Long-term international migration: quality assuring administrative data

Administrative data sources and quality assurance in the production of admin-based long-term international migration estimates published in bi-annual releases.

Contact:
Brendan Georgeson
pop.info@ons.gov.uk
+44 1329 444661

Release date:
16 November 2023

Next release:
To be announced

Table of contents

1. Overview

2. Operational context and administrative data collection

3. Communication with data supply partners

4. Quality assurance principles, standards and checks applied by data suppliers

5. Producer's quality assurance investigations and documentation

6. Cite this article

1 . Overview

This report provides information and assurance about the quality of the data used by the Office for National Statistics (ONS) in its bi-annual releases of admin-based long-term international migration (LTIM) estimates. It covers three datasets:

- Home Office Borders and Immigration (HOBI) data supplied by the Home Office (HO)
- Registration and Population Interaction Database (RAPID) data supplied by the Department for Work and Pensions (DWP)
- Higher Education Statistics Agency (HESA) data supplied by Jisc, the UK's expert body for digital technology and digital resources in higher education, further education and research

LTIM estimates are produced using mainly HOBI and RAPID data, while HESA data are used for making adjustments to the estimates. We use HOBI data to produce migration estimates of non-EU nationals (excluding British nationals); and RAPID and HESA data to produce migration estimates of EU nationals.

There are a number of other data sources that are used in a relatively minor way compared with HOBI, RAPID and HESA, and are not included within this assessment. These are:

- Real Time Information (RTI) data from HM Revenue and Customs linked to HESA Student Record data, to understand employment levels of international students; this is used to adjust HESA enrolments data used in the production of RAPID-based EU migration estimates
- NHS Personal Demographics Service (PDS) - we use this source to distribute international migrants, both EU and non-EU, under the age of 16 years to geographies at local authority level; PDS is used when distributing immigration only

LTIM estimates are produced mainly from administrative (admin) data and our ambition is to move away entirely from the use of survey data. The main area of focus to reach this aim is the estimation of British national migration, which is still derived from the International Passenger Survey (IPS). The quality of the IPS data is covered in Long-Term International Migration QMI. This quality assurance only relates to admin data.

This report covers the quality of the admin data sources that are used to estimate LTIM. It does not cover the wider aspects of quality (such as accuracy or relevance) that contribute to the uncertainty of LTIM estimates. When using admin data to produce timely LTIM estimates (currently five months after the reference period), we apply assumptions to account for the fact that we do not have a full 12 months of information on travel patterns to determine if a person is a long-term migrant. These assumptions introduce uncertainty into our estimates. For more information on these assumptions, see our International migration research, progress update Articles.

This report has been prepared using the UK Statistics Authority's Administrative Data Quality Assurance (QA) Toolkit (PDF, 243KB). The toolkit explains the necessary level of quality assurance based on the data quality risk and level of public interest.

The risk of data quality concerns was judged to be medium as the data may be affected by:

- HESA data – these are delivered with up to a 17-month lag, so timeliness is a concern; however, the Data Futures Programme will improve timeliness for future deliveries

The risk to data quality is mitigated for the following reasons:

- there is regular communication between the ONS, data suppliers, and users
- the admin data are subjected to a series of validation and quality checks by the suppliers (HO, DWP, Jisc) before the datasets are made available to the ONS, and quality checks are also carried out during the processing and analysis stages
- we have processes to quality assure the data when they are supplied, and quality checks built into our analysis

The public interest was judged to be high because:

- we have internal users who use our LTIM estimates to generate a range of National Statistics, including population estimates
- we have many external users, including other government departments (such as the Cabinet Office) and devolved bodies (such as Welsh Government and Scottish Government)
- organisations use the data to inform policy decisions
- local authorities use LTIM estimates to aid planning and resource allocation
- there is interest in LTIM statistics from the public and the media

Therefore, according to the UK Statistics Authority's risk/profile matrix (page 9 of the Administrative Data QA toolkit) (PDF, 243KB), either enhanced (A2) or comprehensive (A3) assurance should be applied to the data. We have chosen to apply enhanced (A2) assurance as the risks in terms of data quality concern are mitigated by several factors described above. Enhanced assurance (A2) means that the statistical producer has evaluated the administrative data QA arrangements and published a fuller description of the assurance.

This report will explore each area of practice in the toolkit: operational context and administrative data collection; communication with data supply partners; QA principles, standards and checks applied by data suppliers; and producer's QA investigations and documentation.

2 . Operational context and administrative data collection

Home Office Borders and Immigration data

Home Office Borders and Immigration (HOBI) data (previously referred to as Exit Checks) are derived from a linked database that combines data from Home Office (HO) systems to build travel histories that consist of an individual's travel into or out of the UK, together with data relating to their immigration status. The HOBI data are compiled from various databases, which are matched so that each individual is allocated a unique identifier, which consists of biographic details and associated events. It enables us to see the types of leave (visas) individuals have held, as well as travel patterns in and out of the UK.

HO supply four files recording the movements of individuals into and out of the UK on a visa. The Office for National Statistics (ONS) uses two of these files to generate the long-term international migration estimates:

- Leave Instance – records details of individuals' visas, entry and exit from the UK and date of birth; supplied quarterly

- Visit Histories – travel histories of individuals in the Leave Instance table, including arrival and departure dates but no personal information; supplied quarterly

In addition to those already provided for non-EU nationals, we have received two deliveries of European Economic Area (EEA) data (including data on those granted leave on the EU Settlement Scheme) to explore its potential use in producing better estimates.

A data specification document provides details on what is to be included in the quarterly and annual data deliveries.

Registration and Population Interaction Database

Access to the Registration and Population Interaction Database (RAPID) is provided to the Office for National Statistics (ONS) via the Department for Work and Pensions (DWP). RAPID captures data for anyone with a National Insurance Number (NINo). It includes people aged nearly 16 years, and over, who were alive at some point after 5 April 2010. RAPID brings together a coherent view of citizens' interactions with DWP, HMRC, and Local Authorities (via Housing Benefits) systems during the tax year. It collates information on individual activities within each tax year to enable a judgement to be made about whether a person is resident in the UK in that year.

There is a single dataset available to the ONS per tax year. The RAPID Migration Dataset uses methodology developed for the ONS to assist with the estimation of international migration to the UK, using data held in the RAPID dataset. Migrants are identified in the RAPID dataset using the Migrant Worker Scan (MWS) dataset. MWS is a record of non-UK nationals who have been issued with a NINo and these data are supplied to the ONS by HMRC on a quarterly basis. The RAPID Migration Dataset includes known and predicted arrivals and departures from the UK and categorises migrants as either short term or long term. Because of the methodology used to create the migrant worker scan (MWS), many children who arrive as migrants will not be captured on the MWS, and therefore are not included in this dataset.

All extracts used in RAPID cover the United Kingdom. RAPID is a dataset that continues to improve as new data sources and derived variables become available. In the most recent release of RAPID, for example, data from the Scottish Government will be included to take into account benefits issued in Scotland.

Higher Education Statistics Agency data

The Higher Education Statistics Agency (HESA) collects, processes, and publishes data about students enrolled at Higher Education (HE) providers in the UK. Data currently supplied to the ONS include:

- HESA Student Record
- HESA Student Alternative Record
- HESA Staff Record
- HESA Estates Management Record

All HE providers must submit data to either the HESA Student Record, HESA Student Alternative Record or the ESFA Individualised Learner Record. HESA data therefore provide excellent coverage of those in HE, including international students residing in England and Wales.

To produce long-term international migration (LTIM) estimates, only the Student Record data are used but in future we will expand this to include the Student Alternative Record data, which the ONS received for the first time in September of this year. From the 2022 to 2023 collection onwards, the "data futures" model, as explained on the HESA website, will bring the requirements of both the Student and Student Alternative records collections, into a new student data collection. Historically, student data were published separately for Alternative providers in England which, under the old terminology, were "privately funded" providers. This explains why the Student Alternative Record is still much smaller than the Student Record (during 2021 to 2022 just over 100,000 students, compared with around 2.75 million on the Student Record). The nature of the Student Alternative Record has evolved to be closely aligned to the Student Record, however, there are a small number of fields that providers submitting to the Student Alternative Record are not required to submit.

The Student Record Data supply includes:

- variables – person-level details of students in university and the courses they are studying
- geography – UK-funded HE providers
- population – "instances" of students registered at HE providers in the UK

For each academic year, the data were until recently delivered at the beginning of the next year, for example 2019 to 2020 data were delivered in January 2021. Data are collected throughout the academic year (1 August to 31 July) and submitted to HESA in October after the academic year ends. HESA then has a validation, collection and quality process that completes in November after collection. HESA passes the data to Jisc who prepare and deliver data to the ONS. Since this process has been in place, HESA has merged with Jisc, which is now the name of the supplier (while the data are still referred to as HESA). For the latest year of Student Record data, HESA's data collection schedule is available on their website.

For those enrolling at the start of the academic year, the lag in data delivery could be up to 17 months. Since 2020 to 2021, however, the ONS has an agreement with Jisc (formerly HESA) for early deliveries of Student Record Data on the condition that we do not publish before them. This brings the delivery date of this item from January or February to December. As part of Jisc's Data Futures Programme, the organisation is looking to move to an "in-year collection" meaning there will be two collections and releases of data per year by 2024 to 2025. This will greatly increase the timeliness of the data, which will improve the accuracy and quality of the overall estimates.

A data specification document provides a breakdown of data items delivered by HESA to the ONS. It includes the name of the dataset, the population of students covered in the dataset, names and descriptions of variables, and details on data delivery. It also specifies additional permitted uses.

3 . Communication with data supply partners

Home Office Borders and Immigration data

We hold regular meetings with the Home Office (HO) to discuss Home Office Borders and Immigration (HOBI) data. The Office for National Statistics (ONS) and HO have a Memorandum of Understanding (MoU) in place to facilitate the safe transfer of data to support joint ambitions to improve UK international migration statistics and statistical research. The MoU allows the data to be used for:

- data linkage projects (in certain, agreed circumstances only)
- studies of migration flows
- estimating migrant stocks
- understanding statistical quality of other admin data sources
- exploratory research on the travel behaviour of foreign nationals in and out of the UK

The ONS shares publications with HO for quality assurance purposes prior to release.

Registration and Population Interaction Database

There is a data sharing agreement (DSA) for non-disclosive data in place between the Department for Work and Pensions (DWP), HM Revenue and Customs (HMRC), Department for Communities Northern Ireland and the ONS, for the sharing of Registration and Population Interaction Database (RAPID) migration data. The DSA covers which data are to be shared and how often, the security measures in place and the retention and disposal policy for the data. The ONS maintains regular meetings and email contact with DWP and discusses any quality issues related to the data prior to the bi-annual publications.

Secure file transfer systems are used to send and receive RAPID migration data between DWP and the ONS. Once received, data are held in a secure area with role-based access. Only aggregated data are shared between DWP and the ONS in the format for publication. The ONS shares publications with DWP for quality assurance purposes prior to release.

Higher Education Statistics Agency data

The Data Growth and Operations (DGO) team of the ONS has meetings with Jisc to cover the datasets they supply. These are usually monthly but are more frequent before an upcoming delivery. Ad hoc emails are also sent between DGO and Jisc outside of these meetings.

A data sharing agreement (DSA), similar to the RAPID data DSA, is in place between Jisc and the ONS for the sharing of HESA data.

4 . Quality assurance principles, standards and checks applied by data suppliers

Home Office Borders and Immigration data

The Office for National Statistics (ONS) keeps documentation on the processes that Home Office Borders and Immigration (HOBI) data go through prior to arrival in the ONS. This includes quality checks completed by the supplier and any other information provided prior to the delivery of the data.

The Home Office (HO) performs regular quality assurance on the data. This is done to generate frequency counts for each non-ID field in the files, and a missingness rate for ID fields. HO then compares these against the counts for the previous extract, to look for any unexpected changes. If there are no issues, HO then passes the extracts and the frequency counts on to the ONS.

In addition to these, where changes to the extract have been agreed with the ONS, HO will normally perform more detailed checks to ensure these changes have been correctly applied. This will vary depending on the nature of the change but could, for example, take the form of spot checks on specific rows or identities within the data and comparing numbers to other sources of data where possible.

The HO's Developments in Exit Checks report describes recent challenges and changes in measuring migration, including changes resulting from the coronavirus (COVID-19) pandemic. The Home Office statistics on exit checks: user guide includes information on methodology and data quality. Until recently, the HO published regular Reports on statistics relating to exit checks with quality metrics for the HOBI data. These reports covered some metrics on coverage, completeness and matching rates for data linkage. Since the year-ending March 2020 report on statistics relating to exit checks, there have been no updates on the quality metrics for the HOBI data.

Prior to transfer from the HO to the ONS, quality checks are conducted by HO and ONS secondees at HO. Secondees from the ONS are able to work directly on HO data. The future ambition is for HO to complete these basic quality assurance (QA) checks, which will allow ONS secondees to focus on more in-depth QA.

Registration and Population Interaction Database

The ONS keeps documentation on the processes that Registration and Population Interaction Database (RAPID) data go through prior to arrival in the ONS. This includes quality checks completed by the supplier and any other information provided prior to the delivery of the data.

The Department for Work and Pensions (DWP) provides annual release documentation of their data processing in the development of RAPID. The latest version is Release 7, which covers data up to 5 April 2023.

RAPID is updated every year, in August. It takes data from existing sources including Customer Information System, Migrant Worker Scan, Child Benefit, Pay as You Earn, Tax Credit, Self-Assessment, Housing Benefit, benefits included as part of the National Benefits Database, and Universal Credit.

DWP derives multiple variables for each data source. This can include renaming variables, correcting any errors identified from analysis, removing null values, and identifying family units.

DWP holds quality information regarding purpose, administration, completeness, coverage, accuracy, and outliers on each data source used in RAPID. The data sources and relevant coverage information are listed below.

RAPID data sources and relevant coverage information

Customer information system

Data are collected by DWP and hold a single record of each person with a National Insurance Number who was alive in 2010 or later. This can include individuals who are no longer alive, but their deaths have not been reported to DWP. It excludes individuals who have not registered for a National Insurance Number.

Migrant worker scan

Holds registration details for people who register for National Insurance purposes. It includes all registrations from 1978. It excludes migrant registrations after state-pension age and child registrations.

Child benefit

Covers information on child benefit claims including details on the main payee and inclusion and exclusion dates of the children involved. Data are reliable from 2008 to 2009 onwards although any awards finishing between 6 April and October 2008 are not included.

Pay As You Earn (P14)

Includes data from 2008 to 2009 to current tax year.

Tax credit

Records for DWP customers and non-DWP customers in receipt of tax credit benefits.

Single housing benefit extract

Data go back to the 2008 to 2009 tax year.

DWP legacy benefits

Includes data on the following benefits:

- State Pension (SP)
- Pension Credit (PC)
- Widows Benefit (WB)
- Bereavement Benefit (BB)
- Bereavement Support Payments (BSP)
- Attendance Allowance (AA)
- Disability Living Allowance (DLA)
- Incapacity Benefit (IB)
- Passported Incapacity Benefit (PIB)
- Severe Disablement Allowance (SDA)
- Industrial Injuries Disablement Benefit (IIDB)
- Industrial Death Benefit (IDB)
- Employment Support Allowance (ESA)
- Invalid Care Allowance (ICA & CA)
- Income Support (IS)
- Job Seekers Allowance (JSA)
- Maternity Allowance (MA)

Personal Independence Payment

Introduced in April 2013, available from this date onwards.

Universal Credit

Only available from January 2013 and partners are only available from August 2015. Northern Ireland data only from 2017. Amounts relate to amount paid not entitlement.

Self-employment

Assesses tax liabilities for the self-employed or any income not captured by PAYE.

Furlough

Covers all Coronavirus Job Retention Scheme and Self-Employed Income Support Scheme claims. Schemes were from March and May 2020 to 30 September 2021 and therefore the coverage for this extract is only within this time period.

Adult Disability Payments and Child Disability Payments

Scottish Government devolved benefits. Data available from 2022 to 2023.

Higher Education Statistics Agency data

The ONS keeps documentation on Higher Education Statistics Agency (HESA) Student Record Data that records what happens to the individual admin data source from the point of data collection outside the ONS to where teams use the data for analysis inside the ONS.

Data supplied to Jisc are subject to an extensive quality assurance process with a range of automated validation checks that are applied to all submissions. Providers first validate the data themselves, as explained on HESA's Validation overview webpage and then Jisc puts the data through quality rules, as shown on HESA's Quality Rules Directory webpage. If the data fail this check, they are returned to the university to be corrected.

Jisc collects data from higher education (HE) providers and provide resources on HESA's Data Collection webpage to support the data collection process, which ensures data are coherent. It is mandatory for HE providers to report data to HE funding and regulatory bodies.

Data are collected by HE providers through enrolment (made online or by post) and the main facilitator of this is applications through the University and Colleges Admissions Service (UCAS). A verification service run by UCAS flags applications for potentially fraudulent activity, missing and misleading information or potential duplicates, for further investigation.

Address data (postcode) are collected at the start of a student's period of study and may not be updated again. Because of the coronavirus (COVID-19) pandemic, HESA issued guidance on their COVID-19 exceptional guidance for collections webpage on how HE providers should collect the required student data. The coronavirus pandemic increased the likelihood that a student's term-time postcode reflected where they intended to reside, rather than where they were actually residing, potentially reducing the accuracy of data about their location.

5 . Producer's quality assurance investigations and documentation

Home Office Borders and Immigration data

Once the Office for National Statistics (ONS) has received the Home Office Borders and Immigration (HOBI) data, the data are quality assured in more detail to ensure they meet ONS requirements and are fit for purpose. On HOBI data this includes:

- validation – verifying the dataset received in the ONS against requirements, metadata, and knowledge of the data
- counts – counting the number of records and looking at the coverage of the data to ensure they are consistent with expectations (such as previous extracts and other data)
- duplication and replication – counting how much duplication or replication there is in the dataset
- missingness – counting the number of nulls, blanks or coded missingness found in each variable in the dataset
- unexpected entries – counting how many unexpected or inappropriate responses there are in the dataset, for example how many special characters are in the dataset, and how many inappropriate dates of birth
- quality of linkage – counting how many links can be made to the previous supply of data via unique IDs or match keys, and summarising how much missingness there is in match key variables
- longitudinal sense checks – counting the top 20 frequencies and comparing all the results from the above checks across years or supplies of data

We use Home Office asylum seeker data to make an adjustment to the HOBI data, removing potential duplicates from the data.

Registration and Population Interaction Database

We look at the quality assessment carried out by DWP on the dataset and consider the impacts on our long-term international migration estimates. The risks of potential errors feeding through to our estimates are deemed to be low. The relevant quality issues assessed include:

- error
- coverage
- timeliness

The completion of the Quality Assurance Error Framework identified some coverage issues within the RAPID Migration dataset. These coverage issues include:

- those under the age of 16 years
- student population
- those who have recently arrived in or departed from the UK and therefore may not meet the activity requirements to be considered long-term international migrants (LTIM)

In the case of those under 16 there is an acknowledged coverage gap in RAPID and therefore those aged under 16 years are completely excluded in the aggregates derived from RAPID and a separate statistical adjustment is made to LTIM estimates, to avoid introducing bias in the estimates.

Statistical adjustments are also calculated separately for students, using the HESA data and historical trends in the RAPID data for the recent arrivals/departures adjustment.

Further details on these can be found in our International migration research, progress update: November 2022.

Higher Education Statistics Agency data

The ONS conducts quality assurance (QA) checks on the Higher Education Statistics Agency (HESA) dataset. This includes QA of the data covering the 2021 to 2022 academic year supplied in December 2022 (for student data) and February 2023 (for other HESA data, such as for staff and estates).

Some additional quality checks were taken on the data during the coronavirus (COVID-19) pandemic, including an in-depth breakdown of term-time postcode missingness during the pandemic. This found that doubling in term-time postcode missingness could be almost entirely attributed to the Open University changing their reporting of the variable. Other findings on changes to collections during COVID-19 were generally as expected.

The Quantitative Quality Indicators (QQI) for the HESA dataset (for 2016-2021) are provided on Worksheet 6 of our QQI dataset.

6 . Cite this article

Office for National Statistics (ONS), released 16 November 2023, ONS website, methodology, Long-term international migration: quality assuring administrative data