# Composite 2020 UK Index of Multiple Deprivation

This repository uses methodology from a paper by [Gary A Abel, Matthew E Barclay, and Rupert A Payne](https://bmjopen.bmj.com/content/6/11/e012750) to create a composite measure of multiple deprivation across the UK.

The data can be downloaded from: <https://mysociety.github.io/composite_uk_imd/>

Update 26/08/2022: A previous version of this dataset had incorrect rankings for Northern Ireland SOAs. [You can read a write-up of this error.](https://mysociety.github.io/composite_uk_imd/analysis/v1_uk_index_error_analysis.html)

## Using this dataset

There are two sets of files depending if you want a GB or UK index (methodological differences explained below). For each set (uk\_index and gb\_index folders), there is a different file for each nation. In each case, the measures of deprivation in the other nations has been integrated into the base scores for that nation.

When in a dataset for a particular nation (for instance, uk\_index/UK\_IMD\_E.csv), the rank order of the original deprivation index is preserved, but it is not for other countries (explained more below).

The composite labels vary by national dataset, but the key columns in the UK England dataset are:

* lsoa - The code for the LSOA/Datazone/SOA
* original\_decile - the original decile in the four different IMD datasets.
* overall\_local\_score - the original deprivation score in the four different IMD datasets.
* UK\_IMD\_E\_score - the transformed deprivation score into the combined index. This will be the same as the original for the base country.
* E\_expanded\_decile - the English deciles, with scores for other nations fitted into the same devisions as the original deciles.
* UK\_IMD\_E\_pop\_decile - A new set of deciles for the whole dataset, so that 10% of the population is in each decile (there will be an uneven number of areas).
* UK\_IMD\_E\_pop\_quintile - A new set of quintiles for the whole dataset, so that 20% of the population is in each quintile (there will be an uneven number of areas).

### Example usage of different kinds of deciles

Given the theoretical basis for national metrics is more sound than the composite index, you should chose the base that better reflects the majority of your data.

If your analysis runs across the whole country, you would use uk\_index\UK\_IMD\_E.csv (as the majority of data is English), and use the UK\_IMD\_E\_pop\_decile column to break the country in ten deciles by population. If a variable is evenly spread amount different levels of deprivation across the whole country, aggregating by this decile type should return a similar value for all deciles.

If you wanted to base the analysis in SIMD, but include a few compariable English examples on the other side of the border, you would use uk\_index\UK\_IMD\_S.csv and the S\_expanded\_decile column. This will reflect the ordering and deciles of the original SIMD dataset, but put areas in other nations into Scottish deciles. Overall, this set of deciles is uneven because extra areas are being squeezed into the divisions made on the basis of national data.

In many cases, the two kinds of deciles are similar if based in English data (only 11% of areas have disagreements between the two kinds of deciles). If unsure, think about how you would explain what is happening in your analysis (e.g. if conceptually exploring UK deprivation, it is best to stick to the population deciles).

### How would I map this data - what variable should I use?

There is no single answer to this question. There are four files that cover the whole UK at the small area level (UK\_IMD\_E.csv, UK\_IMD\_N.csv, UK\_IMD\_S.csv, UK\_IMD\_W.csv), each of which is slightly different. If you are an England-based researcher, for example, and you want to map the whole UK in a comparable way so that you can meaningfully compare patterns of deprivation in East London with patterns of deprivation in the Scottish Highlands, then you would use the UK\_IMD\_E.csv file and make a map using the UK\_IMD\_E\_pop\_decile column in this dataset. This would create a map that puts every LSOA/SOA/DZ in the UK into one of ten deciles (1 = most deprived, 10 = least deprived) so that an equal number of people are in each decile (n.b. this means that there won't be an equal number of small areas in each decile because populations vary by LSOA/SOA/DZ).

Our advice for anyone making a UK deprivation map would be to use the base UK\_IMD file that makes most sense to them from an end user point of view. If you primarily work with SIMD data and want to see how the decile position of a data zone in Scotland might compare to an SOA in Northern Ireland or a LSOA in England and Wales, then the UK\_IMD\_S.csv would likely be most meaningful.

[Alasdair Rae](https://twitter.com/undertheraedar) has created [a single shapefile](https://drive.google.com/drive/folders/1L-C-Ym7VTwRhIRDQUMHPxFvOznMrKfy3) that covers the different sets of small area that will be helpful in cross-border mapping.

### Interpreting scores, ranks, deciles and quintiles

For users new to indexes of deprivation it is not always clear which direction is "worse". This set of indexes follows the general convention that:

* Higher scores show 'worse' deprivation. The most deprived area is the one with the highest score.
* For ranks this flips around, Rank 1 is the most deprived area.
* Deciles and Quintiles follow ranks. Decile 1 and Quintile 1 are the most deprived decile and quintile respectively.

So if plotting a column graph of deprivation deciles from 1-10, the most deprived decile is on the far left. This may differ from the initiative sense that in a measure of deprivation higher numbers mean more deprivation, but this maintains consistency with how deciles are generated by the individual indexes.

Explored in more detail below, but we do not recommend relying too much on raw score in this dataset as the theoretical basis for precise rankings is less robust. The exception to this is the employment score in the UK indexes, and the employment and income scores in the GB indexes, as these are based on similar underlying indicators.

## What are the indexes of multiple deprivation?

For each nation in the UK, the relevant statistics authority has created a ranking of multiple deprivation that ranks small census areas (LSOAs in England/Wales, DZs in Scotland and SOA in Northern Ireland). There is no equivalent ranking of areas across the entire UK.

These indexes are independent and cannot be directly compared. These measures are not directly comparable because the underlying indicators and the method used to combine them are different in each case. Additionally data is collected at different times for each index. It is not directly possible to take a DataZone from Scotland and see where it would fit inside the deprivation measure in England.

However, Abel and colleagues suggests you can fudge this a bit to create a combined index that gives useful results. This repository uses this approach to create an updated index with the national indexes avaliable in Feburary 2021.

## Method

The method described by the [Abel and colleagues paper](https://bmjopen.bmj.com/content/6/11/e012750) depends on the fact that the two domains of deprivation (Income and Employment) were roughly comparable between nations at the time they wrote the paper. Both domains used a single equivalent indicator, meaning the domain score expressed a real comparable value.

Given the overall IMD score, the income score and the employment score, the residuals of a linear regression will leave a lump of 'other' deprivation data that goes beyond the general effect of income/employment. Using the equivalent regression model for a different nation, the income, employment and lump deprivation can be recombined to produce a score that can be compared between nations, or fit into a single index. This looks like the following in practice:

restandardized\_residuals <- (source\_residuals/stdev(source\_residuals)) \* stdev(target\_residuals)  
overall\_score <- target\_income\_coefficent\*source\_income\_score + target\_employment\_coefficent\*source\_employment\_score + restandardized\_residuals

Unfortunately, Northern Ireland's 2017 Multiple Deprivation Measure deviates from the approach other nations use in their income domain, and so the NIMDM2017 Income measure is no longer directly comparable. For the 2020 data, only the employment score is used in the model. This represents a slight loss of information, but as Income and Employment scores are generally correlated this still explains most of the variance (see tables below). A seperate 'GB' ranking for England, Wales and Scotland uses both options.

To account for the different sizes of small area units in different nations (Scottish DataZones contain less people than LSOAs), the deciles and quintiles provided are based on deciles and quintiles of the overall population rather the rank. This means that some deciles have more 'areas' than others, but if you sampled a random pool of the population, you should have even an even division.

## How is this getting over the fundamental issue the indexes aren't comparable?

This is a good question, and the answer is it does assume that actually the scores are roughtly compariable when you bundle them all up. As the [original paper](https://bmjopen.bmj.com/content/6/11/e012750) put it *"[o]ur approach treated all standardised residuals as equivalent between the constituent nations, allowing us to create consistent UK-wide estimates of relative deprivation"*. If you think this is reasonable, the rest works fine.

### Why would it not be reasonable?

As expressed in various ways in the accompaning material to the indexes, there are reasons not to treat indexes are comparable that this method does not avoid:

* This lump of 'other' deprivation is not measuring the same thing in different indexes. Different indicators are combined in different ways in each nation to create the overall score.
* Different nations have different problems the indexes that are deliberately tuned to address (through including specific forms of deprivation that other nations do not), or this arises implicitly from the different results of the principal component analysis (PCA) in different national datasets.
* If one nation has a higher rate of one form of deprivation (for instance, if one generally has higher health deprivation), areas that would considered highly deprived in another context would be rescaled as less deprived.
* In general, the lump is difficult to examine exactly why one area is rated as more deprived as another. While within each national index you can explore domains to understand why one area is ranked above another, this just is not possible to compare in the same way between the lumps.

### Why is it probably OK anyway?

Against this, there are a few reasons to think the scale of the problem is manageable:

* While the employment score is generally only around a quarter of the total contribution to the overall IMD score, many other measures of deprivation are correlated with it, meaning that the features it measures (where there are comparable measures between nations) explain much more than a quarter of the variation (see models below).
* The remainder 'non-employment' lump in all indexes is comparing *similar* if not identical indicators of income, health, education and crime.
* Domain scores are combined using an exponential distribution. A high deprivation score on any domain contained within the lump is likely to be expressed in a high 'other' score. A clear signal of 'some form of high deprivation' should emerge from the 'other' data. This helps address some (but not all) of the issues from different scales of the data.

## When would I want to use the composite index?

For some uses the combined index is probably fine. It should not be used in a public ranking of places as it is unclear if a Scottish Datazone being ranked one level about a Welsh LSOA is theoretically valid or how to explain why it is happening. However, if breaking the result into deciles or mapping against other variables, this reintroduces sufficent fuzziness to that the results are likely to be reasonable. Given the concept of multiple deprivation is already a little fuzzy, in situations where the complete English index is used as a variable it is likely to be fine to use the expanded index.

For instance, our use case is trying to avoid throwing away data generated by service users from non-English nations when analysing patterns of deprivation against specific usage. A fuzzy measure is fine for this.

## Why not just build a proper UK index?

There are some practical issues where the statistics are generated by different statistics authorities, but the broader issue is that designing an index requires designing towards the needs of a user-base. National indexes of multiple deprivation have a set of use cases within local and national governments and the design of the indexes is geared towards that. A combined index has less usefulness locally because it is not tuned towards the specific conceptions of deprivation in each nation. It is annoying from a comparative perspective that the NIMDM2017 changed their measure of Income, but it also creates a better tool for needs in Northern Ireland.

A combined index *is* useful for researchers or organisations running services across the UK, but all of these are likely to have slightly different needs of it. It is likely that getting much beyond this fuzzy measure, the answer is for organisations to construct measures that more precisely deal with the issue at hand. For instance, evaluating access to online services suggests measures of internet access, but this is currently not included in the English index, and is only a small component of others.

Basically: the slightly dodgy composite ranking is probably fine for most purposes, and going beyond it suggests a bigger think about what needs to be measured.

## What options are available

For the UK (England, Scotland, Wales and Northern Ireland), there are versions of the overall ranking using each national model as a base. uk\_index\UK\_IMD\_E.csv contains all four nations rebased onto the English model. Generally you should use the base of the country where the majority of your data is from (see rank integrity below). Chances are this is England, but the other indexes are included for other use case.

If only using data from England, Scotland or Wales, there is a separate GB index that uses both of the comparable Income and Employment domains to create rankings.

## Data analysis

### Regression models

In all cases, the r² shows majority of the variation in the IMD score is explained using the employment score alone. GB models gain a little more usefulness through including the comparable Income score.

#### UK models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Nation | Intercept | Employment score | r² | Residuals SD |
| England | 0.06 | 2.16 | 0.90 | 4.74 |
| Scotland | 0.11 | 2.26 | 0.94 | 3.90 |
| Wales | -3.49 | 2.40 | 0.87 | 5.54 |
| Northern Ireland | -7.84 | 1.40 | 0.88 | 5.11 |

#### GB Models

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Nation | Intercept | Income score | Employment score | r² | Residuals SD |
| England | 0.51 | 1.09 | 0.72 | 0.94 | 3.63 |
| Scotland | 0.28 | 1.03 | 0.92 | 0.97 | 2.82 |
| Wales | -3.73 | 1.24 | 0.53 | 0.94 | 3.87 |

### Rank intergrity is not preserved

National subsets of the overall indexes are not identical to the original indexes. Adjusting the impact of the employment (and in the GB index, income) to match the target index and scaling the remaining residuals affects the order within nation scores as well as their placement in the overall scores. This affects almost all rankings to some degree in transposed indexes. The following tables show the differences in rank when a nation's dataset is ordered by the local IMD score and the overall IMD score of a composite index.

This is why a dataset that is mostly Scottish data should use the Scotland based version, as the theoretical basis of the majority of the rankings is clearer.

#### UK rank displacement (England based)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Northern Ireland | Scotland | Wales |
| count | 869.00 | 6945.00 | 1845.00 |
| mean | 29.58 | 108.93 | 8.61 |
| std | 26.72 | 102.22 | 8.03 |
| min | 1.00 | 1.00 | 0.50 |
| 25% | 9.00 | 31.00 | 2.50 |
| 50% | 23.00 | 81.00 | 6.00 |
| 75% | 42.00 | 154.00 | 12.00 |
| max | 216.00 | 819.00 | 57.00 |
| % of total | 97.60 | 99.60 | 96.60 |

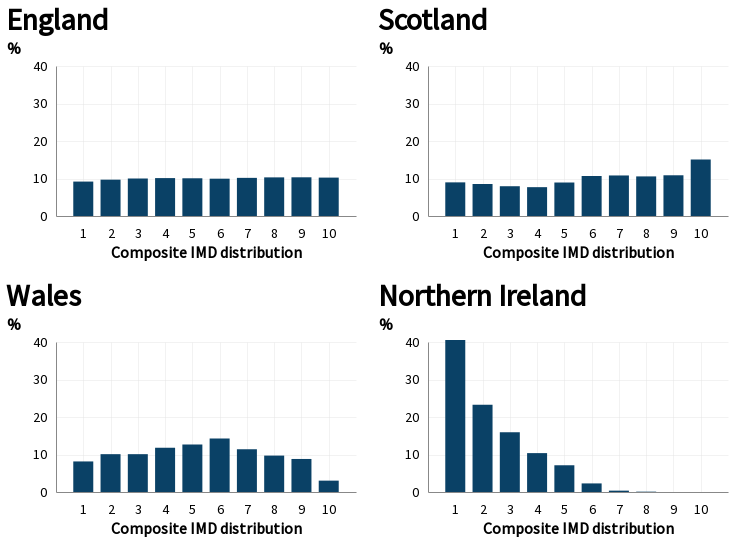
#### GB rank displacement (England based)

|  |  |  |
| --- | --- | --- |
|  | Scotland | Wales |
| count | 6953.00 | 1876.00 |
| mean | 117.22 | 13.12 |
| std | 113.24 | 11.46 |
| min | 1.00 | 0.50 |
| 25% | 31.00 | 4.00 |
| 50% | 81.00 | 9.50 |
| 75% | 170.00 | 19.50 |
| max | 1037.00 | 69.50 |
| % of total | 99.70 | 98.30 |

### Deprivation distribution

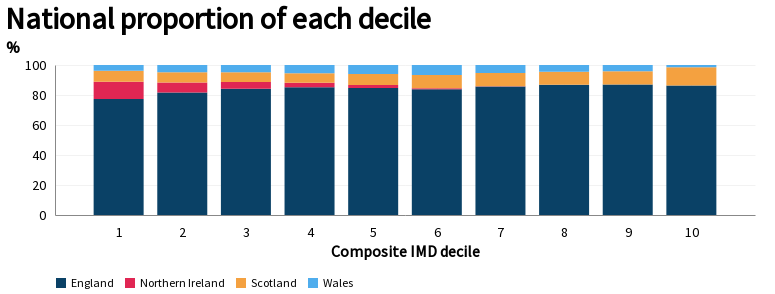
The following table and graph show the distribution of the population by nation and UK IMD decile (England model):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| IMD Decile distribution | E | N | S | W |
| 1 | 9% | 41% | 9% | 8% |
| 2 | 10% | 23% | 9% | 10% |
| 3 | 10% | 16% | 8% | 10% |
| 4 | 10% | 10% | 8% | 12% |
| 5 | 10% | 7% | 9% | 13% |
| 6 | 10% | 2% | 11% | 14% |
| 7 | 10% | 0% | 11% | 11% |
| 8 | 10% | 0% | 11% | 10% |
| 9 | 10% | 0% | 11% | 9% |
| 10 | 10% | 0% | 15% | 3% |



# National composition of each decile

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| nation | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| England | 77% | 82% | 84% | 85% | 85% | 84% | 86% | 87% | 87% | 86% |
| Northern Ireland | 11% | 7% | 5% | 3% | 2% | 1% | 0% | 0% | 0% | 0% |
| Scotland | 7% | 7% | 6% | 6% | 7% | 9% | 9% | 9% | 9% | 12% |
| Wales | 4% | 5% | 5% | 6% | 6% | 7% | 5% | 5% | 4% | 1% |



As in the original paper, England and Scotland have a similar distribution, Wales is on average slightly more deprived, and Northern Ireland is mostly contained in the bottom half of the overall deprivation scale. For comparsion to the previous model, 37% of NI population were in the bottom overall quintile compared to 27% in this model.

The GB distribution (generated using Income and Employment) is similar to the UK distributions:

|  |  |  |  |
| --- | --- | --- | --- |
| IMD Decile distribution | E | S | W |
| 1 | 10% | 10% | 12% |
| 2 | 10% | 9% | 12% |
| 3 | 10% | 8% | 12% |
| 4 | 10% | 8% | 13% |
| 5 | 10% | 9% | 13% |
| 6 | 10% | 11% | 12% |
| 7 | 10% | 11% | 10% |
| 8 | 10% | 10% | 8% |
| 9 | 10% | 11% | 7% |
| 10 | 10% | 14% | 1% |

The concentration of NI SOA in the lower deciles in the UK model means less of other nations appear in this decile. When comparing datasets without NI data, using the GB index will present a cleaner picture between deciles. Although if part of the overall picture is *"we don't have data/have good takeup of services in a part of the UK that is on average more deprived"*, the UK index may still be appropriate.

### Comparison of different national base population deciles

The different processes for different national base files change rankings to the extent that decile divisions are noticeably different between the different files. The table below shows that while there is a high (96%) agreement between the England and Wales deciles, the Wales and Northern Ireland dataset only agree in 65% of cases. This reinforces the need for choosing a base dataset that best reflects the largest set of your data, where the rankings are most theoretically valid.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Nation | E | W | S | N |
| E | - | 96.1% | 84.6% | 61.6% |
| W | 96.1% | - | 80.8% | 64.5% |
| S | 84.6% | 80.8% | - | 52.2% |
| N | 61.6% | 64.5% | 52.2% | - |

## Data sources

* [English Index of Multiple Deprivation 2019](https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019)
* [Scottish Index of Multiple Deprivation 2020](https://www.gov.scot/collections/scottish-index-of-multiple-deprivation-2020/)
* [Welsh Index of Multiple Deprivation (WIMD) 2019](https://gov.wales/welsh-index-multiple-deprivation-full-index-update-ranks-2019)
* [Northern Ireland Multiple Deprivation Measures (NIMDM) 2017](https://www.nisra.gov.uk/statistics/deprivation/northern-ireland-multiple-deprivation-measure-2017-nimdm2017)
* [ONS 2019 population estimates](https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/lowersuperoutputareamidyearpopulationestimates)
* [National Records Scotland 2019 Population Estimates](https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/population/population-estimates)
* [NISRA 2019 2019 population estimates](https://www.nisra.gov.uk/publications/2019-mid-year-population-estimates-northern-ireland)

For WIMD the domain scores in the domains file for Income and Employment are scaled, and you need to use the original single indicator from the stats wales site instead.

The SIMD and NIMDM datasets do not contain the raw score for the combined index. This was regenerated from the domain indexes (create\_missing\_overall\_scores.py) using the following procedure:

scaled\_rank <- reversed\_series/len(series) # scale to [0,1 - 1 is most deprived]  
transformed\_score <- -23\*ln(1 - scaled\_rank\*(1 - exp(-100/23)))

This is then weighted by the correct weight to combine into the complete domain score. Comparing the ranks of this score against the published rank generally agrees, with a number of 1-2 rank differences where there is likely an undocumented rounding step in the original composite process.

## National domain weights

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Domain weights | NI | E | S | W |
| Income | 25 | 22.5 | 28 | 22 |
| Employment | 25 | 22.5 | 28 | 22 |
| Health deprivation | 15 | 13.5 | 14 | 15 |
| Education, skills and training | 15 | 13.5 | 14 | 14 |
| Crime and disorder / Community Safety | 5 | 9.3 | 5 | 5 |
| Barriers to housing and services | - | 9.3 | - |  |
| Access to services / Geographic access to services | 10 | - | 9 | 10 |
| Living Environment | 5 | 9.3 | - | - |
| Physical Environment | - | - | - | 5 |
| Housing | - | - | 2 | 7 |

## Differences between different domains

All four indexes have measures for Income, Employment, Health, Education, and Crime. Income and Employment use roughly equivalent measures in most cases, and the score for these domains represents a real percentile rather than a composite score. Northern Ireland changed its measure in 2017 to make it a proportion of housholds below 60% of NI median rather than a % of households recieving some form of income-assistance benefit like the other nations and so is no longer directly comparable.

All indexes incorporate some measure of difficulty accessing services, generally the physical distance from an area to GPs, shops, etc. This is a domain level score in all but England, where it is one sub-domain in the 'Barriers to housing and services' domain. Wales, Scotland and Northern Ireland incorporate measures of broadband speed into their access to services score. England does not.

The quality of the physical environment is only a domain level measure in Wales. It is one of two sub-domains of Living Enviornment in England and one of three sub-domains of Living Environment in Northern Ireland. Scotland does not include an equivalent domain or sub-domain.

Aspects of housing appear in different domains in different nations. The 'housing' domain in Scotland and Wales includes housing quality and overcrowding. The equivalent overcrowding measure is includes in one of two sub-domains of Living Enviornment in England and one of three sub-domains of Living Environment in Northern Ireland. England uniquely includes a measure of affordability as part of the access to housing subdomain.

### Benefits used to generate Income and Employment domains

#### Income

|  |  |  |  |
| --- | --- | --- | --- |
| Benefit | England | Scotland | Wales |
| Jobseeker’s Allowance | yes | yes | yes |
| Employment and Support Allowance | yes | yes | yes |
| Pension Credit | yes | yes | yes |
| Universal Credit (excluding 'Working - no requirements') | yes | yes | yes |
| Working Tax Credit | yes | yes | yes |
| Child Tax Credit | yes | yes | yes |
| Supported Asylum seekers | yes |  | yes |
| Income Support | yes | yes | yes |

#### Employment

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Benefit | England | Scotland | Wales | Northern Ireland |
| Jobseeker’s Allowance (contrib + income) | yes | yes | yes | yes |
| Employment and Support Allowance (contrib + income) | yes | yes | yes | yes |
| Incapacity Benefit | yes | yes | yes | yes |
| Carer’s Allowance | yes |  |  | yes |
| Severe Disablement Allowance |  | yes |  | yes |
| Universal Credit - 'Searching for work', 'No work requirements' | yes | yes | yes |  |

## Generating index

To recreate the indexes requires python (3.8+) and pipenv:

pipenv install  
pipenv run python create\_missing\_overall\_scores.py  
pipenv run python merge\_indexes.py

Alternatively, the requirements.txt can be used without pipenv.

### Maintenance

The column names and sources files used can be adjusted in country\_indexes\join\_description.yaml.

If you've added or changed the packages used, run the following to update the requirements.txt from pipenv:

pipenv run pipenv\_to\_requirements

The readme file should not be edited directly, instead edit the jupyter notebook and rerun the final cell to update.

## Licencing

Source indexes for each country are licenced as Open Government License (OGL) version 3.0.

Unless otherwise stated, composite datasets are licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

Code and scripts are licenced under a MIT Licence.

## Citation

Future versions may use same URL, but past versions will be included in git history.

Suggested citation if using the dataset, but worth also referencing the original methodology paper:

Parsons, Alex (2021), UK 2020 Composite Index of Multiple Deprivation, https://github.com/mysociety/composite\_uk\_imd