Distr.: General
06 August 2024

English

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Population and Housing Censuses

Twenty-sixth Meeting
Geneva, 2–4 October 2024
Item 5 of the provisional agenda
Transformations in population statistics

Reflections on the United Kingdom's journey through the statistical transformation of population statistics

Note by Office for National Statistics, United Kingdom of Great Britain and Northern Ireland

Summary

At the Office for National Statistics we are transforming how we produce population statistics. At the heart of the new system is a set of modelled demographic accounts. These offer some conceptual continuity, since they are structured around the traditional cohort component method for measuring population change. The new Dynamic Population Model presents new challenges, as it is a statistical modelling approach. This offers transparency and statistical coherence, but requires upskilling and a paradigm shift towards extended use of administrative data. Building organizational and stakeholder understanding is an imperative, as we transition from experimental statistics to National Statistics status.

We are productionizing the new methods, learning and building capability and statistical and organizational resilience and sustainability as we go. With other countries also exploring opportunities to transform their population statistics, this presentation will benefit them and also us from the valuable discussion we hope to have. We will share our reflections and lessons learned so far.

*Prepared by Dominic Webber (presenter), Louisa Blackwell, Ann Blake, Katie Coria, Sarah Crofts, Sophie Gilbert-Johns, Teri Howells, Charmaine Lawson, and Justine McNally.

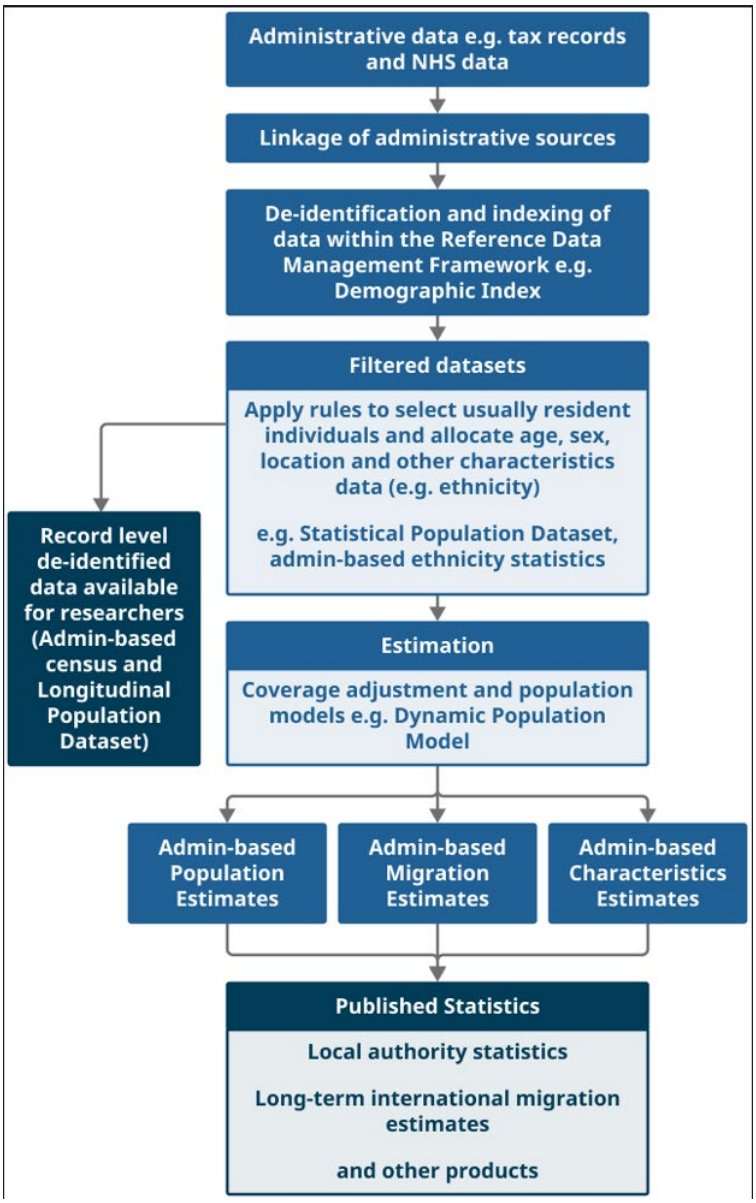
I. Introduction

1. Population estimates are one of the most critical statistical outputs produced by the United Kingdom (UK) Office for National Statistics (ONS). The ONS produces estimates for England and Wales and works alongside the devolved administrations (representing Scotland and Northern Ireland) to produce UK-wide estimates. They are used for funding allocations and to inform decisions by governments and others, such as migration policy. They also underpin many other statistical outputs, as survey weights or denominators. We are being ambitious, working to provide more timely statistics that reflect changes in society and meet the needs of users. We are establishing methods which make greater use of administrative data (collections of data maintained for administrative reasons e.g. registrations, referred to throughout this paper as 'admin' data) to create improved population estimates.
2. In the UK, a survey-based census has taken place every 10 years since 1801 (with the exception of 1941). Since the mid-1800s, the population has been estimated each year by starting with the census, ageing the population on by one year, adding births from birth registrations, subtracting deaths from death registrations, and adding or subtracting net migration. There are further adjustments for specific groups of the population using data from administrative systems, for example using Ministry of Defence data for armed forces and Higher Education Statistics Agency data for students. This method is known as the cohort component method (see our *Population estimates for England and Wales methods guide* for more information). However, every time there is a new census, we find that these population estimates have 'drifted' from the census estimate of the population. This means that the accuracy of population estimates reduces, the further we move away from a census.
3. In recent years, improvements to computing power and the availability of administrative data sources (for example, tax and benefits records, visas issued, and National Health Service data) means we can explore new approaches to estimate the population to a more consistent level of accuracy, in greater detail, more frequently. While we have already been using administrative data to measure the population, for example birth and death registrations, the NHS' Personal Demographics Service and Higher Education Statistics Agency data, our aim is to make greater use of administrative data to improve the quality, timeliness and inclusion of ONS population statistics. This paper sets out the new statistical design of how administrative sources will be used to produce accurate and timely Admin-based Population Estimates (ABPEs).

II. The proposed new system

4. The UK's new design for producing population estimates involves several stages. These, as illustrated in Figure 1, include acquiring and preparing administrative sources for statistical purposes, producing separate estimates of international and internal migration, and then applying statistical modelling techniques including coverage adjustment. The result is a set of demographic accounts for England and Wales, with complete consistency between the levels and the flows.

Figure 1
Overview of statistical design for future UK population and migration statistics



A. Administrative data

5. Building on the Statistics and Registration Service Act 2007, the Digital Economy Act 2017 provides a legal gateway for ONS to access data held by public authorities and commercial undertakings to support the production of official and accredited official statistics, including the census. These data will be accessed for statistical purposes only and personal information will be removed during the processing so that individuals cannot be identified.

6. At the heart of our statistical design is the acquisition and use of a range of data sources to cover the population and its characteristics. For children, for example, we use birth registration and school censuses, while for students, Further and Higher Education datasets are used. His Majesty's Revenue & Customs (HMRC) and Department for Work and Pensions tax and benefits data are used to cover people of working age and pensioners, while Home Office (HO) data are used to cover special populations such as migrants, refugees and asylum seekers. As with the current population statistics system, we also continue to use NHS health registration data, which provides good coverage of the population across all age groups. Further details of datasets used in our research and statistics are provided in our *data source overview report*.

7. It is important that we understand how new data sources can be used to measure the population, such as mobile phone data or Advance Passenger Information data (the data provided to airlines before travel), to improve the quality, timeliness and inclusion of ONS population statistics. Therefore, we work closely with data suppliers to develop our understanding of these data sources (including any changes over time that may impact content and quality). We use a variety of approaches, including working groups and secondments, where ONS staff work together with data experts in the supplier organizations. We currently have secondments in place with the DWP and the Home Office.

B. Linkage and de-identification

8. Using several data sources together relies on having these data sources integrated and accessible in a consistent and secure way. This underpins our statistical design. To create outputs that meet user needs, we need to use the best available linkage methods and have a robust understanding of the linkage quality. Data or record linkage is a method of bringing together information about the same person or address from different sources to create a new, richer dataset. Data linkage is now commonly used for improving data accuracy and quality over time, to allow the reuse of existing data sources for new studies, and to reduce the cost and effort of data collection.

9. Once the data are linked, identifying information (such as a person's name and address) are removed from the data set used in further processing. The Reference Data Management Framework (RDMF) is ONS's model for handling data securely and consistently for linkage purposes. *Our Data Strategy* provides further information on the RDMF and how it fits into ONS's plans to develop data capabilities.

10. The RDMF is made up of five indexes that link and match data on addresses, businesses, classifications, demographics and location. One of these indexes is the Demographic Index (DI) which combines records from health, tax, education, and the birth register to provide an ever-registered population that may include people who are not resident. The DI does not include any demographic or characteristic information, it simply links records and references the data by removing personal identifiers and replacing with a unique ONS identifier for onward use.

11. The methods used are a mixture of deterministic and probabilistic methods. Deterministic is the approach that uses exact matches. For instance, James Andrew Smith, date of birth (DOB) 11 July 1984 is the same on both records, whereas probabilistic methods consider the likelihood of two records being a match based on a set of criteria such as common nicknames e.g. Jim Andrew Smith DOB 11 July 1984. By developing our linkage methods in this way, we can improve both the quality and our understanding of the quality of our data linkage and how it is likely to impact on our statistical outputs.

C. Filtered dataset of the usual resident population

12. The DI on its own simply provides data on people who have ever registered within the administrative systems used. It is not intended to only include the usual resident population that we want to measure.
13. We apply a set of activity-based rules using 'signs of life' across health, education and income sources, to filter records from the DI to approximate the usual resident population at a reference date (for example, the midpoint of the year at 30 June). We also use those rules to help determine demographic information such as age, sex and location. The dataset created through this process is referred to as the Statistical Population Dataset (SPD).
14. People lead different lives and interact with services differently. Therefore there will be variations in how people are recorded on administrative systems. There will be both undercoverage (where people are not included) and overcoverage (where a person may be recorded more than once, or is included when they should not be included; for example, if they have emigrated). Our *Transforming Population Statistics* article (as well as analysis presented in section III. Evaluation of the Admin-based population estimates) shows the need for a robust coverage adjustment approach to produce ABPEs by local authority (LA), age and sex, to the required quality.
15. As well as providing a main input for the ABPEs, the SPD provides a population spine that enables us to link across characteristic attributes such as ethnicity. This has the potential to provide the foundation for estimates of those characteristics once we have completed development of our coverage adjustment process to support this.

D. Admin-based international migration estimates

16. An international migrant is defined as someone who changes their country of usual residence for 12 months or more. To produce estimates of international migration at a UK level we use a combination of data sources and methods, selecting the best data source for each group. The methodology to derive the latest admin-based international migration estimates (ABMEs) uses different administrative data sources and methods from the ABPEs to produce migration estimates at a UK level.
17. To estimate non-EU migration, Home Office Borders and Immigration (HOBI) data provide information about the numbers of people arriving from non-EU countries who require a visa to move to the UK long-term. EU migration estimates are derived from the Department for Work and Pensions' Registration and Population Interaction Database (RAPID). This provides a single coherent view of interactions across the breadth of benefits and earnings datasets for anyone with a National Insurance number (NINo). For British national migration estimates we use International Passenger Survey data, as the complexity associated with identifying British migrants in administrative data means we cannot use such data at this time. However, we are continuing to explore potential sources of data that capture actual migration behaviour. For further information on the ABME methodology, see our *technical user guide*.
18. International migration estimates are produced at the UK level, with further methods required to produce estimates at La level, by single year of age and sex.
19. The ABMEs are an important input to our model for producing ABPEs. This model, alongside births and deaths, takes the population and migration inputs and produces coherent population and migration (stocks and flows) statistics.

E. Internal migration estimates (including cross-border flows)

20. Estimates of internal migration and cross-border flows are also inputs to the admin-based population estimates. Internal migration describes moves between local authorities in England and Wales. Cross-border flows are moves between England and Wales and the rest of the UK, and are agreed with Devolved Administrations.

21. In the current system, we use health data to produce internal migration and cross-border flow estimates. They are derived from the Personal Demographics Service (PDS) data which flags when people change their address on NHS systems (e.g. with their general practitioner). We also link the PDS to Higher Education Statistics Agency (HESA) data to better identify moves made by higher education students to and from places of study, as these moves are less well captured in health data alone.

22. We use the internal migration and cross-border flow estimates produced for the mid-year estimates as the input for the admin-based population estimates (ABPEs). However, to produce timelier provisional ABPEs (six months after the reference period), we have developed alternative internal migration estimates to use the data available at that point. These PDS-based internal migration estimates are scaled using the ratio between previous years' PDS-based estimates and mid-year population estimates of internal migration to ensure consistency in the time series and adjust for moves less well captured by the PDS alone. These internal migration estimates are then updated with HESA data for the updated ABPEs available the following summer.

F. Admin-based population estimates – the Dynamic Population Model

23. The admin-based population estimates (ABPEs) are produced by bringing together a range of administrative and other data sources and applying statistical modelling techniques. The statistical model is referred to as the Dynamic Population Model (DPM) which uses available information on the usual resident population (stocks) and movement into and out of the population (flows) at specific points in time. Similarly to the census-based mid-year estimates, the DPM is also based on the approach set out in our *Population estimates for the UK methods guide*. However, it uses a range of sources to estimate the stock population each year rather than use the decennial census as a baseline.

24. To produce the ABPEs, we start with extracts from administrative systems at specific points in time. To produce mid-year ABPEs, the DPM utilizes the available information on the usual resident population as close to the mid-year reference period as possible. This uses data on stocks in addition to data that show changes in the population over time. The stocks data include data in the SPD, and census-based mid-year estimates. The flows data include births, deaths, international migration (ABMEs), internal migration, and cross-border flows.

25. The DPM balances stocks and flows to produce a coherent set of population and migration estimates. If the information on stocks and flows is not consistent, we use information about the uncertainty of the estimates to help determine the most likely true stocks and flows, (i.e. that the change in population stocks over time is equal to the net flows).

26. The DPM has advantages, particularly its flexibility which will improve the quality of the statistics. It can take account of quality limitations in the underlying data sources and draw strength from across the wide range of sources being used. The statistical models can take account of underlying demographic trends and differing levels of coverage and uncertainty associated with input data. It

can also incorporate other data sources as and when they become available. This could be helpful to address the challenge of some administrative data sources being more reliable to measure particular population groups than others, for example, older people being more likely to interact with healthcare systems, or a local data source could more accurately measure local populations than a national-level data source.

G. Coverage adjustment

27. To produce accurate population estimates, it is important to have unbiased population stock estimates for each year broken down by local authority, single year of age and sex. Coverage adjustment is required to address gaps or duplicated records in the administrative data sources.
28. We are currently exploring methods for coverage adjustment, including using administrative data sources. Work to date has focused on applying Dual System Estimation (DSE) to available administrative data sources as a possible approach. DSE is a well-recognized and established method typically used to ensure that estimates resulting from a census have maximum coverage. It uses a coverage survey following the census to estimate how many people responded. We are currently considering whether a similar method could be applied with different sources of administrative data. Further work could cover approaches such as Multiple System Estimation, use of additional administrative data sources and potentially the use of surveys. In the meantime, we will continue to use census 2021 results to apply coverage adjustment to the ABPEs and will continue consulting with additional methodological experts as we develop our methods.
29. These admin-based estimates will be updated regularly to provide timely estimates on changing populations. The statistical processes involved in producing our new population and migration statistics, alongside their limitations, and plans related to revisions and quality are outlined below. We have set out our workplan and timeline on how we are going to deliver our statistics.

III. Evaluation of the admin-based population estimates

30. The UK ONS has been developing admin-based population estimates for several years. As such, we now have a substantial amount of analysis and evidence to support an assessment of the quality of admin-based population estimates. The evaluation we present here is made possible by the availability of high-quality comparator data - in this case census 2021 which was conducted in England and Wales in March 2021.
31. Figure 2 compares our highest-quality population estimates (i) against a series of alternative population estimates (ii-v). The analysis is conducted at LA geographic level (of which there are 318 in England and Wales.) The full suite of population estimates are:
    i. Census 2021-based mid-year estimates (MYE). Refer to a reference date of 30 June 2021 (mid-year) and were produced by rolling forward from the census 2021 by adding births and people arriving in the country, and removing deaths and people leaving the country; MYE at LA level are also adjusted according to moves into and out of each LA. As they are derived from census data, we consider them to be the best quality information we have on the population in mid-year 2021, and therefore form the baseline against which the remaining population series are compared.
    ii. 2021 rolled-forward MYE. Refer to a reference date of 30 June 2021 and are produced by rolling forward yearly from census 2011. These estimates are indicative of the quality that

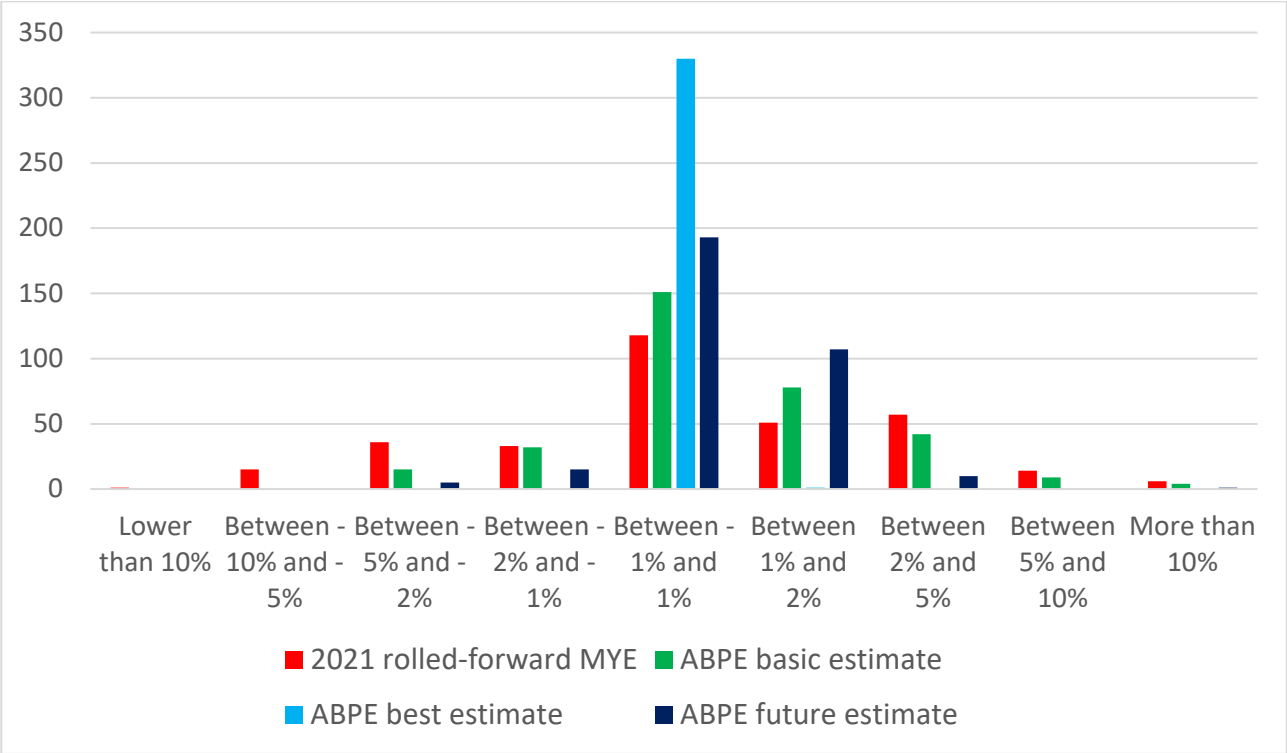
our current official estimates would be if we were no longer able to benchmark to a census every 10 years, and demonstrate the intercensal drift that occurs between censuses.

iii. ABPE basic estimate. These are estimated using the DPM, and do not include census 2021. They are comparable to the 2021 rolled-forward MYE series, but instead are estimated using the DPM with an independent annual admin-based stock (the SPD) to inform the estimation.

iv. ABPE best estimate. This represents our current best estimate derived from our new DPM and uses the best available data, including census 2021.

v. ABPE future estimate. This is indicative of what users might expect from our transformed population estimates derived from the DPM in the future, if there were no longer a census. It is similar to the basic estimate in that it does not include census 2021 as a stock estimate. However, while the basic estimate uses census 2011 to inform the coverage adjustment (and hence suffers from drift in the 2021 analysis), the future estimate uses census 2021 as the coverage adjustment. It essentially highlights an admin-based population estimate without a census but with a very high-quality coverage adjustment which we hope to develop over the coming years (see Next Steps section).

Figure 2
Accuracy of admin-based population estimates compared with census 2021



32. Figure 2 presents the number of LAs in different bands, where for each population measure, the bands represent the proportionate difference from the census 2021-based mid-year estimates. Looking first at the 2021 rolled-forward MYE, we see that this measure has the most spread, indicating the greatest divergence from the census 2021-based mid-year estimates. This neatly demonstrates the extent of intercensal drift that is a prevalent feature of population estimation that relies on rolling forward a census year after year. Contrasting these against the ABPE basic estimate clearly demonstrates the improvement in accuracy that can be achieved when using the DPM and an

independent stock measure. For instance, 151 LAs under the basic estimate lie within +/- 1 per cent of the census 2021-based mid-year estimate, compared with 118 for the 2021 rolled-forward MYE.

33. Clearly, an ABPE that includes a census 2021 stock (best estimate) estimate performs well with all but one LA landing within +/1 per cent of the census 2021-based mid-year. Interestingly, the future estimate, whilst being more spread than the best estimate, still demonstrates good coherence with the census 2021-based mid-year estimates (193 out of 331 LAs within +/1 per cent). This highlights the potential for new transformed population estimates to operate well without the need for a census, albeit reliant on developing a coverage adjustment method, addressing both under- and over-coverage, that performs as well as one derived from a census.

IV. Next steps

34. As this paper demonstrates, we have made significant progress on our transformation journey but there is much more work to do. We have simultaneous ambitions to further the research and enhance the quality of the admin-based population estimates, whilst moving operationally from a research-focused to a production-focused state. This all leads towards ABPEs being our official estimate of the population in 2025, and so entails a significant amount of effort to engage with users and ensure that they understand and trust our new approach.
35. We acknowledge there are several challenges and opportunities that drive our future research plans, including:
    a) Working with data suppliers to mature the processes and systems used to ensure timely and reliable supplies of the data we currently use, including the automation of delivery processes and the strengthening of data sharing agreements.
    b) Investigating new data sources for use in the future; for example, exploring the potential from tax and benefits data and mobile telephone data to improve our internal migration estimates.
    c) Developing our statistical models including a robust coverage adjustment strategy.
    d) Delivering coherence across the full range of population and migration outputs including ABPEs, ABMEs, statistics about households and information about population characteristics.
    e) Providing an admin-based census for approved researchers which will contain de-identified person-level records by age, sex, Lower level super output area (LSOA) and Demographic Index reference.
    f) Developing coherence across our population and migration estimates for England and Wales and across the UK, considering our best approach for producing consistent migration estimates from our population model, with an aim to publish international migration and internal migration and cross-border flows separately.
    g) Ensuring harmonization and consistency of definitions across our data sources and outputs. For example, we are working across the Government Statistical Service to promote the adoption of their standards, ensuring greater usefulness of statistics.
    h) Working with the Office for Statistics Regulation (OSR) on the assessment of our admin-based population estimates and working to ensure that these, and our long-term international migration estimates, meet the standards expected of accredited official statistics by summer 2025.

i) Addressing user feedback including recommendations from the OSR report.

j) User needs are central to us achieving these aims. Understanding how well our statistics meet user needs is essential for informing our workplans. We continuously seek this feedback in various ways such as through our published research, public consultations, conferences, webinars and detailed meetings with stakeholders. If you would like to hear more or get involved let us know by contacting us at pop.info@ons.gov.uk.

V. Conclusion

36. ONS is introducing a demographic accounting system to estimate population in a timely way, to better respond to user needs. We aim to produce mid-year population estimates within six months of the reference date, and update these with more mature information within a year. At the heart of the new system is a dynamic population model (DPM). This model draws strength from a range of data sources including administrative and survey data, incorporating not just those data but also measures of their statistical quality. The model also includes data and system models, transparently using known features of demographic behaviour and data. The outputs will be fully coherent estimates of population counts and changes due to births, deaths and migration. For sub-national estimation this will include internal migration.

37. There is consistency in the theory underpinning the transformed system; our traditional mid-year population estimates and the new, modelled approach both use the cohort component method to estimate and report on population change. By drawing heavily on administrative data we plan to optimize its use. Using multiple sources within a modelling framework allows us to draw strength across sources, and address some of the quality gaps that are, arguably, inevitable, given that our statistical use was not the intended primary purpose for these data. Reducing our reliance on the decennial census means that we can avoid the inter-censal drift from which our traditional population estimates suffer. However, new reliance on administrative sources also brings new challenges, which we describe here in terms of having to adapt to our data suppliers' discontinuities in data collection. Uncertainties in the nature and supply of administrative sources demand methods adjustments during statistical production, removing the comfort of 'business as usual' predictability. This methodological agility, coupled with the complexity of statistical modelling, has meant that as we develop the new statistics we must also build data and methods capability.

38. This paper demonstrates significant progress in developing this statistical design for a transformed population system that builds administrative data at its core. Moreover, the evidence points towards this system being at least good enough to reduce the inter-censal drift, if not offering hope for population estimation that does not need to rely on a traditional census.

39. These are innovative, hopefully ground-breaking new methods. There is no textbook we can draw on to help guide their implementation. For this reason, we keenly draw on the experience of other National Statistical Institutions to share knowledge and learning, and it is in this spirit of transparency and collaboration that we hope that the ONS experience will be of interest to others who are also navigating statistical transformation.