

Beyond the Index: Identifying Distinct Patterns of Deprivation Across England

Table of Contents

1. INTRODUCTION AND AIMS.....	2
2. METHODOLOGY.....	3
2.1. DATA SOURCES.....	3
2.2. TOOLS.....	4
2.3. PREPARATION AND CHECKS.....	5
2.4. ANALYSIS.....	5
3. RESULTS AND DISCUSSION.....	6
3.1. SHAPE OF DOMAIN SCORES.....	6
3.2. RQ1: HOW ARE THE SEVEN DOMAIN SCORES RELATED TO EACH OTHER ACROSS LSOAs?.....	7
3.3. RQ2: CAN LSOAs BE GROUPED INTO A SMALL NUMBER OF DISTINCT DEPRIVATION PROFILES?..	8
3.4. RQ3: HOW ARE THOSE DEPRIVATION PROFILES SPREAD ACROSS THE REGIONS OF ENGLAND?..	10
4. CONCLUSION.....	13
5. REFERENCES.....	15
6. APPENDIX.....	16

1. Introduction and Aims

The English Indices of Deprivation (IMD) 2019 are the official measure of relative deprivation in England. Produced by the Ministry of Housing, Communities and Local Government (MHCLG, 2019), they rank every Lower-layer Super Output Area (LSOA) using information from seven domains, namely income, employment, education, health, crime, barriers to housing and services, and living environment. These domains combine different aspects of deprivation into a single overall score, making the IMD one of the most widely used datasets in research and public policy (Noble et al., 2006). The IMD 2019 suits this question well because it publishes each of the seven domain scores separately alongside the combined index, which makes it possible to test whether the parts agree with the whole.

The idea behind the index is not as simple as putting all types of deprivation into one score. Townsend (1987) argued that deprivation can take different forms, and these should be looked at separately. Noble et al. (2006), whose model is used for the English indices, also follows a similar idea. But this does not mean that all the domains are closely related. An area could have a serious problem in one domain but perform quite well in the others. In that case, the overall score could make the problem look smaller than it actually is.

Other studies have also found that looking at deprivation in different groups or profiles can show things that a single ranking might miss. Salvatore and Grundy (2021) found a link between area deprivation and mental health among older people in the UK. Interestingly, this was not only about money or material conditions. How connected people felt to their neighbourhood also mattered. At the same time, research in this area does not always give the same results. The study itself points out that earlier research has found mixed results. So, any relationship found between an area and deprivation needs to be looked at carefully. Aungkulanon et al. (2017) provide a methodological precedent, using cluster analysis to identify area-level deprivation profiles that a single ranking did not reveal.

This report looks at deprivation in England using the IMD 2019 domain scores. It focuses on three research questions.

RQ1: How are the seven domain scores related to each other across English LSOAs?

RQ2: Can LSOAs be grouped into a small number of distinct deprivation profiles?

RQ3: How are these deprivation profiles spread across the regions of England?

These three questions are connected. First, correlation is used to see which domains are related. Then, clustering is used to group areas with similar patterns. Finally, the groups are compared across different regions. The results show that one domain is very different from the others. This creates a large group of areas where the overall index does not fully show their level of deprivation.

2. Methodology

2.1. Data Sources

The main dataset is the English IMD 2019 from the Consumer Data Research Centre (CDRC, 2019), based on the official English Indices of Deprivation 2019 (MHCLG, 2019). The data comes as a shapefile containing both the deprivation data and the LSOA map boundaries, so the same file supports the analysis and the map. Checked with `names()` and `n_distinct()`, it holds 32,844 LSOAs across 317 local authority districts, matching the official counts. Each LSOA has around 1,500 people (MHCLG, 2019), small enough to map a neighbourhood but large enough to be reliable.

The second dataset is an Office for National Statistics lookup table (ONS, 2019) connecting each local authority district to one of the nine English regions, joined to the IMD data on district code with a left join. Download links for both datasets are in the reference list. The join was verified rather than assumed: no area was left without a region, so all 32,844 LSOAs matched. This mattered because some district boundaries changed during 2019, so a lookup from the wrong year would leave areas unmatched.

Variable	Description
lsoa11cd	LSOA code (unique identifier)
LADcd / LADnm	Local authority district code and name (join key for the region lookup)
IncScore	Income deprivation score (proportion, 0 to 1)
EmpScore	Employment deprivation score (proportion, 0 to 1)
EduScore	Education, skills and training score (0 to 100)
HDDScore	Health deprivation and disability score (standardised)
CriScore	Crime score (standardised)
BHSScore	Barriers to housing and services score
EnvScore	Living environment score
GBScore / WBScore	Sub-domains of barriers: geographical and wider (RQ3)
IMDScore	Overall IMD score (describes clusters only)
geometry	LSOA boundary shapes (used for the map)
Region	English region, joined from the ONS lookup

Table 1. Variables used in the analysis and their description

Table 1 lists the variables used, with names taken from the shapefile using `names()` and descriptions from the technical report (MHCLG, 2019). Table 2 shows the first three rows of the prepared data using `head()` function. We used domain scores rather than ranks or deciles, because ranks only show which area is more deprived, while scores also show by how much.

The overall IMD score was used later to describe the clusters, but not to build them. It is made from the same seven domains, so using it would count them twice.

LSOA	Income	Employment	Education	Health	Crime	Barriers to Housing and Services	Environment
City of London 001A	0.007	0.010	0.024	-1.654	-2.012	29.472	31.873
City of London 001B	0.034	0.027	0.063	-1.115	-2.343	24.412	23.084
City of London 001C	0.086	0.086	5.804	-0.102	-1.032	40.103	40.535

Table 2. First three rows of prepared data

2.2. Tools

All data preparation, analysis and visualisation were carried out in R using VSCode. Data manipulation used the tidyverse package (Wickham et al., 2019), skewness was checked with e1071 (Meyer et al., 2025), and the correlation heatmap was drawn with ggcorrplot (Kassambara, 2026). Cluster analysis was done using the k-means method from the stats package, using the Hartigan and Wong (1979) algorithm. Visualisations were created using ggplot2 (Wickham, 2016), while the sf package (Pebesma, 2018) was used to read the shapefile and create the geographical maps. The full R script is included in the Appendix and can be run from start to finish to reproduce the results used in this report.

2.3. Preparation and checks

The shapefile(.shp) had all 32,844 rows, which matches the published number of English LSOAs. All the variables used in the analysis were checked for missing values. The LSOA codes were also checked for duplicates. No missing values or duplicates were found, so no data needed fixing or removing and the preparation focused on two issues instead. First, the seven domains use very different scales, as shown in Table 2. Income and employment run from 0 to 1, education from 0 to around 99, while health and crime are already standardised around zero and can be negative. This is a problem for methods based on distance, because education would dominate the clustering simply because it has larger values. All seven domains were therefore converted into z-scores before clustering, giving each a mean of 0 and standard deviation of 1.

Second, some of the domains are strongly skewed to the right. The skewness values from the code are 1.34 for employment, 1.22 for education and 1.20 for income. This means that most areas have lower scores, and a small number of areas have very high scores, as shown in Figure 1. Spearman correlation was used to tackle this issue. Spearman correlation uses the ranking of the areas rather than their actual values. It shows how much the long right tails pull the linear measure around. Given the long right tails in Figure 1, it is the safer choice here. A log transform is a common way to reduce this kind of right skew, but it is not possible here. The

health and crime scores are standardised around zero and take negative values, for which a logarithm is undefined. Rank-based Spearman correlation and z-standardisation were therefore used instead, as they handle the skew without requiring the scores to be positive.

A few outputs are quoted directly in the text rather than given a table of their own, specifically the cross-tabulation of the three- and four-cluster solutions, the district-level shares and the two barriers sub-domain averages. Each is a single short output from the script and is reproduced by running that section of the appendix code.

2.4. Analysis

We followed an iterative analysis process in which each stage was motivated by the output of the previous stage, looping back rather than moving in a straight line (O’Neil and Schutt, 2013), so not all the research questions were decided at the start. The first stage was the correlation analysis (RQ1), where barriers to housing and services returned a correlation of only 0.04 with income. Correlation alone could not show how many areas have high deprivation in this domain, which motivated the cluster analysis (RQ2). Table 5 then showed one group mainly affected by this domain, raising the question of where those areas are, answered using region data and spatial analysis of the LSOA boundaries (RQ3).

K-means (Hartigan and Wong, 1979) was used on the standardised scores. It was run with 25 random starts and a fixed random seed, so the appendix code reproduces these results. The within-cluster sum of squares was calculated on a random sample of 5,000 areas to narrow down the number of clusters, since running it on all 32,844 would take far longer, but the final clustering used all 32,844 areas. The sum of squares fell most sharply at $k = 2$, but two clusters would only separate deprived areas from the rest, which the overall IMD score already does. The drops after $k = 4$ shrink to less than half of what the fourth cluster added, so the useful range was three or four. Both were fitted on the full data and cross-tabulated before choosing four clusters, as explained under RQ2 below. Region and cluster are categories, so both were treated as factors in R and were shown as counts and percentages instead of averages.

3. Results and Discussion

3.1. Shape of Domain Scores

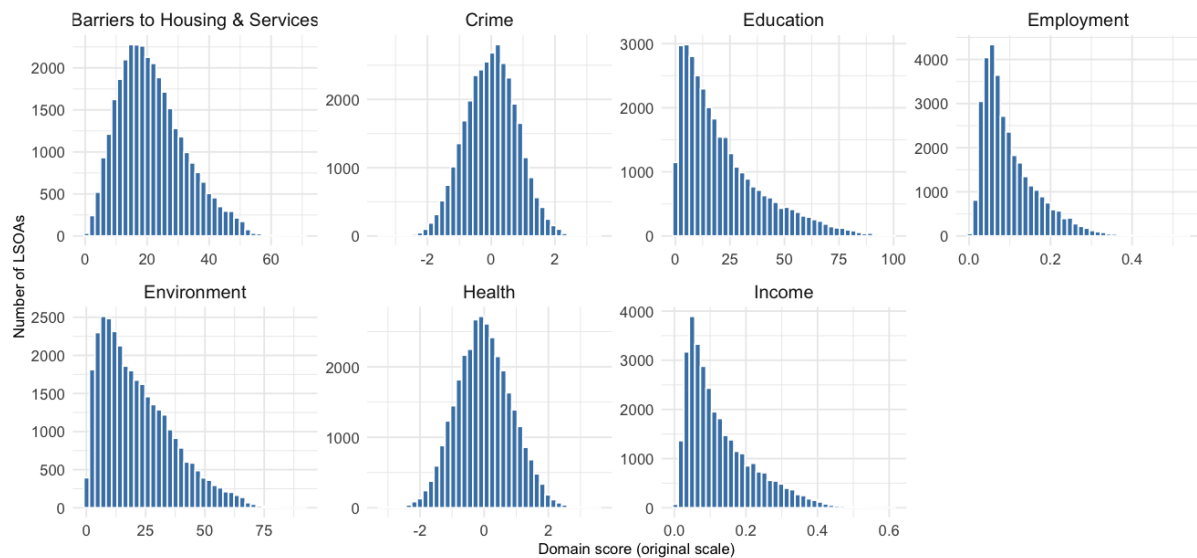


Figure 1. Distribution of each domain score across all 32,844 LSOAs

Figure 1 shows how the scores are spread across each domain, while Table 3 shows the actual values. Income, employment and education have most of their scores at the lower end, with only a few areas having very high scores. The median income score is 0.099 against a maximum of 0.609, around six times higher. Health and crime look more evenly spread because MHCLG publishes them in standardised form (MHCLG, 2019), which is why both average 0.000.

Domain	Min	Median	Mean	Max	Skewness
Income	0.003	0.099	0.128	0.609	1.195
Employment	0.002	0.080	0.100	0.534	1.344
Education	0.013	16.181	21.691	99.446	1.224
Health	-3.215	-0.027	0.000	3.547	0.112
Crime	-3.459	0.016	0.000	3.350	-0.019
Barriers to housing & services	0.483	20.196	21.691	70.456	0.622
Environment	0.126	18.512	21.691	91.602	0.873

Table 3. Summary statistics and skewness of each domain

Note. The mean values of education, barriers to housing and services and living environment agree to seven significant figures (21.6910839, 21.6910808 and 21.6910806 respectively) and appear identical only after rounding to three decimal places.

3.2. RQ1: How are the seven domain scores related to each other across LSOAs?

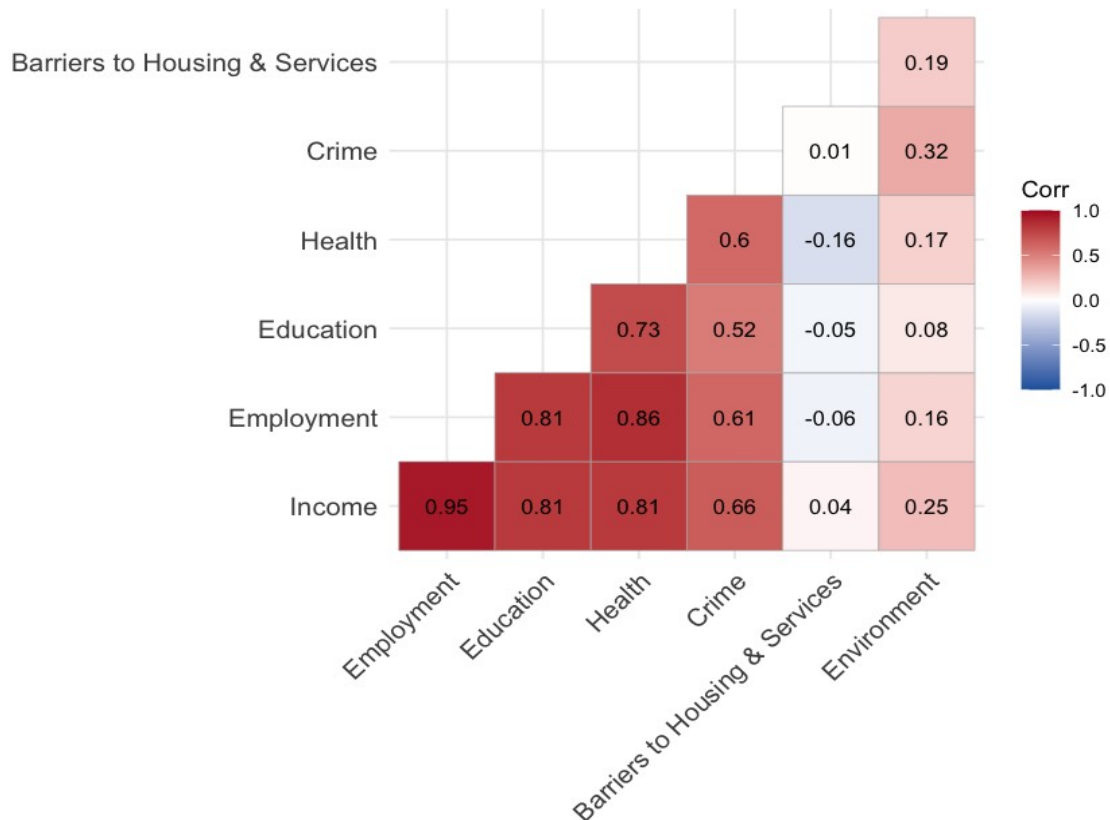


Figure 2. Spearman correlation between the seven domain scores

In Figure 2, we can see all 21 correlations between the domains. Four domains are strongly connected: income, employment, education and health, with correlations between 0.73 and 0.95. Income and employment have the strongest connection at 0.95, meaning they show almost the same pattern across areas. Crime is also related to these domains, but not as strongly, with correlations between 0.52 and 0.66. These four form a common pattern: areas with low income usually also face problems with employment, education and health.

Barriers to housing and services is very different from the other domains. Its correlation with income is only 0.04, with employment it is -0.06, with education -0.05 and with health -0.16. These are all close to zero, so knowing an area's income deprivation reveals almost nothing about its access to housing and services, and the negative figure for health shows the two move in opposite directions. Living environment is also weakly connected, correlating between 0.08 and 0.32 with the others.

RQ1 therefore gives a clear result. Deprivation in England does not follow one single pattern: most domains move together, while barriers to housing and services and living environment both behave very differently. Because the overall IMD score combines all seven, an area with serious access problems but no other deprivation will not stand out. Correlation alone cannot show how many areas are like this, which is why cluster analysis follows.

3.3.RQ2: Can LSOAs be grouped into a small number of distinct deprivation profiles?

K-means cannot decide the number of clusters for itself. The within-cluster sum of squares fell from 36,091 at $k = 1$ to 21,601 at $k = 2$, a drop of 14,490 and easily the sharpest bend in the curve. But two clusters would only separate deprived areas from non-deprived ones, which the overall IMD score already does. The later drops fall from 3,680 ($k = 3$) and 2,675 ($k = 4$) to 1,283 at $k = 5$, so each cluster past the fourth adds far less than the fourth did. That narrows the choice to three or four, and both were fitted on the full data and compared. With $k = 4$, cluster 2 holds 6,360 neighbourhoods that share a similar profile. With $k = 3$, those same neighbourhoods are split almost entirely across two larger mixed groups, 4,110 into one and 2,249 into the other. They stop being a group of their own, so the three-cluster solution loses the pattern already found in RQ1, and $k = 4$ was chosen. K-means numbers its clusters in no particular order, so the four were renamed using their average overall IMD score: cluster 1 is Least deprived, cluster 2 is High-barriers, cluster 3 is Moderately deprived, and cluster 4 is Multiply deprived. Aungkulanon et al. (2017) picked their number of clusters the same way, by asking whether the final groups meant anything rather than relying on a statistical rule alone.

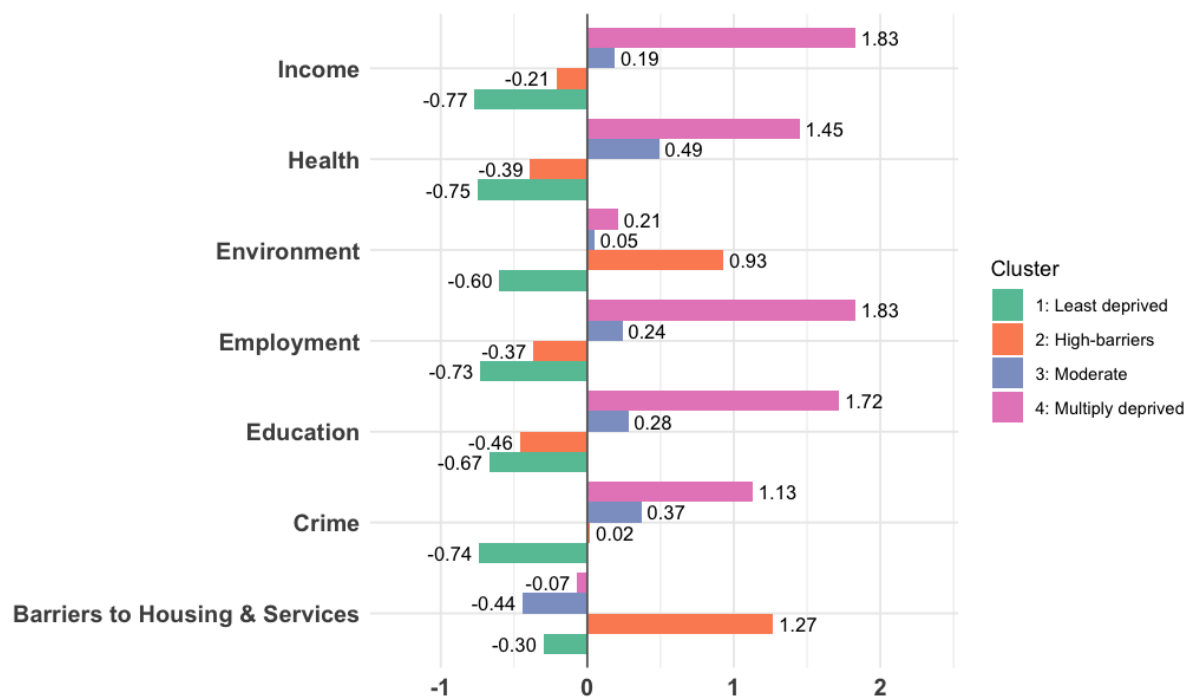


Figure 3. Average standardised score per domain for each cluster

Note: Cluster labels are assigned by reading the profiles, not produced by the algorithm

Cluster	Label	LSOAs	Share of England	Mean IMD score
1	Least deprived	12,397	37.7%	8.6
2	High-barriers	6,360	19.4%	20.6
3	Moderately deprived	9,055	27.6%	24.6
4	Multiply deprived	5,032	15.3%	50.0

Table 4. Cluster sizes, share of England and mean overall IMD score.

Figure 3 and Table 5 give each cluster's profile, and Table 4 gives their sizes.

Cluster	Income	Employment	Education	Health	Crime	Barriers to Housing & Ser.	Environment
1	-0.77	-0.73	-0.67	-0.75	-0.74	-0.30	-0.60
2	-0.21	-0.37	-0.46	-0.39	0.02	1.27	0.93
3	0.19	0.24	0.28	0.49	0.37	-0.44	0.05
4	1.83	1.83	1.72	1.45	1.13	-0.07	0.21

Table 5. Average standardised score per domain per cluster

Cluster 4 is the most deprived group, and its mean IMD score of 50.0 is more than double that of any other cluster.

Cluster 2 is the high-barriers group. Its mean overall IMD score is only 20.6, even though it is far more deprived in barriers to housing and services, so the overall score does not pick this type of deprivation up. It is not a small group either. Almost one in five neighbourhoods in England are in it. The index actually ranks these areas, with a mean score of 20.6, as less deprived than the moderately deprived cluster at 24.6, even though the two face completely different problems rather than the same problem in different amounts. Aungkulanon et al. (2017) found something similar in Thailand, where looking at separate deprivation profiles showed differences that a single overall ranking did not make clear.

3.4.RQ3: How are those deprivation profiles spread across the regions of England?

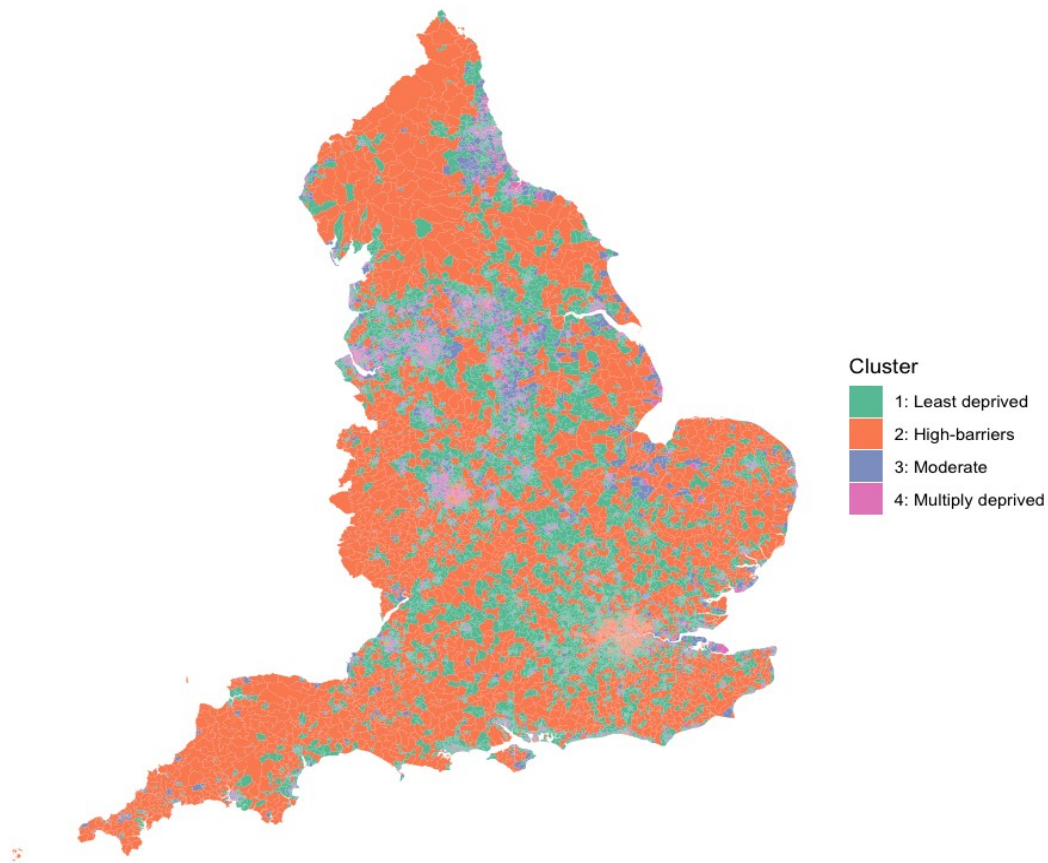


Figure 4. The four clusters across English LSOAs

In Figure 4, we can see where the clusters are, and Figure 5 and Table 6 compare them across the English regions using the ONS lookup. Cluster 4 is mainly in the North (Table 6). It makes up 30.5% of neighbourhoods in the North East, 28.4% in the North West and 26.0% in Yorkshire and The Humber, against only 6.3% in the South East and 4.7% in London. The pattern is sharper still at district level, where more than half the neighbourhoods in several districts fall into cluster 4. These are Knowsley (55.1%), Liverpool (54.4%), Middlesbrough (53.5%), Nottingham (51.1%) and Manchester (50.7%), with Blackpool exactly at half, Kingston upon Hull at 49.4% and Birmingham at 47.7%. In districts like these, multiple deprivation is not limited to a few neighbourhoods but covers a large part of the district. This supports Noble et al. (2006), who argued that deprivation should be measured over smaller areas, since regional averages hide differences like these.

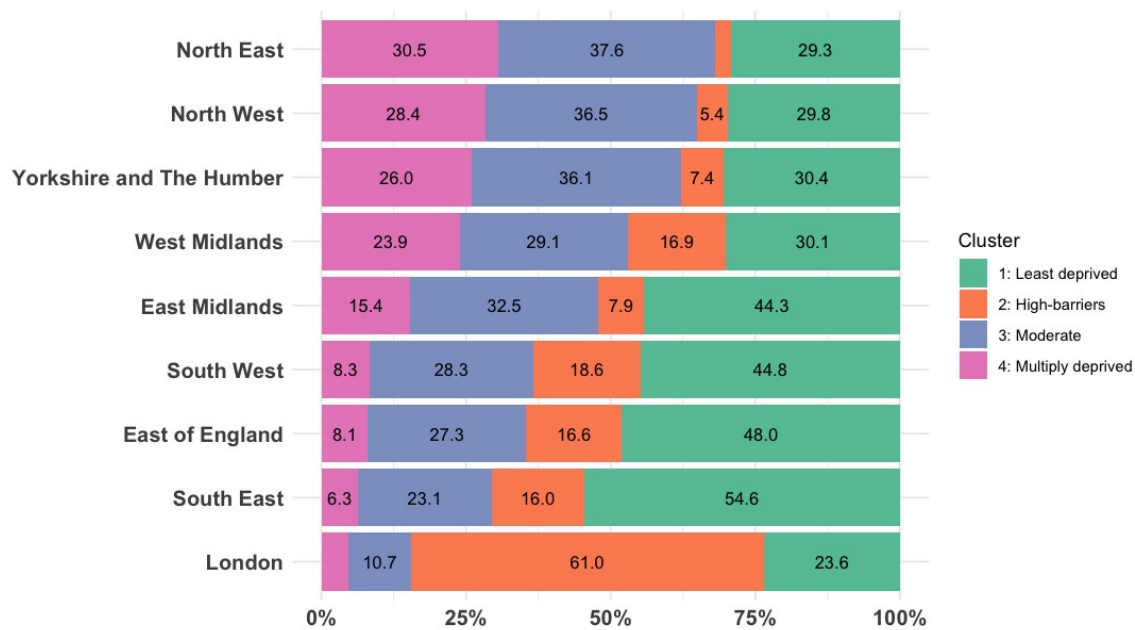


Figure 5. Share of LSOA clusters by Region

Region	1: Least deprived	2: High-barriers	3: Moderate	4: Multiply deprived
North East	29.3	2.6	37.6	30.5
North West	29.8	5.4	36.5	28.4
Yorkshire and The Humber	30.4	7.4	36.1	26.0
West Midlands	30.1	16.9	29.1	23.9
East Midlands	44.3	7.9	32.5	15.4
South West	44.8	18.6	28.3	8.3
East of England	48.0	16.6	27.3	8.1
South East	54.6	16.0	23.1	6.3
London	23.6	61.0	10.7	4.7

Table 6. Percentage of each region's LSOAs in each cluster

Cluster 2 has the most unusual pattern in Table 6. It makes up 61.0% of neighbourhoods in London, but it is also common outside the cities: in the South West (18.6%), the East of England (16.6%), and the West Midlands (16.9%), which mixes both. Since London and rural areas are very different, the two parts of the barriers domain were looked at separately. Both sub-domain scores are standardised so that the national mean of each is zero, which makes the following contrast direct. In London, areas in cluster 2 have an average score of +4.29 for wider barriers, which includes problems such as housing affordability, overcrowding and homelessness (MHCLG, 2019). Their average score for geographical barriers is -0.63. Outside London, the pattern is the opposite. The average score is +1.04 for geographical barriers, which

measures how far people have to travel to reach services such as shops, GP surgeries, schools and post offices (MHCLG, 2019). The average score for wider barriers is -0.47.

The same cluster therefore contains two different problems: access to affordable housing in London, and distance from services in rural areas. This is why the parts of the barriers domain need to be read separately from its overall score. Salvatore and Grundy (2021) make a similar point. They argue that it is important to understand how deprivation affects an area, not just how much deprivation there is. Because the published barriers score averages two components that move in opposite directions in these two settings, it partly cancels itself out, which is a finding about the measurement instrument as much as about places.

4. Conclusion

In this report, we look at deprivation across all 32,844 LSOAs in England. For RQ1, the results show that the domains do not all behave the same way. Income, employment, education and health are strongly related, at 0.73 to 0.95, and crime more weakly at 0.52 to 0.66. Barriers to housing and services is very different, correlating 0.04 with income and -0.16 with health. For RQ2, k-means found four groups. The high-barriers group of 6,360 neighbourhoods is better than average on all four socioeconomic domains but deprived on access to housing and services and on living environment (+0.93). Its mean overall IMD score is still only 20.6, so the overall score does not show the problem. For RQ3, the most deprived group sits mainly in the North East, North West and Yorkshire, and in districts such as Knowsley and Liverpool, more than half the neighbourhoods are in it. The high-barriers group covers 61.0% of London but is also common in rural southern and eastern England, for very different reasons.

There are three main points from these results. First, deprivation in England is not just one single thing. There are different types of deprivation, and looking only at the overall IMD score can hide some of them. Second, 19.4% of neighbourhoods face a type of deprivation the overall score shows poorly, where the problem is access rather than income. The score does more than hide this group. Cluster 2 has a mean score of 20.6 and cluster 3 has 24.6, so the index actually places the high-barriers areas as the less deprived of the two. But cluster 3 is mildly deprived in income, employment, education and health, while cluster 2 is better than average in all four and deprived in access and living environment instead. The score puts them in an order as if they were the same problem in different amounts, but they are not the same problem at all. Third, high barriers to housing and services do not mean the same thing everywhere. Inside this group, London areas score +4.29 on wider barriers, which covers affordability and overcrowding, but -0.63 on geographical barriers, while outside London it is the other way round at -0.47 and +1.04. The two halves therefore have opposite problems, the published barriers score is partly cancelling itself out, and the two need different solutions.

Two main lessons came from this analysis. First, the analysis changed as the data was explored. Clustering was not part of the original plan. It was added after the correlation results showed that one domain behaved very differently from the others. This is the kind of looping back described by O'Neil and Schutt (2013). Second, an overall score should not be taken at face value, because here the IMD score and its seven domains gave different pictures of the same areas. The number of clusters also mattered more than expected. With three clusters instead of four, the 6,360 high-barriers areas are split 4,110 and 2,249 between two other groups and disappear as a group of their own. So, a normal choice would have removed the main finding of this report from view.

There are also some limitations to this analysis. The results are about areas and not the individual people living in them. Drawing conclusions about individuals from area-level patterns would be an ecological fallacy. The analysis also uses data from only one year, so it cannot show how deprivation has changed over time or what actually caused it. K-means also puts every LSOA into exactly one cluster, so the four groups make the patterns easier to read but

simplify the real situation. The analysis only covers England because Scotland, Wales and Northern Ireland measure deprivation differently, so they cannot be directly compared (MHCLG, 2019). With more time, the analysis could be taken further in three ways. Hierarchical clustering could test whether similar groups appear and comparing the 2015 and 2019 releases would show whether areas move between groups over time. Finally, mortality data could be added to see whether areas with high barriers to housing and services also have worse health outcomes that are not clear from their overall IMD score. This would also help to test the findings of Aungkulanon et al. (2017).

5. References

- Aungkulanon, S., Tangcharoensathien, V., Shibuya, K., Bundhamcharoen, K., & Chongsuvivatwong, V. (2017). Area-level socioeconomic deprivation and mortality differentials in Thailand: Results from principal component analysis and cluster analysis. *International Journal for Equity in Health*, 16(1), Article 117. <https://doi.org/10.1186/s12939-017-0613-z>
- Consumer Data Research Centre. (2019). *Index of Multiple Deprivation (IMD): English IMD 2019* [Data set]. <https://data.cdrc.ac.uk/dataset/index-multiple-deprivation-imd>
- Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1), 100–108. <https://doi.org/10.2307/2346830>
- Kassambara, A. (2026). *ggcorrplot: Visualization of a correlation matrix using 'ggplot2'* (Version 0.2.0) [R package]. <https://CRAN.R-project.org/package=ggcorrplot>
- Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2025). *e1071: Misc functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien* (Version 1.7-17) [R package]. <https://doi.org/10.32614/CRAN.package.e1071>
- Ministry of Housing, Communities and Local Government. (2019). *The English indices of deprivation 2019: Technical report*.
- Noble, M., Wright, G., Smith, G., & Dibben, C. (2006). Measuring multiple deprivation at the small-area level. *Environment and Planning A*, 38(1), 169–185. <https://doi.org/10.1068/a37168>
- Office for National Statistics. (2019). *Local authority district to region (April 2019) lookup in England* [Data set]. ONS Open Geography Portal. <https://geoportal.statistics.gov.uk/datasets/ons:local-authority-district-to-region-april-2019-lookup-in-en/about>
- O'Neil, C., & Schutt, R. (2013). *Doing data science: Straight talk from the frontline*. O'Reilly Media.
- Pebesma, E. (2018). Simple features for R: Standardized support for spatial vector data. *The R Journal*, 10(1), 439–446. <https://doi.org/10.32614/RJ-2018-009>
- Salvatore, M. A., & Grundy, E. (2021). Area deprivation, perceived neighbourhood cohesion and mental health at older ages: A cross lagged analysis of UK longitudinal data. *Health & Place*, 67, Article 102470. <https://doi.org/10.1016/j.healthplace.2020.102470>
- Townsend, P. (1987). Deprivation. *Journal of Social Policy*, 16(2), 125–146. <https://doi.org/10.1017/S0047279400020341>
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (2nd ed.). Springer.
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemond, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>

6. Appendix

```
# Importing Libraries =====
library(sf)
library(tidyverse)
library(e1071)
library(ggcorrplot)

# to make the k-means result reproducible
set.seed(1234)

# Figure size for plots displayed (inches).
options(repr.plot.width = 12, repr.plot.height = 6)

# Setting up Data Directory =====
data_dir <- "/Users/srujan/Downloads/Data3R/Data3"

# Loading Data =====
imd_raw <- st_read(file.path(data_dir, "IMD_2019.shp"))

# Data info and Selection =====
nrow(imd_raw)
names(imd_raw)
n_distinct(imd_raw$LADcd)

imd <- select(imd_raw,
  lsoa_code = lsoallcd, lsoa_name = lsoallnm, lad_code = LADcd, lad_name =
  LADnm,
  imd_score = IMDScore, Income = IncScore, Employment = EmpScore,
  Education = EduScore, Health = HDDScore, Crime = CriScore,
  `Barriers to Housing & Services` = BHSScore, Environment = EnvScore,
  barriers_geo = GBScore, barriers_wider = WBScore)

domain_cols <- c("Income", "Employment", "Education", "Health",
  "Crime", "Barriers to Housing & Services", "Environment")

# the data without the map shapes, used for all the tables below
imd_flat <- st_drop_geometry(imd)

# data quality checks: missing values and duplicate codes (both 0)
summarise(imd_flat, across(everything(), ~ sum(is.na(.))))
sum(duplicated(imd_flat$lsoa_code))

# Table 2: Three header rows of the prepared data
table2_data <- select(imd_flat, lsoa_name, all_of(domain_cols))
head(mutate(table2_data, across(where(is.numeric), ~ round(., 3))), 3)

# EDA =====

# one row per area per domain, used for Table 3 and Figure 1
domain_long <- pivot_longer(select(imd_flat, all_of(domain_cols)),
  everything(),
```

```

names_to = "Domain", values_to = "Score")
# Table 3: summary and skewness of each domain in one output
table3_data <- summarise(group_by(domain_long, Domain),
  Min = min(Score), Median = median(Score),
  Mean = mean(Score), Max = max(Score),
  Skewness = skewness(Score))
as.data.frame(mutate(table3_data, across(where(is.numeric), ~ round(.,
3))))

# Figure 1: how each domain is spread
ggplot(domain_long, aes(Score)) +
  geom_histogram(bins = 40, fill = "steelblue", colour = "white") +
  facet_wrap(~ Domain, scales = "free", ncol = 4) +
  labs(x = "Domain score (original scale)", y = "Number of LSOAs") +
  theme_minimal()
ggsave("fig1.png", width = 11, height = 5.2, dpi = 300)

# Correlation =====
cor_matrix <- cor(select(imd_flat, all_of(domain_cols)),
  method = "spearman")
round(cor_matrix, 2)

# Figure 2: heatmap with every coefficient printed on the tile
ggcorrplot(cor_matrix, type = "lower", lab = TRUE, lab_size = 3.5,
  colors = c("#2166AC", "white", "#B2182B"),
  outline.color = "grey70")
ggsave("fig2.png", width = 7, height = 6, dpi = 300)

# Clustering =====
domains_z <- scale(select(imd_flat, all_of(domain_cols)))

# Total WCSS for k = 1 to 8, on a random sample of 5,000
# To narrow down the number of clusters before fitting on the full data
sample_idx <- sample(seq_len(nrow(domains_z)), 5000)
domains_sample <- domains_z[sample_idx, , drop = FALSE]

wss <- numeric(8)
for (k in 1:8) {
  wss[k] <- kmeans(domains_sample, centers = k, nstart = 10, iter.max =
50)$tot.withinss
}
data.frame(k = 1:8, wss = round(wss),
  drop_from_previous = c(NA, round(-diff(wss))))

# two clusters would only split deprived from not deprived, so the
# useful range is 3 or 4; both are fitted on the full data and compared
km3 <- kmeans(domains_z, centers = 3, nstart = 25, iter.max = 50)
km <- kmeans(domains_z, centers = 4, nstart = 25, iter.max = 50)

# how the two solutions line up: under k = 3 the high-barriers group
# splits 4,110 / 2,249 across two groups
table(km3$cluster, km$cluster)

```

```

# renumbering the 4 clusters from least to most deprived
imd$cluster_raw <- km$cluster
cluster_means <- summarise(group_by(st_drop_geometry(imd), cluster_raw),
                           m = mean(imd_score))
cluster_order <- cluster_means$cluster_raw[order(cluster_means$m)]
imd$cluster <- factor(match(imd$cluster_raw, cluster_order),
                     levels = 1:4,
                     labels = c("1: Least deprived", "2: High-barriers",
                                "3: Moderate", "4: Multiply deprived"))

# refresh the flat copy so it carries the cluster column
imd_flat <- st_drop_geometry(imd)

# Table 4: cluster sizes, share of England and mean IMD score
cluster_sizes <- summarise(group_by(imd_flat, cluster),
                           LSOAs = n(),
                           Mean_IMD = round(mean(imd_score), 1))
mutate(cluster_sizes,
       Share_of_England_pct = round(100 * LSOAs / sum(LSOAs), 1))

# Table 5: average z-score per domain per cluster
z_scores <- as_tibble(domains_z)
z_scores$cluster <- imd$cluster
cluster_profiles <- summarise(group_by(z_scores, cluster),
                              across(everything(), ~ round(mean(.), 2)))
cluster_profiles

# Figure 3: the same profiles as a chart, with the values printed on the
bars so they can be read exactly

profiles_long <- pivot_longer(cluster_profiles, -cluster,
                              names_to = "Domain", values_to = "mean_z")
ggplot(profiles_long, aes(Domain, mean_z, fill = cluster)) +
  geom_col(position = position_dodge(width = 0.9)) +
  geom_text(aes(label = sprintf("%.2f", mean_z)),
            position = position_dodge(width = 0.9),
            hjust = ifelse(profiles_long$mean_z >= 0, -0.15, 1.15),
            size = 3.5) +
  geom_hline(yintercept = 0, colour = "grey40") +
  coord_flip() +
  ylim(-1.3, 2.35) +
  scale_fill_brewer(palette = "Set2", name = "Cluster") +
  labs(x = NULL, y = NULL) +
  theme_minimal() +
  theme(
    axis.title.x = element_text(size = 14, face = "bold"),
    axis.title.y = element_text(size = 14, face = "bold"),
    axis.text.x = element_text(size = 12, face = "bold"),
    axis.text.y = element_text(size = 12, face = "bold")
  )
ggsave("fig3.png", width = 8.5, height = 5, dpi = 300)

```

```

# Map and regions =====
# Figure 4: map of the clusters
options(repr.plot.width = 7.5, repr.plot.height = 9)
ggplot(st_simplify(imd, dTolerance = 150)) +
  geom_sf(aes(fill = cluster), colour = NA) +
  scale_fill_brewer(palette = "Set2", name = "Cluster") +
  theme_void()
ggsave("fig4.png", width = 7.5, height = 9, dpi = 300)

# second data source: ONS district to region lookup
lookup <- select(read_csv(file.path(data_dir, "LAD19_RGN19_EN.csv")),
  lad_code = LAD19CD, region = RGN19NM)

imd <- left_join(imd, lookup, by = "lad_code")
imd_flat <- st_drop_geometry(imd)

# join check
sum(is.na(imd_flat$region))

# Table 6: percentage of each region's areas in each cluster
region_counts <- count(imd_flat, region, cluster)
region_pct <- mutate(group_by(region_counts, region),
  pct = round(100 * n / sum(n), 1))
region_table <- pivot_wider(select(ungroup(region_pct), -n),
  names_from = cluster, values_from = pct)
arrange(region_table, desc(`4: Multiply deprived`))

# Figure 5: the same breakdown as a chart, with percentages printed
region_shares <- mutate(group_by(region_counts, region), share = n /
sum(n))
ggplot(region_shares,
  aes(reorder(region, share * (cluster == "4: Multiply deprived")),
    share, fill = cluster)) +
  geom_col() +
  geom_text(aes(label = ifelse(share >= 0.05,
    sprintf("%.1f", 100 * share), "")),
    position = position_stack(vjust = 0.5), size = 3.5) +
  coord_flip() +
  scale_y_continuous(labels = scales::percent) +
  scale_fill_brewer(palette = "Set2", name = "Cluster") +
  labs(x = NULL, y = NULL) +
  theme_minimal() +
  theme(
    axis.title.x = element_text(size = 14, face = "bold"),
    axis.title.y = element_text(size = 14, face = "bold"),
    axis.text.x = element_text(size = 12, face = "bold"),
    axis.text.y = element_text(size = 12, face = "bold")
  )
ggsave("fig5.png", width = 9, height = 5, dpi = 300)

```

```

# districts where cluster 4 covers half or nearly half of all areas
district_stats <- summarise(group_by(imd_flat, lad_name),
                             n = n(),
                             share_c4 = round(
                               100 * mean(cluster == "4: Multiply
deprived"), 1))
head(arrange(filter(district_stats, n >= 50), desc(share_c4)), 8)

# inside cluster 2: London against the rest on the two barriers
# sub-domains
cluster2 <- filter(imd_flat, cluster == "2: High-barriers")
cluster2 <- mutate(cluster2, where = if_else(region == "London",
                                             "London", "Outside London"))
summarise(group_by(cluster2, where),
           mean_geo   = round(mean(barriers_geo), 2),
           mean_wider = round(mean(barriers_wider), 2),
           n = n())

```