Office for National Statistics

Article

Population and migration statistics transformation in England and Wales, population characteristics update: 2023

Assessment of our progress towards producing admin-based estimates of population characteristics.

Contact:
Sonia Carrera and Becky Tinsley
2023Consultation@ons.gov.uk
01329 444972

Release date:
26 June 2023

Next release:
To be announced

Table of contents

1 . Overview of the population characteristics update

- The research and proofs of concept produced to date demonstrate our ability and potential to produce outputs at subnational levels for population characteristics primarily using administrative data, with higher frequency and timeliness than is currently possible in the years between censuses.

- Our ambition within this statistical design is for most characteristics estimates to be based on administrative data, with modelling and integration with surveys where administrative data do not yet allow us to fully meet user needs.

- Future research will include continued partnering with data providers, government analysts and academics to consolidate, widen and strengthen data provision, and developing statistical methods to further enhance quality and coverage, as well as across the Government Statistical Service to produce new outputs to meet user needs.

- Priorities for future research will be informed by evidence on users' needs, gathered through the consultation on the future of population and migration statistics in England and Wales launching 29 June 2023.

2 . Using administrative data to transform statistics on population characteristics

By making greater use of administrative data alongside surveys, we aim to provide more frequent and timely statistics on the characteristics of the population than is currently possible in the years between censuses. This is part of our work to improve inclusivity of data and statistics, described in the recently published Embedding Inclusivity in UK data: 2023 update on implementing Inclusive Data recommendations, available on the UK Statistics Authority webpage.

This article summarises the proofs of concept produced to date showing our ability to meet this aim, and an assessment of what users can expect future research and experimental statistics to offer for characteristics.

This is part of the evidence that underpins our consultation on the future of population and migration statistics in England and Wales, launching 29 June 2023. The consultation informs the National Statistician's forthcoming recommendation on the future of population and migration statistics in England and Wales. The assessment helps users respond to the consultation and should be read alongside other publications (see Section 9: Related links).

3 . Assessment of administrative data availability and research to date for population characteristics

Over the last decade we have been developing a series of proofs of concept to show what users might expect from a transformed population and social statistics system underpinned by administrative data.

Research and proofs of concept

Alongside the development of our transformed population and migration statistics, we have focused on producing research and proofs of concept for a selection of characteristics that users are used to getting from the 10-yearly census, and also on characteristics that have a high user need but are not met by the census, such as income.

Through this work, we have shown the ability to produce outputs at subnational levels, down to Lower layer Super Output Areas (LSOA) for income, ethnicity and housing (excluding tenure), and at local authority level for households. The research also combined income and ethnicity, and ethnicity and housing to show the potential to produce outputs that provide insights across multiple characteristics, as the census does.

We have produced feasibility research on estimates of population by specific times of day based on mobility and on admin-based health statistics for morbidity. We have also published initial yearly estimates of veterans and identified the future developments needed to produce multivariate and longitudinal evidence for this cohort.

For these characteristics, the research shows that we can produce these outputs every year, and the potential of producing them within a year of the reference date (the date that the outputs relate to), providing users with more timely updates than are currently available.

The research has also highlighted where further developments are needed to increase the coverage and quality of estimates. To achieve this, we are reliant on a regular and timely flow of administrative data to agreed quality standards from providers, through continued partnership across government and, where relevant, with other providers to consolidate, widen and strengthen data provision.

This report also describes our current assessment of the potential frequency, timeliness and granularity of statistics for other population characteristics that the census typically provides or for which there is a high user need, including protected characteristics. This is based on our preliminary research and the methods that we will explore in the future, based on the increased understanding of users' needs from the consultation. The following sections also describe some of the methods that we will be exploring to ensure that these outputs are accurate and meet the definitions that best address user needs. While administrative data will be the sole source for many of these characteristics, it is expected that some characteristics will be reliant on surveys or a combination of the two, alongside modelling.

In line with this, we envisage a framework where characteristics estimates can be delivered. The framework consists of a continuum with different mixtures of administrative and survey data sources. Our published proofs of concept have largely been based on administrative data alone, and these sit at one end of the spectrum. Outputs reliant on only survey data (many that are part of our existing statistical portfolio) are at the opposite end. In between, multi-source methods such as statistical modelling, calibration weighting, coverage estimation and small area estimation can be used to bring administrative and survey data together. Our ambition within this statistical design is for ongoing improvements to the estimates delivered within this framework, aiming in the longer term for the majority to be based on administrative data. This means we would need to:

- secure additional administrative data
- improve completeness of those available
- ensure availability on a more frequent and timely basis

This may mean asking departments to collect data on our behalf, provide data more regularly, or provide data with a shorter lag.

The following sections of this article outline in more detail what users can expect our future research and experimental statistics to offer for characteristics, grouped by our current assessment of administrative data available, and progress with the research to date. More details on the sources explored to date for each topic is available in Section 8.

4 . Expected offer for characteristics estimates based on administrative data

Characteristics with high availability of administrative data

For many characteristics we have identified, and in many cases acquired, administrative data that can be used to produce statistics that meet users' needs. For some characteristics, we have already produced research, developed proofs of concept, and even developed experimental estimates.

These characteristics include:

- age
- sex
- ethnicity
- income
- housing characteristics, excluding tenure (these include accommodation type, number of bedrooms, number of bathrooms, number of rooms and "build period")
- health and education

Our expected offer for users for these characteristics would be:

- annual estimates
- produced within a year of the reference period
- at local levels – lower super output area (LSOA)

The research to date has shown the potential for producing statistics down to local levels from administrative data. In some cases, such as for ethnicity, this showed the need to improve population coverage or granularity of the statistics produced. We are exploring opportunities to include additional data sources to improve the population coverage and methods to adjust for:

- missingness
- lag between reference and reporting periods
- definitional differences

In some cases, this may need to be addressed by complementing administrative data with surveys. When new data sources are scoped and before incorporating them in research, we will continue applying robust ethical procedures and follow the advice of the National Statistician's Data Ethics Advisory Committee (NSDEC).

For a number of characteristics, we have identified administrative data that would allow us to produce statistics, or already have access to historical data that need updating to the latest available years. In some cases, we have produced feasibility research, such as for veterans and communal establishments. As part of the next phase of work, we will work with data providers to set out requirements and acquire these data. This will be followed by a development period to assess the quality of the data, develop the statistical methods required to produce statistics, and produce first outputs, to understand their quality and how they compare to existing outputs, such as the census. Our development work will focus on producing research and estimates at local authority level, before increasing the granularity of the outputs at the local level where data allow.

These characteristics include:

- vehicle ownership
- labour market status
- veterans
- household composition
- communal establishments and special populations
- tenure
- disability
- marital or legal partnership status

Our expected offer for users of these characteristics would be:

- annual estimates
- produced within a year of the reference period
- initially at local authority level before developing them for local levels (LSOA)

We are developing a methodological strategy to ensure the resulting statistics are of high quality and meet user needs, especially where gaps in coverage from administrative data currently exist. Although coverage of these topics in administrative data is promising, in a number of cases the data are not complete and cannot currently be used alone to obtain robust estimates. This strategy, therefore, includes development of statistical methods to account for missingness, coverage, integration with other data sources, statistical disclosure control methods and so on. We have already progressed some of this development and will build on this over time, as set out in our Methods for producing multivariate population statistics using administrative and survey sources paper (PDF, 353KB).

Methods to resolve missingness and conflicting records could include imputation, multiple imputation and latent class analysis (MILC), fractional hot deck imputation, and multivariate imputation. An additional approach would be small area estimation techniques, including generalised structure preserving estimation (GSPREE) and regression models. These methods are being developed in consultation with external academics and will require extensive future work. These methods are data dependent and would likely be bespoke for each topic, as the data available will vary in coverage and relevance.

Characteristics with partial availability of administrative data

For some characteristics, the assessment to date has shown that we can expect less readily available administrative data. This might be because:

- it is only collected for some subgroups of the population (coverage)
- the definition of what is collected does not sufficiently meet user needs (relevance)
- it is not collected at all

In these cases, we may initially place more reliance on data that are collected from surveys. This might mean combining survey and administrative data, and future work will focus on developing methods to achieve this. The type of methods that would support these estimates could include small area estimation, imputation, and modelling to combine different data sources. The suitability of these methods will depend on the quality of the data themselves and may vary by characteristic.

Our assessment is that currently this applies to country of birth.

Our expected offer for users for these characteristics would be:

- estimates less frequently than every year
- estimates that are produced less timely than within a year of the reference period
- at local authority level

Characteristics where further research is needed, before we can say more about what our offer would be

Finally, there are some characteristics where we need to conduct further research to define the future offer for users. This is because the evidence available to us about administrative data collected on these characteristics is partial or limited and further investigation is needed to assess their suitability. We will continue working with data providers across government and beyond to further investigate sources available and consider where current collection can be improved and widened.

These characteristics include:

- national identity
- occupation
- religion
- caring responsibilities
- main language
- Welsh language
- National Statistics Socio-economic Classification (NS-SEC)
- sexual orientation
- main language
- gender identity
- pregnancy and maternity

In the future we can continue producing estimates directly from surveys. We expect the Transformed Labour Force Survey (TLFS) to include questions on most protected characteristics and other census-type topics, and will investigate improved estimation approaches to improve the granularity of these survey-based outputs. We are also working with Welsh Government to identify appropriate sources for producing Welsh language statistics in future.

5 . Quality standards and measures for characteristics

Our work so far on quantitative quality standards that admin-based estimates can be compared with has focused on the quality of population estimates produced from the existing system at local authority (LA) level. For more information, see the Bias and Variance quality standards for 2023 recommendation, published by the UK Statistics Authority (PDF, 223KB). Quantitative quality standards for the characteristics estimates for England and Wales will also be required to understand the statistical quality of admin-based estimates relative to current estimates.

To support this work, we will research and share quality standards that can be set for population characteristics based on Census 2021. Our previous characteristics quality standards, set out in the Beyond 2011: Options Report 2 (PDF, 492KB), were based on an approximation of what was planned at that time – for example lower super output area (LSOA) data annually based on a five-year rolling average. As we have refined our intended offering since the Beyond 2011 work, these quality standards will also need to be updated as part of future research. We will produce assessments of standards for both variance and bias. User feedback on what level of bias would be acceptable will be important in setting these standards.

Our research will also need to focus on developing methods to measure statistical uncertainty (both variance and bias) in the administrative data-based characteristics estimates themselves. These quality measures will allow an assessment of the accuracy of outputs based on administrative data, and identification of where users will need to consider a relative prioritisation between accuracy, granularity, timeliness, or frequency.

6 . Future developments

Our research to date has shown the potential to produce admin-based statistics on population characteristics to an increased level of frequency and geographic granularity than is currently possible based on census data. Proofs of concept have demonstrated our ability to produce estimates on multiple characteristics at a time (for example income, and housing, by ethnicity).

Future steps in the development of our offer for statistics on population characteristics will include developing:

- statistical methods to improve quality
- methods to integrate administrative and survey data sources
- statistical disclosure methods for outputs

Other research will include developing quality measures for population characteristics, as well as widening access to data and progressing research to cover those characteristics for which the current offer is less developed. Where possible, to produce outputs at Lower layer Super Output Area (LSOA) level we will also continue to explore where these can be used to build flexible outputs that target user needs for specific geographies, such as coastal areas or commuter belts.

Progress in the research outlined in this article is reliant on the regular and timely flow of administrative data to agreed quality standards from providers. We are working in partnership with departments across government and, where relevant, with other providers to consolidate, widen and strengthen data provision.

Feedback

Feedback from users is important for the future prioritisation of research. To this aim, your input to the consultation on the future of population and migration statistics in England and Wales launching 29 June 2023 will be essential, both to understand which characteristics users will want the next steps of research to focus on, and to gather users' assessments on the relative importance they place on accuracy, granularity, timeliness, detail, or frequency for estimates of population characteristics.

7 . Glossary

Administrative data

Administrative data refer to information collected primarily for administrative reasons (not research). This type of data is collected by government departments and other organisations for registration, transactions, and record-keeping, usually when delivering a service.

Calibration weighting

Calibration weighting is a statistical technique used to compensate for non-response and coverage error, and to ensure internal estimates are consistent with external measures. For example, ensuring that survey estimates for sub-regional populations correspond with the estimated age and sex composition by region.

Communal establishment

A communal establishment is an establishment with full-time or part-time supervision providing residential accommodation, such as student halls of residence, boarding schools, armed forces bases, hospitals, care homes, and prisons.

Lower layer Super Output Area (LSOA)

LSOAs are made up of groups of Output Areas, usually four or five. They comprise between 400 and 1,200 households and have a usually resident population between 1,000 and 3,000 persons.

Small area estimation

Small area estimation methods combine and borrow strength from different data sources in order to obtain robust model-based survey estimates where sample counts are too small for direct survey estimates. They provide a powerful mechanism for bringing information together across sources (typically survey, census and administrative data) and estimating from integrated data.

Statistical modelling

Statistical modelling involves making a set of assumptions about underlying processes that generate data in order to make inferences or to create estimates or predictions. Often a model is fitted to a set of observed data to establish the values of parameters that describe the relationships between variables.

8 . Administrative data sources used in our current assessment

This section provides a summary of the administrative data sources included in our current assessment for producing estimates of population characteristics, broken down by topic and data availability to the Office for National Statistics (ONS).

Age

The following data sources are available to the ONS and used in published research:

- Personal Demographic Service (PDS)
- Higher Education Statistics Agency (HESA)
- English School Census (ESC)
- Welsh School Census (WSC) – also known as Pupil-Level Annual School Census (PLASC)
- Hospital Episode Statistics (HES)
- Emergency Care Dataset (ECDS)
- Department for Work and Pensions (DWP) Customer Information System (CIS)
- DWP Benefits and Income Datasets (BIDS)
- Individualised Learner Record (ILR)

Sex

The following data sources are available to the ONS and used in published research:

- PDS
- HESA
- ESC
- WSC (PLASC)
- HES
- ECDS
- DWP CIS
- DWP BIDS
- ILR

Mobility: internal migration

The following data sources are available to the ONS and used in published research: PDS, and HESA.

The other source where some data are available to the ONS is Her Majesty's Revenue and Customs (HMRC) Frameworks.

Household composition

The sources where some data are available to the ONS are:

- PDS
- ESC
- ILR
- HMRC Child Benefit
- DWP BIDS
- HMRC Frameworks
- Births registrations and notifications
- Marriages and civil partnerships
- Driver and Vehicle Licensing Agency (DVLA) driver data
- HESA

Communal establishments (CEs) and special population groups (SPGs)

The following data sources are available to the ONS and used in published research:

- PDS
- Ministry of Justice (MoJ) prisoners data
- ESC

The sources where some data are available to the ONS are:

- HESA
- WSC (PLASC)
- ILR
- HES
- Patient Episode Database for Wales (PEDW)
- HMRC Frameworks
- HMRC Pay As You Earn (PAYE) Real Time Information (RTI)
- DVLA driver data

The other data source relevant to the topic, but not currently available to the ONS is Adult Social Care Client Level Data (ASC CLD).

Housing characteristics

The data sources available to the ONS and used in published research are Valuation Office Agency (VOA) property attributes, and Energy Performance Certificate (EPC) data.

The other sources where some data are available to the ONS are Health and Safety Executive (HSE) gas safety certificate data, and Council Tax.

The other data source relevant to the topic, but not currently available to the ONS is the utilities company data.

Tenure

The data source available to the ONS and used in published research is the Continuous Recording of Lettings and Sales in social housing in England (CORE).

The other sources where some data are available to the ONS are:

- Tenancy Deposit Protection Scheme (TDPS)
- Zero Deposit
- VOA private rentals data

The other data sources relevant to the topic, but not currently available to the ONS are Financial Conduct Authority (FCA) mortgage data, and Rent Smart Wales.

Vehicle ownership

The source where some data are available to the ONS is the DVLA Vehicle database.

Marital or legal partnership status

The sources where some data are available to the ONS are:

- marriage registrations
- civil partnership registrations
- divorce registrations (includes dissolutions)

Pregnancy and maternity

Data for this topic are not currently collected by the census. The sources where some data are available to the ONS are:

- birth registrations
- birth notifications
- abortion notifications
- death registrations

The other data sources relevant to the topic, but not currently available to the ONS are the Maternity Services Dataset (MSDS), and Community Services Dataset (CSDS).

Ethnicity

The following data sources are available to the ONS and used in published research:

- ESC
- WSC (PLASC)
- Lifelong Learning Wales Record (LLWR)
- ILR
- HESA
- HES
- ECDS
- NHS Talking Therapies
- PEDW
- Emergency Department Dataset (EDDS) Wales
- NHS birth notifications

The other sources where some data are available to the ONS are DWP CIS, and MoJ prisoners data.

The other data sources relevant to the topic, but not currently available to the ONS are:

- General Practice Data for Planning and Research (GPDPR)
- NHS Ethnic Category Information Asset
- MSDS
- CSDS

National identity

The sources where some data are available to the ONS are:

- HESA
- LLWR
- MoJ prisoners data

The other data source relevant to the topic, but not currently available to the ONS is the WSC (PLASC).

Main language

The sources where some data are available to the ONS are:

- PDS
- ESC
- WSC (PLASC)
- HESA
- HES
- ECDS
- PEDW
- MoJ prisoners data

The other data sources relevant to the topic, but not currently available to the ONS are:

- DWP CIS
- HMRC frameworks
- HMRC PAYE RTI
- EDDS

Welsh language

The sources where some data are available to the ONS are:

- WSC (PLASC)
- LLWR
- HESA
- MoJ prisoners data

The other data sources relevant to the topic, but not currently available to the ONS are the School Workforce Annual Census (SWAC), and Wales National Workforce Reporting System.

Religion

The sources where some data are available to the ONS are HESA, and MoJ prisoners data.

The other data sources relevant to the topic, but not currently available to the ONS are NHS Talking Therapies, and GPDPR.

Country of birth

The data source available to the ONS and used in published research is asylum and refugee data.

The other sources where some data are available to the ONS are:

- exit checks
- DWP Registration And Population Interaction Database (RAPID)
- HMRC PAYE RTI linked to Migrant Workers Scan (MWS)
- MWS
- HESA

General health

The data sources available to the ONS and used in published research are:

- PDS
- HES
- General Practice Data for Pandemic Planning and Research (GPDPPR)

The other sources where some data are available to the ONS are:

- ECDS
- PEDW
- EDDS
- NHS Talking Therapies
- DWP BIDS
- ILR
- LLWR
- ESC
- Council Tax

The other data sources relevant to the topic, but not currently available to the ONS are:

- CSDS
- MSDS
- Mental Health Services Dataset (MHSDS)
- GPDPR

Disability

The sources where some data are available to the ONS are:

- National Pupil Database (NPD)
- WSC (PLASC)
- HESA
- PDS
- HES
- ECDS
- PEDW
- EDDS
- NHS Talking Therapies
- DWP BIDS
- ILR
- LLWR
- ESC
- Council Tax

The other data sources relevant to the topic, but not currently available to the ONS are:

- CSDS
- MSDS
- GPDPR

Caring responsibilities

The sources where some data are available to the ONS are:

- DWP BIDS
- HMRC Self-Assessment
- HES
- ECDS
- PEDW
- EDDS
- Council Tax
- HESA

The other data sources relevant to the topic, but not currently available to the ONS are:

- RAPID
- management information collected by local authorities
- CSDS

Income

The data sources available to the ONS and used in published research are:

- HMRC PAYE P14
- HMRC Self-Assessment
- HMRC Tax Credits
- HMRC Child Benefit
- DWP CIS
- DWP BIDS – includes National Benefits Database (NBD), Single Housing Benefit Extract (SHBE), Universal Credit (UC) and Personal Independence Payment (PIP)

The other sources where some data are available to the ONS are HMRC PAYE RTI, and Council Tax.

The other data source relevant to the topic, but not currently available to the ONS is DWP RAPID.

Education (including highest qualification)

The following data sources are available to the ONS and used in published research:

- NPD
- ILR
- LLWR
- HESA

The other data sources relevant to the topic, but not currently available to the ONS are:

- Welsh examinations and assessments datasets
- Wales post 16 education and training
- Educated Other than at School (EOTAS)

Labour market status

The following data sources are available to the ONS and used in published research:

- HMRC PAYE RTI
- HMRC Self-Assessment
- HMRC Tax Credits
- HMRC Child Benefit
- DWP CIS
- DWP BIDS – includes NBD and PIP
- HESA
- ESC
- WSC (PLASC)

Veterans

The data source available to the ONS and used in published research is the Ministry of Defence (MoD) Service Leavers Data (SLD).

Sexual orientation

The sources where some data are available to the ONS are HESA, and NHS Talking Therapies.

The other data source relevant to the topic, but not currently available to the ONS is GPDPR.

Gender identity

The source where some data are available to the ONS is HESA.

The other data source relevant to the topic, but not currently available to the ONS is GPDPR.

Mobility and travel to work

The data sources available to the ONS and used in published research are the National Travel Survey and the National Trip End Model (NTEM) version eight core scenario planning data.

The sources where some data are available to the ONS are:

- financial transaction data
- Labour Force Survey (LFS) and Labour Market Survey (LMS)
- mobile phone data
- Council Tax
- Business Register and Employment Survey (BRES)

9 . Related links

Population and migration statistics transformation in England and Wales, progress update: 2023
Article | Released 26 June 2023
A summary of our research on the future of population and migration statistics in England and Wales, underpinning our consultation on the proposed new system.

Population and migration statistics transformation in England and Wales, technical topic guide: 2023
Methodology | Released 26 June 2023
This topic guide provides further information on topics which are part of the evidence being published to support the public consultation on the National Statistician's forthcoming recommendation on the future of population statistics.

Methods for producing multivariate population statistics using administrative and survey sources (PDF, 353KB)
Methodology | November 2022
This paper provides an outline for the programme of methodological work to produce multivariate population outputs which are primarily based on administrative data but use survey and other data sources to provide robust outputs that account for missingness and other data problems.

10 . Cite this article

Office for National Statistics (ONS), released 26 June 2023, ONS website, article, Population and migration statistics transformation in England and Wales, population characteristics update: 2023