Using hierarchical clustering to explore patterns of deprivation among English local authorities

USING HIERARCHICAL CLUSTERING TO EXPLORE PATTERNS OF DEPRIVATION AMONG ENGLISH LOCAL AUTHORITIES

Steven L. Senior, Specialty Registrar in Public Health, Division of Population Health, Health Services Research & Primary Care, University of Manchester, UK.

Address: Division of Population Health, Health Services Research & Primary Care
Stopford Building
Oxford Road
Manchester
M13 9PG

E-mail: steven.senior@manchester.ac.uk

Phone: 0161 725 1653

Mobile: 07732 208 559

Key words: public health; indicators; social determinants

Word count:

Abstract: 200

Main text: 2,970

ABSTRACT

Background

The English Indices of Multiple Deprivation (IMD) is widely used as a measure of deprivation. However, similarly ranked areas can differ substantially in the underlying domains of deprivation. These domains contain a richer set of data that might be useful for classifying local authorities. Clustering methods offer a set of techniques to identify groups of areas with similar patterns of deprivation.

Methods

Hierarchical agglomerative (i.e. bottom-up) clustering methods were applied to domain scores for 152 upper-tier local authorities. Advances in statistical testing allow clusters to be identified that are unlikely to have arisen from random partitioning of a homogeneous group. The resulting clusters are described in terms of their subdomain scores and basic geographic and demographic characteristics.

Results

Five statistically significant clusters of local authorities were identified. These clusters only partially reflect different levels of overall deprivation. In particular two clusters share similar overall IMD scores, but have contrasting patterns of deprivation.

Conclusion

Hierarchical clustering methods identify five distinct clusters that do not correspond closely to quintiles of deprivation. This approach may help to distinguish between places that face similar underlying challenges, and places that appear similar in terms of overall deprivation scores, but that face different challenges.

INTRODUCTION

The English Indices of Multiple Deprivation (IMD) is a commonly used measure of relative deprivation for geographic areas in England. It is often used for analysing inequalities in health between more and less deprived areas.

The IMD provides an overall score of relative deprivation that is used to rank Lower-Super Output Areas (LSOAs, small geographic areas in England, each with a population of approximately 1,500). The overall score is calculated as a weighted sum of seven domains of deprivation. These domains measure deprivation in: income; employment; health and disability; education, training and skills; crime; access to housing and services; and living environment [1].

In analyses at upper-tier local authority level (the upper-most layer of local government in England), each local authority is typically ranked according to the average score or average rank of its constituent LSOAs, and local authorities are then grouped into quintiles or deciles. This forms the basis for estimating inequalities in health outcomes. For example, the Public Health Outcomes Framework, a widely used tool for understanding variations in health between English local authorities, has an option to view health indicators by decile of deprivation [2].

This method of grouping areas together in quintiles or deciles may group together areas with quite different deprivation profiles. For example, when upper-tier local authorities are ranked by their average 2015 IMD scores, North East Lincolnshire and South Tyneside are ranked as the 25th and 26th most deprived local authorities respectively (out of 152 upper-tier local authorities), both in the second most deprived decile. However, their scores on the domains are different: in terms of average health and disability deprivation South Tyneside is ranked 12th most deprived (in the most deprived decile), while North East Lincolnshire is ranked 66th most deprived (in the fifth most deprived decile). So the use of a single overarching ranking system

may obscure differences in the types of challenges faced by local authorities, and does not make full use of the information contained in the full set of domain scores.

Statistical clustering methods offer a way to find groups of geographic areas that face similar challenges. Bellis et al [3] used k-means clustering on health outcomes data and found that the cluster with the worst health outcomes was concentrated in the ex-industrial areas of the North of England. An alternative to the k-means approach is hierarchical clustering, which produces a set of nested clusters, allowing the researcher to describe the clustering structure in the data. Recent advances in statistical methods make it possible to assess at each 'branch' in the tree, whether the resulting clusters are likely to have arisen by chance or not, allowing the numbers of clusters to be determined by the degree to which the clusters account for the variation in the data [4].

This complements the use of 'nearest neighbours' models, such as that produced by the Chartered Institute of Public Finance and Accounting [5]. This model calculates the Euclidean distance between a given local authority and all other local authorities based on a selection of indicators. However, this approach is designed to find similar local authorities given an initial local authority, rather than grouping all local authorities based on patterns of deprivation.

The aim of this paper is to apply clustering techniques to the domains of deprivation to explore patterns of deprivation among upper-tier local authorities in England. It aims to test the usefulness of unsupervised statistical learning methods in understanding the challenges faced by local authorities and to find out if any resulting clusters are consistent with a single continuum of multiple deprivation, or whether there are clusters that differ more in the pattern of deprivation than in their overall IMD scores.

METHODS

Data sources and processing

IMD data were downloaded from the website of the Ministry of Housing, Communities and Local Government [6]. Geographic data were downloaded from the Office for National Statistics Open Geography Portal (geoportal.statistics.gov.uk/datasets/). Urban rural classification data were downloaded from the Office for National Statistics. Demographic data was downloaded using the FingertipsR package [7].

The IMD summary statistics for upper-tier local authorities include a range of scores that summarise the IMD scores of each local authority's component LSOAs. These include the average LSOA scores in each domain, and the proportion of LSOAs in a given local authority in the lowest 10% nationally. This study uses only the average scores for each local authority for the seven domains of deprivation. Average scores for each local authority were centered and scaled to mean 0 and standard deviation of 1 to prevent the distribution of any variable affecting its weight in the clustering analysis.

Statistical analysis

This study uses a hierarchical agglomerative (i.e. bottom-up) clustering algorithm. The algorithm computes the Euclidean distance between each local authority based on average scores on the seven domains of deprivation. The algorithm identifies the two local authorities that have the lowest distance score and links them in a cluster. The algorithm then repeats this process until one cluster remains containing all the local authorities. The 'complete' linkage method was used, where the distance between two clusters is the distance between their two farthest points.

Statistical analysis was done in R version 3.5.0 [8]. Hierarchical clustering was done using the sigclust2 package [4]. This package enables testing for statistical significance of clustering. At

each node, the algorithm computes the two-mean cluster index, a measure of the extent to which observations within the two clusters vary relative to the overall variation in the data set. The value of the cluster index at each node is then compared to the distribution expected under the null hypothesis that both clusters are drawn from a single gaussian distribution. This produces a p-value for each node in the clustering structure indicating the likelihood that a cluster index that large or larger could have arisen by chance. To be conservative, nodes with p-values less than 0.001 were treated as statistically significant. To correct for multiple testing, the algorithm applies a family-wise error rate correction, which shrinks the critical value at which the clustering at a given node is judged to be statistically significant. If a node has a p-value greater than 0.001, the algorithm does not test nodes below it in the clustering hierarchy.

As the domain scores were used to create the clustering, testing for statistically significant differences between cluster subdomain scores is neither necessary nor appropriate. Several local authority domains and other variables were skewed. For consistency medians and interquartile ranges are reported, and non-parametric statistical tests are used. Cluster geographic and demographic characteristics were compared using Kruskal-Wallis tests. Where the Kruskal-Wallis test for a given cluster characteristic was significant at $p < 0.05$ post-hoc Dunn's tests using Bonferroni correction for multiple comparisons were used to identify statistically significant differences between pairs of clusters.

All R code is available online at https://github.com/stevenlsenior/IMD_clustering.

RESULTS

Hierarchical clustering

Figure 1a shows the clustering structure as a dendrogram. Statistically significant clustering is identified by red branches in the dendrogram, and associated p-values are displayed where the

clustering at that node was statistically significant. Five statistically significant clusters were identified. Figure 1b shows the geographic distribution of the five significant clusters.

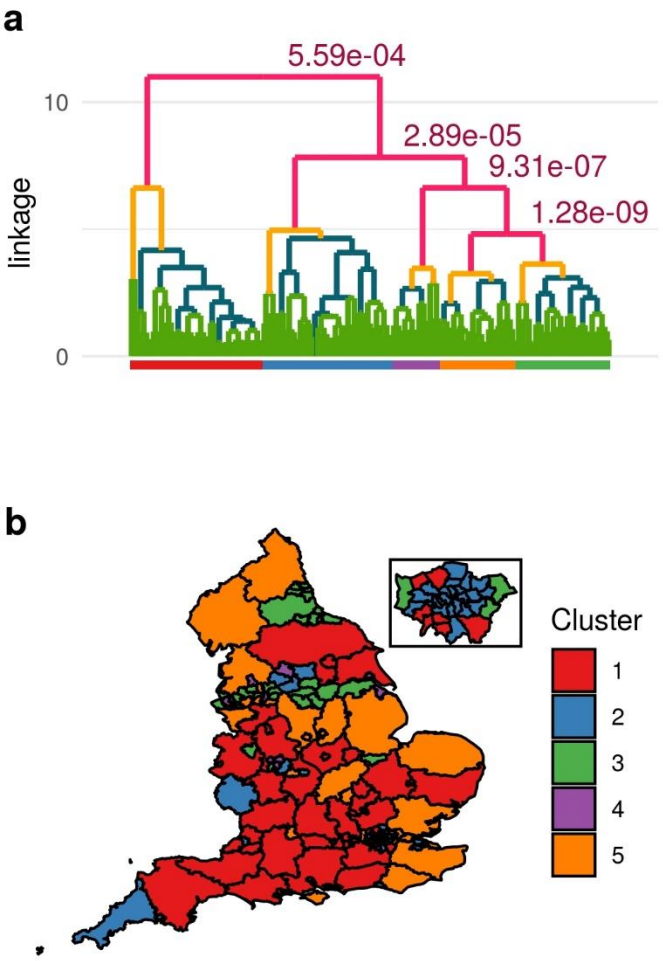

Fig. 1 (a) Clustering dendrogram showing the clustering structure. Nodes at which clustering was statistically significant are indicated by red branches. P-values for significant nodes are presented. Five statistically significant clusters are identified, indicated by the coloured bars at the bottom. FWER - family-wise error rate. (b) Map showing the geographic distribution of the significant clusters identified.

Figure 2 shows the median subdomain z-scores scores for each of the clusters.

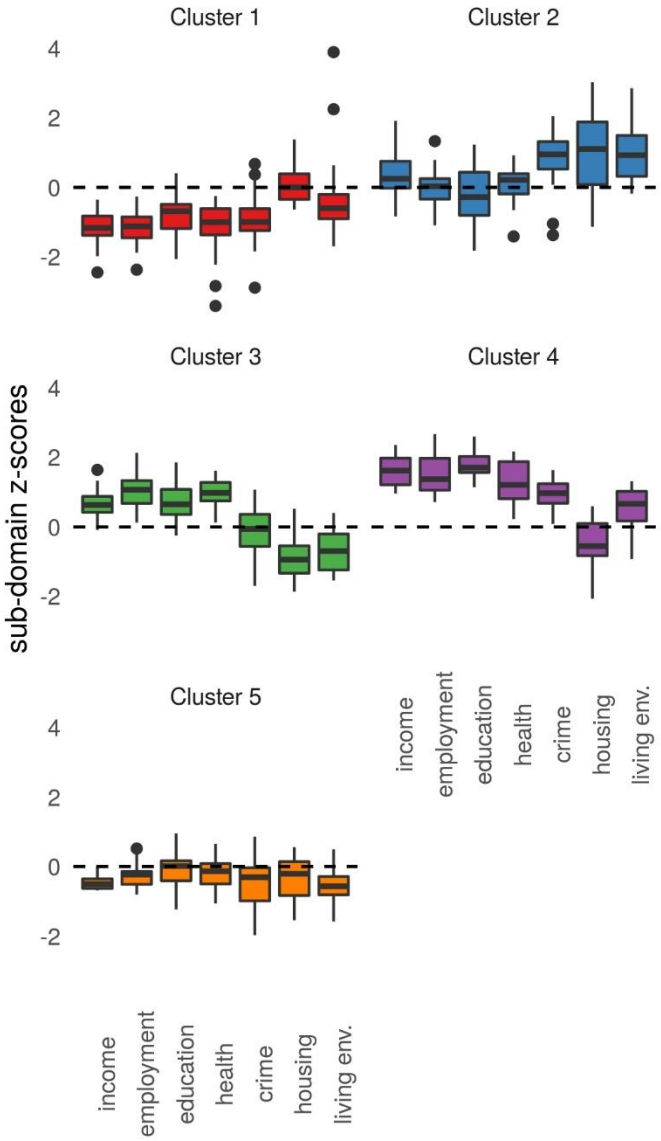

Fig. 2: Boxplots of domain z-scores for the five clusters.

The clusters form a nested structure described by the dendrogram shown in figure 1 panel A. Clusters 2, 3, 4, and 5 are nested within a larger cluster that is distinct from cluster 1. This suggests that there is more difference in terms of patterns of deprivation between cluster 1 (the least deprived cluster) and the remaining clusters. Further down the dendrogram, cluster 2 is separated from clusters 3, 4, and 5. The profiles of IMD subdomain scores in figure 2 suggest

that cluster two has a different pattern of deprivation to clusters 3, 4, and 5. Cluster 4 is then separated from clusters 3 and 5.

Cluster 1, the least deprived overall, contains 42 local authorities. Average scores for local authorities in this cluster are low on all domains of deprivation apart from barriers to housing, which are average. These areas are mainly large rural counties, with some more affluent boroughs of London.

Cluster 2, comprising 41 local authorities, has average deprivation scores only slightly above average for income, health, and employment, and slightly below average deprivation in education, but high levels of deprivation in crime, housing, and living environment. These areas include a mix of London boroughs as well as some more rural areas, such as Kirklees and Cornwall.

Cluster 3 (30 local authorities) has a similar average IMD score to cluster 2 but experiences low levels of housing and environmental deprivation and average crime levels, but high deprivation in income, employment, health, and education and skills. These areas are mainly post-industrial towns concentrated in the North of England.

Cluster 4 is the most deprived overall. Average deprivation scores are high across all domains except housing. This cluster is small, containing only 15 local authorities, mostly deprived towns and cities in the North and Midlands, such as Blackpool, Liverpool, Manchester, and Wolverhampton.

Cluster 5 (24 local authorities), the second least deprived, scores slightly below average across all domains of deprivation, with lower scores in income and environmental deprivation. This

cluster includes a mix of rural and urban local authorities, many county councils, such as Northamptonshire and Lincolnshire, which contain both relatively deprived and relatively affluent areas.

Table 1 shows the distribution of the clusters across quintiles of deprivation. There are extensive overlaps between the clusters' distributions across the quintiles of deprivation. This suggests that the clusters do not correspond to a simple continuum of deprivation. Only cluster 4 is entirely contained within a single quintile - in this case the most deprived quintile. However, cluster 4 only comprises half of the most deprived local authorities. The remainder fall in clusters 2 and 3. Local authorities in cluster 1 fall entirely in the least two deprived quintiles. There is substantial overlap between clusters 2 and 3, with most local authorities in each cluster falling in the three most deprived quintiles.

Table 1: distribution of clusters across quintiles of IMD		**Least Deprived**		**Quintiles of IMD**		**Most Deprived**	
		1	**2**	**3**	**4**	**5**	**total**
Cluster	**1**	31	11	0	0	0	**42**
	2	0	3	16	14	8	**41**
	3	0	0	7	16	7	**30**
	4	0	0	0	0	15	**15**
	5	0	16	8	0	0	**24**
	total	**31**	**30**	**31**	**30**	**30**	**152**

Figure 3 shows how the clusters of local authorities differ in geographic size, age structure, and the proportion of the population that lives in areas classified as rural by the ONS.

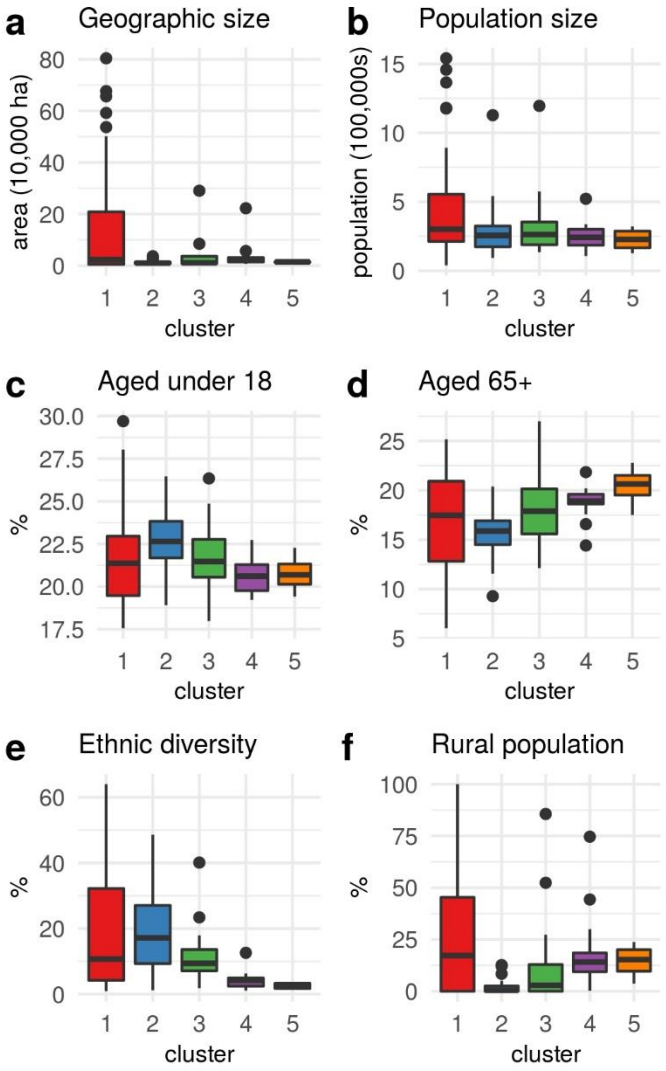

Fig. 3: Characterisation of clusters according to: (a) Geographic size in hectares (source: ONS), (b) Population size (100,000s) (source: PHE; 2016 data), (c) Percentage of population aged under 18 years (source: PHE; 2016 data). (d) Percentage of population aged 65 years and older (source: PHE; 2016 data), (e) Percentage of population from black and minority ethnic backgrounds (source: PHE; 2016 data), and (f) Percentage of population living in rural areas (including large market towns; source: ONS; 2011 data)

Clusters 1 and 5 contain most of the English counties, and as such tend to be significantly larger than clusters 2 and 4 which contain mostly unitary authorities (Kruskall-Wallis statistic 55.59 on

4 degrees of freedom, $p < 0.001$; post-hoc Dunn's tests for pairwise comparisons between cluster 1 and clusters 2 and 4, and between cluster 5 and clusters 2 and 4 all significant at $p < 0.05$).

The same pattern is seen in the proportion of the population who live in rural areas, with clusters 1 and 5 having significantly more of their population in areas classified as rural than clusters 2 and 4 (Kruskall-Wallis statistic 68.2 on 4 degrees of freedom, $p < 0.001$, Dunn's tests for pairwise comparisons between clusters 1 and clusters 2 and 4 and cluster 5 and clusters 2 and 4 significant at $p < 0.001$).

There is some evidence that clusters vary in size of population. (Kruskall-Wallis statistic 11.37 on 4 degrees of freedom, $p = 0.023$). However, in pairwise comparisons, only clusters 3 and 5 differed significantly (Dunn's tests for pairwise comparisons significant at $p = 0.025$).

There was limited evidence of differences in the proportion of the population aged under 18 (Kruskall-Wallis test statistic 9.18 on 4 degrees of freedom, $p = 0.057$), but clusters did appear to differ in the proportion of the population aged 65 and over (Kruskall-Wallis test statistic 61 on 4 degrees of freedom, $p < 0.001$). Cluster 2 had the lowest proportion aged 65 plus (Dunn's test comparisons between cluster 2 and clusters 1, 3, and 5 all significant at $p < 0.05$), followed by cluster 4, which differed significantly from clusters 1 and 5 (post-hoc Dunn's tests significant at $p < 0.001$).

Clusters also varied in ethnic diversity (Kruskall-Wallis test statistic 46.2 on 4 degrees of freedom, $p < 0.001$). Cluster 2 was the most ethnically diverse, with clusters 1, 3, and 5 less ethnically diverse (Dunn's tests for pairwise comparisons between clusters 2 and 1, 3, and 5

significant at $p < 0.001$). Cluster 4 had a higher median ethnic diversity, but pairwise comparisons with other clusters were not statistically significant at the $p < 0.05$ level.

DISCUSSION

Main findings of this study

Hierarchical clustering methods identify five statistically significant clusters of local authorities. These clusters are different groupings from the quintiles of deprivation commonly used. This suggests that hierarchical clustering methods offer an alternative way to explore patterns of deprivation at local authority level. The identification of clusters of local authorities that share patterns of deprivation may help to guide public health and other policy interventions. For example, successful interventions in one local authority might transfer more easily to other areas within the same cluster, as these areas may share similar challenges and contextual features.

Clusters 2 and 3 have similar overall deprivation scores, but have contrasting patterns of deprivation. This suggests that the use of the overall IMD score to group local authority areas obscures some of the information contained in the IMD subdomain scores, and groups together areas that face different challenges. Local authorities in cluster 3 appear to have worse health deprivation than their overall deprivation scores might suggest. These differences may be explained by some of the differences in the characteristics of these clusters of local authorities shown in figure 3. Cluster 2 appears to contain local authorities whose populations are on average more educated, less likely to be retired, and more ethnically diverse, and more urban than those in cluster 3. This suggests that it is important to look at the pattern of deprivation and the social and demographic factors underlying it, not only the overall IMD score.

The five statistically significant clusters identified here display a degree of face-validity (reflecting areas with similar economic histories), while adding value by identifying similarities and groupings that are not captured when local authorities are grouped based on their overall IMD score. This may offer a more nuanced picture of patterns of deprivation among English local authorities.

What is already known on this topic

Bellis et al used k-means clustering to identify five clusters of local authorities based on health outcomes. That study focused on health outcomes or proximal risk factors (such as smoking and unhealthy eating). It found that the cluster with the worst health was concentrated in the North of England. [3] This is consistent with other epidemiological data showing that health tends to be worse in the North even after adjusting for deprivation as measured by the IMD [9,10].

What this study adds

Whereas previous studies have considered patterns of health outcomes or proximal risk factors, this study focuses on important upstream determinants of these outcomes and risk factors. The results presented here may help to explain some of the patterns seen in previous research. For example, the fact that local authorities in cluster 3 are concentrated in Northern England may explain why the least healthy cluster in Bellis et al's study is concentrated in the same areas, and my help to explain the observation that areas in the North of England appear to suffer worse health even after correcting for IMD score [9,10].

The results presented here also suggest that the use of quintiles or deciles of deprivation based on the overall IMD score for benchmarking local authority performance against may be

misleading. The statistical neighbours approach (for example that used by CIPFA) may be better able to identify comparable local authority areas [5].

Limitations of this study

The use of upper-tier local authorities in this analysis is intended to make the results useful for local authority and national policymakers. However, it means that the resulting clustering may be affected by variation in types of local authority. The use of English upper-tier local authorities also limits the sample size to 152. Further analysis using smaller areas such as LSOAs would partly remove variation in population size, and would substantially increase the statistical power of the analysis, potentially allowing for finer distinctions between patterns of deprivation. Such an analysis might also help to reveal similarities between small areas in different local authorities. This could help to spread successful local interventions by finding other areas that share similar challenges and contexts.

Clustering techniques, including the hierarchical clustering methods used here, show substantial sensitivity to small changes in the data and to the choice of linkage and distance measures [11]. As such the results should be treated with caution. The use of a stringent alpha value for identifying statistically significant clusters mitigates the risk of instability affecting the lower branches of the dendrogram, it does not remove the risk that the overall clustering structure is sensitive to small perturbations of the data.

The results here may also be sensitive to the choice of measure for summarising the domain scores at upper-tier local authority level. The average score measure provided by the Office for National Statistics is used here for ease of interpretation. While the summary domain scores at local authority-level are largely normally distributed, at the underlying LSOA-level, some of the

domain scores are skewed, notably the income and employment domains. This may threaten the validity of the average score measure as a summary measure.

An extension to this work would be to test the stability of the clustering structure over time. The recent publication of the 2019 Indices of Deprivation [12] along with the earlier 2010 Indices of Deprivation [13] provide a longitudinal data set with a largely unchanged underlying set of indicators. Analysing the clustering structure over time would test whether the patterns of deprivation described here represent enduring features of English social geography.

REFERENCES

1. Smith T, Noble M, Noble S *et al. The English Indices of Deprivation 2015*. Department for Communities and Local Government, 2015.

2. Public Health England. Public Health Outcomes Framework. 2018.

3. Bellis MA, Jarman I, Downing J *et al.* Using clustering techniques to identify localities with multiple health and social needs. *Health Place* 2012;**18**:138–43.

4. Kimes PK, Liu Y, Neil Hayes D *et al.* Statistical significance for hierarchical clustering. *Biometrics* 2017;**73**:811–21.

5. CIPFA. Nearest Neighbours Model. 2017.

6. Ministry of Housing, Communities, Local Government. English indices of deprivation 2015. *GOVUK* 2015.

7. Fox S, Flowers J, Thelwall S *et al.* fingertipsR: an R package for accessing population health information in England. *Epidemiology* 2017:5.

8. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2008.

9. Kontopantelis E, Buchan I, Webb RT *et al.* Disparities in mortality among 25-44-year-olds in England: a longitudinal, population-based study. *Lancet Public Health* 2018;**3**:e567–75.

10. Whitehead M, McInroy N, Bambra C *et al.* Due North: report of the inquiry on health equity for the North. *Liverpool: University of Liverpool and the Centre for Economic Strategies* 2014.

11. Data clustering: 50 years beyond K-means. *Pattern Recognit Lett* 2010;**31**:651–66.

12. Noble S, McLennan D, Noble M *et al. The English Indices of Deprivation 2019: Research Report*. MHCLG, 2019.

13. Department of Communities and Local Government. *The English Indices of Deprivation 2010*., 2011.