A., Rendine, S., Zei, G., Moroni, A. and Cavalli-Sforza, L. L. (1987) Migration rates of human populations from surname distribution. *Nature*, **329**, 714–716.

5 Jobling, M. A. (2001) In the name of the father: surnames and genetics. *Trends in Genetics*, **17**, 353–357.

6 McEvoy, B. and Bradley, D. G. (2006) Y-Chromosomes and the extent of patrilineal ancestry in Irish surnames. *Human Genetics*, **119**, 212–219.

7 Shriver, M. D. and Kittles, R. A. (2004) Genetic ancestry and the search for personalized genetic histories. *Nature Reviews Genetics*, **5**, 611–618.

8 Cavalli-Sforza, L. L. and Feldman, M. (1981) *Cultural Transmission and Evolution: a quantitative approach*. Princeton: Princeton University Press.

9 Scapoli, C., Mamolini, E., Carrieri, A., Rodriguez-Larralde, A. and Barrai, I. (2007) Surnames in Western Europe: a comparison of the subcontinental populations though isonymy. *Theoretical Population Biology*, **71**, 37–48.

10 Mateos, P. and Tucker, D. K. (2008) Forenames and surnames in Spain in 2004. *Names,* **56**, 165–184.

11 Equifax (2007) ER voter opt-out hits record levels. *Data Strategy*, vol. 7 February, no. 6.

12 Cortina-Borja, M. and Chappas, C. (2006) A stylometric analysis of newspapers, periodicals and newscripts. *Journal of Quantitative Linguistics*, **13**, 285–331.

13 Manni, F., Heeringa, W. and Nerbonne, J. (2006) To what extent are surnames words?: comparing geographical patterns of surname and dialect variation in the Netherlands. *Literary and Linguistic Computing*, **21**, 507.

14 Yule, G. U. (1944) *The Statistical Study of Literary Vocabulary*. Cambridge: Cambridge University Press.

15 Gray, P. G. (1958) Initial letters of surnames. *Applied Statistics*, **7**, 58–59.

16 Tucker, D. K. (2004) The forenames and surnames from the GB 1998 Electoral Roll compared with those from the UK 1881 Census. *Nomina*, **27**, 5–40.

17 Office for National Statistics (2006). *Focus on Ethnicity and Religion*. Internet communication. http://www.statistics.gov.uk/pdfdir/eth1006.pdf

Fiona McElduff, Angie Wade and Mario Cortina Borja are statisticians at the University College London Institute of Child Health; Pablo Mateos is lecturer in Human Geography at the University College London Department of Geograpy. Fiona is doing a PhD on long-tailed discrete distributions supported by an MRC Capacity Building Studentship.