Fitting a multilevel index of segregation in R: using the MLID package

Richard Harris

2017-02-23

Introduction

This tutorial introduces the tools and functions available in the MLID package to fit a multilevel index of dissimilarity, a measure of ethnic or social segregation that captures both of the two principal dimensions of segregation - unevenness and spatial clustering - and looks for scale effects as well as the contributions of particular places to the index value.

To begin, install the package from CRAN by typing

```
install.packages("MLID")
```

Or, for the latest development version:

```
# Needs devtools. Use: install.packages("devtools")
require(devtools)
devtools::install_github("profrichharris/MLID")
```

Next, load the package.

```
require(MLID)
```

About the Index of Dissimilarity

The index of dissimilarity (ID) widely is used in social and demographic research and examines whether the places where one population group are most likely to be located are the places where another group is most likely to be present too. The logic of the index is that if, for example, 1 per cent of population group Y resides in a neighbourhood then, all things being equal, 1 per cent of population group X ought to reside there too. If another neighbourhood is a little bigger and contains 2 per cent of all the Y group then it should contain 2 per cent of all the X group as well. In this way, if the share of the Y group is equal to the share of the X group in each and every neighbourhood then the two populations are said to have an even geographical distribution, described as a situation of ‘no segregation’. However, if wherever the Y population is found, X is not (and vice versa) then there is a situation of ‘complete segregation’.

The ID measures unevenness - how unevenly the two groups are distributed across the study region relative to one another and regardless of how big or small each group is in the total population (all that matters is the share of each group in each neighbrouhood). However, unevenness is only one of the two principal dimensions of segregation. The other is spatial clustering. Although the ID measures the scale of segregation in a numeric sense, giving an amount of segregation, it does not do so in a geographic sense. The classic example is to compare a checkerboard-style pattern of alternating back-white squares with other patterns that have increasing amounts of spatial clustering. In each of the examples below, the ID is the same, showing complete black-white segregation, yet the pattern of spatial clustering is not.[1]

A multilevel index of dissimilarity (MLID) improves upon the standard ID by capturing both the unevenness and the clustering. To see this, run the examples in the MLID package. Note that although the ID value is

always 1.000 the other measures, Pvariance and Holdback, change with the geographical scale of segregation. Those other measures are explained later. All that matters for now is that they are sensitive to the pattern of spatial clustering whereas the standard ID is not.

```
checkerboard()
```

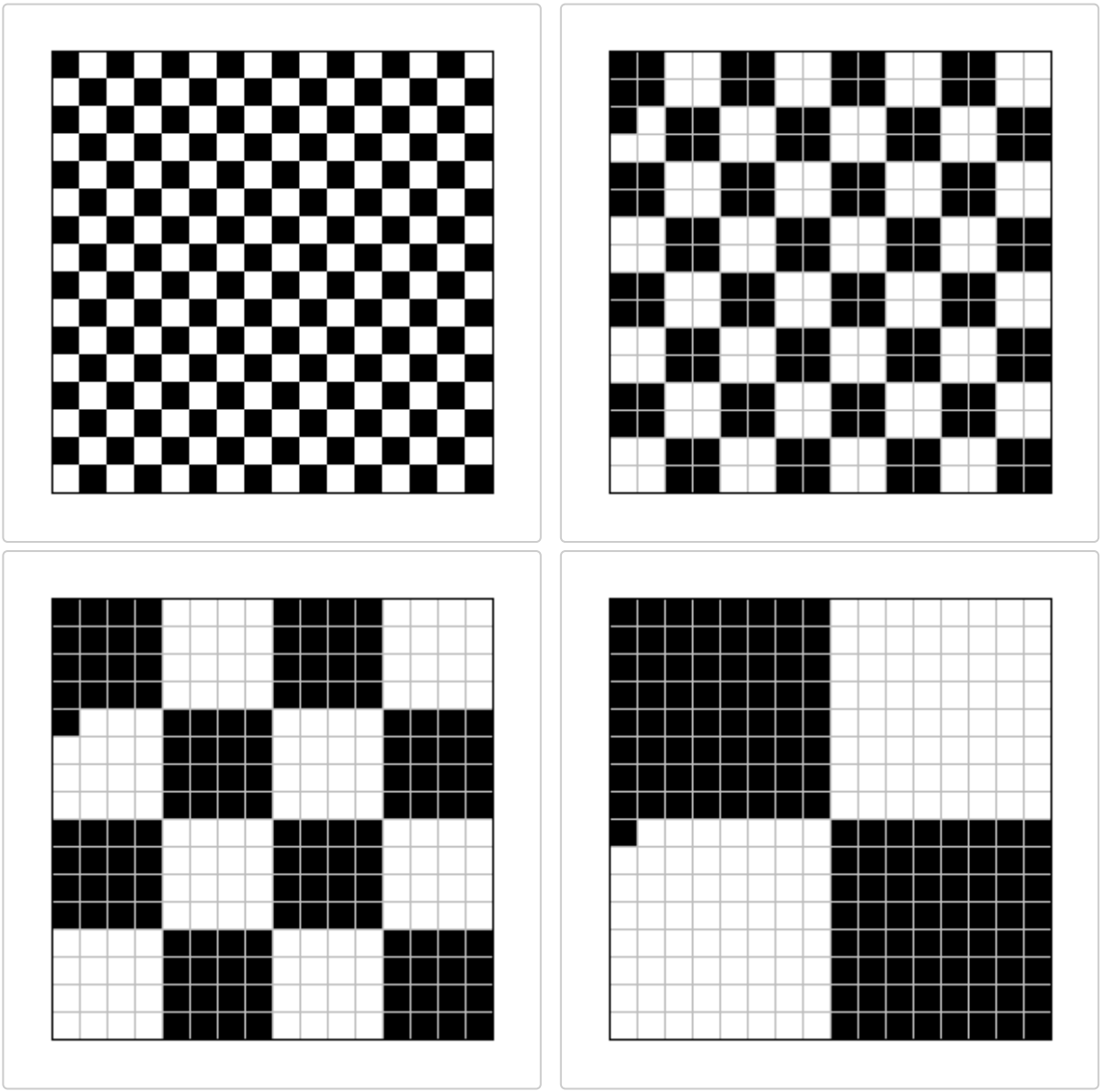

Figure 1. Each of these patterns generates the same ID value yet they represent different degrees of spatial clustering. The multilevel index distinguishes between them.

Calculating the ID and MLID

The index of dissimilarity is calculated as

$$\mathrm{ID} = k \times \sum_i \left| \frac{n_{yi}}{n_{y+}} - \frac{n_{xi}}{n_{x+}} \right|$$

where n_{yi} is the count of population group Y in neighbourhood i, n_{y+} is the total count of Y across all neighbourhoods in the study region ($n_{y+} = \sum_i n_{yi}$), and n_{xi} and n_{x+} are the corresponding values for population group X. Setting the scaling constant to be $k = 0.5$ means that the maximum range for the ID is from 0 to 1.

The index summarises the differences between a set of observed values, $y_i = n_{yi}/n_{y+}$ and what those values would be under an expectation of 'zero segregation', $x_i = n_{xi}/n_{x+}$, which is when the share of the Y population per neighbourhod everywhere is equal to the share of the X population. Substituting y_i and x_i for n_{yi}/n_{y+} and $x_i = n_{xi}/n_{x+}$ in the formula gives

$$\mathrm{ID} = 0.5 \sum_i |y_i - x_i|$$

Writing this within a regression framework,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Setting $\beta_0 = 0$ and $\beta_1 = 1$, and rearranging gives

$$\epsilon_i = y_i - x_i$$

from which the ID can be calculated as

$$\mathrm{ID} = 0.5 \sum_i |\epsilon_i|$$

This shows that the ID is half the sum of the absolute values of the residuals from a regression model where the dependent variable is the share of the Y population per neighbourhood, the intercept is zero and there is an offset, which is the share of the X population.

The multilevel model is achieved by estimating what of the residuals is due to different levels of a geographic hierarchy. For example, for a four level model where neighbourhoods at level i group into districts at level j, those into larger administrative authorities at level k, and then into regions at level l, the residuals can be estimated as

$$\epsilon_i = \hat{\lambda}_i + \hat{\mu}_j + \hat{\nu}_k + \hat{\xi}_l$$

giving

$$\mathrm{ID} = 0.5 \sum_i |\hat{\lambda}_i + \hat{\mu}_j + \hat{\nu}_k + \hat{\xi}_l|$$

The geographical scales of segregation are then explored by looking at the residuals at each level, as the following case study demonstrates

Case Study

Fitting and exploring the standard ID

The data frame

```
data(ethnicities)
```

contains counts of various ethnic groups living in census small areas in England and Wales in 2011. Those small areas are called Output Areas (OAs).

```
head(ethnicities, n = 3)
```

```
##          OA Persons WhiteBrit Irish OtherWhite Mixed Indian Pakistani
## 1 E00000001     194       150     7         18    10      2         0
## 2 E00000003     250       177     2         26     9     17         0
## 3 E00000005     367       254    14         53    12      9         1
##   Bangladeshi Chinese OtherAsian BlkAfrican BlkCaribbean OtherBlk Arab
## 1           0       4          0          0            0        0    0
## 2           3       3          3          3            0        0    0
## 3           0      10          5          2            0        2    0
##   Other      LSOA      MSOA            LAD    RGN
## 1     3 E01000001 E02000001 City of London London
## 2     6 E01000001 E02000001 City of London London
## 3     5 E01000001 E02000001 City of London London
```

To calculate the index of dissimilarity for the residential segregation of the Bangladeshi from the White British, we may use

```
index <- id(ethnicities, vars = c("Bangladeshi", "WhiteBrit"))
index
```

```
## Bangladeshi ~ WhiteBrit
## ID:  0.852
## E(ID):   (not calculated)
```

which generates an ID value of 0.852. The interpretation is that 85.2 per cent of either the Bangladeshi or White British populations would need to move for both to be evenly distributed relative to one another. It seems a lot and reflects the concentration of the Bangladeshi population in particular parts of the country such as London, and especially the Boroughs of Tower Hamlets and Newham within the capital, which the following 'impact' calculations reveal.

```
impx <- impacts(ethnicities, c("Bangladeshi", "WhiteBrit"), c("LAD","RGN"))
head(impx, n = 3)
```

```
## Bangladeshi ~ WhiteBrit

## $LAD
##                pcntID pcntN impact scldMean scldSD scldMin scldMax pNYgtrNX
## Tower Hamlets  10.58  0.41   2566     7.91   6.85   -0.13   30.78     44.5
## Newham          4.82  0.45   1080     3.33   2.41   -0.17   16.57     35.8
## Oldham          2.18  0.39    555     1.51   4.97   -0.33   37.50      7.9
##
## $RGN
##                pcntID pcntN impact scldMean scldSD scldMin scldMax pNYgtrNX
## London         29.70 13.81    215     0.54   2.01   -0.43   30.78      2.9
## West Midlands  10.97  9.88    111     0.03   0.94   -0.69   24.06      1.7
## North West     12.02 12.87     93    -0.05   1.12   -1.01   37.50      0.5
```

The impact calculations take advantage of two things. First, that the census geography is hierarchical and so OAs can be matched to higher-level areas, in this case local authority districts (LADs) and government regions (RGN), as they have been in the data. To confirm this, look again at,

```
head(ethnicities)
```

Second, they use the knowledge that the $\mathrm{ID} \propto \sum_i |\epsilon_i|$ where each ϵ_i is a local value - the difference in the share of the Bangladeshi and the share of the White British populations per OA. Those small areas differences can be summarised by a higher-level geography. For example, the differences in the shares of the two population groups within Tower Hamlets can be summarised as $\sum_i w_i|\epsilon_i|$ where $w_i = 1$ if the OA is located in Tower Hamlets, otherwise 0. As a percentage of the overall ID for England and Wales, OAs within Tower Hamlets contribute,

$$\mathrm{pcntID} = \frac{\sum_i w_i|\epsilon_i|}{\sum_i |\epsilon_i|} \times 100$$

which is 10.58 per cent. This is a disproportionate amount because only 0.41 of all OAs are in Tower Hamlets. Calculating $10.58 \div 0.41 \times 100$ gives the impact of the neighbourhoods within Tower Hamlets upon the overall ID, and is 2566 - i.e. 25.66 times greater than expected. This impact could be because the shares of the Bangladeshi population exceed the shares of the White British or it could be the other way around. It is the former: on average the share of the Bangladeshi population is greater than the share of the White British in Tower Hamlets, shown by the positive value for the mean difference, scldMean. This is calculated as,

$$\mathrm{scldMean} = \frac{\sum_i w_i \epsilon_i}{\sigma_\epsilon \sum_i w_i}$$

which simply is the average difference within Tower Hamlets ($\bar{\epsilon}_k$) , scaled by σ_ϵ, the standard deviation of the ϵ_i.[2] As a rule of thumb, values with a magnitude greater than 2 may be regarded as unusual, which here include Tower Hamlets and Newham. Within both, and especially Tower Hamlets, there is variation from one OA to the next: the standard deviation for the ϵ_i values in Tower Hamlets alone is 6.85 greater than for all of England and Wales. In at least one Tower Hamlets neighbourhood the share of the Bangladeshi population is less than the share of the White British (because the minimum value of ϵ_i, scldMin is negative) and in one it is much greater (the maximum value is 30.78; both the minimum and maximum are scaled by σ_ϵ). Overall, whilst Tower Hamlets is a place within which the Bangladeshi population seems to be disproportionately concentrated relative to the White British, a sense of perspective is deserved: a minority (44.5 per cent) of its neighbourhoods have more Bangladeshi residents than they do White British so the White British are the more prevalent group.

Finding the expected value and aggregating the data

Although in principle the ID ranges from zero ('no segregation') to one ('complete segregation'), in practise, when the two population groups are of very different sizes (and especially when one is small) it is very difficult, if not impossible, for them to be evenly distributed relative to one another. In the current case study, the Bangladeshi group comprise 0.8 of the population of England and Wales, whereas the White British comprise 80.5. With 181408 neighbourhoods to be spread across, there are simply too few Bangladeshis for their distribution to match that of the White British.

An expected value for the ID may be generated that essentially is the value that would arise, on average, if the Bangladeshi and White British populations randomly were assigned to the existing neighbourhoods whilst broadly respecting the population size of each neighbourhood as well as the total number of Bangladeshi and White British overall. The value is obtained by simulation, for which the total population in each neighbourhood should be supplied:[3]

```
index <- id(ethnicities, vars = c("Bangladeshi", "WhiteBrit", "Persons"), expected = TRUE)
index
```

```
## Bangladeshi ~ WhiteBrit
## ID:  0.852
## E(ID):   0.251 (29.5%)
```

In this example, the expected value under randomisation is 0.251 which is 29.5 per cent of the actual ID score. Is that a lot? The answer is a matter of judgment but certainly it is sizable. It suggests that perhaps there are too few of the Bangladeshi population to be analysed at the OA scale.

An option is to take advantage of the census' geographical hierarchy and aggregate the OAs into what are called Lower Level Super Output Areas (LSOAs), calculating and using the population counts for those areas instead. The ethnicities data shows which LSOA each OA belongs to. For example, OA E00000001 is in LSOA E01000001:

```
head(ethnicities, n = 1)
```

```
##          OA Persons WhiteBrit Irish OtherWhite Mixed Indian Pakistani
## 1 E00000001     194       150     7         18    10      2         0
##   Bangladeshi Chinese OtherAsian BlkAfrican BlkCaribbean OtherBlk Arab
## 1           0       4          0          0            0        0    0
##   Other      LSOA      MSOA            LAD    RGN
## 1     3 E01000001 E02000001 City of London London
```

Because the higher-level groupings are known, to aggregate the data and recalculate the ID is simple,

```
aggdata <- sumup(ethnicities, sumby = "LSOA", drop = "OA")
head(aggdata, n = 3)
```

```
##        LSOA Persons WhiteBrit Irish OtherWhite Mixed Indian Pakistani
## 1 E01000001    1465       968    32        237    54     49         3
## 9 E01000002    1436      1024    28        222    54     30         0
## 5 E01000003    1346       813    37        205    55     48         6
##   Bangladeshi Chinese OtherAsian BlkAfrican BlkCaribbean OtherBlk Arab
## 1           3      44         29          6            1        4    6
## 9           0      37         28          0            3        1    3
## 5           9      44         61         23           13        9   13
##   Other      MSOA            LAD    RGN
## 1    28 E02000001 City of London London
## 9     6 E02000001 City of London London
## 5    10 E02000001 City of London London
```

```
index <- id(aggdata, vars = c("Bangladeshi", "WhiteBrit", "Persons"), expected = TRUE)
index
```

```
## Bangladeshi ~ WhiteBrit
## ID:  0.772
## E(ID):   0.11 (14.2%)
```

The ID is now 0.772 with a much smaller expected value of 0.11. The impacts of Tower Hamlets and Newham on the ID remain pronounced.

```
head(impacts(aggdata, vars = c("Bangladeshi", "WhiteBrit"), levels = c("LAD", "RGN")), n = 3)
```

```
## Bangladeshi ~ WhiteBrit
```

```
## $LAD
##               pcntID pcntN impact scldMean scldSD scldMin scldMax pNYgtrNX
```

```
## Tower Hamlets  11.68  0.41   2818     9.18   5.64    0.19   24.48     49.3
## Newham          5.33  0.47   1128     3.68   1.80    0.01   10.26     40.2
## Oldham          2.34  0.41    576     1.70   4.86   -0.34   35.93      8.5
##
## $RGN
##               pcntID pcntN impact scldMean scldSD scldMin scldMax pNYgtrNX
## London         30.29 13.91    218     0.63   2.02   -0.28   24.48      3.0
## West Midlands  11.18 10.03    111     0.04   0.94   -0.43   18.06      2.0
## North West     12.21 12.94     94    -0.05   1.10   -0.50   35.93      0.4
```

Generally, the affect of aggregation is to smooth over some of the variations in the data that may be 'noise' due to the small population size. However, it also risks smoothing out the geographical detail (which is why the ID tends to decrease with aggregation) so there is a cost as well as a benefit.

Fitting a multilevel index

So far we have been fitting the standard index of dissimilarity, aggregating the local differences in the shares of the populations into higher-level geographies and thereby assessing the contributions of those higher-level geographies on the overall ID. The methods take advantage of the hierarchical structure of the data but there is nothing specifically multilevel about them in the sense of multilevel modelling. They are useful for identifying places that contribute most to the ID score but not on separating out the scale effects due to each level *net* of the other levels. To achieve the latter, we need to handle the regression residuals in a different way using a multilevel model.

As an example, consider a four level model with LLOAs as the base, Middle Level Super Output Areas (MSOAs) as the next level up, LADs above that and finally regions. To fit the model:

```
index <- id(aggdata, vars = c("Bangladeshi", "WhiteBrit"), levels = c("MSOA","LAD","RGN"))
index
```

```
## Bangladeshi ~ WhiteBrit
## ID:  0.772
## E(ID):   (not calculated)
##
##      Pvariance Holdback
## Base     25.41    -7.44
## MSOA     32.42    -9.02
## LAD      35.92    -5.09
## RGN       6.25   -14.96
```

The ID value is unchanged and we have chosen to omit the expected value because it is known already. What we do now have are the Pvariance and Holdback scores at each of the levels.

The Pvariance is the percentage of the total variance due to each level. For example,

$$\text{Pvariance}_{Base} = \frac{\hat{\sigma}_i}{\hat{\sigma}_i + \hat{\sigma}_j + \hat{\sigma}_k + \hat{\sigma}_l} \times 100$$

It is a measure of spatial clustering, of the pattern of segregation. What the results show is that the pattern of Bangladeshi - White British residential segregation is primarily at the LAD and MSOA scales.

The Holdback scores are different. They consider what the change would be to the ID score if the effect on it due to the level was set to zero. For example, holding back the regional effect reduces the ID by -14.96 per cent. The holdback score at that level is calculated as,

$$\text{Holdback}_{\,l} = \frac{(0.5 \sum_i |\hat{\lambda}_i + \hat{\mu}_j + \hat{\nu}_k + 0|) - \text{ID}}{\text{ID}} \times 100$$

It may seem odd that in the results above the Holdback score is greatest where Pvariance is least, at the regional scale. However, there is no contradiction because they measure different things. The model is additive so any uplift (or decrease) in segregation that is due to the regional scale applies to all the neighbourhoods within that region, whereas the change due to the LAD scale, for example, is restricted to the smaller sub-group of neighbourhoods that are in that local authority. It is entirely possible for the proportion of the variance to be small at the higher levels but for the differences between places at those levels to still have a strong cumulative effect upon all the lower levels to which they must be added. This will be picked-up on by the holdback scores.

From the initial analysis it may be suspected that the LAD variance is being driven by Tower Hamlets and Newham. This is confirmed by plotting the residuals at each level and their confidence intervals using 'caterpillar plots':[4]

```
ci <- confint(index)
catplot(ci, grid = FALSE)
```

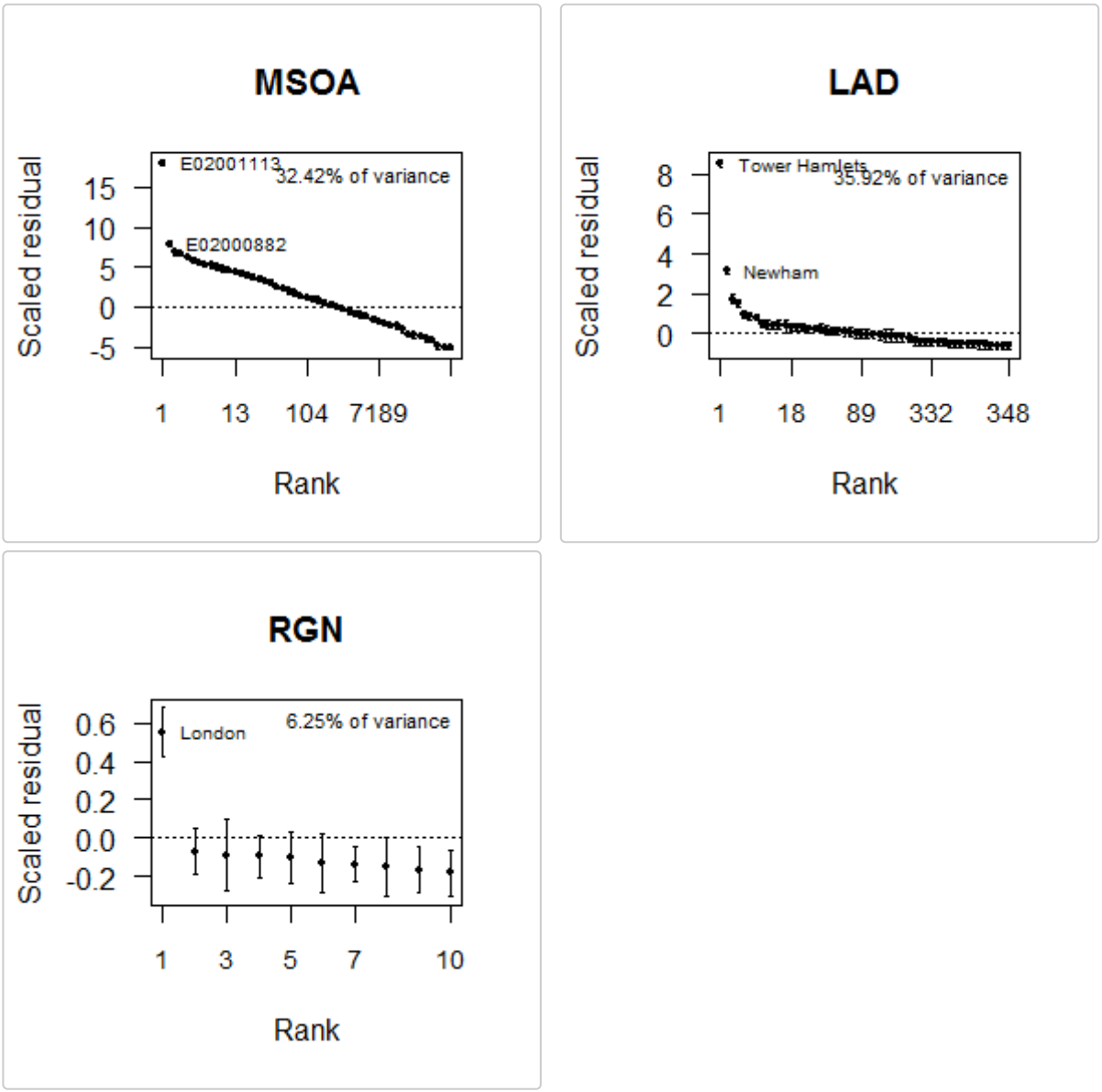

Figure 2. Caterpillar plots of the residuals at each level of the model

To aid the interpretability of the plots, the residuals are scaled by the standard error of the residuals from the OLS estimate of the index (by σ_ϵ). Tower Hamlets and Newham clearly are different from other LADs, with a statistically significant residual difference between the share of the Bangladeshi population and the share of the White British population due to their LAD effect. At the MSOA level, E02001113 stands apart

from the rest; at the regional level so does London but looking at the value for the scaled residual it does not seem especially significant.[5]

Considering the effect of particular places upon the index

What would happen to the ID if the effect of Tower Hamlets and Newham, were omitted? Let's find out.

```
prd <- effect(index, places = c("Tower Hamlets", "Newham"))
prd
```

```
## Bangladeshi ~ WhiteBrit
##          ID1   ID2   ID3
## before 0.772 0.772 0.772
## after  0.703 0.131 0.365
##
## R-squared: 0.381    Impact 1918
##
## Note:
## ID1: if the residual effect of Tower Hamlets Newham is set to zero
## ID2: if the shares of Bangladeshi & WhiteBrit are equal everywhere except Tower Hamlets Newham
## ID3: the index value for Tower Hamlets Newham alone
```

The function evaluates what the value of the ID would be under three different scenarios to give an indication of the effects of the named places upon the current ID. The first is if the LAD level residual effects (ξ_l) were set to zero for Tower Hamlets and Newham, i.e. if

$$\text{ID} = 0.5 \sum_i |\hat{\lambda}_i + \hat{\mu}_j + \hat{\nu}_k + w_l \, \hat{\xi}_l|$$

where $w_l = 0$ for Tower Hamlets and Newham, and 1 for every other LAD. In the present example, it reduces the ID from 0.772 to 0.703, which is 91.1 per cent of the original value.

The second is what the ID would be if the shares of the two population groups were equal, $\epsilon_i = y_i - x_i = 0$, everywhere except Tower Hamlets and Newham. The result is an ID score of 0.131, meaning that Tower Hamlets' and Newham's neighbourhoods contribute $0.131 \div 0.772 \times 100$ per cent of the total ID, which is 17 per cent. Measured in relation to the percentage of neighbourhoods that are in Tower Hamlets and Newham gives an impact score of 1918 - 19.18 times greater than expected.

The third calculates the standard ID *only* for Tower Hamlets and Newham; that is, if all but the data for those two places are omitted from the calculation. The resulting ID is 0.365. Within Tower Hamlets and Newham the Bangladeshi and White British populations are more evenly distributed than they are across the whole of England and Wales. However, this finding is based on what remains of the two groups when everyone living outside of Tower Hamlets and Newham is excluded. It remains the cases that larger shares of the White British are found outside of Tower Hamlets and Newham whereas larger shares of the Bangladeshis are found within them.

A final measure is the R-square of 0.381. This is the proportion of the variation in $\epsilon_i = y_i - x_i$, i.e. the base level differences in the shares of the Bangladeshi and White British populations, that can be attributed to those places being in Tower Hamlets or Newham. It is a sizable proportion. It seems that Tower Hamlets and Newham are having a strong impact on the ID.

Extending the analysis, we can examine the effect of Tower Hamlets, Newham and the MSOA E02001113 upon the index,

```
effect(index, places = c("Tower Hamlets", "Newham", "E02001113"))
```

```
## Bangladeshi ~ WhiteBrit
##          ID1   ID2   ID3
## before 0.772 0.772 0.772
## after  0.697 0.139 0.373
##
## R-squared: 0.45    Impact 1996
##
## Note:
## ID1: if the residual effect of Tower Hamlets Newham E02001113 is set to zero
## ID2: if the shares of Bangladeshi & WhiteBrit are equal everywhere except Tower Hamlets Newham
E02001113
## ID3: the index value for Tower Hamlets Newham E02001113 alone
```

E02001113 appears to be a residential area of Oldham in North West England located by Royal Oldham Hospital and containing high numbers of Bangladeshis: view map. Like Tower Hamlets and Newham it too appears to be an 'outlier' with an unusually high share of the Bangladeshi population.

```
aggdata[aggdata$MSOA == "E02001113",]
```

```
##            LSOA Persons WhiteBrit Irish OtherWhite Mixed Indian Pakistani
## 26226 E01005349    2424        54     7          6    23     12        84
## 26248 E01005351    1986       237     3         13    13     20        53
## 26233 E01005352    1449       356     2          9    16     15        54
## 26237 E01005353    1913       385     5         47    37     24        80
## 26242 E01005354    1764       617     5         20    32     50        74
##       Bangladeshi Chinese OtherAsian BlkAfrican BlkCaribbean OtherBlk Arab
## 26226        2190       5         17          8            4        4    0
## 26248        1605       1          2         16            5       17    0
## 26233         953       6          7         15            9        5    0
## 26237        1176       2         32         90           10       11    0
## 26242         904      11         13         20           10        2    4
##       Other      MSOA    LAD        RGN
## 26226    10 E02001113 Oldham North West
## 26248     1 E02001113 Oldham North West
## 26233     2 E02001113 Oldham North West
## 26237    13 E02001113 Oldham North West
## 26242     2 E02001113 Oldham North West
```

Refitting the multilevel index

Having identified the 'outliers', a next step is to refit the multilevel index with Tower Hamlets, Newham and E02001113 omitted.

```
newindex <- id(aggdata, vars = c("Bangladeshi", "WhiteBrit"), levels = c("MSOA","LAD","RGN"),
omit = c("Tower Hamlets", "Newham", "E02001113"))
newindex
```

```
## Bangladeshi ~ WhiteBrit
## ID:  0.733
## E(ID):   (not calculated)
##
##      Pvariance Holdback
## Base     37.36   -10.87
```

```
## MSOA     46.59   -14.11
## LAD      11.62    -4.21
## RGN       4.44   -16.05
```

The ID increases slightly from 0.772 to 0.733 but the more interesting change is in the measure of spatial clustering, Pvariance. This has changed from

```
attr(index, "variance")
```

```
##  Base  MSOA   LAD   RGN
## 25.41 32.42 35.92  6.25
```

to

```
attr(newindex, "variance")
```

```
##  Base  MSOA   LAD   RGN
## 37.36 46.59 11.62  4.44
```

which is an increase/decrease of

```
attr(newindex, "variance") - attr(index, "variance")
```

```
##   Base   MSOA    LAD    RGN
##  11.95  14.17 -24.30  -1.81
```

What it reveals is a ‘step down’ from the LAD to the MSOA and LSOA (Base) scales.

Overall, the following observations may be drawn:

- The residential segregation of the Bangladeshi from the White British is high across England and Wales (although actually it decreased from the 2001 to the 2011 Census)
- The scale of segregation is highest at the local authority (LAD) scale
- That is because of the effects of Tower Hamlets and Newham
- Omitting Tower Hamlets and Newham (and also MSOA E02001113) leaves the dominant scales of segregation as the MSOA and LSOA levels

Closing Comments

Within the segregation literature there has been a movement away from measuring ethnic segregation at a single scale and using traditional indices, to treating segregation as a multiscale phenomenon about which measurement at a range of scales will shed knowledge. That literature has been the inspiration for this work. Amongst the contributions, several authors have promoted multilevel modelling as a way of looking at segregation at multiple scales of a geographic hierarchy simultaneously. The MLID package takes forward the approach by outlining a multilevel index of dissimilarity that combines the advantages of using a widely-understood index with a means to identify scale effects in a way that is computationally fast to estimate and easily fitted in R.

Acknowledgements

My thanks to Dewi Owen for thoughtful observations and comments, and for good company

The package development was funded partly under the ESRC's Urban Big Data Centre, grant ES/L011921/1.

References

Harris R 2017 Measuring the scales of segregation: Looking at the residential separation of White British and other school children in England using a multilevel index of dissimilarity, *Transactions of the Institute of British Geographers* in press

see also:

Jones K Johnston R Manley D Owen D and Charlton C 2015 Ethnic Residential Segregation: A Multilevel Multigroup Multiscale Approach Exemplified by London in 2011 *Demography* 52 1995-2019

Leckie G and Goldstein H 2015 A multilevel modelling approach to measuring changing patterns of ethnic composition and segregation among London secondary schools 2001–2010 *Journal of the Royal Statistical Society Series A* 178 405-424

Leckie G Pillinger R Jones K and Goldstein H 2012 Multilevel modelling of Social Segregation *Journal of Educational and Behavioral Statistics* 37 3-30

Manley D Johnston R Jones K and Owen D 2015 Macro- Meso- and Microscale Segregation: Modeling Changing Ethnic Residential Patterns in Auckland New Zealand 2001-2013 *Annals of the Association of American Geographers* 105 951-967

Owen D 2015 Measuring residential segregation in England and Wales: a model-based approach Unpublished PhD thesis School of Geographical Sciences, University of Bristol

---

1. The 'stray' cell in examples 2-4 is to allow the model to be fitted. With it, the model correctly identifies that some of the variation remains at the base level.↩
2. Specifically, the standard error of the residuals from the regression used to fit the model↩
3. If it isn't supplied, it will be estimated as the sum of the X and Y populations per neighbourhood, and will generate a warning.↩
4. The width of the confidence interval is adjusted for a test of difference between two means (see *Statistical Rules of Thumb* by Gerald van Belle, 2011, eq 2.18). A 95 per cent confidence interval, for example, extends to 1.39 times the standard error around the mean and not 1.96.↩
5. The caterpillar plots employ what might be considered to be intelligent plotting in that only a maximum of 50 residuals are shown on each plot. These are the 10 highest and lowest ranked residuals and then a sample of 30 from the remaining residuals, chosen as the ones with values that differ most from the residuals that precede them by ranking. In this way, the plots aim to preserve the tails of the ranked distribution as well as the most important break points in-between.↩