Institute for Environmental Analytics

Technical Report

Auxiliary data for the UK hi-res climate data format of the Earth Observation Climate Information System (EOCIS)

7 March 2024

V8.0

Technical Report

Auxiliary data for the UK hi-res climate data format of the Earth Observation Climate Information System (EOCIS)

Authors: Guy Griffiths, Mathieu Roesch, Ben Lloyd-Hughes, Maria Noguer

Institute for Environmental Analytics (IEA)
Philip Lyle Building, University of Reading
Whiteknights Campus,
Reading RG6 6BX
United Kingdom

Reviewers: Chris Merchant and Niall McCarroll

Department of Meteorology,
University of Reading

Version History

Version	Reason for Issue	Status	Date of Issue
V1.0	Internal IEA draft (structure)	Draft	Sep 2023
V2.0	Internal IEA draft (initial auxiliary files)	Draft	End of Oct 2023
V3.0	Internal Draft for PM review	Draft	8 Nov 2023
V4.0	Draft for EOCIS team review	Draft	10 November 2023
V5.0	Final Draft for IEA PM review	Draft	27 November 2023
V6.0	Final Draft for EOCIS team review ahead of acceptance by EOCIS and submission of Final Report	Final Draft	29 November 2023
Final	Addressing comments from EOSIS	Final	30 November 2023
V8.0	Draft for EOCIS team review (addition of D.12 and D.13 (a, b, c))	Draft	7 March 2024

Contents

1. Introduction

The Earth Observation Climate Information System (EOCIS) is a collaborative project that brings together UK research-community expertise to create and make available high quality, trustworthy climate information based on measurements of Earth's environments from space. EOCIS is led by the National Centre of Earth Observation (NCEO) and is funded by the National Environment Research Council, part of UK Research and Innovation, with the funding provided by the Department for Science, Innovation and Technology.

EOCIS is transforming climate data into higher-level climate information that advances both climate science and informed practical action in response to the challenge of climate change.

To support this, the Institute for Environmental Analytics (IEA) was contracted to create categorical geographical information data files for the UK to enable the effective translation of climate data into new forms of actionable information.

The auxiliary datasets that the IEA have created are:

- [D1.1] A classification variable distinguishing land and permanent water
- [D1.2] A classification variable that identifies every grid cell with tags for the devolved nation of the UK (also Eire, France, etc)
- [D1.3] A classification variable that identifies every grid cell with tags for the county / council / unitary authority / metropolitan or London borough
- [D1.4] A classification variable that identifies every grid cell with tags for the parish / community / town council
- [D1.5] A classification variable that identifies every grid cell with tags for The UK postcode sector
- [D1.6] A classification variable that identifies every grid cell with tags for appropriate administrative boundaries relating to the National Health Service
- [D1.7] A classification variable that identifies every grid cell with tags appropriate administrative boundaries relating to the Fire Service
- [D1.8] The dominant land classification from the Centre for Ecology and Hydrology 2021
- [D1.9] The nearest valid UN-system land classification from the ESA Land Cover CCI medium-resolution land cover product, for each available year (2001 to 2020)
- [D1.10] The built and paved area fractions for a recent year or years
- [D1.11] Presence of roads, railway tracks and transmission network
- [D1.12] Elevation and slope
- [D1.13] Socioeconomic data of population, income, and educational attainment

These datasets have been created following the specific format and nature of the Climate information at Hi-res for the UK (CHUK) grid, as specified by NCEO.

This Technical Report is divided into sections according to each of the auxiliary files created. A section defining some of the terminology is also included.

.

2. Terminology

General Terms:

- CHUK - Climate information at Hi-res for the UK
- Centroid - The geometric centre of a polygon, computed mathematically from the locations of all the vertices defining the polygon. Sometimes centroids are estimated visually.
- Office for National Statistics (ONS) - The Office for National Statistics (ONS) is responsible for producing a wide range of economic and social statistics. It also carries out the Census of Population for England and Wales
- Ordnance Survey - Ordnance Survey is the national mapping agency for Great Britain. Northern Ireland is responsible for producing its own mapping. The Land & Property Services (LPS) in Northern Ireland has incorporated the Ordnance Survey of Northern Ireland (OSNI) into its organisation and produces mapping with the OSNI branding.

Terms related to D1.3 (county / council / unitary authority)

- Counties - Counties were formerly administrative units across the whole UK. However, due to various administrative restructurings, the only administrative areas still referred to as counties are the non-metropolitan (shire) counties of England.
- Councils - In Scotland the counties and burghs were replaced by the two-tier local government system of Regional councils and District councils, which in turn were replaced in 1996 by the current single-tier system of 32 councils.
- Local Authority (LA) - Local Authority (LA) is a generic term for any level of local government in the UK. In geographic terms LAs therefore include English counties, non-metropolitan districts, metropolitan districts, unitary authorities and London boroughs; Welsh unitary authorities; Scottish council areas; and Northern Irish district council areas.
- Unitary Authority (UA) - Unitary Authorities (UAs) are areas with a single tier of local government. In practice, the term is usually only applied to the 22 UAs established across the whole of Wales in 1996 and, the 56 UAs established in parts of England between 1995 and 2009. However, the London boroughs and metropolitan districts in England, and the 32 council areas in Scotland and district council areas in Northern Ireland are all also served by single-tier (unitary) administrations.

Terms related to D1.4 (parish, community, etc)

- Civil parish - In England, a civil parish is a type of administrative parish used for local government. It is a territorial designation which is the lowest tier of local government.

- Non-civil parish – In England, an unparished area is an area that is not covered by a civil parish. Most urbanised districts of England are either entirely or partly unparished. Many towns and some cities in otherwise rural districts are also unparished areas and therefore no longer have a town council or city council, and are instead directly managed by a higher local authority such as a district or county council
- Community council areas - Community (Welsh: *cymuned*) is a division of land in Wales that forms the lowest tier of local government in Wales. Welsh communities are analogous to civil parishes in England. There are 878 communities in Wales.
- Community Councils - In Scotland community councils are subdivisions of council areas. These are the Scottish equivalent of English civil parishes.
- Wards – Wards are the smallest administrative unit in Northern Ireland. Wards are used to create constituencies for local government authorities, the Northern Ireland Assembly and the House of Commons of the United Kingdom.

Terms related to D1.6 (National Health Services areas):

- England – NHS England regions – There are 7 regions (East England, London, Midlands, North East and Yorkshire, North West, South East and South West. They support the development of integrated care boards.
    - Integrated care boards in England - A statutory NHS organisation responsible for developing a plan for meeting the health needs of the population, managing the NHS budget and arranging for the provision of health services in the Integrated Care System area
- Scotland - NHS Health Board Areas (HBAs) (Scotland) - Delivery of frontline healthcare services in Scotland are the responsibility of 14 regional National Health Service (NHS) Boards that report to the Scottish Government
- Wales - Local Health Boards (Wales) - Boundaries of the NHS Local health Boards in Wales. The seven Local Health Boards (LHBs) in Wales now plan, secure and deliver healthcare services in their areas, replacing the 22 LHBs and the 7 NHS Trusts which together performed these functions in the past.
- Northern Ireland - Department of Health (DoH) Trust Boundaries - There are a total of 5 Health and Social Care Trusts (HSCT) in Northern Ireland; Belfast, Northern, Southern, South Eastern and Western. Each HSCT is managed directly by a board of directors which has corporate responsibility for its operation. The trusts are responsible for the delivery of responsive and effective health and care services and for the ownership and management of hospitals and other establishments and facilities. This health trust boundary dataset is derived from Ordnance Survey Northern Ireland (OSNI) 1993 local government district boundary dataset.

3. Common data elements and processing

3.1. Processing

The CHUK grid consists of a 100m x 100m grid over the whole of the British Isles, an area approximately 1,000km x 1,500km. This gives a total of ~150 million grid cells which each need to be processed, and for many of the datasets, this is not feasible in a timely manner on a standard desktop machine. Fortunately, the algorithms we are using treat each cell independently, and so each job can be split up and run in parallel.

This is the approach we took when generating every dataset. The basic procedure was (i) to split the input CHUK grid along the y axis (this is the most efficient way given the storage order on disk), (ii) process each chunk separately, and then (iii) re-join the output files along the y-axis into a single output dataset. The exact number of splits depended on the algorithm being used – in general we used 48 splits to run on our 48-core machine, but there were some algorithms which became more efficient with smaller input grids, in which cases as many as 2-3,000 splits (5-8 y-lines per run) were used.

3.2. Metadata

For every variable, we saved the following attributes:

- `grid_mapping`: This is used to relate the x/y and lat/lon co-ordinates to concrete Coordinate Reference System (CRS) definitions
- `coordinates`: This specifies that the 2d co-ordinate variables lat and lon, should be used to reference real-world co-ordinates
- `standard_name`: The closest standard_name from the CF[1] standard-name table, or an appropriate suggestion if no close matches exist
- `long_name`: A longer, more descriptive name of the quantity being described
- `source`: A very brief description of the source data and algorithm used to generate this variable
- `flag_values/flag_meanings` (where appropriate): For datasets which are categorising something into discrete classes (e.g. countries, parishes, postcodes, etc.), these list all possible numerical values of the variable, and the category names which correspond to each of those values respectively. For example, for countries:
    - `flag_values = 0s, 1s, 2s, 3s, 4s, 5s, 6s, 7s, 8s, 9s ;`

[1] CF – Climate and Forecast (CF) metadata conventions are conventions for the description of Earth sciences data, intended to promote the processing and sharing of data files.

    - `flag_meanings = "None England Scotland Wales Northern_Ireland Eire France Belgium Isle_of_Man Channel_Islands " ;`

- `_FillValue` (where appropriate): For datasets where there is “no data” value which specifically corresponds to “data lacking here” (as opposed to “positive determination of a data value which is outside our interesting categories”) a _FillValue attribute is used to mark the absence of data.

On a per-dataset basis, we saved the following attributes, with corresponding values:

- `title`: "EOCIS Auxiliary Data - Dataset name"
- `summary`: <a brief description of what this dataset provides>
- `uuid`: A UUID5 string generated from the OID namespace combined with the dataset title (as above), and version.
- `institution`: "EOCIS CHUK"
- `date_created`: The date the dataset was finalised at this version.
- `version`: "1.0"
- `license`: "Creative Commons Licence by attribution (https://creativecommons.org/licenses/by/4.0/)"
- `comment`: "Technical documentation describing how this data can be found in the accompanying technical documentation, which should have been provided with this data. A copy of the technical documentation can be found at https://eocis.org/"
- `spatial_resolution`: "100 m"
- `Convention`: "CF-1.10"
- `creator_url`: "https://the-iea.org"
- `creator_name`: “The Institute for Environmental Analytics"
- `creator_email`: “tech@the-iea.org"
- `creator_processing_institution`: “University of Reading"
- `publisher_url`: “https://eocis.org"
- `publisher_name`: “EOCIS"
- `publisher_email`: “EOCIS@reading.ac.uk"
- `acknowledgement`: “Funded by the Natural Environment Research Council [NERC grant reference number NE/X019071/1, “UK EO Climate Information Service”]"

4. Land / water classification [D1.1]

4.1. Requirement

To determine, for each cell of the CHUK grid, whether that cell represents:

- Permanent land
- Permanent freshwater
- Permanent saltwater
- Mixed, mostly land
- Mixed, mostly water, which is primarily freshwater
- Mixed, mostly water, which is primarily saltwater

Additionally, we need to determine, for every cell which is categorised as permanent freshwater, whether that cell is part of a lake, and whether a river passes through the cell.

4.2. Data source

- GB land cover, 10m - https://catalogue.ceh.ac.uk/documents/a22baa7c-5809-4a02-87e0-3cf87d4e223a
- NI land cover, 10m - https://catalogue.ceh.ac.uk/documents/e44ae9bd-fa32-4aab-9524-fbb11d34a20a
- UKCEH (UK Centre for Ecology and Hydrology) Lakes database - https://data.catchmentbasedapproach.org/datasets/theriverstrust::uk-lakes-lakes-portal-ukceh/explore

Note that the two land cover data sources for this are provided in separate files, on different grids. To combine the two datasets into a single input dataset, GDAL[2] was used with the following commands:

- `gdalwarp -s_srs EPSG:29903 -t_srs EPSG:27700 nilcm10m2021.tif ni-gbgrid.tif`
- `gdal_merge.py -o uklcm10m2021.tif -n 255 -a_nodata 255 gblcm10m2021.tif ni-gbgrid.tif`

The first command reprojects the Northern Irish data onto the OSGB grid, and the second command merges the two files into a single UK land cover dataset, making sure to use the value 255 for no data, since this is expected from both files. This same dataset is used for deliverable

[2] GDAL is a translator library for raster and vector geospatial data formats.

Urban / suburban areas, and a similar process is used for the data for Land classification (albeit with a different no data value).

The output file, `uklcm10m2021.tif`, is delivered as part of the input datasets, but this process can be used to create new input data when future UKCEH products are released.

The UKCEH lakes database was downloaded from the above link and reprojected into EPSG:4326 using QGIS. No further modification was made to the dataset.

The creation of this dataset also requires the use of dataset *Devolved nations of the UK*, to determine the presence of oceans, as described below.

4.3. Auxiliary file creation

For this deliverable, 3 variables were created in separate datasets and then merged into a single NetCDF file using a simple python/xarray script. Here we describe the process for creating each variable.

4.3.1. Land-Water mask

The algorithm for determining whether a CHUK grid cell is land or water was as follows:

- If the cell is entirely outside the bounds of the UKCEH grid, we test the same cell in dataset *Devolved nations of the UK*. If it falls within a country, it is assumed to be permanent land. If not, it is assumed to be permanent saltwater (ocean).
- Otherwise, we read the 10x10 pixels in the UKCEH grid which fall within this CHUK cell, and take the following steps:
    - If all pixels are 255 (no data) and dataset *Devolved nations of the UK* shows no data, we classify as “permanent saltwater” (i.e. ocean).
    - We count the number of pixels in the UKCEH sample in each of the categories of freshwater (14), saltwater/no data (13 or 255), and other land data (any other classification). Saltwater is conflated with no data, because in the vast majority of cases this is what it represents (since the UKCEH data is clipped to the land)
    - If there are no water pixels and some land pixels, we classify as “permanent land”.
    - If there are no land pixels, we classify as either “permanent freshwater” or “permanent saltwater”, depending on which has the majority (freshwater wins in a tie).
    - If there are as many or more land pixels than water, we classify as “mixed, mostly land”.
    - If there are more water pixels than land, we classify as either “mixed, mostly freshwater” or “mixed, mostly saltwater”, depending on which has the majority (freshwater wins in a tie).

4.3.2. Lake flag

To set the lake flag, the following process was undertaken:

- For every CHUK cell, take the above-calculated value, and if (and only if) it is classed as "permanent freshwater", test the centroid of the cell to see if it falls within any of the lakes in the UKCEH lakes dataset. If so it is flagged as "lake", otherwise it is flagged as "no data".

4.3.3. River flag

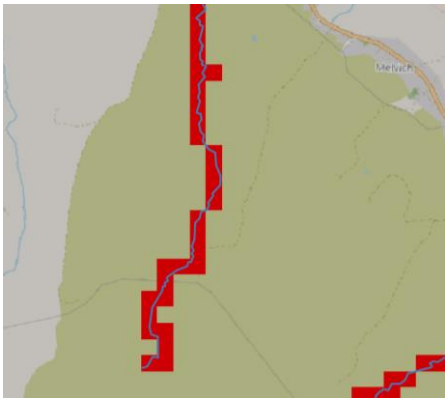
Data sources:

- GB: https://www.ordnancesurvey.co.uk/products/os-open-rivers#:~:text=A%20free%20dataset%20showing%20the,rivers%2C%20tidal%20estuaries%20and%20canals.
- NI: https://land.copernicus.eu/en/products/eu-hydro

The core algorithm is shared with other auxiliary files based on linear features:

- It loads the CHUK grid. In reality, this is a subregion of the grid, as described in the Processing section. This is particularly important since this algorithm's efficiency increases dramatically with smaller subregions.
- A shapely Polygon is created from the subregion and a function is used to get the list of geometries. This function excludes any geometries shorter than 2.5m (for 100m grid), and keeps any that intersects or is covered by the Polygon (by doing so, we reduce the total number of geometries to test each cell against – this is where the efficiency increases come from).
- At this point we have a subregion CHUK grid and a list of features that we know interact with the area covered by the grid. The script repeats then this process for each grid cell:
    - Create a shapely Polygon from the cell and test each geometry against it:
        - If a geometry is completely covered by the cell, the cell is flagged. A flag value of 0 means there is no rivers passing/covered by this cell, a value of 1 means there is at least one.
        - If a geometry intersects with the Polygon perimeter, it breaks the line into segments and test if at least one of them is covered by (inside) the cell and that this segment length is > 2.5m. If so, the cell is flagged.

An image using a sample from the 100m result is displayed bellow. The blue lines are rivers, the red squares are 100m cells. If you look at the river in the picture below, you can see it appears to jump diagonally from one cell to another, because the length of river in the intermediate cell is less than 2.5m, meaning the cell isn't flagged, which is the expected behaviour.



4.4. Quality control

Once created, the dataset was loaded into QGIS alongside the UKCEH land cover and the lakes dataset. Visual inspection of multiple lakes was undertaken to ensure that the resulting EOCIS data was consistent with the input datasets, and this was found to be the case.

4.5. Maintenance

This dataset's validity depends on the validity of the UKCEH 2021 land cover and lakes products. In case of an update to UKCEH, this dataset can be updated by following the data generation steps outlined above and re-running the appropriate scripts. The OS Data for rivers is updated every 6 months.

5. Devolved nations of the UK [D1.2]

5.1. Requirement

To determine, for every cell in the CHUK grid, whether that grid cell belongs to:

- England
- Scotland
- Wales
- Northern Ireland
- Eire
- France
- Belgium
- Isle of Man
- Channel Islands
- Outside country borders

5.2. Data source

- GB country boundaries - https://geoportal.statistics.gov.uk/datasets/ons::countries-december-2022-boundaries-gb-bfe/explore
- NI boundary - https://www.data.gov.uk/dataset/d3ca9d44-a7eb-4380-86cb-0cc28e1f1b27/osni-open-data-largescale-boundaries-ni-outline
- Ireland boundary - https://data.gov.ie/dataset/province-boundary/resource/ebe32ca4-b837-4ae4-8338-4d9d9c927c9d
- France boundary - https://maps.princeton.edu/catalog/stanford-dw125xh0996
- Belgium boundary - https://maps.princeton.edu/catalog/stanford-hs337qd4914
- Isle of Man / Channel Islands boundaries - https://www.data.gov.uk/dataset/d310b2c5-5253-4bc2-a78d-f8240293119d/boundary-line (note that this is actually a dataset of all polling regions of the UK and dependencies, but we exclude all but the IoM and Channel Islands from it as part of the processing)

The six datasets above were loaded into QGIS for processing into a single shapefile which can be used to determine the country of each CHUK cell. The following changes/modifications to each layer were made before combining into a single shapefile:

- Conversion to EPSG:4326
- Setting the country name attribute to match that of the GB boundaries – "CTRY22NM"
- Accurate Ireland boundaries were not available. Instead, the 4 provinces of Ireland were used from the above source and combined into a single Eire entity, using QGIS's merge tool. It was also a few metres off (probably due to co-ordinate conversion) and was moved manually to match the NI boundary as well as possible. Note that there was still a small gap along the border, but this is accounted for by the processing algorithm.

The resulting shapefile, `d1.2-countries.shp`, is delivered as part of the input datasets, and should only need regenerating in the event of country borders changing. Countries within the shapefile are ordered as "England, Scotland, Wales, Northern Ireland, Eire, France, Belgium, Isle of Man, Channel Islands" and this order will be followed in the assignation of the `flag_values` attribute.

5.3. Auxiliary file creation

It is worth noting that the ONS data is clipped to the extent of the realm, which means there are a few oddities in the shape of the country, most notably in the Bristol Channel. Generally speaking these take the form of protrusions into the surrounding ocean.

The boundary protrusion in the Bristol Channel

To remove these discrepancies, we follow an iterative process of:

- Generate the countries on the CHUK grid, taking the ONS data as truth
- Generate dataset *Land classification*. Because D1.8 uses the UKCEH dataset, which is clipped to the coastline, this will classify such pixels as "no data". Most sea-pixels will be classified as "saltwater" by using this dataset (D1.2).
- Now re-generate the countries on the CHUK grid, taking anything which is a) within one of the countries of the UK and b) classed as "no data" in D1.8 to be outside the country boundaries.
- We now have a corrected version of this dataset, which can be used when generating all of the other datasets (including re-generation of D1.8 using this new country data).

Iterative process aside, the basic algorithm for generating this dataset is fairly simple:

- First, we step through every country in the shapefile, assigning a unique number to its name.
- Now for every CHUK cell, take the centroid, and test it against each of the country shapes. If it is present in one, store the unique number associated with that country in the data array, otherwise store the value for "None" (-1).
- If we have land cover data, and we are in England, Scotland, Wales, or Northern Ireland, and yet our land cover has the value of "no data", assume this is a discrepancy due to ONS boundaries, and mark as "no data".

However, near the Eire/NI border, we will end up with some erroneous "no data" pixels using this method, because the borders do not quite touch (see above). To remove these spurious pixels, we take the following approach:

- If a pixel is within the range [-8.2, -6] longitude and [54, 55.1] latitude and has the value "no data", check all 8 surrounding pixels. If more than 70% are in Eire or NI, this pixel is classed as "Eire" (since the Northern Ireland border is the more accurate).

5.4. Quality control

For the Irish border, every single pixel was checked visually to ensure that no "no data" pixels were missed by the corrective part of the algorithm.

Other unusual areas (e.g. the Bristol Channel protrusion) were checked visually, and the various boundaries were checked visually against the original vector boundary data to ensure that they matched correctly.

5.5. Maintenance

This dataset's validity will remain until any country boundaries change. Maintenance would involve updating the country shapefile with updated boundaries and re-running the data generation code. However, it is worth noting that this dataset is used in the generation of D1.1, D1.3, D1.4, D1.5, D1.6, D1.7, D1.8, and D1.10, and so if this changes, it will mean updating all auxiliary datasets.

6. County / Council / Unitary authorities [D1.3]

6.1. Requirement

To determine, for every CHUK grid cell, which county, or most appropriate equivalent, it lies within.

6.2. Data source

UK counties and national equivalents - https://geoportal.statistics.gov.uk/datasets/ons::counties-and-unitary-authorities-may-2023-boundaries-uk-bfe/explore

All required data was present in this single dataset. It was loaded into QGIS for initial visual inspection, and to convert to EPSG:4326, but was otherwise unchanged.

The resulting dataset is included in the input datasets.

6.3. Auxiliary file creation

This dataset is the first to follow a standard procedure which is also used for the following datasets:

- Parish / community / town councils
- Postcode sectors
- National Health Service boundaries
- Fire Service boundaries

In all cases, the aim was to test the centroid of each CHUK cell against a vector dataset of boundaries, to determine which region the cell is part of. The algorithm is described here for counties.

- First, loop through the list of all counties in the vector dataset, mapping their names to unique numbers, starting from 1. This allows us to generate the `flag_values` and `flag_meanings` attributes, and allows us to store the data numerically.
- At the same time, calculate the maximum length of county name. For convenience, we also want to store the county data as strings, so users who simply want to extract the county name do not have to perform any mapping between the `flag_values` and `flag_meanings` attributes. Since a 2-dimension array of strings in NetCDF is stored as a 3-dimensional array of characters, knowing the maximum name length allows us to store the data efficiently.

- Now every cell in the CHUK grid is iterated over, and
    - If the equivalent cell in the Devolved nations of the UK dataset is “no data”, set this cell as “no data”. Otherwise:
    - Test the centroid of the cell against each of the county definitions. As soon as it is found to be present in a county, the appropriate values are stored, and the algorithm moves to the next cell. If the cell is not determined to be in any of the county definitions, the “no data” value is stored instead.

6.4. Quality control

Note that there is a degree of QA built into the algorithm of this dataset – no county can appear outside of a country. However, in addition to this, and a standard visual inspection alongside the county definitions, we also check each cell to ensure that a single county does not appear in multiple countries.

6.5. Maintenance

The only maintenance this dataset will require is if the county boundaries change, or if the country definitions have changed. At which point the dataset should be regenerated.

7. Parish / community / town councils [D1.4]

7.1. Requirement

To determine, for every CHUK grid cell, which parish, or most appropriate equivalent, it lies within.

7.2. Data source

- England & Wales parishes and non-civil parish areas - https://geoportal.statistics.gov.uk/datasts/ons::parishes-and-non-civil-parished-areas-december-2022-ew-bfe-v3-2/explore
- Scottish community council boundaries - https://data.spatialhub.scot/dataset/community_council_boundaries-is
- NI electoral wards - https://www.data.gov.uk/dataset/a7802ada-9d95-4f37-bc40-1063b5bc4506/osni-open-data-largescale-boundaries-wards-2012

The 3 datasets above were loaded into QGIS to be converted to EPSG:4326 and merged together using the merge tool. The parish name attribute was renamed to "`parish_nam`" (10 char max limit in shapefiles, realised after editing) for consistency across the 3 datasets.

The resulting dataset is included in the input datasets.

7.3. Auxiliary file creation

The same method used for the *County / Council / Unitary authorities* was applied here. However, some key points to note:

- There are over 12,000 parishes in the dataset. This means that the `flag_values` and `flag_meanings` attributes are extremely long. Since these flags were never designed for human readability, this is not a problem.
- A larger issue with such a large number of parishes is that checking whether a cell centroid falls into any of them can be a very slow process. To get around this, some initial spatial indexing is performed. We divide the input CHUK grid into discrete boxes, and when initially checking through the parishes to map their names to numbers, we also determine which of these larger boxes each parish may fall into (with a simple bounding box check). This allows us to quickly discard the majority of parishes when checking each grid cell, and speeds the process up immensely.

7.4. Quality control

Note that there is a degree of QA built into the algorithm of this dataset – no parish can appear outside of a country. However, in addition to this, and a standard visual inspection alongside the county definitions, we also check each cell to ensure that a single parish does not appear in multiple counties.

7.5. Maintenance

The only maintenance this dataset will require is if the parish boundaries change, or if the country definitions have changed. At which point the dataset should be regenerated.

8. Postcode sectors [D1.5]

8.1. Requirement

To determine to which sector postcodes each CHUK grid cell belongs to.

8.2. Data source

We decided to use Doogal archive available here: https://www.doogal.co.uk/UKPostcodes

Note: we included the script we used to download the files in the code repository.

8.3. Auxiliary file creation

A brief note about the download. The script uses BeautifulSoup4 / bs4 and navigates through the website pages. It excludes the sector postcodes which are "no longer in use" but an additional check using QGIS was made - if QGIS can read the file and it has shapes and was inside the CHUK grid coverage, the .kml was kept.

It uses the same algorithm as previously described:

- There are over 10,000 sector postcodes in the dataset. The `flag_values` is the sector postcodes.
    - Note: the format of the value has space between area and sector, as described here: https://assets.publishing.service.gov.uk/media/5a81ebbded915d74e6234d42/Appendix_C_ILR_2017_to_2018_v1_Published_28April17.pdf
- Based on visual inspection, the dataset does not seem to contain any gap between sector (we are not expecting in-land cell without a postcode). However, some postcodes shapes overlap. Given that this is not a desired behaviour, and because we can't have the ground truth easily, we have a "first come first serve" rule meaning that the first postcodes that is detected to match with the cell, will give its value to the cell (the postcodes in the shapefile are stored by their "Name" in alphabetical order).

8.4. Quality control

We noticed that the boundaries of postcode sectors in the source data extended beyond the coastal boundaries. To mitigate this, we used the countries NetCDF that we have already created to mask this. The image below shows the source data sectors in grey and superimposed the postcodes sectors (various colours) from the final D1.5 NetCDF on the 100m CHUK grid.

8.5. Maintenance

Even though we know that the dataset has been updated in the past (due to the presence of "no longer in use" on page description), we don't know how often. Also, as discussed, we decided to use Doogal as an acceptable default but if other more reliable sources are available, we recommend using those instead.

9. National Health Service boundaries 9 [D1.6]

9.1. Requirement

To determine, for every CHUK grid cell, which NHS Health Board, or most appropriate equivalent, it lies within.

9.2. Data source

- England health boards - https://geoportal.statistics.gov.uk/datasets/ons::nhs-england-regions-july-2022-en-bfe-1/explore
- England integrated care boards - https://geoportal.statistics.gov.uk/datasets/ons::integrated-care-boards-april-2023-en-bfe-2/explore
- Wales health boards - https://geoportal.statistics.gov.uk/datasets/ons::local-health-boards-april-2022-wa-bfe-1/explore
- Scotland health boards - https://spatialdata.gov.scot/geonetwork/srv/api/records/f12c3826-4b4b-40e6-bf4f-77b9ed01dc14
- NI DoH trust boundaries - https://admin.opendatani.gov.uk/dataset/department-of-health-trust-boundaries/resource/645f8eef-8813-47a9-bb1e-a4932ada721a

All of these input datasets were processed to ensure it was in the EPSG:4326 projection. The integrated care board (ICB) dataset was left as-is, and the other 4 datasets were combined into a single shapefile using the QGIS merge feature. The health board name attribute was renamed to "`nhs_name`" for consistency.

Both of these datasets are delivered as part of the input datasets.

9.3. Auxiliary file creation

The same method used for the County / Council / Unitary authorities was applied here. The key difference is that two separate variables were defined – the health boundaries, and the integrated care boards (ICBs) for England-only. These were processed entirely separately, using the same algorithm, and then combined into a single file afterwards.

9.4. Quality control

Along with standard visual inspection, and the built-in QA ensuring that no values can appear outside of a defined country, we have run QA to ensure that no ICB is present in more than one health board.

9.5. Maintenance

The only maintenance this dataset will require is if the health board boundaries change, or if the country definitions have changed. At which point the dataset should be regenerated.

10. Fire Service boundaries [D1.7]

10.1. Requirement

To determine, for every CHUK grid cell, which Fire and Rescue Authority, or most appropriate equivalent, it lies within.

10.2. Data source

- England & Wales Fire and Rescue Authorities - https://geoportal.statistics.gov.uk/datasets/ons::fire-and-rescue-authorities-december-2022-ew-bfe/explore
- Scottish Fire and Rescue Services - https://spatialdata.gov.scot/geonetwork/srv/api/records/e7a1bb0c-68c0-46e4-a794-63280644d660
- NI – These is a single authority in Northern Ireland, and so we simply use the country definition from Devolved nations of the UK.

The above 3 datasets were loaded into QGIS for visual inspection, converted into EPSG:4326, and merged to a single layer. The name of the attribute was changed to "`fire_name`". The resulting dataset is delivered as part of the input datasets.

10.3. Auxiliary file creation

The same method used for *the County / Council / Unitary authorities* was applied here.

10.4. Quality control

Standard visual inspection, alongside the built in QA excluding values which are not within a defined country, were performed, but no additional QA was undertaken on this dataset.

10.5. Maintenance

The only maintenance this dataset will require is if the fire boundaries change, or if the country definitions have changed. At which point the dataset should be regenerated.

11. Land classification [D1.8]

11.1. Requirement

To determine, for each cell of the CHUK grid, what is the most likely land classification of that area.

11.2. Data source

- GB land cover, 25m - https://catalogue.ceh.ac.uk/documents/a1f85307-cad7-4e32-a445-84410efdfa70
- NI land cover, 25m - https://catalogue.ceh.ac.uk/documents/f3310fe1-a6ea-4cdd-b9f6-f7fc66e4652e

Note that the two land cover data sources for this are provided in separate files, on different grids. To combine the two datasets into a single input dataset, GDAL was used with the following commands:

- `gdalwarp -s_srs EPSG:29903 -t_srs EPSG:27700 nilcm25m2021.tif ni-gbgrid.tif`
- `gdal_merge.py -o uklcm25m2021.tif -n 0 -a_nodata 0 gblcm25m2021.tif ni-gbgrid.tif`

The first command reprojects the Northern Irish data onto the OSGB grid, and the second command merges the two files into a single UK land cover dataset, making sure to use the value 0 for nodata, since this is expected from both files. Note that this is different from the nodata value using in

12. Common data elements and processing

12.1. Processing

The CHUK grid consists of a 100m x 100m grid over the whole of the British Isles, an area approximately 1,000km x 1,500km. This gives a total of ~150 million grid cells which each need to be processed, and for many of the datasets, this is not feasible in a timely manner on a standard desktop machine. Fortunately, the algorithms we are using treat each cell independently, and so each job can be split up and run in parallel.

This is the approach we took when generating every dataset. The basic procedure was (i) to split the input CHUK grid along the y axis (this is the most efficient way given the storage order on disk), (ii) process each chunk separately, and then (iii) re-join the output files along the y-axis into a single output dataset. The exact number of splits depended on the algorithm being used – in general we used 48 splits to run on our 48-core machine, but there were some algorithms which became more efficient with smaller input grids, in which cases as many as 2-3,000 splits (5-8 y-lines per run) were used.

12.2. Metadata

For every variable, we saved the following attributes:

- `grid_mapping`: This is used to relate the x/y and lat/lon co-ordinates to concrete Coordinate Reference System (CRS) definitions
- `coordinates`: This specifies that the 2d co-ordinate variables lat and lon, should be used to reference real-world co-ordinates
- `standard_name`: The closest standard_name from the CF standard-name table, or an appropriate suggestion if no close matches exist
- `long_name`: A longer, more descriptive name of the quantity being described
- `source`: A very brief description of the source data and algorithm used to generate this variable
- `flag_values/flag_meanings` (where appropriate): For datasets which are categorising something into discrete classes (e.g. countries, parishes, postcodes, etc.), these list all possible numerical values of the variable, and the category names which correspond to each of those values respectively. For example, for countries:
    - `flag_values = 0s, 1s, 2s, 3s, 4s, 5s, 6s, 7s, 8s, 9s ;`
    - `flag_meanings = "None England Scotland Wales Northern_Ireland Eire France Belgium Isle_of_Man Channel_Islands " ;`
- `_FillValue` (where appropriate): For datasets where there is "no data" value which specifically corresponds to "data lacking here" (as opposed to "positive determination of

a data value which is outside our interesting categories") a _FillValue attribute is used to mark the absence of data.

On a per-dataset basis, we saved the following attributes, with corresponding values:

- `title`: "EOCIS Auxiliary Data - Dataset name"
- `summary`: <a brief description of what this dataset provides>
- `uuid`: A UUID5 string generated from the OID namespace combined with the dataset title (as above), and version.
- `institution`: "EOCIS CHUK"
- `date_created`: The date the dataset was finalised at this version.
- `version`: "1.0"
- `license`: "Creative Commons Licence by attribution (https://creativecommons.org/licenses/by/4.0/)"
- `comment`: "Technical documentation describing how this data can be found in the accompanying technical documentation, which should have been provided with this data. A copy of the technical documentation can be found at https://eocis.org/"
- `spatial_resolution`: "100 m"
- `Convention`: "CF-1.10"
- `creator_url`: "https://the-iea.org"
- `creator_name`: "The Institute for Environmental Analytics"
- `creator_email`: "tech@the-iea.org"
- `creator_processing_institution`: "University of Reading"
- `publisher_url`: "https://eocis.org"
- `publisher_name`: "EOCIS"
- `publisher_email`: "EOCIS@reading.ac.uk"
- `acknowledgement`: "Funded by the Natural Environment Research Council [NERC grant reference number NE/X019071/1, "UK EO Climate Information Service"]"

Land / water classification. This comes from the UKCEH data – we are not choosing a different value, we are choosing to maintain the no data value supplied by UKCEH in each case.

The output file, `uklcm25m2021.tif`, is delivered as part of the input datasets, but this process can be used to create new input data when future UKCEH products are released.

12.3. Auxiliary file creation

Because the CHUK grid was designed so that the corners coincide with points on the UKCEH grid, this dataset requires taking a majority vote of the 16 pixels of the UKCEH grid which fall into each CHUK cell to determine land classification. However, to obtain better results in the result of a tie, the algorithm is somewhat more complex:

- Read the 16 pixels of the UKCEH grid, and count how many there are of each value. If one pixel value (i.e. land cover type) has a majority, store this value in the data array and move onto the next CHUK cell.
- If there are multiple pixels with the same count, we calculate a secondary wider category for each pixel, such that the following land cover classes are grouped:
    - No data (0), Saltwater (13), Freshwater (14)
    - Deciduous woodland (1), Coniferous woodland(2)
    - Arable (3), Improved grassland (4), Neutral grassland (5), Calcareous grassland (6), Acid grassland (7), Fen (8), Heather (9), Heather grassland (10), Bog (11), Saltmarsh (19)
    - Inland rock (12), Supralittoral rock (15), Supralittoral sediment (18), Littoral rock (17), Littoral sediment (18)
    - Urban (20), Suburban (21)
- We then count the number of pixels in each of these groups. If one of these groups has a majority, we pick the class within the group which had the most pixels.
- If this does not resolve the tie, we use whichever tied land class pixel is closest to the start of the flattened array of 16 pixels. This will give preference to the North-Westernmost of the pixels, which avoids giving bias to a particular land class.
- If we have decided, based on a majority vote, that this pixel does not fall under one of the UKCEH land cover classes (i.e. it has a majority of no data (0) pixels), we test dataset *Devolved nations of the UK*. If the CHUK cell is outside all countries, mark it as saltwater. Otherwise it is a land pixel, but of an unknown type, and is marked as no data (-1).

12.4. Quality control

During the development of the algorithm, many individual tied cases were inspected manually to make sure they were behaving as expected. This has given us a great degree of confidence in the algorithm's behaviour for these edge cases.

Visual inspection of this dataset plotted against the UKCEH land cover product was also undertaken in multiple locations to ensure the correct functioning of the algorithm.

12.5. Maintenance

This dataset will only need updating when a new UKCEH land cover product is released, or if the country definitions change.

13. Land classification (2001 to 2020) [D1.9]

13.1. Requirement

To determine, for every cell in the CHUK grid, what European Space Agency (ESA) land cover class it belonged to for each of the years 2001 to 2020.

13.2. Data source

https://cds.climate.copernicus.eu/cdsapp#!/dataset/satellite-land-cover?tab=overview

For more details about the different class (flag values) between global and regional class, please check the Appendix A of the Product User Guide and Specification document for the source (https://datastore.copernicus-climate.eu/documents/satellite-land-cover/D5.3.1_PUGS_ICDR_LC_v2.1.x_PRODUCTS_v1.1.pdf)

In summary, the CCI-LC[3] maps are described by 22 classes, but they can also be described by a more detailed legend, called "level 2" or "regional". This level 2 legend makes use of more accurate and regional information – where available – to define more LCCS classifiers and so to reach a higher level of detail in the legend. This regional legend has therefore more classes which are listed in Appendix A. Global classes are numbered as multiples of 10, and associated regional classes are integer values within that bound of 10. For example, the global class "sparse vegetation" is assigned value 150, and the regional classes of "sparse tree" and "sparse shrub" are 151, and 152 respectively. Regional classes are not present all over the world since they were not properly discriminated at the global scale, however there are regional classes present in the area covered by the CHUK grid.

13.3. Auxiliary file creation

The download was made using the Climate Data Store (CDS) web GUI. The algorithm behaves as follow:

- Load the CHUK grid file and create a spatial subregion.
- Use this subregion with buffer on all sides (so we don't lose potential information on the sides) to subset the Land Classification file – LC_subset.
- Create a numpy array with extreme values with the same shape as the subregion.
- Loops for each point of the CHUK subregion:
    - For each 4 corners of the cell, pick the nearest points in LC_subset to get all cells of interest. This will return anything from 1 to 4 unique cells. They do not need

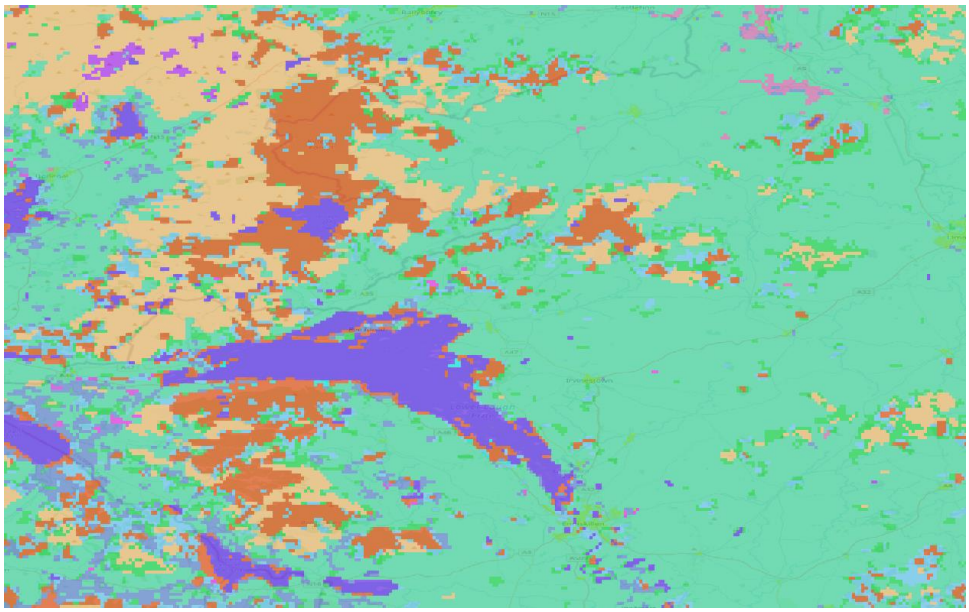
[3] CCI-LC: Climate Change Initiative – Land Cover

to be distinguished, and the algorithm works perfectly well when the same cell is selected multiple times. There are now three potential resolutions:

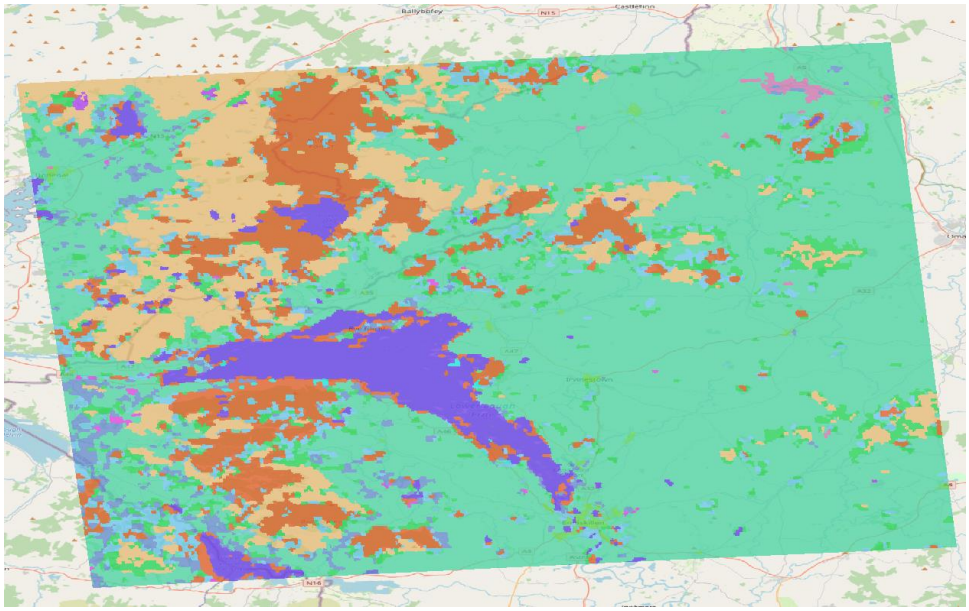
- Option A: if the number of unique values is equal to 1 (i.e., if all the points in the pool have the same value/code), the CHUK cell will get this code.
- Option B: if all values of LC have the same count (i.e. 4 different LC values, or 2 pairs of matching LC values), it will check the distances between the CHUK cell and each of the nearest points from the ESA grid. The closest ESA cell centroid to the CHUK cell centroid is given priority.
- Option C: if Option A and Option B didn't happen (else condition), we know we can find a winner in the pool so the value/code that appears the most often will be given to the CHUK cell.

13.4. Quality control

A visual inspection comparing the LC 300m input dataset:



with the output over the same area on the 100m CHUK grid:



shows that the two datasets are equivalent. Close manual inspection and comparison of a number of smaller areas was performed.

13.5. Maintenance

Unfortunately, the dataset seems to be discontinued as the last year available is 2020 and to our knowledge this won't change.

14. Urban / suburban areas [D1.10]

14.1. Requirement

To determine, for every CHUK grid cell, what fraction of that cell is covered by suburban and urban land.

14.2. Data source

The UKCEH 10m land cover product was used for this dataset, exactly as described in

Common data elements and processing

14.3. Processing

The CHUK grid consists of a 100m x 100m grid over the whole of the British Isles, an area approximately 1,000km x 1,500km. This gives a total of ~150 million grid cells which each need to be processed, and for many of the datasets, this is not feasible in a timely manner on a standard desktop machine. Fortunately, the algorithms we are using treat each cell independently, and so each job can be split up and run in parallel.

This is the approach we took when generating every dataset. The basic procedure was (i) to split the input CHUK grid along the y axis (this is the most efficient way given the storage order on disk), (ii) process each chunk separately, and then (iii) re-join the output files along the y-axis into a single output dataset. The exact number of splits depended on the algorithm being used – in general we used 48 splits to run on our 48-core machine, but there were some algorithms which became more efficient with smaller input grids, in which cases as many as 2-3,000 splits (5-8 y-lines per run) were used.

14.4. Metadata

For every variable, we saved the following attributes:

- `grid_mapping`: This is used to relate the x/y and lat/lon co-ordinates to concrete Coordinate Reference System (CRS) definitions
- `coordinates`: This specifies that the 2d co-ordinate variables lat and lon, should be used to reference real-world co-ordinates
- `standard_name`: The closest standard_name from the CF standard-name table, or an appropriate suggestion if no close matches exist
- `long_name`: A longer, more descriptive name of the quantity being described
- `source`: A very brief description of the source data and algorithm used to generate this variable
- `flag_values/flag_meanings` (where appropriate): For datasets which are categorising something into discrete classes (e.g. countries, parishes, postcodes, etc.), these list all possible numerical values of the variable, and the category names which correspond to each of those values respectively. For example, for countries:
    - `flag_values = 0s, 1s, 2s, 3s, 4s, 5s, 6s, 7s, 8s, 9s ;`
    - `flag_meanings = "None England Scotland Wales Northern_Ireland Eire France Belgium Isle_of_Man Channel_Islands " ;`
- `_FillValue` (where appropriate): For datasets where there is "no data" value which specifically corresponds to "data lacking here" (as opposed to "positive determination of a data value which is outside our interesting categories") a _FillValue attribute is used to mark the absence of data.

On a per-dataset basis, we saved the following attributes, with corresponding values:

- `title`: "EOCIS Auxiliary Data - Dataset name"
- `summary`: <a brief description of what this dataset provides>
- `uuid`: A UUID5 string generated from the OID namespace combined with the dataset title (as above), and version.
- `institution`: "EOCIS CHUK"
- `date_created`: The date the dataset was finalised at this version.
- `version`: "1.0"
- `license`: "Creative Commons Licence by attribution (https://creativecommons.org/licenses/by/4.0/)"
- `comment`: "Technical documentation describing how this data can be found in the accompanying technical documentation, which should have been provided with this data. A copy of the technical documentation can be found at https://eocis.org/"
- `spatial_resolution`: "100 m"
- `Convention`: "CF-1.10"
- `creator_url`: "https://the-iea.org"
- `creator_name`: “The Institute for Environmental Analytics"
- `creator_email`: “tech@the-iea.org"
- `creator_processing_institution`: “University of Reading"
- `publisher_url`: “https://eocis.org"
- `publisher_name`: “EOCIS"
- `publisher_email`: “EOCIS@reading.ac.uk"
- `acknowledgement`: “Funded by the Natural Environment Research Council [NERC grant reference number NE/X019071/1, “UK EO Climate Information Service”]"

Land / water classification

14.5. Auxiliary file creation

The algorithm used to generate this dataset was straightforward. For each CHUK grid cell:

- If this CHUK cell reads “no data” in Devolved nations of the UK, mark both urban and suburban values as zero and move on.
- Read the 10x10 pixels of the UKCEH data which fall into the CHUK grid cell.
- Sum up the pixels classified as urban (20), suburban (21), and “not no data” (i.e. 1-19)
- Store the fraction of urban and suburban in each cell, defined as the number of (sub)urban pixels divided by the total number of data pixels.

14.6. Quality control

Standard visual inspection, alongside the built in QA excluding values which are not within a defined country, were performed, but no additional QA was undertaken on this dataset.

14.7. Maintenance

This dataset will remain valid for as long as the 2021 UKCEH land cover values are considered valid. If they change, this dataset should be recalculated. Note though that there is a good chance that output data will remain the same, since land cover changes are often small and incremental, and may not be picked up by the majority resampling we are using.

15. Roads [D.11a]

15.1. Requirement

To flag each CHUK cell with a value determining which type of road that is passing inside it.

15.2. Data source

Data source:

- GB (OS Data): https://www.ordnancesurvey.co.uk/products/os-open-roads
- NI: https://admin.opendatani.gov.uk/dataset/osni-open-data-50k-transport-transport-lines/resource/25462321-e67b-4d67-8a6c-02cd9f3ca9a3

Both shapefiles have been merged using QGIS. Their fields don't have the same name and the value differ slightly, but their meaning is the same (for example: "A_CLASS" for NI == "A Road" for GB). The main roads we decided to keep are Motorways, class A and dual carriage. This last category is mostly due to the NI dataset having their geometries with only one type, while on the GB OS Dataset, roads have function and type.

The goal was to have what appears to be the main roads and to have the least number of isolated segments.

15.3. Auxiliary file creation

The core script used is the same as other script handling line features, so for more detail on this part see the "River flag" part of D1.1. In this case, the main difference is we want to retain the category of road information into the final auxiliary file. There is also a _str variables which contain the worded type of road passing in each cell. The script behaves as follow:

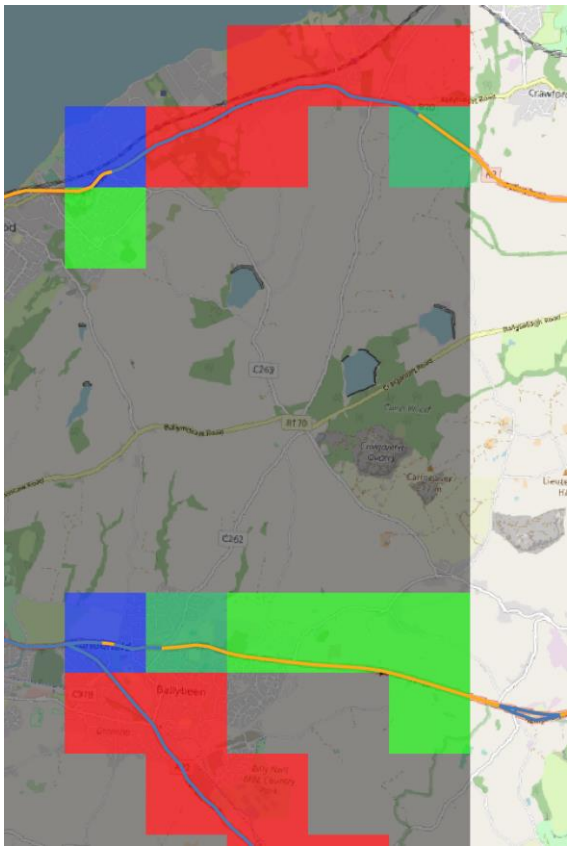
- We loop through each cell, and we compare it with a subset of line features. When a road passes inside a cell, and it passes the minimum length check (2.5m, as per River flag section), it gives its code to the final value.
- There are 3 classes of road present in the dataset – Motorways and A-roads which are present everywhere. However, in Northern Ireland, roads are categorised as being either motorway, A-road, OR dual carriageway. Therefore, this 3rd category of "dual carriageway" only applies to NI.
- The CF-standards for defining independent categories dictates that we use a bitmask specified in the `flag_masks` attribute to store such data. As such, motorways are assigned the code 1 (binary 001), A-roads the code 2 (binary 010), and dual carriageways the code 4 (binary 100). So, for example, a cell containing both a motorway and a dual carriageway (but no A-road), would be assigned the value 5 (binary 101) such that all the appropriate bits are set in the binary representation of the number.

15.4. Quality control

As described in the introduction to this section, the goal was to have the main roads with a minimal number of isolated segment (by visual inspection in QGIS), the focus was not on having the exact type/function or road. For example, in the GB dataset both motorways and A-roads can be dual carriageways, whereas in the NI dataset this is considered to be a separate class of road.

Another observation is the impact of road length on the final code, as we can see in the image:

- blue cell: orange road AND blue road
- light green cell: orange road only
- red cell: blue road only
- dark green cell: orange road, and:
    - detects blue road with length-checking step OFF
    - correctly ignores blue road with length-checking step ON



15.5. Maintenance

The OS Data for roads is updated every April and November. The update frequency for NI roads is unknown.

16. Railways [D1.11b]

16.1. Requirement

To flag each CHUK cell with the type of railway(s) passing through it.

16.2. Data source

To retrieve the source data we used the QuickOSM plugin on QGIS allowing us to query OpenStreetMap via its popular Overpass API. An example of command used was:

[out:xml] [timeout:25];

{{geocodeArea:"Northern Ireland,UK"}} -> .area_0;

way["railway"="rail"](area.area_0);

(._;>;);

out body;

Due to the number of features, we had to split into countries.

16.3. Auxiliary file creation

The input files here consist of line features, therefore it uses the same core script as the rivers or roads. The main difference is on the categories.

- None = 0; meaning there is no power: line shape passing by the cell.
- not_electrified = 1: any line with value not in the list below.
- Electrified = 2; ["contact_line", "rail", "4th_rail", "ground-level power supply", "yes"]

Note that any pre-existing rules has been applied:

- If a line, or its inner segment from a cell, is shorter than 2.5m this line is excluded.

16.4. Quality control

We compared sample of data with the superposition of shapefile in QGIS.

16.5. Maintenance

OpenStreetMap data is continually updated, so this dataset should be regenerated whenever significant enough changes have been made to the railway network.

17. Transmission line and substation [D1.11c]

17.1. Requirement

To flag each CHUK cell with the highest voltage transmission line and substation/switch present.

17.2. Data source

To retrieve the source data we used the QuickOSM plugin on QGIS allowing us to query OpenStreetMap via its popular Overpass API. An example of the command used to retrieve power line data was:

```
[out:xml] [timeout:25];
{{geocodeArea:United Kingdom}} -> .area_0;
way["power"="line"](area.area_0);
(._;>;);
out body;
```

The substation data was mostly used as-is, but reprojection into OSGB grid was performed, as was conversion of some erroneous line features to polygons.

17.3. Auxiliary file creation

Transmission line

The input files here consist of line features, and hence uses the same core script as used to generate the rivers and roads data. A key difference is that if more than one voltage of power line appears within a cell, we only wish to record the highest voltage.

The voltages we record are:

- 400kV
- 275kV
- 132kV
- 66kV
- 33kV
- Other, voltage power lines

Substation/switch

To test for the presence of a substation in a cell, the algorithm was as follows. For each CHUK cell:

- Loop through the substation polygons (with some suitable spatial-filtering for efficiency)
- If the substation has a known voltage, and we already have stored a higher known voltage for this cell, skip to the next polygon
- Test every node of the substation polygon until we find one which is within the cell
- Store the voltage of the substation in the output data array. If the voltage is unknown, this is stored as “unknown”.

17.4. Quality control

We compared sample of data with the superposition of shapefile in QGIS.

17.5. Maintenance

OpenStreetMap data is continually updated, but generally the UK power network is fairly static. This dataset will need to be regenerated when the underlying physical network changes significantly.

18. Elevation [D1.12]

18.1. Requirement

To determine, for every CHUK grid cell, the elevation above sea-level, as well as the gradient and direction of the slope at the cell centre.

18.2. Data source

The Copernicus GLO-30 dataset was used as the basis for elevation. It is a Digital Surface Model (DSM), and includes buildings, infrastructure, and vegetation. The site describing the data is available at https://spacedata.copernicus.eu/collections/copernicus-digital-elevation-model

To download the data, we accessed the tiled data files at

https://prism-dem-open.copernicus.eu/pd-desk-open-access/prismDownload/COP-DEM_GLO-30-DGED__2023_1/ -

there are files there with the name format Copernicus_DSM_10_NYY_00_WXX_00.tar, where YY and XX are the latitude and longitude co-ordinates of the grid (each tile is 1˚x1˚).

18.3. Auxiliary file creation

The algorithm used to generate this dataset was straightforward. For each CHUK grid cell:

- If this CHUK cell reads “no data” in Devolved nations of the UK, mark elevation as NO DATA
- Extract the elevation value from the Copernicus data at the CHUK cell centre

Once all elevation values have been calculated, we use the numpy library to calculate the gradient and direction of slope using second-order central differences.

18.4. Quality control

Standard visual inspection, alongside the built in QA excluding values which are not within a defined country, were performed, but no additional QA was undertaken on this dataset.

18.5. Maintenance

This dataset will remain valid for as long as the Copernicus DSM is considered valid. If they change, this dataset should be recalculated. The source dataset will be maintained until 2026.

19. Income [D1.13a]

19.1. Requirement

To determine, for every CHUK grid cell, approximate household income in this region, at as fine spatial-resolution as possible.

Because of devolution, England & Wales, Scotland, and Northern Ireland each collect slightly different statistics, each on their own statistical units. This is the case for much of the socioeconomic data (all elements of D1.13).

As a result, we have agreed to calculate the following for income, based on data availability:

- Gross-disposable household income (GDHI) per-capita at ITL3 (UK replacement for NUTS3, following the exit from the European Union) level across the whole of the UK, from 1997 to 2021.
- Household income on Middle-layer Super Output Areas (MSOAs, the defined regions on which census and other data is collated for small area statistics) for England & Wales in 2014, 2018, and 2020, and household income on Scottish Data Zones (DZ, the equivalently-sized region for Scotland) for 2014 and 2018.
- Generally speaking, the MSOA/DZ data should be used when possible, since it considers actual income (more common for comparison) at a considerably higher spatial resolution. However, if higher time resolution is needed, or Northern Irish data is important, the ITL3 data should be used.
- The ITL3 and MSOA/DZ codes for each CHUK grid cell.

19.2. Data source

The data sources used for this dataset were:

- ITL3 boundary definitions - https://geoportal.statistics.gov.uk/datasets/c769b68ed2f34da7a936da425cf6d853_0/explore
- ITL3 GDHI estimates - https://www.ons.gov.uk/economy/regionalaccounts/grossdisposablehouseholdincome/datasets/regionalgrossdisposablehouseholdincomegdhi
- MSOA boundaries - https://geoportal.statistics.gov.uk/datasets/e3b9e3e551d54ff5a41d47c46a9fd1fb_0/explore
    - Note that these are the boundaries for 2011. There are also boundaries available for 2021, but since all of the household income data is prior to 2021 they are not appropriate.

- MSOA income estimates - https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/datasets/smallareaincomeestimatesformiddlelayersuperoutputareasenglandandwales
- DZ boundaries - https://spatialdata.gov.scot/geonetwork/srv/eng/catalog.search#/metadata/7d3e8709-98fa-4d71-867c-d5c8293823f2
- DZ income estimates - https://www.gov.scot/collections/local-level-household-income-estimates/

19.3. Auxiliary file creation

The algorithm used to generate this dataset was as follows. For each CHUK grid cell:

- If this CHUK cell reads "no data" in Devolved nations of the UK, mark everything as NO DATA
- Determine which ITL3 region the cell centre is in.
- Look up the corresponding value for GDHI for this ITL3 region, for all years, and store it.
- Determine which MSOA or DZ region the cell centre is in.
- Look up the corresponding value for household income for the relevant region, for all appropriate years.
- Store the final results as one file per year for ITL3, and one file per year for MSOA/DZ, with a single-value time axis in each.

19.4. Quality control

QA on this dataset was primarily visual, and consisted of checking each dataset to ensure that every ITL3/MSOA/DZ region had data for every time point.

19.5. Maintenance

No maintenance is needed to keep the data as-is, but once new data is released by the various national authorities, additional years of income data will become available and could be added by updating the income datasets. For MSOA data, this may start using the 2021 MSOA boundaries, and this will need to be taken into account.

20. Population Density [D1.13b]

20.1. Requirement

To determine, for every CHUK grid cell, approximate population density in this region, at as fine spatial-resolution as possible.

Population data for the UK is collected similarly to income data – i.e. it is slightly different for each devolved nation of the UK. As a result, and for better interoperability with the income data, we aim to collect data on the same spatial and time scales as in Income [D1.13a].

As a result, we have agreed to calculate the following for population density, based on data availability:

- Population density at the ITL3 level
    - Complete for 2012-2019
    - Complete for England & Wales from 2002-2019
    - Incomplete England & Wales data from 1997-2002
- Population density for England &Wales at MSOA level for years 2014, 2018, and 2020
- Population density for Scotland at DZ level, for years 2014, and 2018

20.2. Data source

The data sources used for this dataset were:

- The same ITL3, MSOA, and DZ boundaries as in section 19.2. However, since we need population density, and data is published as raw population/region, we needed to process these datasets to include the area of each region. This was done using standard functions in QGIS, and a variable named "area" was added to the shapefiles, containing the area, in m^2, of each region.
- ITL3-level data: https://ec.europa.eu/eurostat/web/population-demography/demography-population-stock-balance/database
    - This dataset needed some processing. Firstly, it is actually at NUTS3 rather than ITL3 level, but since NUTS3 is (currently) a superset of ITL3, with different identifiers, this is not a problem. Any region not starting "UK" was removed from the data, and the IDs had "UK" swapped for "TL", to fit the naming format (e.g. "UKC11" changes to "TLC11" etc). Secondly, the data is separated by age group and sex. Anything which was not "TOTAL" (for the age group) was removed, and the values for both sexes were added to get the total population in each ITL3 region.
- MSOA population estimates: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/middlesuperoutputareamidyearpopulationestimates

- DZ population estimates: https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/population/population-estimates/small-area-population-estimates-2011-data-zone-based/time-series

20.3. Auxiliary file creation

The algorithm used to generate this dataset was as follows. For each CHUK grid cell:

- Find the ITL3 region which the cell centre falls within
- Look up the population for that region, and divide it by the area in km^2 to get people/km^2
- Similarly for MSOA and DZ regions.

20.4. Quality control

QA on this dataset was primarily visual, and consisted of checking each dataset to ensure that every ITL3/MSOA/DZ region had data for every time point (where this was expected – i.e. not in the ITL3 data.)

20.5. Maintenance

No maintenance is needed to keep the data as-is, but once new data is released by the various national authorities, additional years of income data will become available and could be added by updating the income datasets. For MSOA data, this may start using the 2021 MSOA boundaries, and this will need to be taken into account.

21. Educational Attainment [D1.13c]

21.1. Requirement

To determine, for every CHUK grid cell, the most recent available data on level of educational attainment in this region, at as fine spatial-resolution as possible.

Again, these data are collected differently by different nations of the UK, but here the data can be extracted directly from census data, which is available for England & Wales, Scotland, and Northern Ireland individually.

For England & Wales, the data are available at MSOA level from the 2021 census, for Scotland at DZ level from the 2011 census, and Northern Ireland at Northern Irish Data Zone (NIDZ) level from the 2021 census. For reference, the MSOA/DZ/NIDZ level is also included in the data.

Educational attainment in the UK censuses is divided into 6 classes, which are roughly:

- No qualifications: no formal qualifications
- Level 1: one to four GCSE passes (grade A* to C or grade 4 and above) and any other GCSEs at other grades, or equivalent qualification
- Level 2: five or more GCSE passes (grade A* to C or grade 4 and above) or equivalent qualifications
- Apprenticeships
- Level 3: two or more A Levels or equivalent qualifications
- Level 4 or above: Higher National Certificate, Higher National Diploma, Bachelor's degree, or post-graduate qualifications

The data records the number of people for whom a particular level is their highest level of qualification (i.e. someone with a Level 3 qualification is not included in Level 1 or Level 2, even if they have such qualifications), and are recorded for over-16s.

However, Scotland does not include data on apprenticeships, and has a different educational system and ways of classifying it, which leads to an artificially-inflated number of Level 1.

21.2. Data source

The data sources used for this dataset were:

- MSOA boundaries - https://geoportal.statistics.gov.uk/datasets/4039b4fe06a845f2b3b67fdbbe2dae0e_0/explore
    - Note that these are the boundaries for 2021, since we are using 2021 census data. This gives some slight variation in boundaries compared to the income and population density data.

- MSOA educational attainment - https://www.ons.gov.uk/filters/94f23ccb-c03f-4ddc-ae6c-07c1db29dc38/dimensions
- DZ boundaries (as previously) - https://spatialdata.gov.scot/geonetwork/srv/eng/catalog.search#/metadata/7d3e8709-98fa-4d71-867c-d5c8293823f2
- DZ educational attainment - https://www.scotlandscensus.gov.uk/search-the-census#/topics/location/SNS2011?title=SNS%20Data%20Zone%202011
- NIDZ boundaries - https://www.nisra.gov.uk/publications/geography-data-zone-boundaries-gis-format
- NIDZ educational attainment - https://build.nisra.gov.uk/en/custom/data?d=PEOPLE&v=DZ21&v=HIGHEST_QUALIFICATION

21.3. Auxiliary file creation

The algorithm used to generate this dataset was as follows. For each CHUK grid cell:

- If this CHUK cell reads “no data” in Devolved nations of the UK, mark everything as NO DATA
- Determine which MSOA/DZ/NIDZ region the cell centre is in.
- Read the number of people with each level of qualification, from the appropriate data source.
- Convert each number to a fraction of the total, such that the metric recorded is “fraction of over-16s in this region with qualification level X”
- Store each level of qualification in a separate variable.

21.4. Quality control

QA on this dataset was primarily visual, and consisted of checking each dataset to ensure that every MSOA/DZ/NIDZ region had data, with the exception of apprenticeship data in Scotland.

21.5. Maintenance

The Scottish data is only available for the 2011 census, and it may be wise to update this when the 2021 data becomes available (currently planned for release in 2024). This may require modification to the code, since it is likely that their data output format will have changed somewhat since 2011. In addition, the Scottish DataZones may need to be updated as well to their 2021 boundary definitions.

Institute for Environmental Analytics

For further information, please contact Colin McKinnon by email at c.mckinnon@the-iea.org or call +44 (0)7896 315331

https//www.the-iea.org