Producing admin-based household estimates – research to date and plans

Alison Morgan, Ann Blake, and Sally Mylles

9th May 2024

1. Key Messages of the Paper

Purpose

This paper provides a summary of research to date on producing statistics on households from administrative data. It also sets out current areas of research and the development of a longer-term research workplan to move from counts to estimates.

Key Asks of MARP

- Is there anything that we have missed that panel members feel should be included in the research workplan?
- Do MARP want to see the full research workplan at the next meeting?

2. Executive summary

Work to date on producing statistics on households from administrative data has focused on the production of a record-level dataset that aims to assign people to addresses, through using Unique Property Reference Number (UPRN) address identifiers. **As the grouping of people is based on UPRNs rather than the Census definition of a household, outputs from this dataset have occupied addresses as the statistical units.** Outputs have been produced on the number of occupied addresses and the number of occupied addresses by size. These outputs were raw counts produced directly from the dataset.

We are now aiming to develop and implement a method to produce admin-based household estimates where the methodology adjusts for coverage issues, linkage error and UPRN mis-recording. We are aiming for the estimates to be coherent with the admin-based population estimates and of sufficient quality to meet user needs (including addressing the definition to be used) and be labelled as official statistics in development. To get to this point, we are working to produce a research workplan by the end of August that will set out the programme of research and user engagement for the next few years.

Scope

The scope of the work in this paper is statistics on the number, size (number of people), and composition of households. Statistics on families, which report on the

relationships between members of the household, are currently out of scope but will need to be addressed in future to fully meet the user need expressed below. People living in communal establishments need to be considered as part of the work to produce statistics on living arrangements, including households, but the production of estimates of communal establishment populations is also out of scope for this paper. Statistics about the characteristics of people within households, also expressed as part of the user need, is out of scope here but will be given consideration as part of the wider ONS Future of Population and Migration Statistics (FPMS) programme.

3. Background

The Census is currently the main data source for local-level statistics on the number, size, and composition of households. During intercensal periods, statistics on households are produced using the Labour Force Survey but there are restrictions on the geographic breakdowns that can be produced due to sample sizes. As part of the future of population and migration statistics programme, we are exploring the feasibility of producing statistics on households using administrative data.

User needs

In 2023, ONS ran a consultation on the future of population statistics in England and Wales. A summary of consultation responses has not yet been published but we have reviewed the responses and extracted information on user needs for statistics on households. Respondents stated a need for estimates of the number and size of households at various levels of geography. They also stated a need for estimates of various kinds of living arrangements and family and household compositions, again for low-level geographies. This included information on specific types of households such as:

- single parents
- older people living alone
- multi-generational households
- co-residents
- blended families
- concealed households
- houses of multiple occupancy

Respondents also stated a need for information on relationships between household members and highlighted the importance of being able to use data on households in combination with information on personal characteristics and housing. Work to gather use cases and information about policy needs is ongoing and will continue to direct the development of the research workplan. This will include engagement already underway across government around the need for data and definitions on households more broadly.

Definitions

The definition of a household used for the census is “one person living alone or a group of people (not necessarily related) living at the same address who share cooking facilities and share a living room or sitting room or dining area”. In administrative data, it is not possible to identify whether people living at an address have shared rooms and cooking facilities. Outputs produced directly from administrative data therefore use occupied addresses as the statistical units rather than households. Using Census 2021 data, we have estimated that 0.31% of occupied addresses contain multiple households. The impact of the difference in statistical unit should therefore be minimal but will affect some population groups and geographic areas more than others.

Our ambition is to produce statistics that meet user needs. When deciding which statistical unit to produce statistics for in future, we will need to balance what is possible from administrative data against what users require. When responding to the consultation, users were generally negative about a potential move from using households as the statistical unit to using occupied addresses. As part of the research, we will consider what definition(s) can be reached via the estimation process. We will also conduct further engagement with users to direct the way in which we develop the statistics using definitions that would best meet user needs for statistics on how people live.

4. Work to date

Admin-based living arrangements dataset

The work to date has focused on creating a record-level dataset, called the admin-based living arrangements dataset (ABLAD). This is created as follows:

- Use the Address Index Matching Service (AIMS) to add a Unique Property Reference Number (UPRN) variable to all administrative data sources containing full address information (note: this is completed by Data Engineering as part of the initial data processing)
- Link the UPRNs from the Personal Demographic Service (PDS), English School Census (ESC), Individualised Learner Record (ILR) and Lifelong Learning Wales Record (LLWR) to the Statistical Population Dataset (SPD) using unique identifiers
- If an individual appears multiple times in a data source, select one UPRN per data source based on date information and information on the quality of the UPRN match
- After selecting one UPRN per data source, if an individual appears across multiple data sources, use a dataset hierarchy of broadly ESC, LLWR, ILR, PDS (with some additional aged-based rules) to select one UPRN per person (note: this hierarchy was constructed based on findings from record-level comparisons with Census 2021 data)
- Link on the Address Frame using the UPRN to add geography variables, a variable that classifies UPRNs into households and communal establishments, and an establishment type variable

- Link on Higher Education Statistics Agency (HESA), Ministry of Justice (MoJ) and ESC data to identify students in halls of residence, prisoners and pupils boarding in English state schools, and update the address and establishment type to reflect that they are living in a communal establishment

The ABLAD is then split into four sub-datasets:

- If address type is Household, put these records into the admin-based occupied address dataset (ABOAD). In the 2023 ABLAD, 53,691,604 records (95.2%) went into the ABOAD
- If address type is Communal Establishment, put these records into the admin-based communal establishment dataset (ABCED). In the 2023 ABLAD, 1,015,847 records (1.8%) went into the ABCED
- If the record has a UPRN but no address type, put it into a 'non-address frame UPRN' dataset. In the 2023 ABLAD, 1,503,637 records (2.7%) had a UPRN but no address type
- If the record does not have a UPRN, put it into a 'UPRN-less' dataset. In the 2023 ABLD, 193,238 records (0.3%) did not have a UPRN

The following outputs have been produced from the ABOAD:

- Number of occupied addresses
- Number of occupied addresses by size (number of people)

All outputs have been produced directly from the record-level dataset, without any adjustment for coverage issues, linkage errors or UPRN mis-recording.

Prior review

A paper outlining the research to date on this project has already been reviewed by MaRAG and colleagues working in the Methodology and Census addressing teams. The feedback and our actions/planned actions can be summarised as follows:

	Feedback – suggested actions	Progress
1	Increase understanding of user needs for statistics on households, communal establishments, and special population groups	Review of consultation responses completed, and initial conversations started on further user engagement
2	Assess the impact of using occupied addresses as the statistical unit instead of households	Analysis completed on the characteristics of people living in multi-household addresses
3	Outline plans to monitor the quality of administrative data sources and processing going forward	Quality paper being written (see below)
4	Increase transparency around data processing and the decisions and assumptions made	Internal documentation in progress and plans for reporting on quality included in the quality paper
5	Assess the impact of decisions and assumptions on the final outputs e.g. via sensitivity analysis	UPRN selection method being reviewed, further record-level Census comparisons being scoped

		out and sensitivity analysis included in the quality paper
6	Continue to build understanding of where errors could happen and the scale of them, including producing quantitative measures at the level of the final outputs	Quality paper to cover potential sources of error at all stages of the production process, how they will be assessed, and proposed quality measures
7	Conduct analysis of the people whose UPRN is neither an occupied address nor a communal establishment	Completed
8	Conduct comparisons over time once methods are stable	Included in the quality paper
9	Acquire the HMO register and planning application data	Data requirements submitted
10	Focus on composition rather than attempting to identify relationships/families	Families out of scope

International literature review

We have conducted a review of international approaches to producing statistics on households using administrative data. The challenge of what to use as a base for the statistics was a common theme, with statistics traditionally produced on a household basis but this being challenging in an admin data context. From the literature review, we did not identify any international approaches that were directly transferable to a UK context, but we did identify the following research avenues for further exploration:

- Using council tax, electoral register, and utilities data to identify whether addresses are occupied
- Using a model-based approach for placing people into addresses
- Using relationship data to identify connections between people
- Using graph theory to group people

The feasibility of these, including data availability and suitability in a UK context is still to be confirmed.

5. Current work and short-term plans

Current work and plans for the next four months are focused on three key areas: improving the ABLAD, scoping of estimation methods and quality.

Improving the ABLAD

We are exploring the following options for improving the ABLAD:

- Incorporating additional data sources
- Testing alternative approaches for dealing with multiple recorded UPRNs for a person and placing people into addresses

- Developing and implementing categories for producing outputs on occupied address composition (which cover the number of adults and children at the address, and number of residents by broad age group and sex, not the relationships between residents)

Estimation

As already outlined, outputs from the ABLAD so far have been produced directly from the record-level dataset, without any adjustment for coverage issues, linkage errors or UPRN mis-recording. Moving forward, rather than simply producing counts, our aim is to produce estimates that are statistically robust, coherent with the admin-based population estimates and are of sufficient quality to meet user needs (including addressing the definition to be used) and be labelled as official statistics in development.

We have conducted a review of methods options for producing household estimates, including liaising with Methodology colleagues in ONS. The methods that have been suggested are:

- Weighting the ABOAD to the admin-based population estimates (ABPEs), where the weights could be based on occupied address distributions from the ABOAD, a statistical model, or combinatorial optimisation
- Creating a Dynamic Population model for households
- Statistical modelling using a Bayesian hierarchical population model (B-Pop)
- Statistical modelling using a Structure Preserving Estimator (SPREE)

We are currently writing a paper outlining each method. Once finalised, the options will be discussed with the internal technical working group and then a recommendation proposed by the group on which method or methods to progress. The recommendation will need to balance the potential to meet user needs, maturity, and complexity of each method, and the ONS resource and external support that would be required.

Quality

We have written a paper that recommends quality assurance and quality measures for each stage of the process for producing the admin-based living arrangements dataset and in future, the admin-based household estimates. These include:

- Working with the product owners to improve quality reporting for, and the quality of, the data inputs and linkages
- Peer review of code
- Expert review of methods
- Automated checks in code
- Quality assurance checklists
- Aggregate-level and record-level comparisons over time and with other data sources
- Sensitivity analysis
- Bias and representativity indicators
- Measures of variance and precision

- Documentation and reporting

The latest version of the ABLAD has recently been produced for 2021, 2022 and 2023, with good practice around code already followed. We are now moving on to implement some of these methods and measures to assess the quality of the record-level datasets. This will form a baseline for any future changes to the ABLAD methodology to be compared with and will help us to understand the issues that we are aiming to fix via dataset improvements and the estimation methodology. Where some of the recommended quality measures are more complex, we will incorporate them in the longer-term plans.

6. Long-term research workplan

We are working towards a deadline of the end of August for a full research workplan that will inform how we produce admin-based household estimates that can be published as official statistics in development. We envisage that the research workplan will extend out to 2026 and will outline the tasks, timelines and resource required. It will cover plans for further research to improve the ABLAD, further engagement with users, and work to implement the recommendations on estimation methods and quality. In advance of the August deadline, we are planning to get the research workplan reviewed by MaRAG. We are also currently proposing to bring the full research workplan to the next MARP meeting but would like confirmation from panel members on whether you would like to review it.

Annexe 1: Definitions

Address	An address is a collection of information, presented in a mostly fixed format, used to give the location of a place. A place can be any kind of building, or object that might not have a 'normal' address – such as a bus shelter or an electricity sub-station for example.
Occupied address	An occupied address is a unique property reference number (UPRN) on the Address Frame which has been successfully linked to at least one individual in the Statistical Population Dataset.
Household	One person living alone or a group of people (not necessarily related) living at the same address who share cooking facilities and share a living room or sitting room or dining area. A household can consist of a single family, more than one family or no families in the case of a group of unrelated people.
Communal Establishment (CE)	A place providing managed residential accommodation. 'Managed' here means full-time or part-time supervision of the accommodation, such as care homes, student halls of residence, hospitals, or prisons.

UPRN	A Unique Property Reference Number is a unique identifier for every addressable location in the UK they provide every place with a consistent identifier throughout its lifecycle, from planning through to demolition.