UK Biobank

Deriving the grid coordinates

Version 2.3

http://www.ukbiobank.ac.uk/

May 2025

This manual details the procedure for the derivation of the grid coordinate data.

Contents

1. Introduction 2
2. Location at assessment 2
3. Location at birth 3
4. Reference system 4
5. Possible applications of the grid coordinate data 4
6. Appendix: Formerly available data 6

1. Introduction

Many research projects require some level of understanding of the participant's residential location. Address details (including postcode) are not disclosed to researchers owing to their potential identifying nature; only designated members of staff have access to this information.

In order to provide researchers with information on residential location, UK Biobank has included grid coordinate data fields, available in three levels of resolution: 100m, 1km and 10km grid postings. All location data is now restricted and requires approval. Researchers who require grid coordinate data should request those at a lower resolution (i.e. 1 km or 10km), which will be of sufficient granularity for most research purposes. Grid coordinates to a 100m resolution will only be released for specific research projects that can demonstrate that they require this level of detail.

In addition, administrative geolocation fields such as local authority districts and lower/middle layer super output areas are available in Category 703. Please see Resource 1406 for information about these fields.

Only one type of geographic fields will be released to any project. Administrative location fields, and/or any future geographic information that may be derived, will not be released to any project which has access to the grid coordinate data.

Three types of residential grid coordinates are currently available: Residential location at assessment (see Category 100024), Residential location at birth (see Category 100072), and Residential location history (see Category 150). A separate document describes the process for deriving the location history; please see Resource 2060.

2. Location at assessment

The software used to geocode the UK Biobank address, originally at Leicester University and later internally within UK Biobank, was Experian QAS Batch (formerly known as the QAS

QuickAddress Batch). The QAS Batch software geocodes address records by verifying them against the official postal address files for the relevant country. Cleaned records are then assigned a match result based on the accuracy of the original address. Only results marked as Verified Correct, or Good Full Match (match code “R9”) were accepted.

This software package, using the Experian DataPlus dataset “United Kingdom Location Essential” (GBRGEO), can determine precise (1m) geocoordinates for each verified address.

In a small number of cases where no 1m geocoding was available (missing data), a less accurate geocoding has been used as the basis for deriving the 100m, 1km and 10km coordinates. These are also obtained from the QAS Batch software but using the Royal Mail Postcode Address File (PAF) rather than the extended dataset.
For further information about the data cleansing and matching processes used by the QAS software please see the QAS Batch API Guide provided by Experian, 2019, available at https://www.edq.com/globalassets/documentation/bat_api.pdf and the UK data guide (2024), available at https://docs.experianaperture.io/address-validation/global-datasets/asset/dg_gbr.pdf [both accessed March 2024].

The higher-resolution (100m) coordinates are published in Fields 22686 (Easting) and 22687 (Northing), and lower-resolution coordinates are available in Fields 22688 (Easting) and 22689 (Northing) for 1km, and in Fields 30077 (Easting) and 30078 (Northing) for 10km.

The 100m coordinates were derived from the 1m coordinates output by QAS Batch, by dividing by 100, truncating any decimal part, and multiplying by 100, and a similar process was used for the 1km and 10km coordinates. For example, a 1m-resolution coordinate of 567890 would result in a 100m-resolution coordinate of 567800, a 1km-resolution coordinate of 567000 and a 10km-resolution coordinate of 560000.

3. Location at birth

Participants who were born in England, Scotland or Wales were asked which town or district they first lived in. The interviewer selected an entry from a fixed list of (~43,000) options or could enter free text if no match was found.

These options, organised in a tree structure, were each linked to a fixed geolocation, and the East and North coordinates were entered in Fields 130 and 129 respectively.

Where a free-text entry was used, the coordinates are entered as -1 ("Location could not be mapped").

4. Reference system

The grid coordinate data are provided in the British National Grid (i.e. OSBS1936) projection. OSGB1936 is the Ordnance Survey National Grid geographic reference system (not latitude and longitude). The EPSG code for this projection is 27700.

The grid measurements refer to easting and northing with a reference point near the Isles of Sicily. The data is usually projected in the units of meters, and there are no negative values (as there are in the case of latitude/longitude convention).

5. Possible applications of the grid coordinate data

These data can be easily imported and projected in GIS software, such as ArcGIS. The north and east coordinates for each participant can be represented as a set of points on the map. These can then be grouped and colour-coded by age band, gender, lifestyle factors and other relevant outcome measures, in order to be visually inspected. The number of participants residing in a particular region of UK can be also determined. An example of visualising the data in ArcGIS is shown in figure 5.1.

These coordinates can also be imported into any statistical software such as SAS, STATA, SPSS or R, and some mathematical algorithm describing the spatial polygon(s) to study the population can be derived. Some statistical packages, such as SAS, have mapping functions that are already built-in.

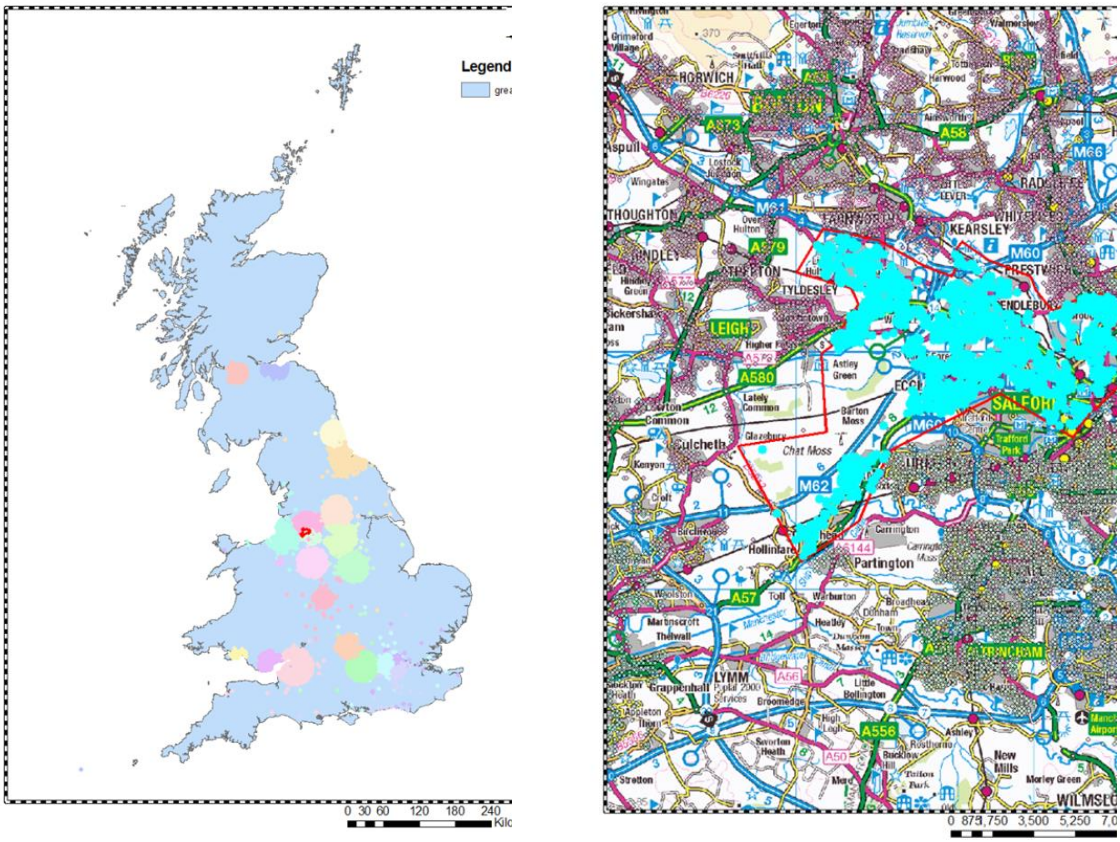

Figure 5.1 An example of visualisation of the grid coordinate data in the ArcGIS software. Left: UK Biobank population density by the assessment centre; Right: Determination of the number of participants from the Salford area. A polygon (in red) defining the Salford city boundary was drawn on the 1:50 000 OS map. The points, representing participants taking part in the UK Biobank study, were selected based on the graphics and counted.

6. Appendix: Formerly available data

Data on the participants' home location at the time of each assessment was originally provided in Fields 20033/20034 (100m resolution) and 20074/20075 (1km resolution). These grid coordinates were based on the participants' postcodes as held by UK Biobank, or provided at the assessment visit (if different). Since a single postcode area can span multiple 100m grid squares, this data was inherently less accurate than if it had been based on the full address. A number of additional issues have been identified with this data including some missing data, some errors in the postcodes used, and the incorrect derivation of the 1km coordinates by rounding the 100m coordinates instead of truncating them (rounding down).

In 2014 an attempt was made to obtain more accurate geocodings for participants' baseline location. The full address of each participant was used to derive a set of 1m grid coordinates, by the Small Area Health Statistics Unit at the University of Leicester. It was later determined that the addresses provided by the UK Biobank for this exercise included some participants' current address (at the date of extraction, thought to be September 2013), instead of the baseline address. Approximately 5% of the addresses were affected by this error.

In 2022 an exercise was undertaken to correct those geolocations where an incorrect address was provided, or where the originally provided address could not be geocoded. The full addresses were supplied to the University of Leicester for geocoding at a 1m resolution.

These coordinates, combined with the [correct] original 1m coordinates received in 2014 were used as the basis for a revised set of 100m **baseline** grid coordinates, first published in November 2022, in Fields 22686 (Easting) and 22687 (Northing). Where no 1m geocoding was available (missing data), the previously published values from Fields 20033/20034 were used.

The previous values from Fields 20033 and 20034, i.e. geocodings which are based on postcode alone, were also used to populate the new Fields 22686 and 22687 for the repeat assessment, imaging, and first repeat imaging visit. This meant that in a small number of cases, a participant's home location coordinates for the later instances did not match their baseline coordinates, even though the participant's address did not change between visits.

In 2023 UK Biobank geocoded the address of each participant at these later assessment visits and (in 2024) updated Fields 22686/22687 so that the values for later instances are now consistent with the baseline assessment data.