8 : Name concentration and dispersion and how these vary by name, region and level of geographical granularity

9 : 'Adjusted' variations – enabling comparisons over names and places

Daryl Lloyd

As discussed in paper 2, the main level of geography that our work has covered (except for Kevin Schürer's work) has been at the postcode area. Richard's later paper (number 20) will deal with fine-scale patterns within the constituency of Falmouth and Camborne, in Cornwall, and will demonstrate that interesting patterns can emerge even at this level.

Surname indices

As with many geographical phenomena it is not realistic to use raw values as a data source. If we were to compare, for instance, the number of *Lloyds* in Birmingham (B) and Llandrindod Wells (LD) it is immediately apparent that in 1881 there were far more in B (1,407) than in LD (792). However, this is not a fair comparison – the base population in B was over 15 times larger than that of LD (see Table 2 in paper 2) – quite simply, we would expect there to be far more *Lloyds* in B than LD as there are more people who could possibly have that name.

Therefore to allow comparisons we are forced to use an index value, which compares the number of people with a given name to the number of people we would expect with that given, based on the background population.

The formula to calculate this is relatively simple:

$$Si = \frac{S_{LT}}{\left({}^{S_{NT}}\!/\!{}_{P_{NT}} \right) * P_{LT}} * 100$$

Where *Si* is the calculated surname index, any *S* is to do with a given surname and any *P* relates to the base population. The subscripts are defined as *L*, *N* and *T*, which mean *local*, *national* and *total* respectively.

This equation, when calculated for each name, produces a value centred on 100. An index score of 100 indicates that there are as many people with a given surname as we would expect, whilst anything scoring higher tells us that there are more than we would expect. The opposite is also true – anything scoring under 100 gives an indication of there being fewer individuals of the given surname than should be based on the basic background count.

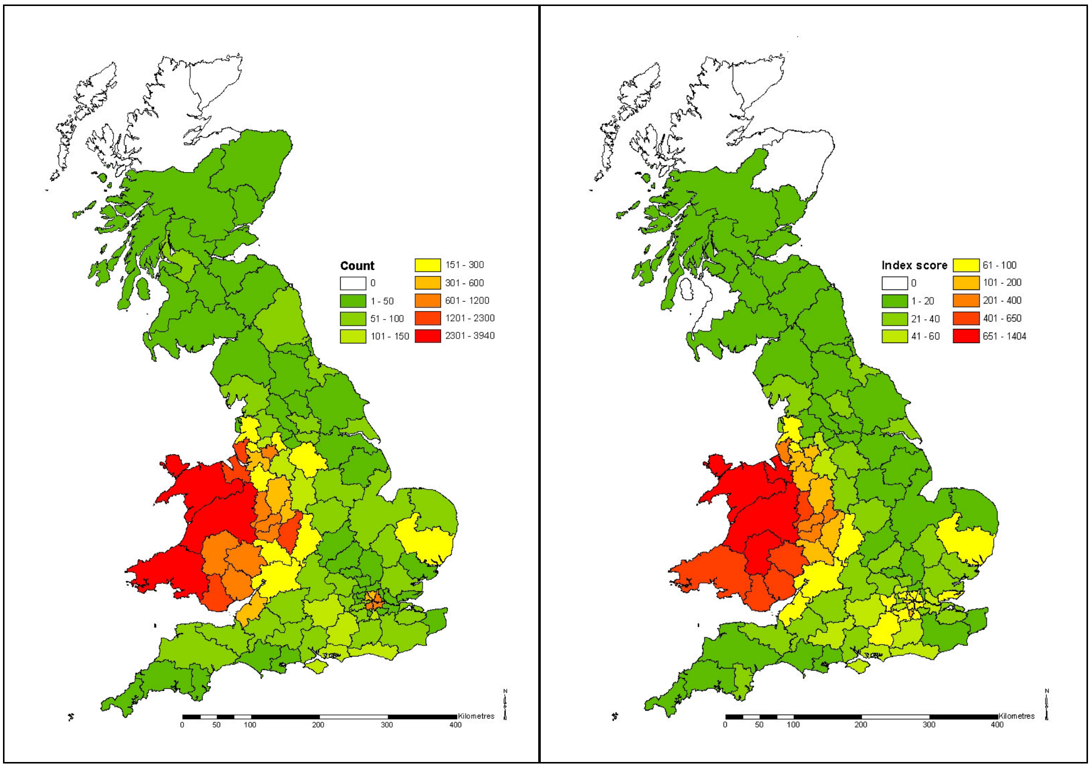

Figure 1: Left – Map of counts of *Lloyds* in 1881; Right – Map of index score of *Lloyds* in 1881 (the 'missing' Scottish data for AB and KA is as a result of rounding)

Concentration and dispersal of names

Once we have index scores it becomes possible to compare names across the country, irrespective of the base population size of each postcode area. In Figure 1 the left hand map shows that *Lloyd* was more common throughout Wales than England or Scotland in 1881. It also, in addition, suggests that there are rather a lot of people with that name in Birmingham and London and Liverpool. Once this has been converted an index though, all these cities' peaks subside (though are still partially evident in Birmingham and Liverpool at least). Not immediately notable, at least from the maps, is the change in LD. In terms of raw numbers, there are a number of other areas which have more *Lloyds* than LD does, yet when we convert the data to indices there is nowhere in the country that scores higher than LD. This, therefore, relates back to Richard Webber's paper on the epicentre of surnames (paper 7).

Each name, once mapped like this, will have its own unique pattern across the country. In some cases, for instance *Smith* (see Figure 2) have relatively little pattern. Wales and the far south-west show up as being below what is expected, but only by a maximum of 50 per cent down. Equally, the areas which are over-represented are rarely even double the expected rates.

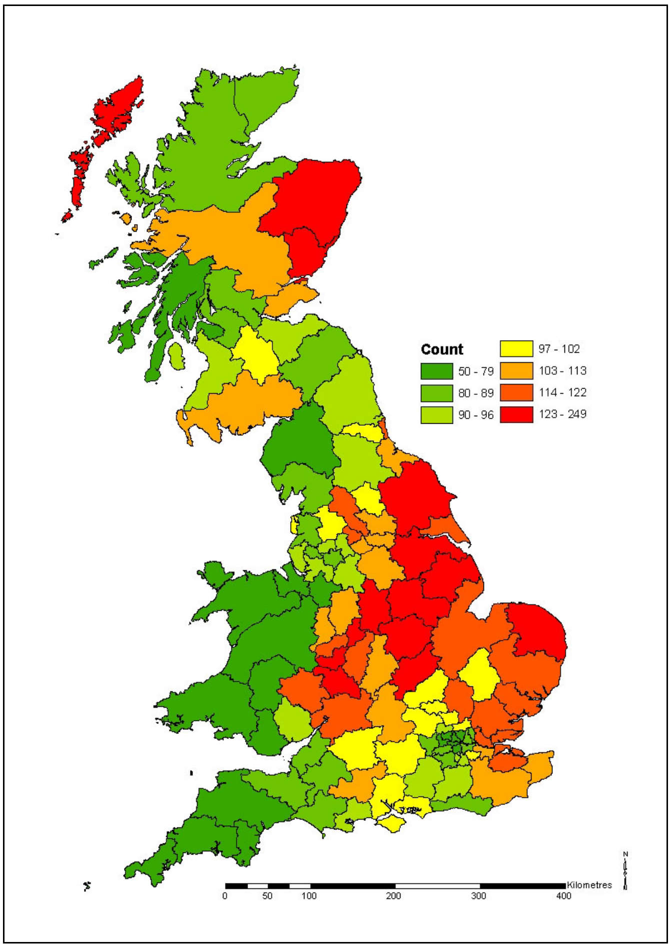

Figure 2: Index scores of Smith.

Smith, therefore, shows a relatively high level of dispersal across the country – there are some deviations from predicted scores, but these are rarely very high. Even *Lloyd*, with its obvious Wales-orientated distribution (which it has in common with many other Welsh names such as *Edwards* or *Jones*) can be said to be fairly well dispersed, with a degree of concentration in one location. On the other hand, there are still some names which have a very localised distribution. As example of this, Figure 3 shows the distributions of the names *Midgley* (left) and *Illingworth* (right) in 1998. In both cases large areas of the country have no individuals at all with the names, yet in the areas around Halifax and Bradford their peak index values come out at 2,284 and 2,709 respectively. So whereas at its peak *Smith* was only 2 ½ times more common than expected, these names are more like 22-25 times more common than would be predicted.

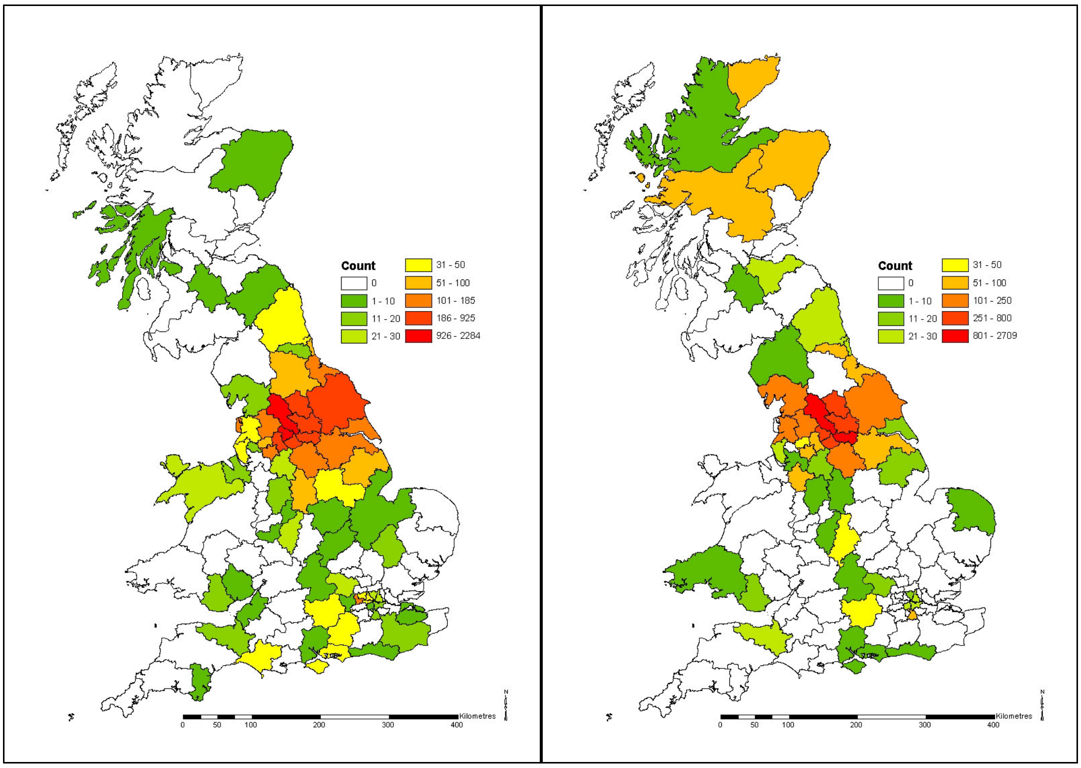

Figure 3: Left – The 1998 index distribution of Midgley; Right – The 1998 index distribution of Illingworth

A similar approach can be taken with smaller areas than postcode areas – a good example can be found in paper 20 on Cornish migration, where Richard Webber looks at Cornish names in the constituency of Falmouth and Camborne. In addition to this, much of Kevin Schürer's work has been with the Census Parishes from the 1881 Census. What could by hypothesised is that names such as the ones shown in Figure 3, which seem highly concentrated at the level of postcode areas, maybe even more concentrated at finer scales. This may also be more applicable depending on the year – it is more likely that a localised name based in one small area would have been less dispersed in 1881 than it is today as the levels of migration over recent years are considerably higher than they were 120 years ago.

Mapping by typology / classification

Later on in the session, Richard Webber will talk about how it is possible to classify names into families of similar names (for instance toponyms, metonyms or foreign). For many of the GB98 surnames, this has been carried out down to three levels, so, for instance, *Lloyd* is classified as (in descending order) 'Name, Forename, Welsh. By combining the index values from each of the names in each category together (weighting them by the constitute surnames' size) and new index value can be produced for the category. As these new indices can also be mapped for the classes, it is possible to compare this to the constitute names, which will help to give an indication if the classification is correct. If, therefore, the combination of a group of names which are perceived to be regional is much more dispersed than the original names, that it suggests that the classification was incorrect on construction.

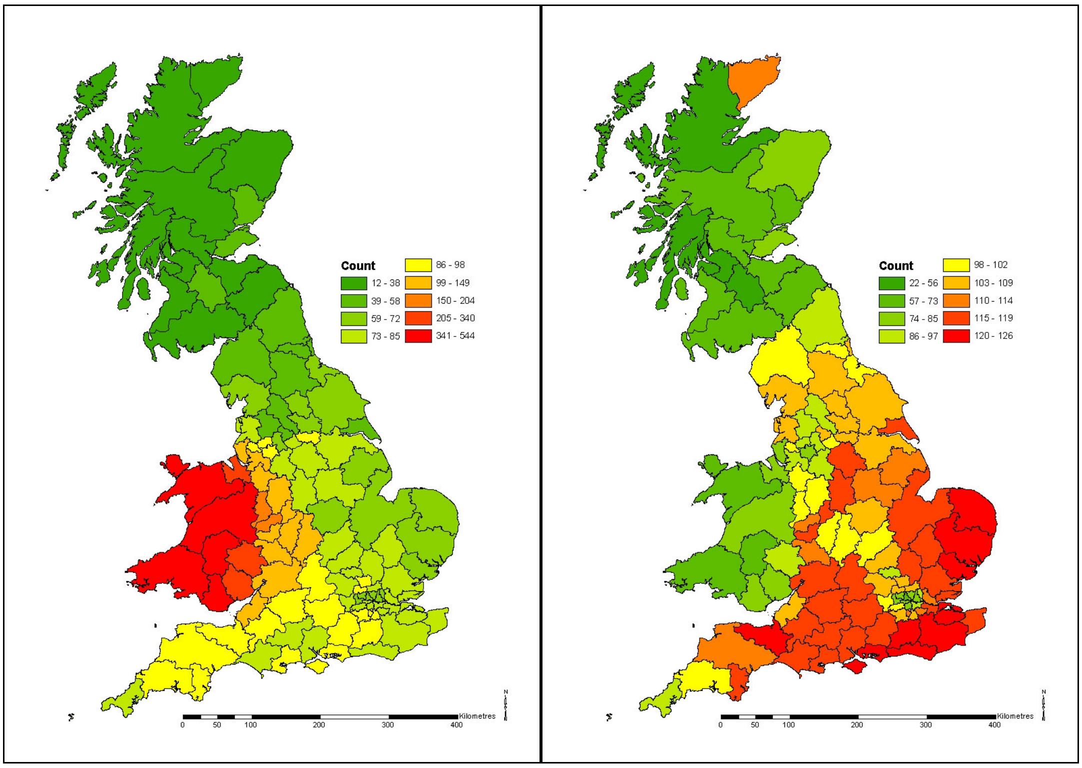

Figure 4: Left – Distribution of the 1998 "Name, Forename, Welsh" classification; Right – Distribution of the 1998 "Metonym, Er" classification

Figure 4 left demonstrates that the typology into which *Lloyd* is classified shows a very similar pattern to the names with in it. There are some slight differences but these are mainly as a result of the difference in range of index scores (the typology maximum index score is 544, but for *Lloyd* it is 3,940). It is a relatively easy task to produce a map of the differing distribution between any name and its typology by normalising both patterns and calculating the difference. This would allow one to see if the name is less or more concentrate than the other names which have been similar classified. This stands out very clearly in Figure 5, as the red areas are the postcodes where the *Lloyd* distribution is greater than the distribution of all the names in the same typology.

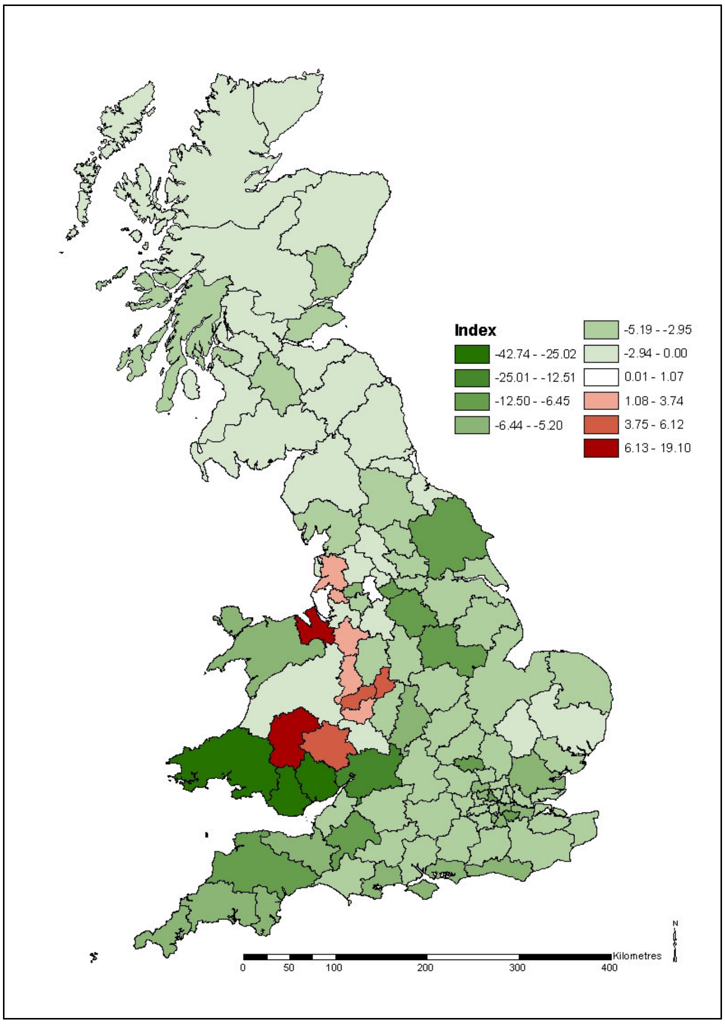

Figure 5: The difference in index score (normalised) between the classification group "*Name, Forename, Welsh*" and the surname *Lloyd*

As a counterpoint to the "*Name, Forename, Welsh*" category, which is clearly fairly concentrated on, unsurprisingly, Wales, the right-hand map shows a much more disperse pattern. Again, there is some bias towards the rural areas of England, but the peak index value is very low, only 126, and the lowest value is 22, meaning that nowhere has particularly more "*Metonym, Er*" type names than would be expected.

Variation by base sizes

When studying the GB98 data, it becomes obvious that there is a relationship between indices and the respective surname and population size. Names with large number of occurrences, such as *Smith* or *Jones*, contain relatively little variation in their index values, whilst small names do.

Smith, for example, has a very low standard deviation of 25 across the 120 postcode areas, with a minimum index of 50 and maximum of 249. *Brown* (st. dev. 31), *Taylor* (st. dev. 28) and *Wilson* (st. dev. 28) all have very similar patterns. At the other end of the scale things are very different. Names such as *Brydon* (with a total population of 1,102) have much higher standard deviations, typically coming in at 200-350, with ranges more like 0 to 800-3,000. In the case of the very smallest names standard deviations in excess of 600 are not unusual, and some of the highest index scores are in the region of 9,000.

	Large <----- Postcode Area -----> Small						Islands
Small <---- Surname ----> Large	66.97649	104.0547	123.3617	45.72743	71.38715	87.36426	114.5136
	79.96946	66.90146	89.39307	103.6835	96.37499	137.7514	525.3124
	108.7157	90.64302	127.1007	134.6313	135.2451	191.8378	1139.261
	143.4714	145.0478	171.4751	187.0547	208.5915	237.4454	1038.858
	229.7391	245.4744	279.06	310.8743	354.5259	419.6614	1586.275

Table 1: Standard deviations of surname indices clustered by surname size and postcode area size.

This pattern can additional be represented by clustering groups of similar sized surnames and postcode areas together in a grid, and viewing the standard deviation across all of the clusters together. This can be seen in Table 1 where both the surnames and postcode areas have been clustered together to represent 20/25 per cent of the population in each direction. Therefore towards the top left there are relatively few unique surnames and postcode areas, whereas moving down and across the grid the number of names and postcodes falling within each cluster increases. The only variation on this is that the final column gives the standard deviation for the most unusual postcode areas in the country: the three island areas in Scotland (KW – Kirkwall, Orkney Islands, ZE – Lerwick, Shetland Islands, and HS – Harris, Outer Hebrides) and the two Central London areas (WC and EC).

Furthermore, this grid is statistically valid, and there is a correlation significant to the 0.01 confidence level that there is negative relationship between the deviation from an index score of 100 and the base postal area and surname total populations.

This variation adds some bias to viewing clusters of surnames. A 'cluster' of small names is far more likely to appear to be significant than a 'cluster' of large names. Equally, the size of the postcode area in which the cluster may appear also partly dictates the relative significance of the clusters. This is, therefore, a classic example of the modifiable areal unit problem (MAUP) occurring in two dimensions – both the physical areal units making up the postcode area as well as the non-areal units defining the surname size. As a result of this it becomes very difficult to say whether one surname is more or less regionally clustered than any other, as this is a function of the surname and postcode area size as well as the distribution of the surname.

Standardisation of surname deviations

To account for this variation, we have taken a route of predicting deviation from the expected index score (i.e. difference from 100). Through the use of a regression model it is possible to predict how great this deviation should be, given the postcode area size and the total surname size, and compare this to the actually difference between the index score and 100.

To construct this a ten per cent sample of the population (in terms of surname / postcode area combinations) was taken, with a weighting by surname total. This meant that large names, such as Smith and Jones, were very well represented in the model, with each possible combination being used. Very small names, on the other hand, had a very small chance of being selected. This allowed the model to be biased towards the more stable and less various larger names over the smaller names with greater levels of variation.

A basic linear regression model was used, with the natural logarithms of surname size and postcode area population used as the independent variables, and the natural logarithm of the index deviation from 100 used as the dependent variable. Using natural logarithms rather than original values is important as the data are not normally distributed, and the natural logarithm helps account for this.

The r^2 score for the regression is not particularly high (0.127), but this is not too important, as if it were then it would demonstrate that this is no reason to carry out such a standardisation. If the residuals were very small and the predicated deviation always came out very close to the actual deviation then there would be no variation by base size in the first place.

The final equation produced is:

Expected deviation = exp(((-0.194 * LN(surname count) + (-0.176 * LN (postcode area population)) + 7.698)

Application of standardised calculations

The equation outlined above allows us to predict how different any given surname size / postcode area size index value will be from 100. By carrying this through an additional stage, it is possible to use this predicted deviation with the actual deviation from 100 to produce a standard deviation figure, which can be used in place of an index score.

By carrying out the significance of deviation calculation:

$$SigDev = \frac{IS - 100}{PD}$$

where IS is the Index Score and PD is the predicated deviation, a new value is produced. If this value falls between -1 and +1, then the value is declared to be within normal parameters – i.e. there are roughly as many individuals with the interested name in the postcode area, given the base counts. If, however, the SigDev exceeds -1 or +1 then the name is either under-represented or over-represented beyond that which we would expect given our base parameters.

It is much easier to follow this through a worked example:

Take the surname *Webber* in the postcode area of Torquay (TQ).

Total number of GB *Webbers* (1998): 9,768

Total population in TQ: 231,372

Index score for *Webber* in TQ: 580

Predicated index deviation: exp(((-0.194 * LN(9768) + (-0.176 * LN (231372)) + 7.698) = 38.2

Significance of deviation: $\dfrac{580-100}{38.2}=+12.8$

In this case it is clear that the number of *Webbers* living in TQ is considerably higher than that which we would expect given both the national level of *Webbers* and also the fact that it is a relatively small name.

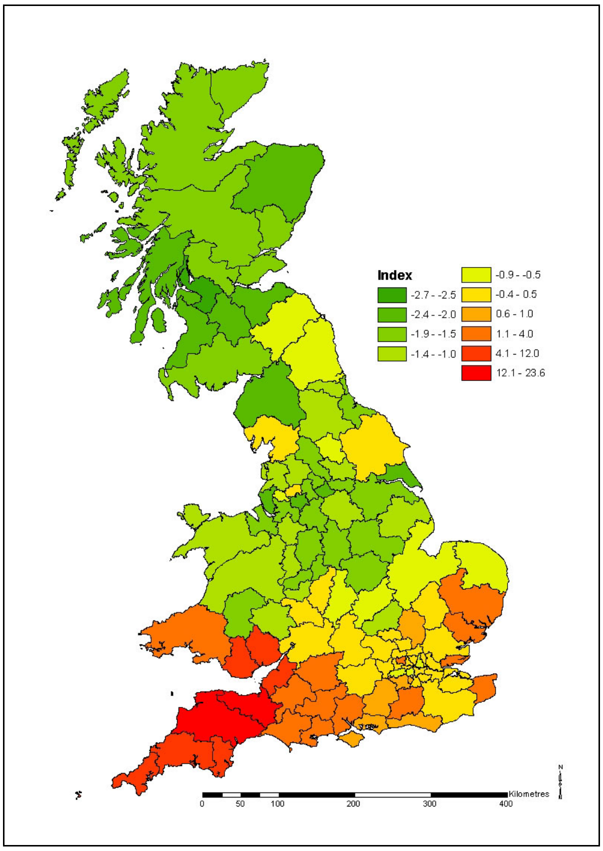

Figure 6: Adjusted and standardised significant deviation scores for Webber in 1998

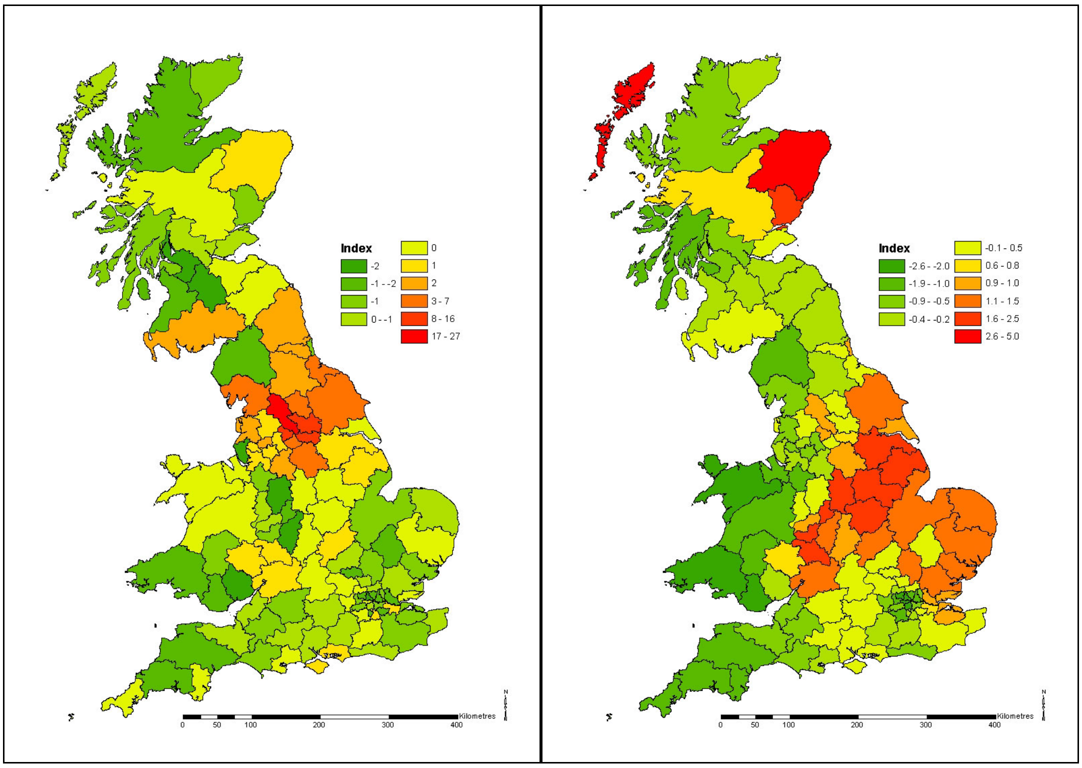

Figure 7: Left – Adjusted 1998 scores for Illingworth; Right – Adjusted 1998 scores for Smith

Surname peaks and centroids

Once we have derived these standard scores, this allows us to revisit the work demonstrated in paper 7 by Richard Webber. We might find that a surname's old peak location moves – i.e. the area with the highest standardised adjusted score is not the same as the location with the highest index score.

It is possible to count how many surnames have their peak in any given postcode area, and map this. Those area which have many surnames with their peaks within the postcode area can be said to have more propensity towards have localised names. By calculating this off both raw index values in 1998 (as shown in Figure 8 left) and comparing this to this calculated off the adjusted values (Figure 8 right), we can clearly see the areas which come out differently (Figure 9 left). Equally we can use this to compare how things have changed over time (Figure 9 right). The actual data for this is provided in Table 2.

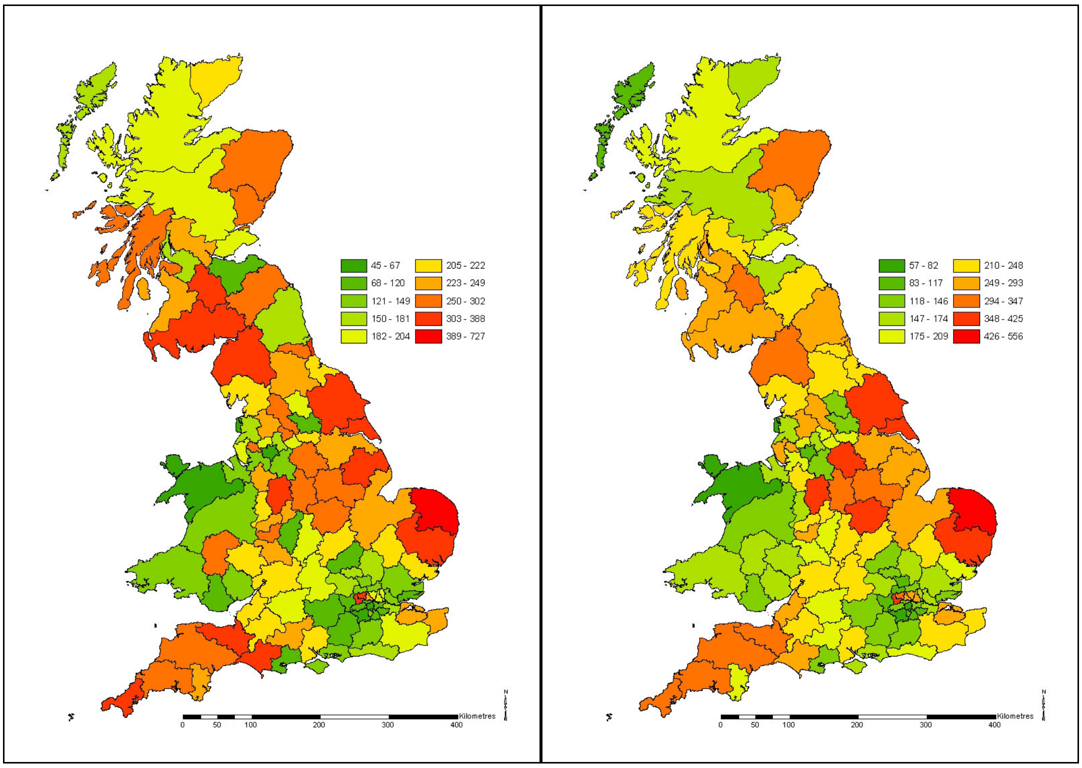

Figure 8: Left – Postcode areas with the number of peak locations of surnames (using 1998 indices); Right – Postcode area with the number of peak locations of surname (using 1998 adjusted values)

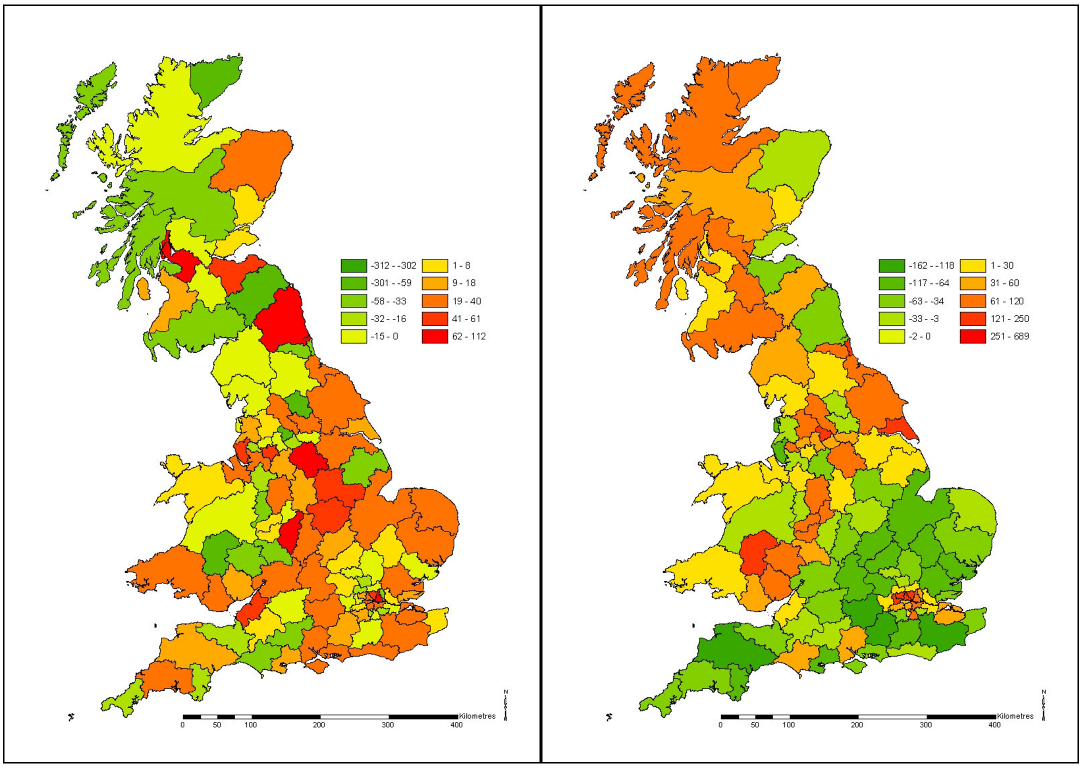

Figure 9: Left – Difference between number of peaks in each postcode area between adjusted values and indices in 1998 (+ve are where there are more resulting from adjusted values); Right – The change between 1881 and 1998 in number of peaks in each postcode area using adjusted values (+ve are where there are more in 1998)

Conclusion

Each individual name, or group of names in their typologies, have a unique spatial pattern throughout the country. In some cases this patter is highly centralised (focussed on one or more points), though others are much more widespread.

In addition to this, a correlation exists in the GB 1998 data between the national size of a given surname, the base population of any postcode area, and its deviation from expected index score (i.e. 100). This results in small names and / or small postcode areas having unusually large or small index values, and thereby suggests that these smaller surnames are more likely to be more unusual than larger names.

To solve this a regression has been produced, which predicts the expected index score deviation from 100, which can then be compared to the actual deviation. Where the difference between the expected and actual deviation is high, then it is possible to state that surname *does* have some unusual characteristic in that postcode area, which is something inherent in the name, rather than as a by-product of its total size.

Postal Area		**Total names for which postal area has highest concentrations**			**Adults 15+ (est) 2002**
		1998 'index'	**1998 'deviations'**	**1881 'deviations'**	
AB	Aberdeen	266	299	289	378871
AL	St. Albans	140	111	149	186846
B	Birmingham	99	181	129	1440771
BA	Bath	221	224	243	332604
BB	Blackburn	249	257	184	370762
BD	Bradford	267	293	183	424784
BH	Bournemouth	114	132	213	443538
BL	Bolton	202	193	143	297356
BN	Brighton	181	207	186	641865
BR	Bromley	111	92	114	240479
BS	Bristol	218	270	191	737262
CA	Carlisle	327	322	290	254884
CB	Cambridge	212	217	286	329601
CF	Cardiff	94	123	58	790946
CH	Chester	137	164	123	531678
CM	Chelmsford	145	171	229	495804
CO	Colchester	210	207	290	327123
CR	Croydon	116	112	48	306312
CT	Canterbury	226	229	269	379266
CV	Coventry	188	220	237	627219
CW	Crewe	238	214	232	241306
DA	Dartford	158	157	130	322719

DD	Dundee	263	266	254	219083
DE	Derby	284	295	269	565314
DG	Dumfries	320	274	246	119802
DH	Durham	267	229	165	253410
DL	Darlington	235	231	221	289284
DN	Doncaster	226	252	222	575568
DT	Dorchester	318	263	259	168283
DY	Dudley	281	286	211	327275
E	London E	198	277	84	611021
EC	London EC	703	391	118	24805
EH	Edinburgh	106	155	153	679274
EN	Enfield	141	137	133	260876
EX	Exeter	301	317	463	430818
FK	Falkirk	243	236	142	211714
FY	Blackpool	67	57	133	239525
G	Glasgow	168	280	159	979402
GL	Gloucester	220	245	271	470383
GU	Guildford	109	127	235	568849
HA	Harrow	336	347	137	352286
HD	Huddersfield	203	187	151	202319
HG	Harrogate	195	132	212	113097
HP	Hemel Hempstead	157	161	229	375958
HR	Hereford	207	163	136	134807
HS	Harris	164	117	77	22040
HU	Hull	371	382	154	355137
HX	Halifax	266	192	144	121514
IG	Ilford	196	161	153	226538
IP	Ipswich	360	385	403	452327
IV	Inverness	199	193	102	162467
KA	Kilmarnock	244	256	231	303651
KT	Kingston-upon-Thames	45	58	101	425055
KW	Kirkwall	209	150	133	40889
KY	Kirkcaldy	189	197	198	284943
L	Liverpool	200	261	273	693443
LA	Lancaster	222	220	219	273062
LD	Llandrindod Wells	269	164	59	40640
LE	Leicester	301	361	307	743257
LL	Llandudno	62	68	35	423418
LN	Lincoln	328	281	312	215936
LS	Leeds	98	136	119	615659
LU	Luton	160	146	178	243225
M	Manchester	60	109	53	873448
ME	Medway	233	263	196	444169
MK	Milton Keynes	120	124	190	367286
ML	Motherwell	315	307	225	303810
N	London N	201	256	20	607067
NE	Newcastle upon Tyne	179	264	213	941595
NG	Nottingham	259	316	295	898113

NN	Northampton	216	226	280	478558
NP	Newport	147	160	53	384119
NR	Norwich	517	556	528	570079
NW	London NW	229	269	22	411635
OL	Oldham	165	162	121	357075
OX	Oxford	202	226	295	488104
PA	Paisley	277	241	187	269173
PE	Peterborough	232	272	297	657833
PH	Perth	199	162	151	127476
PL	Plymouth	299	331	359	434443
PO	Portsmouth	140	171	189	647434
PR	Preston	153	165	171	413660
RG	Reading	98	121	216	597901
RH	Redhill	145	142	221	399659
RM	Romford	145	151	143	375602
S	Sheffield	298	370	224	1080103
SA	Swansea	126	155	104	576487
SE	London SE	106	145	18	689059
SG	Stevenage	167	162	224	307355
SK	Stockport	128	141	185	495151
SL	Slough	120	114	114	278481
SM	Sutton	141	110	119	166284
SN	Swindon	204	202	256	337165
SO	Southampton	214	244	180	521825
SP	Salisbury	228	189	236	180018
SR	Sunderland	328	291	153	208636
SS	Southend-on-Sea	112	129	164	413126
ST	Stoke-on-Trent	336	359	250	520973
SW	London SW	46	82	12	705446
SY	Shrewsbury	133	129	138	262212
TA	Taunton	353	336	393	250590
TD	Galashiels	302	241	247	88707
TF	Telford	219	174	228	157859
TN	Tunbridge Wells	204	233	331	523101
TQ	Torquay	226	203	308	231372
TR	Truro	355	335	403	229427
TS	Cleveland	208	248	96	483428
TW	Twickenham	114	122	80	372632
UB	Southall	388	382	138	272737
W	London W	101	135	75	448373
WA	Warrington	149	181	174	482023
WC	London WC	727	425	38	31077
WD	Watford	127	97	117	200977
WF	Wakefield	211	209	177	396482
WN	Wigan	283	257	206	248415
WR	Worcester	230	197	189	232093
WS	Walsall	249	241	145	342823
WV	Wolverhampton	230	231	139	306965
YO	York	381	410	298	438855
ZE	Lerwick	210	131	89	17820

	Grand Total	26035	26035	22691	47464666

Table 2: The number of surnames with their peaks in the postcode areas. Provided for 1998 and 1881 and uses standard index scores and the adjusted values.