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: deprivation; unsupervised learning; public health intelligence

Word count:

Abstract: 258

Main text: 3,087

ABSTRACT

Background

The English Indices of Multiple Deprivation (IMD) is widely used as a measure of deprivation of geographic areas in analyses of health inequalities between places. However, similarly ranked areas can differ substantially in the underlying subdomains and indicators that are used to calculate the IMD score. These subdomains and indicators 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. This could offer insights into areas that face similar challenges.

Methods

Hierarchical agglomerative (i.e. bottom-up) clustering methods were applied to sub-domain scores for 152 upper-tier local authorities. Recent 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 represented local authorities that were:

(i) Most deprived, predominantly urban;

(ii) Least deprived, predominantly rural;

(iii) Less deprived, rural;

(iv) Deprived, high crime, high barriers to housing; and

(v) Deprived, low education, poor employment, poor health.

Conclusion

Hierarchical clustering methods identify five distinct clusters that do not correspond closely to quintiles of deprivation. These methods can be used to draw on the richer set of information contained in the IMD subdomains and may help to identify places that face similar challenges, and places that appear similar in terms of IMD 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 subdomains of deprivation. These subdomains measure deprivation in: income; employment; health and disability; education, training and skills; crime; access to housing and services; and living environment. Each of the subdomains is made from one or more indicators. Where more than one indicator is used, indicators are combined using weights determined by factor analysis to create the subdomain score. The weights used to combine the subdomains into the overall score were determined using a range of methods including previous research that the income and employment domains should be weighted more heavily, as well as empirical research on public preferences and preferences revealed by relative levels of government spending on each domain (Smith et al. 2015).

Summary IMD scores are available for larger geographic areas, such as upper-tier local authorities (the upper-most layer of local government in England). For each upper-tier local authority, a number of different ways of combining the deprivation of its LSOAs are used, including average IMD score, average rank, and the proportion of population living in the 10% most deprived LSOAs (Smith et al. 2015).

In analyses at local authority level, 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 then 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 (Public Health England 2018).

While not wholly arbitrary, the 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 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 subdomains 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). As a result, 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 rich information contained in the full set of sub-domain scores.

Statistical clustering methods offer a way to find groups of geographic areas that face similar challenges. For example, Bellis et al (2012) 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 without arbitrarily specifying a number of clusters to be found. 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 a clustering structure to be determined without the researcher pre-determining how many clusters will be found (Kimes et al. 2017).

This complements the use of 'nearest neighbours' models, such as that produced by the Chartered Institute of Public Finance and Accounting (CIPFA 2017). This model calculates the Euclidean distance between a given local authority and all other local authorities based on a selection of indicators. 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 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 (Ministry of Housing, Communities and Local Government 2015). 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 (Fox et al. 2017).

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 of LSOA scores in each sub-domain, and the proportion of LSOAs in a given local authority in the lowest 10% nationally. For simplicity, this study uses only the average scores for each local authority for the seven subdomains of deprivation. Average scores for each local authority were centered and scaled to mean 0 and standard deviation of 1. This prevents 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 subdomains of deprivation. The algorithm then identifies the two local authorities that have the lowest distance score and links them in a cluster. The algorithm then repeats this process until only 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 (R Development Core Team 2008). Hierarchical clustering was done using the sigclust2 package (Kimes et al. 2017). 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 (small values indicating tighter clustering). 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. Using this approach, the number of clusters identified is determined by the degree of clustering, rather than being decided in advance by the researcher, or being based on an arbitrary cut-off point on the clustering tree.

As the sub-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 subdomains 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.

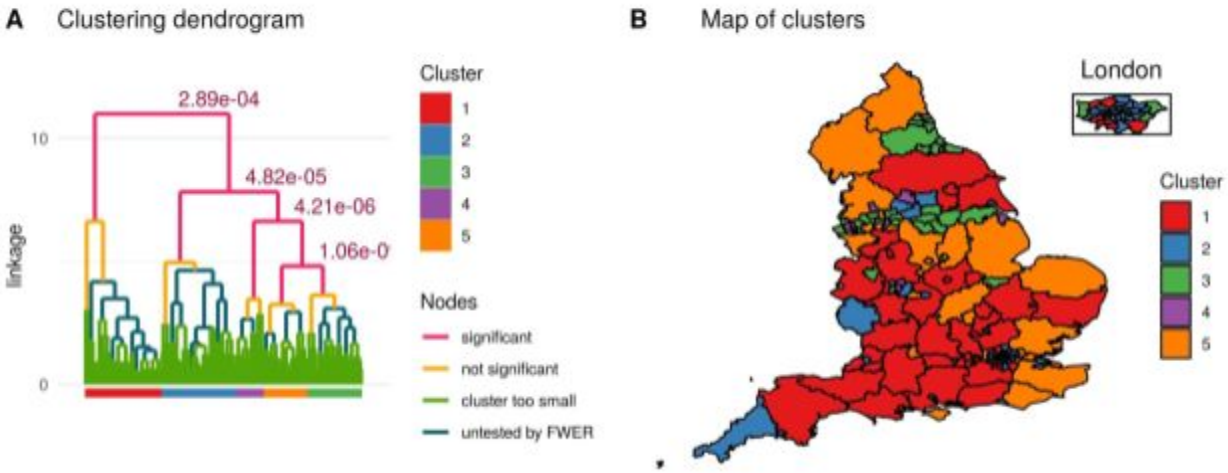
All R code is available online (see supplementary methods).

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.

Figure 1: Heirarchical clustering of local authorities by subdomains of deprivation



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. Cluster colours correspond to those in panel A.

Figure 2 and table 1 show the median subdomain z-scores scores for each of the clusters. Table 2 shows how the local authorities are distributed across the clusters and quintiles of IMD scores.

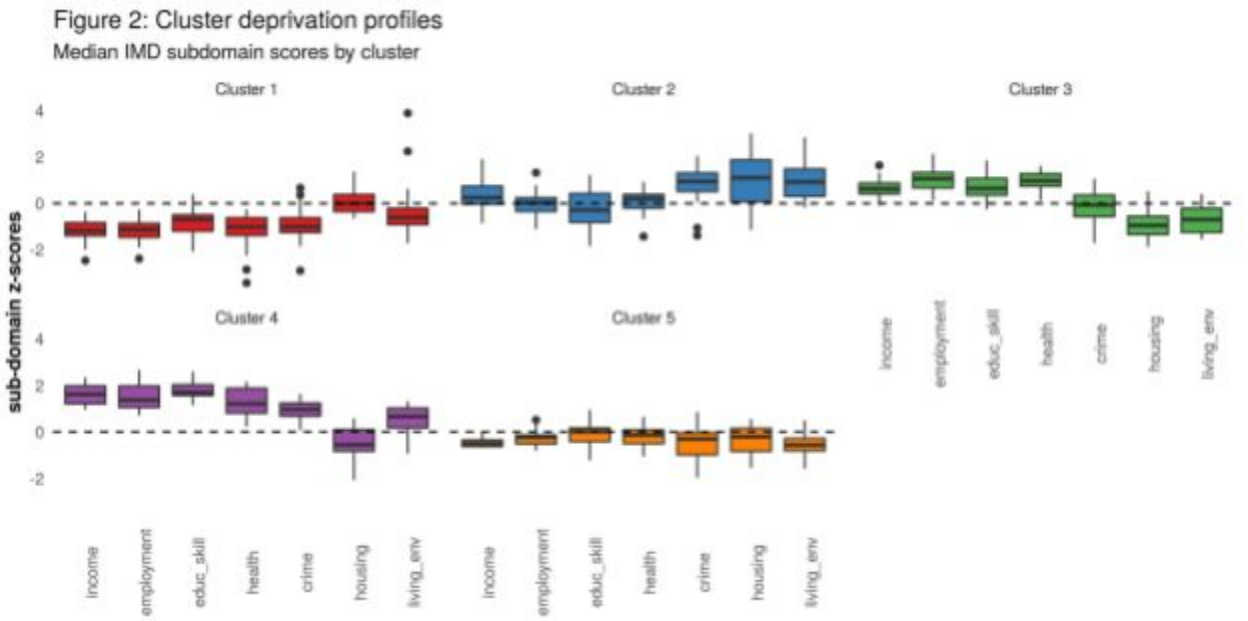


Table 1: cluster subdomain scores

Cluster	n	Median IMD score	Subdomain z-scores: median (interquartile range)						
			Income	Health	Crime	Employment	Education and skills	Housing	Living environment
1	42	13.81 (3.59)	-1.17 (0.56)	-1.12 (0.6)	-0.69 (0.7)	-1 (0.76)	-1 (0.63)	0 (0.73)	-0.6 (0.7)
2	41	26.64 (5.01)	0.24 (0.78)	0.03 (0.59)	-0.28 (1.24)	0.21 (0.58)	0.95 (0.79)	1.1 (1.8)	0.92 (1.18)
3	30	28.27 (3.95)	0.63 (0.46)	1.07 (0.66)	0.65 (0.72)	0.98 (0.53)	-0.06 (0.93)	-0.94 (0.79)	-0.69 (1.04)
4	15	34.61 (7.64)	1.62 (0.77)	1.38 (0.92)	1.71 (0.46)	1.21 (1.07)	0.97 (0.56)	-0.55 (0.93)	0.66 (0.85)
5	24	18.84 (2.76)	-0.5 (0.28)	-0.23 (0.37)	0.02 (0.57)	-0.14 (0.58)	-0.31 (0.95)	-0.21 (0.98)	-0.57 (0.53)

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. Looking 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, which are the last to be separated.

Cluster 1, the least deprived overall, contains 42 local authorities. Average scores for local authorities in this cluster are low on all subdomains of deprivation apart from barriers to housing,

which are around 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 subdomains except housing. This cluster is relatively small, containing only 15 local authorities, mostly deprived towns and cities in the North and Midlands, such as Blackpool, Liverpool, Manchester, and Wolverhampton.

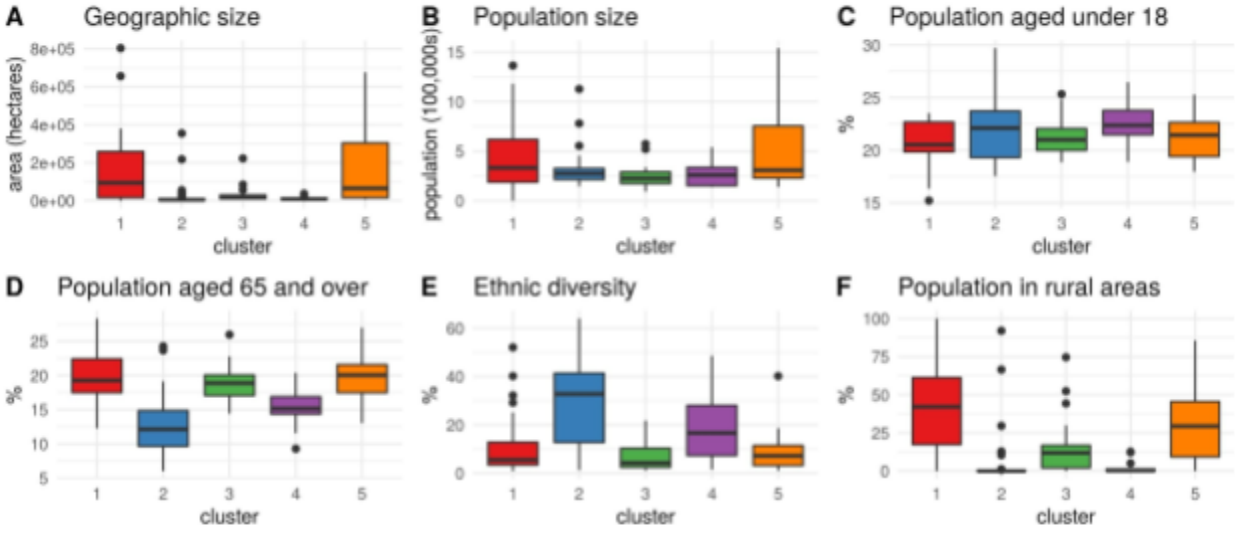
Cluster 5 (24 local authorities), the second least deprived, scores slightly below average across all subdomains of deprivation, with lower scores in income and environmental deprivation. This cluster includes a mix of rural and urban local authorities, many county councils, such as Northamptonshire and Lincolnshire, which contain both relatively deprived and relatively affluent areas.

Table 2 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 multiple 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 extensive overlap between clusters 2 and 3, with most local authorities in each cluster falling in IMD quintiles 3, 4, and 5 (i.e. the three most deprived quintiles).

Table 2: distribution of clusters across quintiles of IMD		**Least**		**Quintiles of IMD**		**Most**	
		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. Table 3 reports the results of Kruskall-Wallis tests for statistically significant differences between clusters.

Figure 3: Cluster characteristics



Characterisation of clusters according to geographic size and demographic characteristics.
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).
F: Percentage of population living in rural areas (including large market towns; source: ONS; 2011 data)

Table 3: Cluster characteristics

Variable	Cluster median (interquartile range)					H	df	p-value
	1	2	3	4	5			
Area (10,000 HA)	9.41 (24.32)	0.43 (0.5)	1.55 (1.84)	0.86 (0.41)	6.48 (28.89)	55.59	4	< 0.001
Total population (100,000s)	3.33 (4.31)	2.74 (1.11)	2.28 (1.16)	2.6 (1.83)	3.09 (5.25)	11.37	4	0.023
% aged under 18	20.54 (2.82)	22.12 (4.42)	21.01 (2.07)	22.37 (2.31)	21.44 (3.24)	9.18	4	0.057
% aged 65 and over	19.27 (4.93)	12.12 (5.21)	18.87 (2.99)	15.15 (2.53)	20.03 (4.1)	61	4	< 0.001
% ethnic minority	5.5 (9.4)	32.8 (28.7)	4.25 (7.85)	16.6 (20.75)	7.15 (8.25)	46.2	4	< 0.001
% living in rural areas	42.13 (43.92)	0 (0.17)	11.87 (14.68)	0.21 (1.23)	29.41 (36.25)	68.2	4	< 0.001

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

Key findings

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 other contextual features.

While three of the clusters reported here appear to reflect groups of local authorities with low, medium, and high levels of deprivation across all IMD subdomains, 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 2 tend to have higher than average deprivation scores for crime, housing, and living environments, but average deprivation scores for the other IMD subdomains. Cluster 3 has the opposite pattern: higher than average scores on income, employment, education and health, but average scores on crime, housing, and living environment deprivation. Importantly for public health policy and practice, local authorities in cluster 3 appear to have worse health deprivation than their overall deprivation scores might predict. These differences may be explained by some of the differences in the characteristics of these clusters of local authorities shown in figure 3. Cluster two 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 fact that local authorities in cluster 3 are concentrated in Northern England may help to explain the observation that areas in the North of England appear to suffer worse health even after correcting for IMD score (Whitehead et al. 2014; Kontopantelis et al. 2018). This suggests that the construction of the IMD may not be accurately capturing deprivation that affects health, and may under-weight deprivation in Northern areas of England.

The results presented here suggest that the use of quintiles or deciles of deprivation 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 (CIPFA 2017).

Strengths and limitations

Strengths

The combination of hierarchical clustering methods and an approach to identifying statistically significant clustering provides a less arbitrary approach to grouping local authorities. 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 offers a more nuanced picture of patterns of deprivation among local authorities in England.

Limitations

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 the large 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.

It should also be noted that the results may be sensitive to the choice of parameters in the linkage algorithm and decisions about which variables to use. The decision to use only average scores for the seven subdomains is also likely to have influenced the resulting clustering structure. The results are also sensitive to the choice of distance and linkage methods. Further work is needed to provide a sound basis for these analytical choices.

REFERENCES

Bellis MA, Jarman I, Downing J, et al (2012) Using clustering techniques to identify localities with multiple health and social needs. Health Place 18:138–143

CIPFA (2017) Nearest Neighbours Model. https://www.cipfastats.net/resources/nearestneighbours/. Accessed 20 Jul 2019

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

Kimes PK, Liu Y, Neil Hayes D, Marron JS (2017) Statistical significance for hierarchical clustering. Biometrics 73:811–821

Kontopantelis E, Buchan I, Webb RT, et al (2018) Disparities in mortality among 25-44-year-olds in England: a longitudinal, population-based study. Lancet Public Health 3:e567–e575

Ministry of Housing, Communities, Local Government (2015) English indices of deprivation 2015. In: GOV.UK. https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015. Accessed 9 Jan 2019

Public Health England (2018) Public Health Outcomes Framework. https://fingertips.phe.org.uk/profile/public-health-outcomes-framework/. Accessed 9 Jan 2019

R Development Core Team (2008) R: A Language and Environment for Statistical Computing. {R Foundation for Statistical Computing}

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

Whitehead M, McInroy N, Bambra C, Others (2014) Due North: report of the inquiry on health equity for the North. Liverpool: University of Liverpool and the Centre for Economic Strategies