National Records of Scotland

Administrative Data Based Population Estimates, Scotland 2016

Quality Assurance of Administrative Datasets

Published on 17 November 2020

Disclaimer: The Administrative Data Based Population Estimates are **not the OFFICIAL STATISTICS** for Population Estimates for Scotland. The Official Statistics can be found at the statistics and data section of National Records of Scotland's website.

Contents

1. Disclaimer

The Administrative Data Based Population Estimates are statistical research outputs. These estimates **should not** be considered as a replacement for the National Statistics publication: Mid-Year Population Estimates for Scotland.

2. Introduction

This document summarises how the quality of the administrative data based population estimates is affected by the source datasets considering the business rules used to combine them. This document also provides details about how the data sources were quality assured prior to linkage to ensure they were suitable for this project.

This information supports our compliance with the UK Statistics Authority and the Office for Statistics Regulation's Code of Practice for Statistics. In particular this document provides evidence against the first and third principles within the Quality pillar of the Code of Practice which are listed below:

> Principle Q1 - "Statistics should be based on the most appropriate data to meet intended uses. The impact of any data limitations for use should be assessed, minimised and explained."
>
> Principle Q3 - "Producers of statistics and data should explain clearly how they assure themselves that statistics and data are accurate, reliable, coherent and timely."

The quality assurance arrangements that are required for statistics compiled using administrative data for compliance with the Code of Practice were clarified in a regulatory standard issued by the UKSA in January 2015. The information in this standard was supported by an Administrative Data Quality Assurance Toolkit to provide guidance for statistical producers.

3. Overall quality of the administrative data based population estimates

The administrative data based population estimates have been produced by linking a variety of datasets. How these datasets are used is dictated by a series of business rules which are used to define which individuals are included in the administrative data based population estimates. These business rules mean that certain datasets have greater importance to the creation of the administrative data based population estimates and therefore have a greater impact on quality. Full details are described in the Methodology Report.

The dataset that is of greatest importance to the administrative data based population estimates is the National Health Service Central Register (NHSCR). The NHSCR has the greatest impact as all individuals included in the administrative data based population estimates must exist on the NHSCR, apart from those aged zero. Zero year olds will also be included if they appear in the birth registration data without appearing on the NHSCR, as there can be a small time lag between these two datasets being updated.

As the NHSCR contains everyone who has registered with a GP in Scotland at some point and everyone born in Scotland since 1939, there are many records for people who are no longer part of Scotland's population. Therefore filter rules are used to reduce the dataset to those who are alive and still appear to be living in Scotland. The quality of the administrative data based population estimates therefore rely on this subset of NHSCR records having good coverage of Scotland's population.

The NHSCR still has some over-coverage even after filtering. This can happen when people move abroad and do not de-register from their GP. Therefore the other datasets are used to provide additional evidence that an individual from the NHSCR should be retained in the administrative data based population estimates, or removed.

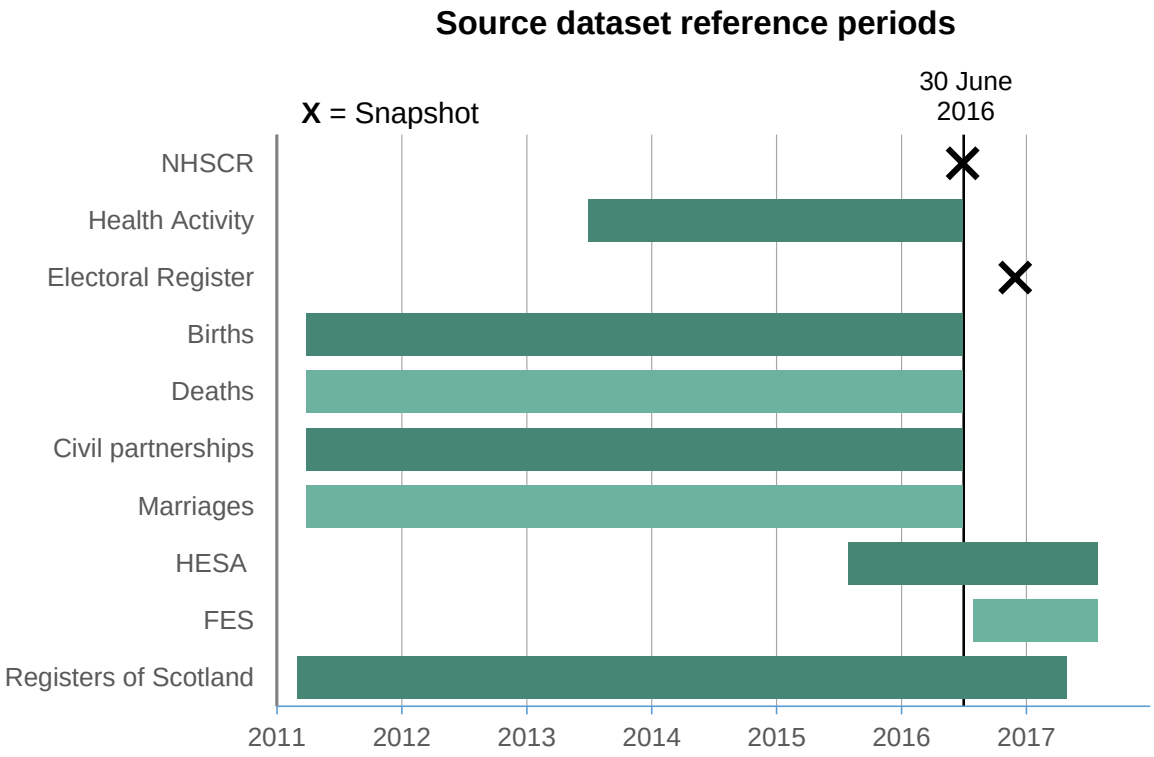
Strengths

- By requiring that individuals appear in more than one dataset, the over-coverage of the NHSCR is mitigated to some extent.
- Similarly, for datasets other than the NHSCR and birth registrations, the impact of any potential over-coverage will be reduced as this will be mitigated by the person also having to appear on the NHSCR as alive and in Scotland.
- These estimates are produced from a dataset of de-identified data at individual level rather than being produced from aggregate counts. This has the potential to allow more accurate migration information to be produced by linking data across different years.

Limitations

There are several reasons why someone who is part of Scotland's population may be missing from the administrative data based population estimates. These include:

- Any individuals who have not registered with a GP in Scotland will not be included as they will not be part of the subset of filtered NHSCR records. This would be the case even if the individual appears in the other datasets.
- Some people who are part of Scotland's population will appear on the NHSCR but will not be present in any of the other datasets. These people will be removed through the application of the business rules.
- Linkage is not perfect, and therefore inconsistencies with how an individual's data is recorded between datasets will mean that some links are missed. These inconsistencies could be caused by errors during data collection, or by the individual providing different information for each data collection. This could lead to a person being excluded from the administrative data based population estimates as they appear to be on the NHSCR but no other datasets when this is not actually the case.
- Similarly, differences in the reference period for each dataset, shown in the following diagram, will lead to some inconsistencies in the data. Again this will mean links between datasets could be missed if the information about a person changes, over this time, for example if they move home or change their name.



Individuals may also be included in the administrative data based population estimates when they should be removed.

- If an individual has not informed their GP that they have moved out of Scotland, they could still be included. For example, if someone moved to France in December 2015 without updating their GP, they may still appear on the NHSCR as living in Scotland. They may also appear on other datasets such as the Health Activity dataset if they had used health services between June 2013 and June 2016 and therefore be included in the administrative data based population estimates.

Risk/Profile Assessment

The matrix below reflects the levels of risk of data quality concerns and the public interest profile of the administrative data based population estimates. These have been determined by a review undertaken by the NRS Administrative Data team using the information contained within the Office for Statistics Regulation's Administrative Data Quality Assurance Toolkit.

Level of risk of quality concerns	**Public interest profile**		
	Low	**Medium**	**High**
Low	Statistics of low quality concern and low public interest. [A1]	Statistics of low quality concern and medium public interest. [A1/A2]	Statistics of low quality concern and high public interest. [A1/A2]
Medium	Statistics of medium data quality concern and low public interest. [A1/A2]	Statistics of medium quality concern and medium public interest. [A2]	Statistics of medium quality concern and high public interest. [A2/A3]
High	Statistics of high data quality concern and low public interest. [A1/A2/A3]	Statistics of high quality concern and medium public interest. [A3]	Statistics of high quality concern and high public interest. [A3]

*A1/A2/A3 – definitions supplied Office for Statistics Regulation's Administrative Data Quality Assurance Toolkit.

The Public Interest profile has been set as “medium” for the following reasons:

- One of the objectives of the Administrative Data Based Population Estimates is to support future recommendations for the census beyond 2022.

- There is a strong interest in the viability of administrative data based population estimates to maximise the use of all available data sources to provide accurate and timely evidence to measure our population.

The risk of quality concerns has been set to “medium” for the following reasons:

- The administrative data based population estimates have produced figures that are broadly comparable at Scotland level with the official mid-year population estimates. These results are encouraging however we are aware that future improvements to the methodology and possible additional datasets are required to further improve the quality of the estimates. This is discussed in the main publication and the methodology report.
- Several administrative datasets are provided by external data suppliers. This means that the data could be subject to change from year to year depending on requirements of the data for that supplier. We will continue to communicate with data suppliers to understand the data they provide and how any changes could impact this project.

4. Source dataset information

National Health Service Central Register (NHSCR)

Data supplier	National Records of Scotland (NHSCR)
Supplier info	National Records of Scotland (NRS) is a Non Ministerial Office of the Scottish Government. The purpose of NRS is to collect, preserve and produce information about Scotland's people and history and make it available to inform current and future generations. The NHSCR branch of NRS is responsible for maintaining the NHSCR, an electronic demographic database of all people born in Scotland, died in Scotland and those who have ever registered with a GP in Scotland.
Data type (counts or unit records)	Unit records
Data content	The following variables are included at an individual record level: • First name • Middle name • Last name • Previous names • Sex • Birthdate • Birth country • Death date • NHS Number (Scottish, England/Wales and Northern Irish numbers) • Person ID • Postcode • Date postcode was recorded • Posting (indicates which health board the person has registered to a GP in)
Time period covered	Extract as at 30 June 2016
Supply schedule	Annually
Use of data	Production of administrative data based population estimates as statistical research

NHSCR: Background information

The NHSCR is an electronic index for:

- every patient registered, now or in the past, with a Scottish general medical practitioner (GP);
- everyone born in Scotland since 30 September 1939, who have not been registered with a Scottish GP;
- patients formerly registered with a Scottish GP, who died after 29 September 1939.

The main purpose of the register is to permit the efficient movement of patient's medical record envelopes when they:

- transfer between Scottish Health Boards and health authorities in the rest of the UK;
- leave the country;
- join the Armed Forces (or are dependants of Armed Forces personnel).

The key inputs into the NHSCR are:

- Births (in Scotland)
- Deaths (from across the UK)
- GP Registration (within Scotland) – ‘migration’ into Scotland
- GP Registration (within the rest of the UK) – ‘migration’ out of Scotland

NHSCR: Data supply and communication

The data provided is done so annually under the terms of a data sharing agreement and includes record level data for a selection of variables as defined in a data sharing agreement.

The data is sent to the admin data team by the NHSCR team via approved NRS data transfer procedures as agreed in a data sharing agreement.

NHSCR: Quality assurance undertaken by data supplier

The data entered by staff is regularly scrutinised. Supervisors check 5% of the work undertaken by staff each day to identify any potential training issues. These records are randomly selected based on subject matter, taking into account new areas of work, trends or concerns previously identified. This also helps the NHSCR to meet its service level agreement with the Scottish Government, NHS National Services Scotland which requires an accuracy level of 97%, which is currently being achieved.

As well as this, the NHSCR team undertake a variety of data quality initiatives on an annual/bi-annual basis where staff investigate the population of different variables in the register and to correct duplicates. These initiatives are carried out relatively

frequently as they target areas of known concern and the findings are generally kept internal to the NHSCR team. These data quality initiatives include:

- investigating records where no death has been recorded for a person aged over 110 years old. In the majority of cases a death is traced (these are usually deaths that were missed at the time, usually from the 1970s or 1980s before the NHSCR was computerised) and the record is updated to reflect this.
- checking records where the postings variable is blank. This allows us to be confident that all records that should have a posting do. Where no posting exists it is usually for persons who are born in Scotland but they never registered with a Scottish NHS GP.
- populating records that do not have a Community Health Index (CHI) number[1] either with the CHI number if one exists or with a flag to show that there is not a CHI number for that record.

Extracts of the NHSCR are used by various statistical teams across the National Records of Scotland for a variety of purposes. NHSCR also collects feedback from these users of the NHSCR extracts where anomalies are identified and investigates these anomalies so a resolution or explanation can be found.

NHSCR: Quality assurance undertaken by the NRS Admin Data team

Once the admin data team receive the data, a number of data consistency and validation checks are performed, including:

- Checking the proportion of missing values for variables.
- Checking the validity of postcodes
- Checking the distribution of the population across different council areas and comparing this to previous years and/or existing population estimates.
- Checking the distribution of the day and month elements of dates of birth
- Checking the age distribution of the population.
- Checking that variables that should be unique are unique.

These checks are largely programmed with the output flagging up any anomalies, although analysts do also look at a small sample of records to spot any issues.

If these checks suggest the data may need to be amended/adjusted then the potential issues are communicated with the data supplier so the register can be amended if appropriate. However in this case these checks did not identify any issues with the data so this was not required.

[1] https://www.ndc.scot.nhs.uk/Dictionary-A-Z/Definitions/index.asp?ID=128&Title=CHI%20Number

NHSCR: Strengths and limitations

The strengths and limitations in the table below relate to the use of the dataset for administrative data based population estimates rather than the original purpose of the data.

Strengths	Limitations
• NHSCR is a comprehensive source of record level data that covers the vast majority of the population. • The data contains all of the variables used to link with other data sources (name, date of birth, postcode and sex)	• Mostly does not include address information beyond postcode. There is a Unique Property Reference Number (UPRN) variable, however this variable is completed for less than 25 per cent of records. • It does not pick-up people who leave the UK (unless they informed their GP) leading to some inflation in the register • Moves within Scotland cannot be picked up until the patient registers with a new GP. As a result some people will be recorded in the wrong area. Particularly an issue among younger adult males. • There will be a lag in recent migrants into Scotland appearing on the NHSCR as they will only appear when registering with a GP. • There is a delay in new born babies appearing in NHSCR with a postcode (and posting) until they are registered with a GP.

Health Activity

Data Supplier:	Public Health Scotland (PHS)
Supplier info:	Public Health Scotland is Scotland's lead national agency for improving and protecting the health and wellbeing of all of Scotland's people. PHS's vision is of a Scotland where everybody thrives. PHS's focus is on increasing healthy life expectancy and reducing premature mortality. To do this, they use data, intelligence and a place-based approach to lead and deliver Scotland's public health priorities.
Data type (counts or unit records)	Unit records
Data Content:	The following variables are included at an individual record level: • Unique ID • Surname • First Forename • Second Forename • Previous Surname • Date of Birth • Sex • Patient Structured Address • Full Patient Postcode • General Practitioner Practice Postcode • Marital Status (where information is available) • Ethnic group (where information is available) • Transfer out flags • Last Interaction – month and years of last interaction with health service • Encrypted CHI number
Time Period Covered	Data extract at 30 June 2016, with 'Last Interaction' variable covering previous 3 years
Supply Schedule:	Twice a year
Use of Data:	Production of administrative data based population estimates as statistical research

Health Activity: Background information

The Community Health Index (CHI) is the main linking key which is used in Scotland for health care purposes. The register exists to ensure that patients can be uniquely identified, and that all information pertaining to a patient's health is available to providers of care. No single body has responsibility for CHI; the data controllers for CHI are the 14 National Health Service (NHS) Boards. An extract called the Health Activity Dataset was created for this project by PHS. No individual health data was supplied, only an activity flag of last time they used a NHS service.

The variable of interest for this project is 'Last Interaction'. This variable reports date of an individual's last engagement with a health practitioner as reported through Primary Care and Secondary Care datasets. The Primary Care dataset is based on Dental, Screening Services (Abdominal Aortic Aneurysm (AAA) and Bowel), and Prescribing through Community Pharmacist and Dispensing Contractors delivering primary care across Scotland; the Secondary Care dataset reports on Day Case and Outpatient Hospital appointment, Maternity and Mental Health episodes, Accident & Emergency. That date provides an up-to-date population register that can help confirm population estimates in any time period. This variable is sent directly to our secure processing site by PHS, with unique identifiable key for subsequent linking as per their data processing agreement.

Health Activity: Data supply and communication

The data is provided twice a year under the terms of a data sharing agreement. The data is sent by PHS to the administrative data team via approved NRS data transfer procedures as agreed in a data sharing agreement.

The 2016 Health Activity dataset data comprises of two files: 5,600,000 records in a Primary Care dataset and 3,700,000 in the Secondary Care dataset

This data was supplied to NRS in 2017 but due to NRS project delays, it was not processed until June 2020. At this point, it was discovered that the linking key that allow the 'Last Interaction' variable to be matched was not available. Though NRS and PHS have been working closely to update this oversight, it has been over the period when COVID-19 lockdown restrictions have been in place. This has caused IT issues due to the size of the datasets. The organisations decided to proceed without this variable at the time. It was felt that the other datasets in the project could be used to help with activity. It is our hope that the next publication will contain the revised variable and this will allow us to assess the importance of this variable to the methodology.

Health Activity: Quality assurance undertaken by data supplier

PHS perform internal quality assurance processes before sharing data. General data management includes checks on *completeness* and *timeliness,* with dataset specific checks as set out below.

- *Completeness*
  NHS data providers will know how complete their Scottish Morbidity Record (SMR) datasets are and the extent of any backlog. SMR data is expected to be received by PHS 6 weeks following the end of the month of discharge or clinic date. Historically, this target has been achieved with a national return of 98% or higher. Data Support and Monitoring – SMR Completeness

- *Timeliness*
  The Scottish Government target for SMR submission to ISD is 6 weeks (42 days) following discharge/transfer/death or clinic attendance. ISD calculates timeliness as data received 6 weeks following the end of month of discharge/transfer/death or clinic attendance, tracking any backlog as well as highlighting number of records that were submitted after the 6-week target. Data Support and Monitoring – SMR Timeliness

- Quarterly trends in Inpatient and Day Case Activity provides completion data on SMR01 (relating to acute and general hospital discharges) and SMR00 (new and return outpatient attendances). Acute and general hospital discharges are estimated to be 98% complete for 2015/16 and 96% complete for quarter ending Sept 2016. The completeness figures quoted are for new outpatient attendances are 99% complete for 2015/16.

- Five main entries from the Scottish Morbidity Record (SMR) datasets feed into the Health Activity dataset, namely:

  - SMR00 Outpatients
  - SMR01 General Acute Inpatients/Day Cases
  - SMR02 Maternity Inpatients/Day Cases
  - SMR04 Mental Health Inpatients/Day Cases
  - SMR06 Cancer Registrations

  Validation is either carried out locally and prior to submission to PHS or centrally at PHS. A set of validation rules is carried out by the data provider, where checks may generate:

  - Errors where the information recorded is missing, invalid or fails to conform to a logical sequence of events, or
  - Queries where the information recorded appears to be infeasible but is found to be to be correct.

  Automatic checks are made to see if a record already exists with the same or similar DOB, Name, Gender, and Address. Validation on address is performed by looking up Quick Address Software (QAS). PHS rely on users who have update access to enter address information correctly, with address changes triggered by patients through GP system or added by hospitals for new patients not yet registered with a GP. The National Health Service Central

Register (NHSCR) is used to update the main PHS records on changes/embarks from Scottish Health Boards, but NHSCR is not involved in addressing of PHS records i.e. they are independent of one another insofar as data entry is concerned.

- Quality assurance measures are in place for data that is sourced from other Primary care providers (Dentist, Community Pharmacist and Dispensing Contractors). For example, in the case of dental treatment, registration and participation data is collected. At September 2016, 3.5 million (72%) of registered patients had seen an NHS dentist within the last two years. Two types of checks are made as payment verification of the General Dental Services (GDS) payment database (see PSD - Dental). The Level 1 checks cover a range of validations to ensure the payment claims submitted meet the criteria laid down within the Statement of Dental Remuneration. Checks include search for duplicates and overall reassurance that the information is correct with regards to the claims being made by the dentist. These are run against all records before being accepted onto the database. They check the quality of the data held in the database. In 2015/16, 3.5% of claims submitted were returned for clarification as they did not meet the specified criteria/were duplicates etc. Levels 2-4 checks are designed to determine fraudulent claims by dentists.

  For Pharmacists and Dispensing Contractors the data is sourced from a payment system and routine monthly checks are carried out by Practitioner Services on a random sample of approximately 5% of prescription payments. These check all data captured for payment and the accuracy of the payment calculation and have a target accuracy of 98% which is routinely met. Data that is captured but is not mandatory for payment purposes can be of lower quality; principally this includes the prescriber code which links a prescription back to the individual prescriber (e.g. GP) and their organisation (e.g. practice or NHS Board). https://www.isdscotland.org/Health-Topics/Prescribing-and-Medicines/_docs/Publication-of-Prescribing-Data-Summary-Report-v6.pdf

Health Activity: Quality assurance undertaken by NRS Admin Data team

Once the Admin Data team receive the data, a number of data consistency and validation checks are performed on Health Activity dataset data prior to de-identification and transfer to safe haven. Those checks include:

- Checking the proportion of missing values for variables.
- Checking the validity of forename, surname and postcodes.
- Sense checking the number of records by single year of age compared to published information (Mid-Year Estimates for 2016).

These checks provide additional information to NRS team when linking data to produce population estimates in the safe haven.

Health Activity: Strengths and limitations

The strengths and limitations in the table below relate to the use of the dataset for administrative data based population estimates rather than the original purpose of the data.

Strengths	Limitations
• Health Activity dataset is a comprehensive source of record level data that covers the vast majority of Scotland's population • High quality data administered by PHS. Also able to use an active flag that gives us a time indication for interaction with the Health service.	• Moves within Scotland cannot be picked up until the patient registers with a new GP. As a result, some people will be recorded in the wrong area. Particularly an issue among younger adult males. • Due to the number of datasets being used to create this dataset there may be a small percentage that are not linked correctly. • The 2016 dataset could not be linked to Last Interaction variable but we do know that there was an interaction within the last 3 years.

Scottish Pupil Census (SPC)

Data Supplier	Scottish Government: Education Analytical Services (EAS)
Supplier info	EAS provides data on school pupils through an annual pupil census that captures characteristics of pupils. This QAAD is based on the data from the census that took place in September 2016. The SPC forms part of 'Summary statistics for schools in Scotland', an annual publication that describes the education system in terms of the number of schools and pupils, the types and sizes of schools and classes they learn in, and some characteristics of the pupils.
Data type (counts or unit records)	Unit records
Data content	The Pupil Census covers all publicly funded schools in Scotland (local authority and grant-aided). Pupils in this census are those recorded by a LA as being on the roll of the school except those in full time education at another institution. Variables on pupil record (as requested in data sharing agreement): • Scottish Candidate Number (SCN) • Home postcode • Gender • Date of Birth as DD/MM/YYYY (10) • Ethnicity (self-identified from categories used in 2011 Census) • National Identity (self-identified from categories used in 2011 Census) • English as Second Language • School seed code
Time period covered	Collect in September 2016 for 2016/17 academic year
Supply schedule	Annual
Use of data	Production of administrative data based population estimates as statistical research

SPC: Background information

Data is collected from all Local Authority and Grant-aided schools and school centres. Submission of data by local authorities is mandatory. Local Authority (LA) management systems that are the source of data, using ScotXed, and as such prove to be a strong driver in ensuring data are correct.

SPC: Data supply and communication

The data is provided annually under the terms of data sharing agreement and includes record level data for a selection of variables as defined in a data sharing agreement for every pupil based on unique identifier of SCN. Although data sharing agreement included pupil's home address, that data is not transferred to ScotXed and is therefore not included in our cut. Address data is captured through home postcode to support geographic analyses of cases.

SPC: Quality assurance undertaken by data supplier

LA systems for supplying data have built in validation checks using the procXed Data Collection System; validation checks agreed with data providers are regularly updated, and Head Teachers sign off summary tables that are used.

SPC: Quality assurance undertaken by the NRS Admin Data team

Once the admin data team receive the data, a number of data consistency and validation checks are performed, including:

- Validation of postcode data to identify missing or incomplete postcodes. This check also identified a small number of out-of-country postcodes, reflecting potential cross-border residency.
- Checking the age distribution of 'pupils', noting out-of-age cases including adult learners and those beyond compulsory education phase. Primary analysis will be completed on core 'pupil' population, dropping children under 3 and adult learners over 19 years. A further option is to reduce to compulsory education ages of 5 to 16 year age groups where returns are likely to be more representative of the population, but at the moment all ages have been included to serve as an activity flag.
- Checking for duplicate 'pupils' where SCN is registered with more than one school, retaining one record per 'pupil'. In the absence of data to signify main school / educational establishment, all bar one record was randomly retained for population and household estimation purposes.
- Validating the distribution of LA population data by single year of age, against relevant Mid-Year Estimates (MYE) of population[2]. This check

[2] Mid-2016 Population Estimates for Scotland from the National Records of Scotland at https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/population/population-estimates/mid-year-population-estimates/archive/mid-2016/list-of-tables .

provides evidence of an under count, attributable in part to private education and home schooling of learners in those age groups. At a national level the under-count amounts to an average of 6.5% of the MYE for ages 5 through 16. Excluding those end points that are themselves less representative of the population due to variable school starting age and school leavers, the average under-count reduces to 4.4% of MYE.

The validation checks of SPC data will be discussed with EAS to consider whether anything can be done to improve the data for administrative data based population estimates. This may result in revisiting the data sharing agreement to get an extra variable to reduce duplication.

SPC: Strengths and limitations

The strengths and limitations in the table below relate to the use of the dataset for administrative data based population estimates rather than the original purpose of the data.

Strengths	Limitations
• SPC data is a comprehensive source of record level data that covers the vast majority of school age population • High quality data administered by LA through ScotXed and EAS division of Scottish Government. • Data includes home postcode making SPC a good dataset for creating/ confirming or validating administrative household estimates. • SPC is an annual data collection that the Scottish Government has run for decades and it is classified as a National Statistics Publication.	• Full address information is not collected by EAS –only having postcode may limit linking exercise. • Lack of any flag to signify main school is a limitation of the data where SCN is associated with multiple centres – explore options to have this included in any future data sharing agreement for SPC data. • No information on independent sector, home schooling, hospital education etc. as out of the scope of this data collection.

Higher Education Statistics Agency (HESA)

Data supplier	Higher Education Statistics Agency (HESA)
Supplier info	HESA are the experts in UK higher education data. They collect, assure and disseminate data about higher education (HE) in the UK on behalf of their Statutory Customers. HESA works with HE providers in each of the four nations of the United Kingdom, collaborating with them to collect and curate one of the world's leading HE data sources.
Data type	Unit records
Data content	The following variables are included at an individual record level: • Forename(s) • Surname • Surname at 16 if different from above • Sex • Birthdate • Nationality • Term-time postcode • Unique Identifiers (Unique Learner Number, Scottish Candidate Number, HESA Unique Student Identifier) • Postcode of permanent home address • Date studies started • Date studies ended • UKPRN (UK Provider Reference Number - for establishment registered at) • Expected Length of study • Year of student instance • Year of course • Location of study • Suspension of active study flag The population covered in this data is all students studying at Scottish higher education providers (including The Open University) and Scottish domiciled students studying at higher education providers in England, Wales and Northern Ireland.
Supply schedule	Annually
Time period covered	2015-16 and 2016-17 academic years
Use of data	Production of administrative data based population estimates as statistical research

HESA: Background information

The HESA Student record has been collected since 1994/95 from subscribing Higher Education Providers (HEPs) throughout the devolved administrations of the United Kingdom. The data collected as part of the Student record is used extensively by various stakeholders and is fundamental in the formulation of:

- Funding
- Performance Indicators
- Publications (including UNISTATS)
- League tables

The aggregated figures from this data are used by HESA in their annual National Statistics publication 'Higher Education Statistics for the UK', links for the relevant years for the data used here are below:

2015-16: https://www.hesa.ac.uk/news/12-01-2017/sfr242-student-enrolments-and-qualifications

2016-17: https://www.hesa.ac.uk/news/11-01-2018/sfr247-higher-education-student-statistics

HESA's Quality Report (link below) provides some additional information on uses of student data in the 'Relevance' section.

https://www.hesa.ac.uk/about/regulation/official-statistics/quality-report

For the years covered in this report, the Student record collects individualised data about students active during the reporting period. The reporting period is from 01 August year 1 to 31 July year 2, for example, the 2015/16 Student record was collected in respect of the activity which took place between 01 August 2015 and 31 July 2016. Students who are studying overseas or who come to the UK for a period of less than 8 consecutive weeks during their programme of study are not included in the Student record.

HESA: Data supply and communication

The data is supplied by Higher Education providers to HESA via a secure web-based transfer system created and maintained by HESA. The data supplied are subject to an extensive quality assurance process.

The data provided to NRS by HESA is shared under the terms of a data sharing agreement. The data includes record level data for a selection of variables for all students studying at Scottish higher education providers (including The Open University) and Scottish domiciled students studying at higher education providers in England, Wales and Northern Ireland.

HESA publish extensive information about the collection of the data, the validation process used and any known issues with the data on their website.

For 2015/16 this information is found at:

https://www.hesa.ac.uk/collection/c15051

https://www.hesa.ac.uk/collection/c15051/support-guides

For 2016/17 this information is found at:

https://www.hesa.ac.uk/collection/c16051

https://www.hesa.ac.uk/collection/c16051/support-guides

The list of known issues and anomalies is found at:

https://www.hesa.ac.uk/support/data-intelligence

HESA: Quality assurance undertaken by data supplier

HESA produce a student record quality report[3] that explains how they assure themselves that the data is accurate, reliable, coherent and timely.

As mentioned in the 'Data supply and communication' section, HESA has developed extensive quality assurance procedures and runs a range of automated validation checks (quality rules) against all submissions from data providers. When submitting final data the provider must pass various rules that ensure the data is in the correct format and does not trigger any validation errors. In the situation that correct data still triggers these validation errors, the provider must contact HESA to provide an explanation.

These rules[4] include, but are not limited to:

- checking unique identifiers are valid by using a checksum
- providing a warning when personal information submitted for a student does not match the previously sent information for the student.
- only allowing dates of birth to be in a certain range if date of birth is provided
- showing an error if it appears that forename and surname have been transposed compared to the last year's submission.
- warning if more than 2% of students have 'other' recorded for sex incase this is due to a systematic error.
- error if all students have been returned with the same sex code as a range of codes is expected

[3] https://www.hesa.ac.uk/about/regulation/official-statistics/quality-report
[4] 2015/16: https://www.hesa.ac.uk/collection/c15051/quality-rules
2016/17: https://www.hesa.ac.uk/collection/c16051/quality-rules

- warning or error if the number of students have the same term-time postcode without being marked as living in halls of residence exceeds specified thresholds
- a postcode must be recorded for all UK domiciled students

Data Quality Analysts at HESA then examine the data to ensure the submission is credible. This is an iterative process during which providers may need to submit and review several times before signing off the data to ensure the final submission is credible.

HESA: Quality assurance undertaken by the NRS Admin Data team

Once the admin data team receive the data, a number of data consistency and validation checks are performed, including:

- Checking the proportion of missing values for variables
- Checking that variables are in the expected formats and values
- Checking the validity of postcodes
- Comparing the data with similar data received in previous years and with published data about students in Scotland to check that trends and patterns appear to be correct.
- Checking the distribution of the day and month elements of dates of birth
- Checking the age distribution of the population.
- Removing duplicate records where identical information is recorded (this can occur if an individual enrols on multiple courses in the academic year).

If these checks suggest the data may need to be amended/adjusted then the potential issues are communicated with the data supplier so the data can be amended if appropriate. However, this was not required after checking these data sets.

HESA: Strengths and limitations

The strengths and limitations in the table below relate to the use of the dataset for administrative data based population estimates rather than the original purpose of the data.

Strengths	Limitations
• A considerable proportion of records in this data are for young adults who can be difficult to identify in other datasets. This dataset should therefore be particularly valuable in improving estimates of young adults. • As the data includes term-time and home postcode, it may be able to resolve issues where postcodes differ for one individual in other datasets. • Contains some previous surname information so have an improved chance of making links where surname has changed. • Extensive validation process by the data supplier and HESA to make the data as complete as possible.	• There is a lag in being able to receive the data. For example 2016/17 data is only available in early 2018. This could therefore impact on when the most up-to-date population estimates could be published. • Only provides data on a specific subset of the population. Even in the age groups where this data will be most beneficial (i.e. young adults) there will be a considerable proportion of the population that will not appear here if they did not attend higher education.

Further Education data (FES)

Data supplier	Scottish Funding Council (SFC)
Supplier info	The SFC is a Non Departmental Public Body of the Scottish Government. The SFC invests around £1.8 billion a year in Scotland's 19 universities and 26 colleges (within 13 college regions) for learning and teaching, skills development, research and innovation, staff, buildings and equipment.
Data type	Unit records
Data content	The following variables are included at an individual record level: • Forename(s) • Surname • Sex • Birthdate • Nationality • Religion • Ethnicity • Does the student have a disability • Pre-study domicile • Postcode of permanent home location (pre-study domicile of student) • Student Matriculation Number • Date studies started • Date studies ended • College attended • Mode of attendance
Supply schedule	Annually
Time period covered	2016/17 academic year
Use of data	Production of administrative data based population estimates as statistical research

FES: Background information

The SFC collect data about students enrolled on Further Education and Higher Education programmes in Scotland's colleges[5] in order to allocate funding and assess the performance of colleges against the outcome agreements.

The FES 2 dataset contains information about the student's enrolled on college programmes. Full student FES 2 details are required for all SFC fundable programmes and non-fundable Employability Fund (SDS) programmes as long as the student has attended at least once. Individuals may appear in this dataset multiple times as a record is submitted for each programme a person is enrolled on.

FES: Data supply and communication

The data provided is done so annually under the terms of a data sharing agreement. When data is received any queries regarding the data are discussed so that the Admin Data team have a full understanding of the data and if there are any reasons for changes from previous years data.

FES: Quality assurance undertaken by data supplier

There are three Management Information System software suppliers in the college sector (Capita, Tribal and Civica) and they annually update college management information systems (MIS) to the latest Further Education Statistical (FES) guidance published by SFC[6]. They in turn will mirror many of the code lists within FES in to the college MIS and build in internal validation and error checks prior to files being uploaded to SFC's FES Data Portal.

The student records are submitted by colleges to SFC via the Further Education Statistics (FES) system (the Data Portal). This is an automated and 'live' data capture and record system which encompasses around 300 built-in iterative validation checks to ensure the data is correct and credible. Only when the data has passed will SFC permit the data to be used for analysis. In addition to checks performed by SFC, every college Principal must also sign off the data as a true and accurate record for their college. The SFC analytical team also conducts data quality visits to ensure the student records submitted by colleges are accurate and comparable across the sector. Aggregations of the FES data are then used to produce National Statistics publication 'College Performance Indicators', the link below is for the 2016/17 version of this publication:

http://www.sfc.ac.uk/publications-statistics/statistical-publications/statistical-publications-2018/SFCST022018.aspx

[5] With the exception of students on Higher Education programmes at Scotland's Rural College or colleges that are part of the University of the Highlands and Islands. Information on these students is sent to the Higher Education Statistics Agency

[6] Guidance Notes for FES 2 2016/17 can be found at: http://www.sfc.ac.uk/publications-statistics/guidance/guidance-2016/SFCGD032016.aspx

In producing population estimates, the variables used to link the datasets are of particular importance. Extra information about the validation of these variables, beyond checking they are valid values, from the data suppliers is provided below:

Names - There are no specific validation to check that individual names are correct. However any errors will be usually be corrected by students throughout their time studying at a college. However it is possible that names will differ from official names, i.e. Jim instead of James, however this can be accounted for to some extent in linkage methodology used in the overall project.

Postcodes - A significant proportion of students provide postcode information at application stage where applicants enter the postcode and then choose their address from a list. This will minimise errors in postcodes entered, however generally no proof of postcode is required.

Date of Birth - If a student applies for student funding the date of birth is checked when the funding application is being processed. Otherwise the date of birth provided by the student is taken on trust.

Sex - Colleges receive this information from students. In some cases colleges are finding that it is becoming slightly more common for students to provide different sex (and name) information than what they had recorded at school. However there is no suggestion that this is an error.

FES: Quality assurance undertaken by the NRS Admin Data team

Once the admin data team receive the data, a number of data consistency and validation checks are performed, including:

- Checking the proportion of missing values for variables
- Checking that variables are in the expected formats and values
- Checking the validity of postcodes
- Comparing the data with similar data received in previous years and investigating when there appear to be significant changes.
- Checking the distribution of the day and month elements of dates of birth
- Checking the age distribution of the population.
- Removing duplicate records where identical information is recorded

If these checks suggest the data may need to be amended/adjusted then the potential issues are communicated with the data supplier so the data can be amended if appropriate.

FES: Strengths and limitations

The strengths and limitations in the table below relate to the use of the dataset for administrative data based population estimates rather than the original purpose of the data.

Strengths	Limitations
• Could be useful data source for young adults who are not as likely to update other their information in other data sources. • Validation processes performed by colleges and the SFC, so data is credible. • Students unlikely to be missed as colleges will want to receive the correct funding allocation. • Data feeds into a National Statistics publication. • Contains all the variables used when linking to other datasets.	• There is a lag in being able to receive the data. For example 2016/17 data is only available in early 2018. • Only provides data on a specific subset of the population. Even in the age groups where this data will be most beneficial (i.e. young adults) there will be a considerable proportion of the population that will not appear here. • Postcode information is from pre-study, so may not match other datasets where a student may have provided a postcode for their term-time address.

Residential Sales

Data supplier:	Registers of Scotland (RoS)
Supplier info:	Registers of Scotland is the public body responsible for keeping public registers of land, property, and other legal documents in Scotland.
Data type (counts or unit records)	Unit records
Data content:	Current and Historic Residential Sales data (including those not for full market value), The data includes: • Application Date • Title Number • Application Number • Price Paid • Property Address • Date of Entry • Granter Name • Granter Address • Applicant Name • Applicant Address
Supply schedule:	One-off supply
Time period covered:	Sales from 1 March 2011 to 30 April 2017
Use of Data:	Production of administrative data based population estimates as statistical research

Residential Sales: Background information

Registers of Scotland (RoS) collects administrative data in the process of fulfilling the Keeper of the Registers of Scotland's (the Keeper) statutory duties to manage, control and maintain the various public registers under RoS' remit. The main purpose is to populate the Land Register, documenting and protecting the legal rights of the owner/tenant/third parties. It is used to maintain an open and public property register clearly showing the details for each title registered within and its corresponding cadastral map. The information registered in the Land Register is covered by the Keeper's warranty, which means that the Keeper may be liable to pay compensation for any inaccuracies in the register that are subsequently rectified.

This data provided by RoS contains all residential property transactions is Scotland that have been sent to the Registers of Scotland for registration in the time period. The data collection process map for this data is included in the RoS QAAD report produced in support of the use of this data in the production of the UK House Price Index.[7]

Among a number of quality checks used by RoS as part of the process of extracting the data, intelligent e-Forms are used to allow RoS Intake staff to enter the 8 digit reference number on submitted forms to automatically input all form information directly into the Land Register System (LRS) which minimises the risk of typing errors when inputting information.

As a further measure to minimise the risk of inaccuracies on forms the submitting agent is sent a link to view the final title sheet and cadastral plan upon completion of the registration process. This enables the agent to check and verify everything has been registered correctly. If there are any inaccuracies highlighted to RoS, the agent has a duty to rectify the Land Register to ensure the integrity of the register under the Land Registration Act (Scotland) 2012.

Residential Sales: Data supply and communication

The data was provided on a one off basis after NRS purchased the data extract from RoS in accordance to the terms and conditions and schedule agreed between both parties.

Residential Sales: Quality assurance undertaken by data supplier

RoS do a variety of checks on the data used in the UK House Price Index on a regular basis. The full quality assurance process has been included in the RoS QAAD report produced in support of the use of this data in the production of the UK House Price Index.[7]

The main checks included in this process are to:

[7] https://www.gov.uk/government/publications/quality-assurance-of-administrative-data-in-the-uk-house-price-index/registers-of-scotland-data

- check price paid, particularly for high (>£75,000) and low (<£20,000) value properties
- check date of entry
- populating blank post town entries

Similar checks are now untaken on the data provided by RoS to NRS. However, prior to 2015 RoS did not undertake any QA on the All Sales (Land Values) data, from which the residential sales data provided to NRS was extracted. This was a raw data product and was not separately quality assured prior to release in the way that the market value residential sales data (Sales for Consideration data) was. It is market value residential sales data that is used in the UK House Price Index and RoS House Price Statistics. Since February 2015, the price paid, date of entry and address fields within the All Sales data are updated as part of the same quality assurance process that is undertaken on the market value residential sales data.

RoS also produced a Strengths and Weaknesses document for the UK House Price Index QAAD that highlights key features as determined when the residential sales data was used to create the House Price Index and RoS Quarterly House Price Statistics.

Residential Sales: Quality assurance undertaken by the NRS Admin Data team

The Admin Data team received the data as four separate files that were linked by application number. Files were merged to support administrative data based population estimates.

- Checking that variables are in expected formats and value ranges.
- Checking the validity of postcodes and assigning UPRN where possible for subsequent data linkage.

A degree of standardisation was required on Address Fields for property purchased to be linked to unique property reference number from Scottish Address Directory (SAD). Particular challenges were evident when dealing with sub-buildings and ordering of address components for Flats and Apartments, where there did not appear to be any uniform practices in line with SAD format to correctly identify UPRN from the address fields provided.

Residential Sales: Strengths and limitations

The strengths and limitations in the table below relate to the use of the dataset for administrative data based population estimates rather than the original purpose of the data.

Strengths	Limitations
• The data has a broad scope because all types of sales are sent to RoS for registration, i.e. cash sales, mortgage sales, high and low value sales, etc. • The data covers the whole of Scotland. • From December 2014, electronic application forms allow for the automatic input of form data into the registration systems, reducing the risk of data input error.	• RoS dataset only contains details on properties sold during the reporting period i.e. it will not be a complete list of properties in Scotland. • RoS dataset is linked to the purchaser(s), but the named person(s) may not be occupants of the property where rental arrangements are in place. • RoS cannot include transactions where the sale has not been sent to them for registration. It is not possible to estimate the volume of such transactions, but the majority of sales will be submitted to benefit from the state guarantee of title and warranty offered by RoS. • There is often a time lag between completion of the sale of the property and the solicitor submitting the application for registration. The majority (90-95%) of sales are submitted within 2 to 8 weeks, but RoS has no control over any time lags caused by the conveyancing process.

Vital Events

Data supplier	National Records of Scotland (Vital Events)
Supplier info	National Records of Scotland (NRS) is a Non Ministerial Office of the Scottish Government. The purpose of NRS is to collect, preserve and produce information about Scotland's people and history and make it available to inform current and future generations. The Vital Events branch of NRS produces statistics about the births, deaths, marriages and civil partnerships that are registered in Scotland.
Data type (counts or unit records)	Unit records
Data content	Birth, death, marriage and civil partnership registration records at individual level. Variables included: **Birth Registration data** • First name • Last name • Date of birth • Sex • Address • Postcode • Date of registration • Father's name • Father's date of birth • Father's address and postcode • Mother's name • Mother's date of birth • Mother's address and postcode. **Marriage and Civil Partnership Registration data** • Date of marriage/civil partnership • Date of registration For each party: • Name • Date of birth • Country of birth • Country of residence • Marital status • Sex • Usual address and postcode.

	Death registration data • Deceased's name • Deceased's date of birth • Deceased's sex • Deceased's usual residence address and postcode • Deceased's date of death. • Informant's name • Informant's relationship to deceased • Informant's address and postcode.
Time period covered	27 March 2011 to 30 June 2016
Supply schedule	Annually
Use of data	Production of administrative data based population estimates as statistical research

Vital Events: Background information

Every birth, death, marriage and civil partnership that occurs in Scotland must be registered by law.[8,9,10]

For a birth or death to be registered the registrar must be satisfied that the event has occurred. For births, evidence of the event usually takes the form of the informant (usually the mother) providing a card issued by the hospital or midwife who was present at the birth. For deaths this usually takes the form of a Medical Certificate of Cause of Death completed by the medical practitioner who certified the death, this certificate is usually given to the deceased's family. These documents are retained by the registrar upon registration of the events to prevent the birth or death being registered again.

Registrars are asked to take all possible measures to ensure no births or deaths fail to be registered. To do this registrars work with local medical establishments, midwives and funeral directors to identify any missed events. When it becomes known that a birth or death has not been registered in the prescribed time for registering these events, there are processes in place to rectify this.

For marriages and civil partnerships the registration of the event is an essential step in a legal marriage or civil partnership taking place. Therefore it is not possible for

[8] Registration of Births, Deaths and Marriages (Scotland) Act 1965
[9] Marriage (Scotland) Act 1977
[10] Civil Partnership Act 2004

these events to occur without being registered. This also removes the risk of these events being registered multiple times.

The data collected is usually input directly to the NRS Forward Electronic Register (FER) computer system as the registrar asks the informant(s) a standard sequence of questions. The computer system will warn the registrar of errors or apparent omissions and warn them of this. The informant(s) and the registrar then read through a printed copy of the record which should pick up any typing errors.

The record is then locked, however corrections can be made if an error is discovered in the future. In every year since 2007, around 97% of records have been created error free, so for individual variables the error rate will be even lower.[11]

There is further scrutiny from NRS examiners who check the information that NRS knows from experience is most likely to contain errors. And corrections are made if necessary.

More details of this process are provided on the NRS website: https://www.nrscotland.gov.uk/files//statistics/vital-events/quality-data-obtained-from-registration-of-ve.pdf

Vital Events: Data supply and communication

The data provided is done so annually under the terms of a data sharing agreement and includes record level data for a selection of variables as defined in a data sharing agreement for every registered birth, death, marriage or civil partnership in the previous year. The data is sent by the Vital Events team to the Administrative Data team via approved NRS data transfer procedures as agreed in a data sharing agreement.

The Administrative Data team have close links with the Vital Events team as they are both in the same organisation and work within the same building. NRS Vital Events have close links with the NRS Registration team, who in turn have close links with registration offices across Scotland. These close working relationships mean that any data quality issues, or planned changes in data collection, are considered in advance and any issues can be considered before the data is used. All parties involved in collecting and processing the data sit within NRS. The Administrative Data team and the Vital Events team all sit within the Statistical Services area of NRS. This means one person has oversight of both areas which further improves the already good links between the teams.

[11] Chapter 9 of the Registrar General's Annual Review of Demographic Trends - 2017: https://www.nrscotland.gov.uk/files//statistics/rgar/2017/rgar17-corrected-08-04-19.pdf

Vital Events: Quality assurance undertaken by data supplier

The data from the FER system is passed to the NRS Vital Events statistical database. Here the Vital Events team do further checks on the data. These checks include:

- Looking for any differences in the number of events in the statistical database and the FER. Where there are differences this is investigated to identify a) records that are missing from the statistical database and b) records that should be deleted from the statistical database. These are corrected in the database following the investigation.
- In FER codes are allocated by the registrar for certain variables such as country of residence. The Vital Events computer system highlights and corrects errors in these codes, and Vital Events staff also aim to identify and correct any anomalies. In addition, quality checks are carried out on each record by the Vital Events branch staff supervisor.

More details of this process are provided on the NRS website: https://www.nrscotland.gov.uk/files//statistics/vital-events/checking-quality-nrs-statistical-data-on-ve.pdf

Vital Events: Quality assurance undertaken by the NRS Admin Data team

Once the Admin Data team receive the data, a number of data consistency and validation checks are performed, including:

- Checking the proportion of missing values for variables.
- Checking the validity of postcodes
- Sense checking the number of records compared to previous years and published information.
- Checking that variables are in expected formats and value ranges.

If these checks raise any questions then this is discussed with the Vital Events team to find an explanation or a solution.

Vital Events: Strengths and limitations

The strengths and limitations in the table below relate to the use of the dataset for administrative data based population estimates rather than the original purpose of the data.

Strengths	Limitations
• Near complete coverage of these vital events occurring in Scotland due to the legal requirement of registration and the steps taken to get full coverage. • Well-defined process for collecting and quality assuring data which will minimise errors. • These datasets are the data source for National Statistics publications published by the National Records of Scotland.	• Events including residents of other countries are included if the event occurs in Scotland. This could lead to additional people being included in the population estimates if not identified. • Events involving residents of Scotland that occur outside of Scotland are not included in the data.

Register of Electors

Data supplier	Electoral Registration Officers in Scotland
Supplier info	The Electoral Registration Officer (ERO) is an official appointed by the local authority to prepare and maintain the Register of Electors.
Data type	Unit records
Data content	The following variables are included at an individual record level: • Forename(s) • Surname • Date of Attainment (Date someone turns 18 if they are under 18). • Address and Postcode • UPRN (for 7 of the 18 areas) • Elector Number (A unique identifier in the dataset) • Franchise (used to show which list of electors the person is registered on e.g. parliamentary, local government, European parliament. Also indicates where someone is an overseas voter)
Supply schedule	Annually
Time period covered	As at 1 December 2016
Use of data	Production of administrative based population estimates as statistical research

Register of Electors: Background information

The Register of Electors contains details of everyone who has registered to vote. It is used to determine who can vote at elections while the Register is in force. A new Register is published at least once a year[12], normally no later than 1st December. A revised version may be published at other times if, for example, major changes are made to the Register in the course of the year.

Individuals are able to be added to the register at any time and are encouraged to do so throughout the year. However the Electoral Registration Officers also have a legal

[12] Details of 14 & 15 year olds who are attainers on the local government register in Scotland are not published and are therefore not in the data set provided to NRS

requirement[13] to run an annual canvass where forms are sent out to every household to help identify any changes that need to be made to the Register. There is also a legal requirement to take specified steps to follow up any non-response to the annual canvass, including issuing two reminders and a personal visit.[14]

By law, a person who is requested for information during the annual canvass must provide the information. in Scotland, there is a criminal penalty of up to £1,000 for failing to provide the requested information, or £5,000 for providing false information.

Another factor that affects the coverage of the data are upcoming elections, as they act as a prompt for people who want to vote to update their details.

There were Scottish Parliamentary elections in 2016 which would have helped to encourage people to ensure their details are up to date so they would be able to vote.

Register of Electors: Data supply and communication

The data provided is done so annually under the terms of a data sharing agreement.

When data is received any queries regarding the data are discussed so that the Admin Data team have a full understanding of the data.

Register of Electors: Quality assurance undertaken by data supplier

The Register is updated monthly between January and September to add new electors and to deal with address changes etc. This procedure is suspended thereafter to allow the annual canvass of households to take place and time for preparation of the new Register. Forms are issued to each household, requesting details of eligible residents. The information obtained during the canvass then helps EROs to identify changes that need to be followed up.

The sections below give some detail of checks performed when updating the register to add, amend or remove an individual from the register.

Checks for new applications

When the ERO receives an application from someone to be added to the register there are a variety of checks. Most relevant for the purposes of producing population estimates are the checks on someone's identity and their address.

- Verification of identify - to verify someone's identify the information they provide is compared to DWP records. If the person's identity cannot be verified against DWP records then local data sources may be used instead. If they still cannot be verified then the application enters an exception process then the individual is

[13] Representation of the People Regulations (Scotland) 2001
[14] Section 9A(2) Representation of the People Act 1983 and Regulation 32ZB 2001 Regulations, Representation of the People Regulations (Scotland) 2001

asked to provide documentary evidence such as a passport or driving licence. If they cannot provide this information then they must get their application attested.

- Residence - among the other requirements to be registered, the ERO must be satisfied that that the individual is resident at the address in the application. If the ERO is not satisfied they can ask for further information and put the application on hold until this is provided.

Amendments to name on existing records

Electors can apply to change their name when already registered. To do so they must provide documentary evidence of the name change. If unable to do so they must provide their date of birth and National Insurance number as part of the application.

Deletions from the register

As well as adding new people to the register, someone who is no longer eligible must be removed to prevent inflation of the register. A person who is registered stays registered unless and until the ERO determines that:

- the person was not entitled to be registered in respect of the address
- the person has ceased to be resident at the address or has otherwise ceased to satisfy the conditions for registration
- the person was registered as the result of an application for registration made by someone else or the person's entry has been altered as the result of an application for a change of name made by someone else.

Examples of when a record is deleted are if the ERO receives a death certificate for an individual or receiving notification from two different sources that the elector is no longer eligible.

Records are also deleted when an ERO is notified that someone has made an application to join the Electoral Register in another area which has been allowed by the ERO in that area, and there is information to indicate that the individual no longer resides at the original address.

Address database

The EROs also have to ensure that their address database is up-to-date, particularly prior to the annual canvass. There is guidance to support EROs in how to do this, however each ERO will have differing procedures depending on the systems they have access to and to handle issues that are particular to their area. Generally the address information comes from the relevant Assessor's council tax valuation list (CTVL) or local authority Corporate Address Gazetteer (CAG) and updated on a regular bases (weekly/monthly).

These updates occur when the CTVL or CAG are updated with properties being added, amended or removed. If the ERO receives information to suggest that an address could be incorrect in some way, it is checked against the Assessor's records or CAG and then amended if necessary.

Register of Electors: Quality assurance undertaken by the NRS Admin Data team

The NRS Admin Data team looked at a study by the Electoral Commission which considered the accuracy and completeness[15] of the Electoral Registers[16]. While this study was not carried out on the 2016 Register, the findings from the 2015 and 2018 studies provide some indication of what can be expected. In the 2015 study the local government register in Scotland was found to be 85% complete and 91% accurate. The study also suggests that young adults, private renters and those who have not lived at their current address for more than one year are most likely to be missing.

Once the admin data team receive the data, a number of data consistency and validation checks are performed, including:

- Checking the proportion of missing values for variables
- Checking that variables are in the expected formats and values
- Checking the validity of postcodes
- Comparing the data with similar data received in previous years and investigating when there appear to be significant changes.
- Checking the distribution of the day and month elements of dates of birth
- Checking the age distribution of the population.
- Removing duplicate records where identical information is recorded

If these checks suggest the data may need to be amended/adjusted then the potential issues are communicated with the data supplier so the data can be amended if appropriate.

Following these checks some small amendments were made to improve the data for the purpose of producing administrative data based population estimates as statistical research, however these did not require the involvement of the data supplier.

[15] Accuracy looks at the number of false entries on the electoral registers and completeness measures whether those eligible to be registered are on the registers.
[16] https://www.electoralcommission.org.uk/who-we-are-and-what-we-do/our-views-and-research/our-research/accuracy-and-completeness-electoral-registers

Register of Electors: Strengths and limitations

The strengths and limitations in the table below relate to the use of the dataset for administrative data based population estimates rather than the original purpose of the data.

Strengths	Limitations
• A large proportion of the adult population in Scotland will be included in the data. The Electoral Commission estimate of completeness in 2015 was 86% and 2019 was 83%. • Identity is verified when applying to be on the register, minimising false entries. • Data provider has legal requirements to meet regarding how the data is maintained and updated. • The risk of receiving a fine for not providing the information, or providing false information, should improve data quality received from individuals. • The data also captures some information on people who have moved abroad, but are registered as overseas voters. This movement may not have been captured elsewhere. • The Unique Property Reference Number (UPRN) is provided on the Electoral Register for some areas, and in all cases full address information is provided. Meaning 94% of records are assigned a UPRN. • In 2016 there was a Scottish Parliament election in May so people are more likely to have updated their details compared to a year when no elections occur.	• The Register was published at 1 December 2016 while our estimates are mid-year. There will be a mismatch in where some individuals are due to this time difference. • The Register does not include sex for any records, and date of birth can only be derived for a small number of records where someone is yet to turn 18. • Unable to identify where someone is born on 29th February 2000 as there is not a 29th February 2018 for them to turn 18 on. • No coverage on children as they are not eligible to vote. • There are some subsets of the population where there is an increased probability of not appearing on the register. These include young adults, homeless, private renters and those who have not lived at their current address for more than one year.

5. Risk/Profile matrix for source datasets

This section contains a risk/profile matrix for each data source. The matrix reflects the levels of risk of data quality concerns in using these datasets for this work and the public interest profile of the administrative data based population estimates. These have been determined by a review undertaken by the NRS Admin Data team using the information contained within the Office for Statistics Regulation's Administrative Data Quality Assurance Toolkit.

For each data source the Public Interest profile has been set to a default value of "medium" for the following reasons:

- One of the objectives of the Administrative Data Based Population Estimates is to support future recommendations for the census beyond 2022.

- There is a strong interest in the viability of administrative data based population estimates to maximise the use of all available data sources to provide accurate and timely evidence to measure our population.

National Health Service Central Register (NHSCR)

Level of risk of quality concerns	Public interest profile		
	Low	**Medium**	**High**
Low	Statistics of low quality concern and low public interest. [A1]	Statistics of low quality concern and medium public interest. [A1/A2]	Statistics of low quality concern and high public interest. [A1/A2]
Medium	Statistics of medium data quality concern and low public interest. [A1/A2]	Statistics of medium quality concern and medium public interest. [A2]	Statistics of medium quality concern and high public interest. [A2/A3]
High	Statistics of high data quality concern and low public interest. [A1/A2/A3]	Statistics of high quality concern and medium public interest. [A3]	Statistics of high quality concern and high public interest. [A3]

*A1/A2/A3 – definitions supplied Office for Statistics Regulation's Administrative Data Quality Assurance Toolkit.

Justification for Risk of Quality Concerns score

The risk of quality concerns has been set to “low” for the following reasons:

- These are inevitable issues given the nature of the data collection and cannot be avoided by the supplier such as inflation when people don’t update details if leaving Scotland and lag in recent migrants appearing on the register. However as these are known issues they can be accounted for when using the data.
- The risk of quality concerns is reduced due to the service level agreement to have at least 97% accuracy that is being met.
- This is further reduced as the NHSCR team have a variety of data quality initiatives that are undertaken on a regular basis to mitigate these data quality issues.
- The NHSCR team and the Admin Data team both fall in the Statistical Services division of NRS and both report to the same Director. This means that there is an increased awareness of issues each other may be facing and the impact this may could have on the other party. We can therefore be confident that we will be made aware of any changes that would have an impact on how this data is used.

Health Activity

Level of risk of quality concerns	Public interest profile		
	Low	**Medium**	**High**
Low	Statistics of low quality concern and low public interest. [A1]	Statistics of low quality concern and medium public interest. [A1/A2]	Statistics of low quality concern and high public interest. [A1/A2]
Medium	Statistics of medium data quality concern and low public interest. [A1/A2]	Statistics of medium quality concern and medium public interest. [A2]	Statistics of medium quality concern and high public interest. [A2/A3]
High	Statistics of high data quality concern and low public interest. [A1/A2/A3]	Statistics of high quality concern and medium public interest. [A3]	Statistics of high quality concern and high public interest. [A3]

*A1/A2/A3 – definitions supplied Office for Statistics Regulation's Administrative Data Quality Assurance Toolkit.

Justification for Risk of Quality Concerns score

The risk of quality concerns has been set to “Medium” for the following reasons:

- While there are some limitations to the data, knowing where under- and over-coverage needs to be addressed means it can be accounted for when using the data.

- It was known that each subject in the dataset had an interaction in the last three years. ‘Last interaction’ information was subsequently provided, but was not included in this initial linking methodology and analysis. That variable will allow future analyses to make use of the time since the latest interaction for each subject.

- The complex nature of Health Activity Dataset that is dependent on multiple sources with varying levels of internal quality assurance measures.

Scottish Pupil Census (SPC)

Level of risk of quality concerns	Public interest profile		
	Low	**Medium**	**High**
Low	Statistics of low quality concern and low public interest. [A1]	Statistics of low quality concern and medium public interest. [A1/A2]	Statistics of low quality concern and high public interest. [A1/A2]
Medium	Statistics of medium data quality concern and low public interest. [A1/A2]	Statistics of medium quality concern and medium public interest. [A2]	Statistics of medium quality concern and high public interest. [A2/A3]
High	Statistics of high data quality concern and low public interest. [A1/A2/A3]	Statistics of high quality concern and medium public interest. [A3]	Statistics of high quality concern and high public interest. [A3]

*A1/A2/A3 – definitions supplied Office for Statistics Regulation's Administrative Data Quality Assurance Toolkit.

Justification for Risk of Quality Concerns score

The risk of quality concerns has been set to “low” for the following reasons:

- There is a clear agreement about what data will be provided through SPC, when, how, and by whom. The producers adhere to quality standards and meet the statistical needs for this judgement to be of low risk.

Higher Education Statistics Agency (HESA)

Level of risk of quality concerns	Public interest profile		
	Low	**Medium**	**High**
Low	Statistics of low quality concern and low public interest. [A1]	Statistics of low quality concern and medium public interest. [A1/A2]	Statistics of low quality concern and high public interest. [A1/A2]
Medium	Statistics of medium data quality concern and low public interest. [A1/A2]	Statistics of medium quality concern and medium public interest. [A2]	Statistics of medium quality concern and high public interest. [A2/A3]
High	Statistics of high data quality concern and low public interest. [A1/A2/A3]	Statistics of high quality concern and medium public interest. [A3]	Statistics of high quality concern and high public interest. [A3]

*A1/A2/A3 – definitions supplied Office for Statistics Regulation's Administrative Data Quality Assurance Toolkit.

Justification for Risk of Quality Concerns score

The risk of quality concerns has been set to “low” for the following reasons:

- There is a well-documented validation process used by HESA to maximise data quality.
- The quality of the variables that are most important to us for the admin mid-year population estimates is likely to be high as students will be motivated to ensure that the provider holds the correct information for them.
- It is unlikely that higher education students are missing from the data as the data providers will benefit from having full coverage of their students as this data is used for funding purposes. Many students will also receive student loans where there is a requirement for them to be registered with their HE provider.

Further Education data (FES)

Level of risk of quality concerns	Public interest profile		
	Low	**Medium**	**High**
Low	Statistics of low quality concern and low public interest. [A1]	Statistics of low quality concern and medium public interest. [A1/A2]	Statistics of low quality concern and high public interest. [A1/A2]
Medium	Statistics of medium data quality concern and low public interest. [A1/A2]	Statistics of medium quality concern and medium public interest. [A2]	Statistics of medium quality concern and high public interest. [A2/A3]
High	Statistics of high data quality concern and low public interest. [A1/A2/A3]	Statistics of high quality concern and medium public interest. [A3]	Statistics of high quality concern and high public interest. [A3]

*A1/A2/A3 – definitions supplied Office for Statistics Regulation's Administrative Data Quality Assurance Toolkit.

Justification for Risk of Quality Concerns score

The risk of quality concerns has been set to “low” for the following reasons:

- There are numerous validation checks performed by both the colleges and the SFC to ensure the data is credible.
- The quality of the name variables are likely to be high as students will be motivated to ensure that the provider holds the correct information for them and there was nothing to indicate an issue with these variables.
- It is unlikely that many higher education students are missing from the data as the data providers will benefit from having full coverage of their students as this data is used for funding purposes.
- For a small proportion of the data default dates of birth and postcodes appear to have been used. However there is not a clear way of identifying if this is the case or not. This will make it more difficult to confidently link these records to other datasets increasing the chance of us missing links. However as this dataset is likely to provide extra evidence of someone’s existence rather than being the primary evidence that they are in Scotland, the quality risk remains low.

Residential Sales

Level of risk of quality concerns	Public interest profile		
	Low	**Medium**	**High**
Low	Statistics of low quality concern and low public interest. [A1]	Statistics of low quality concern and medium public interest. [A1/A2]	Statistics of low quality concern and high public interest. [A1/A2]
Medium	Statistics of medium data quality concern and low public interest. [A1/A2]	Statistics of medium quality concern and medium public interest. [A2]	Statistics of medium quality concern and high public interest. [A2/A3]
High	Statistics of high data quality concern and low public interest. [A1/A2/A3]	Statistics of high quality concern and medium public interest. [A3]	Statistics of high quality concern and high public interest. [A3]

*A1/A2/A3 – definitions supplied Office for Statistics Regulation's Administrative Data Quality Assurance Toolkit.

Justification for Risk of Quality Concerns score

The risk of quality concerns has been set to “medium” for the following reasons:

- The interests of those registering the sale of property with RoS and the checks that are in place mean that there is a low risk of inaccurate information. However there are limitations to the data, namely that there is no way of ensuring sales are registered in a timely manner, or at all, in a minority of cases. While there are some limitations to the data, knowing individuals are in the system at a point in time is beneficial, but this can only ever account for owner-occupiers and overlooks the rest of the population.

Vital Events

Level of risk of quality concerns	Public interest profile		
	Low	**Medium**	**High**
Low	Statistics of low quality concern and low public interest. [A1]	Statistics of low quality concern and medium public interest. [A1/A2]	Statistics of low quality concern and high public interest. [A1/A2]
Medium	Statistics of medium data quality concern and low public interest. [A1/A2]	Statistics of medium quality concern and medium public interest. [A2]	Statistics of medium quality concern and high public interest. [A2/A3]
High	Statistics of high data quality concern and low public interest. [A1/A2/A3]	Statistics of high quality concern and medium public interest. [A3]	Statistics of high quality concern and high public interest. [A3]

*A1/A2/A3 – definitions supplied Office for Statistics Regulation's Administrative Data Quality Assurance Toolkit.

Justification for Risk of Quality Concerns score

The risk of quality concerns has been set to “low” for the following reasons:

- The legal requirement to register these vital events means that there is increased incentive for all parties to ensure information is accurate
- there are robust processes in place for collection and quality assurance of this data.

Register of Electors

Level of risk of quality concerns	Public interest profile		
	Low	**Medium**	**High**
Low	Statistics of low quality concern and low public interest. [A1]	Statistics of low quality concern and medium public interest. [A1/A2]	Statistics of low quality concern and high public interest. [A1/A2]
Medium	Statistics of medium data quality concern and low public interest. [A1/A2]	Statistics of medium quality concern and medium public interest. [A2]	Statistics of medium quality concern and high public interest. [A2/A3]
High	Statistics of high data quality concern and low public interest. [A1/A2/A3]	Statistics of high quality concern and medium public interest. [A3]	Statistics of high quality concern and high public interest. [A3]

*A1/A2/A3 – definitions supplied Office for Statistics Regulation's Administrative Data Quality Assurance Toolkit.

Justification for Risk of Quality Concerns score

The risk of quality concerns has been set to “low” for the following reasons:

- There are well defined procedures for verifying the identity of individuals on the register. Due to this, along with the potential legal ramifications of providing false information, the vast majority of records can be expected to be correct.
- The annual canvass, along with procedures for removing records, should minimise inflation of the register.
- While children are not included, other data sources can be used to identify these.
- There are subsets of adult population that appear to be less likely to appear in the Electoral Register but as this information is being combined with other information it provided a very good indication of recent address.

6. Background notes

Background

This document supports the Statistical Research publication Administrative Data Based Population Estimates, Scotland 2016.

Methodology

The Administrative Data Based Population Estimates Methodology Report provides more detail on the methodology, as well as information on the quality of the data and known uses of the data.

Future developments

We intend to continue developing the methodology for producing administrative data based population estimates based on the learnings from producing these estimates.

Following the publication of Administrative Data Based Population Estimates, Scotland 2016, we wish to discuss the findings of this research with as many users as possible. If you have any comments or would like to be involved in stakeholder events, then please register your interest under demography at http://www.gov.scot/scotstat.

7. Notes on statistical publications

Statistical Research

This publication presents statistical research and the methodology is still under development. We welcome any feedback from users on ways in which the methodology or data sources may be developed to improve the quality of these statistics in future years.

National Records of Scotland

We, the National Records of Scotland, are a non-ministerial department of the devolved Scottish Administration. Our aim is to provide relevant and reliable information, analysis and advice that meets the needs of government, business and the people of Scotland. We do this as follows:

Preserving the past – We look after Scotland's national archives so that they are available for current and future generations, and we make available important information for family history.

Recording the present – At our network of local offices, we register births, marriages, civil partnerships, deaths, divorces and adoptions in Scotland.

Informing the future – We are responsible for the Census of Population in Scotland which we use, with other sources of information, to produce statistics on the population and households.

You can get other detailed statistics that we have produced from the Statistics section of our website. Scottish Census statistics are available on the Scotland's Census website.

We also provide information about future publications on our website. If you would like us to tell you about future statistical publications, you can register your interest on the Scottish Government ScotStat website.

You can also follow us on twitter @NatRecordsScot

Enquiries and suggestions

Please get in touch if you need any further information, or have any suggestions for improvement.

Lead Statistician: Lindsay Bennison

Statistics Customer Services telephone: (0131) 314 4299

E-mail: statisticscustomerservices@nrscotland.gov.uk

For media enquiries, please contact: scotlandscensus@nrscotland.gov.uk