Integrated statistical design for the transformed population and social statistics system- Bayesian methods for demographic estimation

Louisa Blackwell and Duncan Elliot, Integrated Statistical Design, ONS

John Bryant, Bayesian Demography Limited

1. Purpose

This paper describes a collaborative research project to develop a prototype demographic accounting model for the population of England and Wales.

The research is led by the ONS Social Statistics Transformation, Analysis and Research Directorate, in collaboration with ONS Methodology, the Universities of Southampton and Warwick and with John Bryant of Bayesian Demography Limited in New Zealand.

This paper describes:

i. What we currently do in measuring population and social statistics

ii. Background to why a demographic account is needed for the integrated statistical design

iii. How it fits into the wider programme of statistical transformation

iv. Research on other international practices

v. The developmental stages of our approach

vi. The different options we are exploring to develop the methods

vii. The technical and statistical challenges that we will address

This research will provide evidence to support the 2023 Recommendation to the National Statistician on the future of traditional censuses.

We invite members to

i. comment on the proposed approach

ii. to advise on the proposed methods

iii. contribute to the development through iterative review and assurance

2. Background

Current methods in measuring population and social statistics at ONS

The current ONS population estimation of Mid-year estimates relies heavily on the decennial Census. It suffers from increasing bias the further we move away from the Census date, which is explained in the uncertainty measures published alongside the measures. The process relies on the aggregate cohort component method, updating the Census population each year with natural changes (births and deaths) and with estimated in-flows and out-flows of international migrants. Sub-national estimates include estimates of internal migration. These estimates are produced annually, with a lag at 1 year, and further granular information published at a later date.

Bayesian Demographic Accounts

In our previous paper on integrated statistical design (202_EAP132) we outlined how different elements of research across ONS worked together, however we had not outlined a statistical design to help meet the key challenges listed.

Further challenges around timely estimation at subnational level in response to the COVID-19 Pandemic have emerged over the past 18 months, steering us to consider population measures when we are not in a 'steady state'.

Current Subnational population estimation requires bringing together a range of multiple, inconsistent datasets, with reliance on ad-hoc adjustments and unknown levels of uncertainty around these adjusted estimates.

To meet these range of challenges, we are exploring Bayesian Demographic Accounts to achieve

- timely sub-national population estimates
- fully coherent estimates across the different components of population through time

A demographic account is an internally consistent set of demographic stocks and flows. Options for estimating a demographic account include ad-hoc adjustments, which is the most common approach. Producing demographic accounts is the end goal. We are exploring the Bayesian framework for producing demographic accounts, whereby regularities of the demographic system are captured in system models. This allows expertise and understanding of demographic trends to be used in the framework, instead of ad hoc adjustments, to allow for coherent, plausible estimation.

Bayesian methods allow the combination of this prior information about the population (in system models), with evidence from different data sources in measuring population change.

The Bayesian demographic accounts will allow us to use information from

- existing data/estimates for component of population change
- information around the uncertainty relating to these data/estimates
- information about demographic trends

to produce coherent estimates from the combination of these sources.

A more detailed description of the Bayesian demographic accounts methods can be found in section 3.

Prototype Bayesian demographic accounts have been developed in Statistics New Zealand.
Please see Annexe A for approaches by other NSIs that were considered.

ONS transformation of population and social statistics

ONS has been researching, since before the 2011 Census, methods for producing admin-based population estimates. Research is ongoing, into understanding the effects that time lags in administrative sources have on statistical error, and on refining methods for selecting administrative records to include in population estimates.

The ambition is to produce timely population estimates that do not suffer from the drift inherent in the current Census-based approach. Monthly published population estimates are envisaged, to inform policy decision-making. In the interests of accuracy and coherence, it is essential that the components of population change, natural changes, and migration, align with the time series of population stock estimates.

In the existing methods, estimates are informed by their plausibility. Demographers do this informally, making adjustments to the estimates based on expert judgement. In Bayesian demographic accounts, this is encoded in system models. In existing methods, estimates are informed by knowledge of the strengths and weaknesses of the input datasets. In Bayesian demographic accounts, this knowledge is embedded in the data models. Bayesian demographic accounts allow us to integrate incomplete information and balance the errors in components for different elements of the population statistics system. This research involves knowledge and skills development to support implementation of the accounting framework in a sustainable way within ONS.

We see demographic accounts as the statistical scaffold around which other demographic assets in the transformed statistical system will be built. This includes a rolled-forward 2021 Census cohort to provide granular multivariate outputs, longitudinal outcomes based on characteristics derived from the 2021 Census and administrative sources and admin-based population estimates constrained to this longitudinal accounting framework.

Impacts of the COVID-19 Pandemic on population estimation

The COVID-19 Pandemic highlighted an urgent need for timely estimates of the population. Vaccination planning and distribution estimates of infection rates and population at risk, for example, all demand timely and accurate estimates of the population.

The ONS uses the internationally recognised definition of usual residence which refers to residents of the UK who stayed or intended to stay in the UK for 12 months or more, or who had a permanent UK address and were outside and intended to be outside the UK for less than 12 months. The admin-based population estimates seek to capture the population in line with the ‘usual resident’ definition and annual cycle of population estimation.

While the ambition is to capture usual residence, arguably the sources we are using are really capturing a mixture of usually resident and present population, as behaviours are changing and 'usual residence' implies intentions that are both realisable and stable enough to be reliably measured.

Due to the pandemic, many former usual residents are living abroad whilst remotely working in the UK. Others may have been overseas for an extended period while on a UK furlough, unable to return. Usual residence is a fluid concept for many during the pandemic and challenging to measure statistically. When using administrative data we might imply usual residence when someone interacts with an administrative system, but this is essentially an approximation.

.

We need timely estimates of different target populations that reflect and report population movement and change in population stocks to meet our urgent policy requirements. The demographic accounts approach allows us to develop pre-and post-COVID models to reflect different levels of statistical uncertainty, and different prior information based on behavioural evidence that may not be seen in some data measures.

Pre-COVID models reflect population stocks and flows before December 2019, when disruption to recent historic trends was driven by Brexit. The pre-COVID models measure the usually resident population. We begin to detect COVID-related changes to migration from December 2019. Our post-COVID models reflect both changes in behaviour, particularly migration and mortality, and changes to the quality of data sources due to the pandemic-related disruption. They measure a hybrid of usually resident and de facto population. Measures of international migration continue to refer to long-term migrants, in that short-term migrations of less than 1 year are not included. Given the volatility of the pandemic, availability of data and the necessary policy response, international migration estimates rely heavily on assumptions, which have been validated by experts through a Delphi process (see ONS 2021 for details).

Increased reliance on administrative data for estimation

The ONS commitment to maximise use of administrative data for population statistics predates the COVID-19 Pandemic. The transformation has been accelerated by it. When the International Passenger Survey (IPS) was stood down in March 2020[1], using administrative evidence on migration became a necessity. State-space modelling extended the IPS time series with blended administrative sources, producing estimates for quarters one and two 2020, published in April 2021. Development work is underway to incorporate further sources to generate admin-based migration estimates (ABMEs).

Predictive modelling of international migration with timely administrative sources for the period being estimated is key to supporting the production of timely population estimates. This is because of the time lags inherent in observing migration under our current definition; we only know that someone is an immigrant after 12 months elapse following entry or 12 months have elapsed following the departure of an emigrant.

3. Bayesian Demographic Accounts developed in Statistics New Zealand

[1] The IPS was reinstated in January 2021.

We are using and adapting the methods developed by John Bryant and Patrick Graham in Statistics New Zealand Junni Zhang at Peking University. At the heart of the demographic accounting system is the cohort component formula:

$$Populationatt + 1 = Populationatt + entriesinperiod - exitsinperiod$$

Thus we begin with a measure of population stock at time *t*. We add births and immigrants and deduct deaths and emigrants. Combined with intelligence about demographic behaviour and uncertainty measures, the outputs from the model are our best estimates of each component of change and the population stock at time *t+1,*

The estimation framework is summarised in Figure 1. The rectangular box represents the demographic account. It includes population components that are not directly observed. Y_1 has unobserved population counts. $Y_2,\ldots,Y_K$ are counts for the other unobserved components, including births, deaths, migrations.

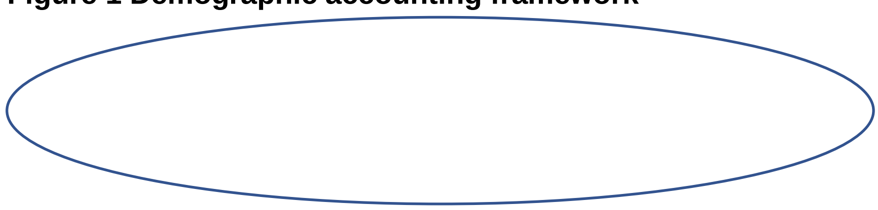
System models

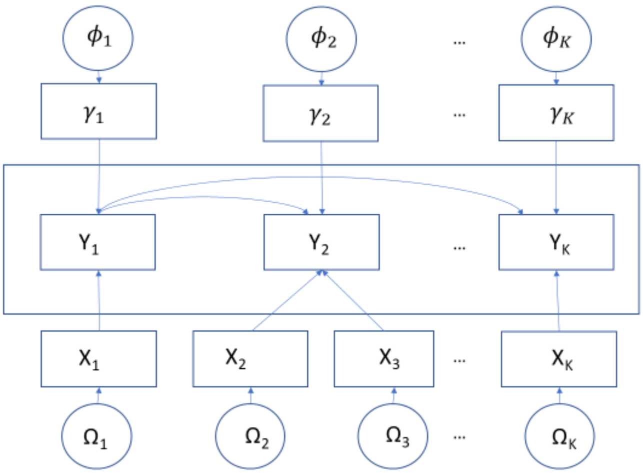
The system models are represented by the ellipse in Figure 1. Each of the components $Y_1,\ldots,Y_K$ can be associated with arrays of underlying rates or means, $\gamma_1,\ldots,\gamma_K$. For instance, if deaths were Poisson distributed, then γ would specify the mortality rate. Each γ has an associated prior model, that reflects our knowledge about γ prior to observing the data and the parameters for this prior model are denoted by φ. In our models we are distinguishing between pre-COVID and post-COVID models.

The system models facilitate the sharing of information across different time periods and birth cohorts, for instance it is reasonable to assume that the mortality of a ninety-one-year-old is related to that of a ninety-year-old. The system models allow this information to be shared as well as the incorporation of information regarding the set of plausible values. Currently the system models are restricted to the options provided by existing **demest** R packages and mostly use dynamic linear models to share information across cohorts and across years.

Priors may be strongly substantive, ‘informative priors’ which heavily influence the model, for example stipulating that age profiles in the population stock will be smooth (sharing the same structure as local trend time series models, only here we would smooth over age, not time). They can also be weaker, ‘uninformative priors’ such as ‘this quantity is a positive integer’. They will be fully documented, with sensitivity test results.

Figure 1 Demographic accounting framework





Adapted from Bryant, J. and Zhang, J. (2020) pages 219-220.

Data models

Within the Bayesian demographic accounting framework, data models attempt to capture biases, coverage errors, lags and other imperfections in the measurement processes related to a dataset. Allowing for one or more unreliable datasets gives us much more flexibility than requiring a single perfect dataset. In return for this extra flexibility, we need to specify explicit data models. In other words, for each dataset, we need to set out a model describing the probability of observing the dataset given the true values (Bryant and Zhang, 2020, p182).

Datasets $X_1,\ldots,X_K$ are the arrays of observed counts for each underlying true count component. There can be more than one dataset for each component. The observed counts are assumed to be related to the unobserved true counts; this relationship is captured by the data model. This data model is parameterised by $\Omega_1,\ldots,\Omega_K$, which describe how the data are generated (see Figure 1).

Potential benefits of this approach

A critical element in the development of the demographic account is model checking, where cross-referencing the data, inferences and substantive implications of the model allows us to assess its quality.

Bayesian demographic accounts incorporate statistical models of underlying demographic rates and of the input dataset allowing model checking to be built into ongoing statistical production processes. The model can be used to forecast the contents of the input datasets during the coming year. Once input data are obtained, the forecast and data can be compared. Any systematic errors, or errors that fell outside pre-specified tolerances, could be used to help diagnose weaknesses in the model or in the processes generating the input data, and help guide future improvements. This process of continuous evaluation and improvement of model and data could help prevent deterioration in the quality of the estimates in the absence of the resets provided by a traditional census.

If we succeed in developing the methods for the England and Wales context at appropriate levels of granularity then this approach offers the following advantages over existing methods:

i. Random variation in the demographic series and in the measurement of these series are properly accounted for in the Bayesian accounts
ii. The method produces detailed measures of statistical uncertainty
iii. Extra datasets and dimensions can be easily added
iv. The system accommodates missing or irregular data
v. The accounts can be used for projection, allowing timely estimates of the population
vi. The outputs will be coherent, across components and over time
vii. Methods are transparent and reproducible

(Bryant and Graham, 2019).

We are making concrete plans to get research reviewed and quality assured as it matures.

4. Guiding Principles

a. Timeliness

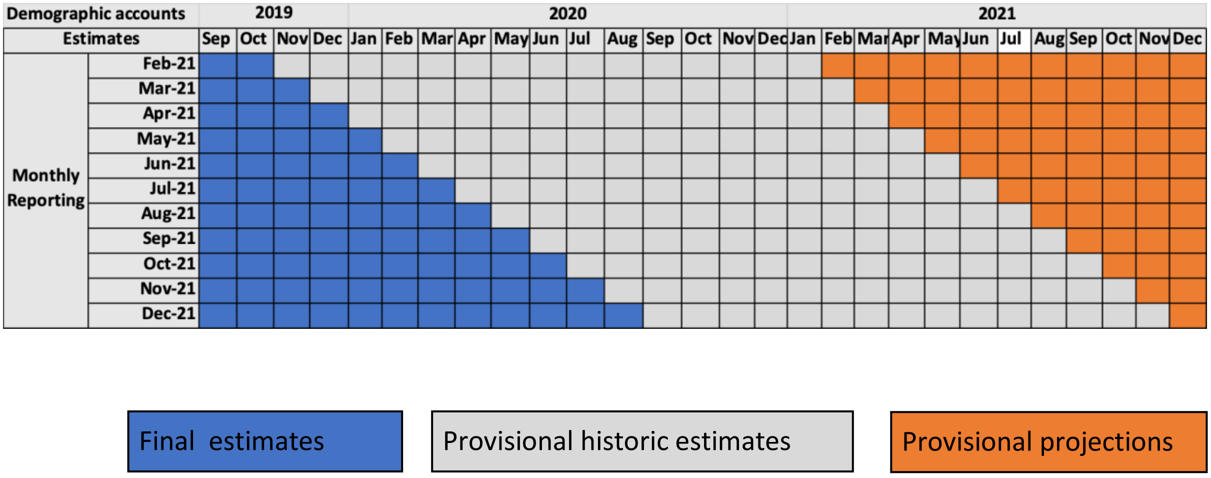
The aim of the demographic accounting model is to deliver population statistics that are coherent and timely. The COVID-19 Pandemic has complicated population estimation, as a direct result of population movement, often driven by family reunification and curtailed by policy restrictions on travel.

Further time-related issues arise from the transition to administrative-based estimation. This is best illustrated with reference to our methods for measuring international migration, described above and in the Census Research Assurance Group paper 'Methodology to estimate international migration in the absence of the International Passenger Survey' paper by Rogers, Elliott, Webber, Bangs and Georgeson on 6 July 2020. To identify long-term migrants, we need to observe their post-travel residence for 12 months. Given the time it takes to receive and process data, this implies as an absolute minimum a 15-month lag in the availability of immigration figures for time reference point *t.* This time lag could be longer given delays in data.

To produce timely statistics, we need to project international (and internal) migration from historic trends. Figure 2 describes how this would work in practice. For international migration we are still developing the methods to produce administrative microdata for migrants, so we are currently using admin-based models at the aggregate level. Figure 2 implies that population reporting in February 2021 would include final estimates to October 2019 and provisional estimates after that, with a 15-month lag.

Figure 2 Estimation window for admin-based Demographic Accounts at time *t*= February 2021*

- Assumes a 15- month lag in administrative data availability



b. Granularity

The ambition for this research is to initially, as a proof of concept, generate accounts of population stocks, births, deaths, international and internal migration by single year of age and sex at the local authority level for each calendar month.

We are currently seeking to produce data inputs to the Demographic Accounts that are aligned in terms of mid-year reference periods. Working with single years of age avoids some of the complexities of aligning data series across age groups (see Wilson and Rees 1973) but uses data that are often suppressed in outputs for disclosure control reasons (for example small counts of the very old).

c. Population model components.

Final estimates will be based on counts and estimates of population components derived in the transformed population statistics system.

Provisional estimates will use historic time series, projected using timely admin-based indicators of change. This will include weekly births and deaths and admin-based international migration estimates using Exit Checks and Passenger information delivered by Migration Statistics Division. For sub-national accounts we need to research and develop methods for estimating timely internal migration flows.

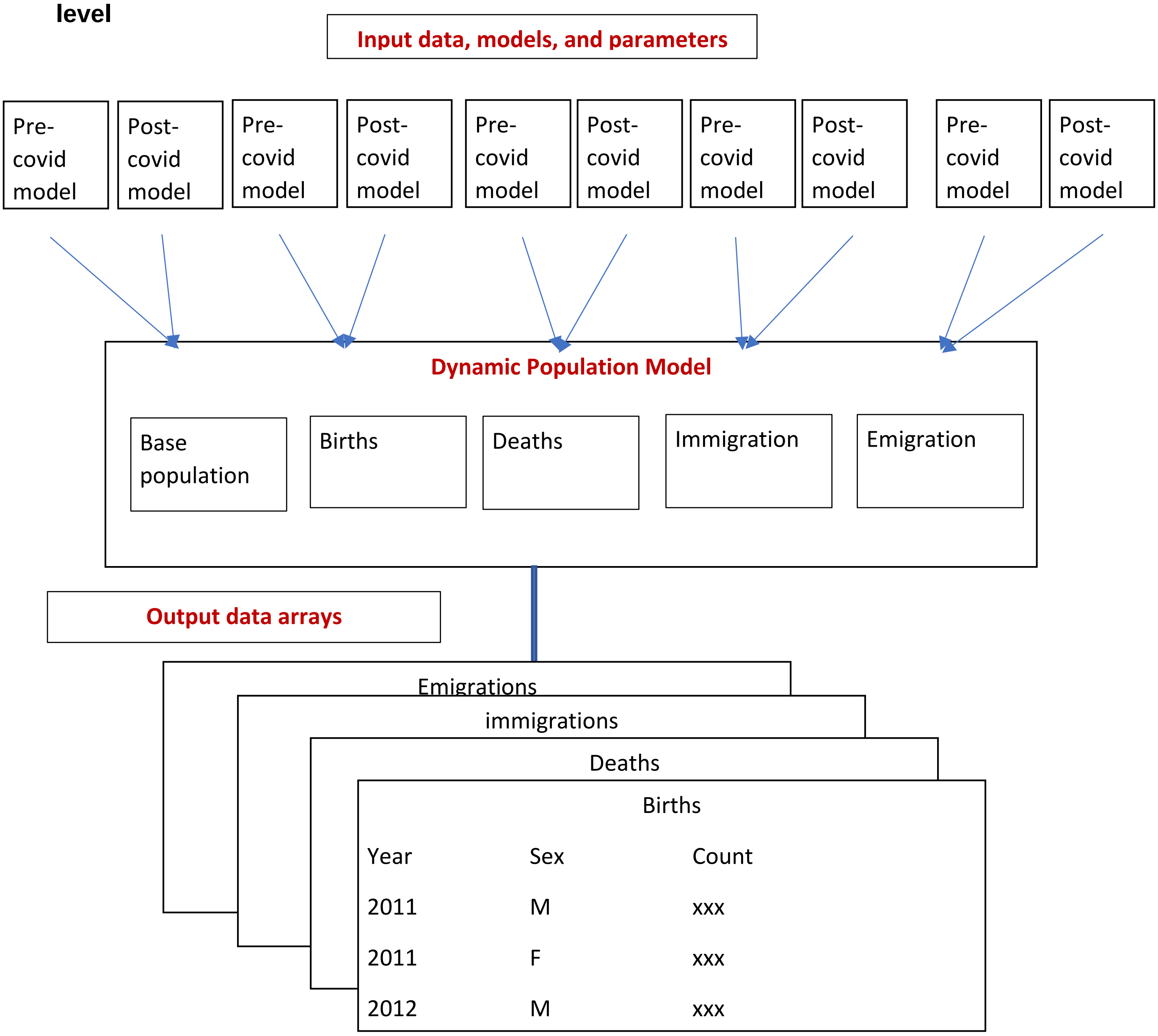
The national-level model is summarised in Figure 3.

d. Iterative development

We will begin with annual estimates at the national level (England and Wales). We will progress to re-running the model for lower geographies (local authority level). Current population estimates make a range of adjustments for ‘special populations. These can be population groups who typically do not ‘age in place’, such as students, school boarders and foreign armed forces. These groups are subject to replacement so are a revolving age-specific population. Others may age in place but have atypical interactions with our administrative systems and so require special attention. These include residents in communal establishments (for example prisons and care homes) and armed forces populations. The ultimate ambition is to produce monthly accounts.

Incremental modelling approach for demographic accounts					
Iteration	Lowest Geography	Age single year	Sex male, female	Accounting period	Components
1. Proof of concept- annual	E&W			2011-2021	Stock, International migration, Cross-border flows, natural change
3. Local Authority level- annual	Local Authority			2011-2021	Stock, International migration, Cross-border flows, natural change, inter-LA migration
4. LA with special populations- annual	Local authority			2011-2021	Stock, International migration, Cross-border flows, natural change, inter-regional migration, special populations
5. LA with special populations- monthly estimates	Local authority			2011-2021	Stock, International migration, Cross-border flows, natural change, inter-regional migration, special population

Figure 3 Overview of the Demographic Accounts model at the England and Wales level



e. Model specification guided by ONS demographers

We have established an Integrated Statistical Design Working Group with representatives from all teams within ONS who produce the components of the population estimation system. The purpose is to share and develop new methods under the Transformation Programme so that we can approach the modelling with this knowledge built in, by design. We will derive efficiencies, for example in the multiple use of source data.

The more challenging aspects of population estimation, such as the special population adjustments, need to be incorporated into the demographic accounts and this requires us to draw on the knowledge and expertise within our ONS demographic community, though the Working Group and the ONS Demographic Methods Expert Group.

5. Computational issues: Bayesian bottleneck

The algorithm used to estimate the demographic account in Statistics New Zealand in the R package **demest** is relatively slow. Obtaining high-quality estimates at the national level can take several hours. Obtaining estimates for 343 local authorities using **demest** is not practical. Speeding up the calculations is therefore a major priority of the project.

Using existing estimation tools

There are existing general-purpose tools for doing fast Bayesian inference, such as Stan and INLA. In general, however, demographic accounts have several special features that make them poor candidates for estimation using these tools, including the large number of unknowns, the large number of constraints, and the fact that they consist of discrete counts, rather than continuous values (Stan requires continuous variables).

a. A two-stage approximation of fully Bayesian estimation

The fully Bayesian approach to estimating a demographic account is to simultaneously estimate (i) counts of births, deaths, migration and population and (ii) the birth rates, death rates and migration rates underlying these counts. While this approach provides the most accurate point estimates and measures of uncertainty, it can be slow. In the UK case it should be possible, at least for initial estimates, to use a faster approximation. In the UK, the original input data are likely to be sufficiently high quality that estimates of birth rates, death rates and migration rates obtained from them are likely to be *relatively* accurate. Rather than simultaneously updating these initial estimates, we can treat them as given, and use them to guide the estimation of counts. This will allow us to make early progress, but the intention is to later revisit these assumptions.

b. Estimating counts independently within each birth cohort

In a demographic account that includes age, the basic accounting identity:

$$Populationatt + 1 = Populationatt + entriesinperiod - exitsinperiod$$

Is applied separately within each cohort. (Demographers refer to this as a Cohort Component Approach). Birth cohorts are independent in the sense that a person cannot migrate from one birth cohort to another one. In most applications, this theoretical independence does not, however, allow counts of population, births, deaths, and migration to be estimated separately for each cohort. Data sources typically aggregate over cohorts, and cohorts are linked through birth (women in many cohorts give birth to children born during any given year).

In the UK case, however, separate estimation is in fact feasible. The input data are sufficiently detailed that they do not aggregate over cohorts, and the births data are sufficiently accurate that births can be taken as fixed (which breaks the dependency between cohorts).

Estimating counts of population, deaths, and migration one cohort at a time is much easier than estimating for all cohorts simultaneously. Among other things, it allows us to use relatively simple 0models for each cohort.

We are currently working on a prototype system for estimating accounts using the independent-cohort approach. We have successfully built and tested a prototype using data disaggregated by age, sex, and 5 regions. This now needs further elaboration, for example properly separating different migration flows.

The current methods are mostly based on Metropolis-Hastings and encounter challenges in such levels of disaggregation. Such challenges are not new to the Bayesian community and there exist many general-purpose solutions we have utilised. We are using the Sequential Monte Carlo (SMC), which is well suited to producing monthly estimates which is an ambition of our work. Sequential Monte Carlo can be thought of as approximating the desired quantity through particles which each have an associated weight and then summing those particles appropriately.

There are many packages that already implement SMC or particle filtering in R, among them pomp, Nimble, and RcppSMC. We have explored the use of such packages for two main reasons, firstly it would speed up development and secondly these packages use C++ for the computationally intensive steps. Unfortunately, although we produced a simple protype that delivered considerable improvements, we found the packages too restrictive and have opted for a more tailored approach. The benefits of this approach are that we have more control over the algorithms used and the diagnostics produced. The greater flexibility of the algorithms allows us to develop methods that are more appropriate for our problem and to incorporate expert knowledge more easily. The more tailored approach allows us to include specific diagnostics of interest, for both model fitting and checking the performance of the sampler. However, this does increase the amount of code that needs to be written.

At present, we have implemented the new independent-cohort approach entirely within the two-step approximate framework outlined above. In the short term, we will continue developing within this framework, to the point where we have a working prototype for the UK. However, it should be possible, if the independent-cohort approach proves to be successful, to imbed it within a fully Bayesian framework.

6. Next steps

This research is at an early, formative stage. We have parallel workstreams that are:

- preparing and standardising the input datasets
- developing a first iteration at the England and Wales level
- we will then research to improve data models, for example incorporating special populations
- researching methods to speed up computation
- researching platform requirements for research and operational implementation
- examining the impact of COVID-19 on data supplies and demographic estimates
- Working on a plain-English account of Bayesian demographic accounting to make this methodology understandable for the widest possible audience

Once we have the first, national model, we will work with the Integrated Statistical Design Working Group to consider how to integrate internal migration into the accounts, including timely projections and reflecting the changed behaviours and data models for 2020. This includes the potential use of mobile phone data for timely signals.

References

Bryant, J. and Graham, P. (2013), A Bayesian method for deriving population statistics from multiple imperfect data sources, Proceedings from the 59th ISI World Statistics Congress, 25-30 August 2013, Hong Kong (Session IPS027).

Bryant, J. and Zhang, J.L. (2020) Bayesian Demographic Estimation and Forecasting, Chapman and Hall/ CRC Press, Boca Raton

Byron, R. (1978), The Estimation of Large Social Account Matrices, Journal of the Royal Statistical Society A, vol. 141(3), pp. 359-367.

Di Zio, M., Fortini, M. & Zardetto, D. (2018). Reconciling Estimates of Demographic Stocks and Flows through Balancing Methods, The Italian National Institute of Statistics.

Kantas, N., Doucet, A., Singh, S. S., Maciejowski, J., & Chopin, N. (2015). On particle methods for parameter estimation in state-space models. *Statistical science*, *30*(3), 328-351.

Office for National Statistics (2021) Using statistical modelling to estimate UK international migration, Newport, Wales. https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/usingstatisticalmodellingtoestimateukinternationalmigration

Raymer, J., Bai, X., Liu, N. & Wilson, T. (2020). Reconciliation of Australian Demographic Data to Study Immigrant Population Change Across Space and Time. SPATIAL DEMOGRAPHY, 8 (2), pp.123-153. https://doi.org/10.1007/s40980-020-00060-9.

Stone, R., Champernowne, D.G. & Meade, J.E. (1942), The Precision of National Income Estimates, Review of Economic Studies, vol. 9 (2), pp. 111-125.

van de Schoot, R., Depaoli, S., King, R. *et al.* Bayesian statistics and modelling. *Nat Rev Methods Primers* **1,** 1 (2021). https://doi.org/10.1038/s43586-020-00001-2

Wilson, A. G., and P. H. Rees. ‘Population Statistics and Spatial Demographic Accounts.’ Journal of the Royal Statistical Society. Series D (The Statistician), vol. 23, no. 3/4, 1974, pp. 229–257. JSTOR, www.jstor.org/stable/2987582. Accessed 16 Aug. 2021.

Annexe A Review of International Methods and Practice

Reconciliation of Australian demographic data to study immigrant population change across space and time

Raymer *et al.* (2020) developed a methodology for reconciling differences between population stock data with international migration flows and other demographic processes. The method aims to assess the contribution of the international migration component through creating a complete, consistent, and detailed account of spatial population change in Australia. The Australian model uses data on population stocks, fertility, mortality, internal and international migration. These are broken down by age, sex, geographic area, and place of birth between 1981 and 2016.

The method uses a multiregional demographic accounting model to identify discrepancies between the sources of population change data and the corresponding change found in official population stock estimates. These discrepancies are then reconciled by adjusting the reported immigration and emigration total flows and the resulting figures are then assessed by age, sex, birthplace, and subnational area.

First, a cohort-component demographic accounting model is applied for each birthplace and sex group at the regional level to calculate the error terms and demographic components of change. These error terms are distributed to immigration and emigration at a regional level. A multiregional cohort-component prediction model is then used to validate the reconciled data by how well it matches the estimated resident population and various time points.

For each of the 18 overseas-born populations, the following demographic accounting equation was used to calculate error terms by period, region, and sex:

$$\epsilon_{y,i,k}^{(t,t+5]} = P_{y,i,k}^{t} + P_{y,i,k}^{t+5} + I_{y,i,k}^{(t,t+5]} + O_{y,i,k}^{(t,t+5]} + IM_{y,i,k}^{(t,t+5]} + EM_{y,i,k}^{(t,t+5]}$$

where ε denotes the error term, P denotes the population stock, D denotes deaths, I denotes in-migration, O denotes out-migration, IM denotes immigration, and EM denotes emigration. The subscript y denotes sex, i denotes one of the 11 regions in Australia, k denotes birthplace, and t denotes a specific census year from 1981 to 2011.

Contrary to our purposes, the Australian model assumes that error terms are attributed to misreporting in the immigration and emigration data. This is due to the assumption that international migration is the largest component of change for the 18 immigrant populations of interest. In reality, error could come from any of the demographic components. In the UK context, international migration is not always the largest source of population change, with internal migration often outweighing international migration. As we know that there are coverage errors in all base-data components, we will need to take a more refined approach in which we balance the error of each component individually.

Reconciling Estimates of Demographic Stocks and Flows through Balancing Methods

The Italian National Institute of Statistics (Istat) is investigating a move from a traditional population statistics production model to one integrating as much administrative and survey data as possible. The estimates from this production process however need to be consistent with available information about demographic events, fulfilling the demographic balancing equation (DBE). However, as each component of the DBE will be estimated independently under this new system, each with associated sampling and non-sampling errors, methods are required to obtain consistent final estimates.

The method selected by Istat aims to simultaneously adjust both the initial estimates of population size and the rough civil registry figures which cover births, deaths and migration. The DBE is also used to jointly enforce the time consistency of estimated population counts and consistency between population and migration statistics that relate to different geographic areas. Istat have identified increasing computational complexity as optimisation moves to lower geographic areas, so uses NUTS 3 regions (provinces) as a trade-off.

To solve this issue, Istat are investigating methods used for balancing large systems in national accounts. Given initial estimates the aggregates entering the DBE for a given geographic area, they search for final estimates that are balanced to satisfy the DBE's and are as close as possible to the initial estimates. Therefore, the function to be minimised is the appropriate distance metric between initial and final estimates, while the constraints acting on these final estimates are the area-level DBE's. They adopt a weighted distance metric, wherein the data inputs that are assessed as more reliable tend to be changed less. This method draws on Stone *et al*. (1942), who suggest a closed form solution derived from the generalised least-squares method that explicitly recognises the impact of measurement errors on initial estimates, suggesting that less reliable initial estimates should undergo larger adjustments. However, this solution was found to be extremely computationally demanding and therefore infeasible. As a viable alternative, Istat used a derivation of the iterative constrained optimisation approach proposed by Byron (1978), which exploits the Conjugate Gradient algorithm. This approach was found to be far more computationally efficient, even for very large matrices.

The method assumes that the following are available:

- Initial estimates of population size, broken down by region, sex and age classes at times t and t+1
- Initial estimates of the natural increase between t and t+1
- Estimates of the migration flows and net migration matrices both internal and external between time t and t+1
- A measure of reliability for each of the above. This can be statistically derived or from an assessment by subject matter experts.

The method developed uses the Stone-Byron reconciling procedure to constrain the data to the demographic balancing equation and uses this to reveal any inconsistencies in the raw figures. Indicators of data coherence and degree of adjustment of the raw data required are then computed, allowing assessment of the individual raw data streams.

The study and subsequent assurance simulation study found that as well as gaining consistency between estimates, the constrained estimates exhibit lower bias and variance when compared to raw figures. Additionally, the accuracy gain appeared robust against misspecification of reliability weights. However, the model assumes that population counts at time t are of high quality and naturally finds that unbiased estimates induce balanced estimates. The further this bias is stretched, the greater the task of the model to find balance. This method should therefore work well with a robust population base at time t, such as a Census. However, the further the estimates get from this point of relative certainty, the more constraining will be needed by the model. This is likely to be case in the England and Wales context, as we move further from Census year. Further testing is required on how this model will operate under greater uncertainty.

Annexe B Markov chain and Monte Carlo.

The idea behind Markov chain Monte Carlo is to construct a Markov chain whose stationary distribution is the posterior distribution then we can derive approximate summaries of the posterior distribution from our sample of the stationary distribution. Further detail can be found in

Please refer to van de Schoot et al (2021) for a review of Bayesian computation and Kantas et al (2015) for sequential Monte Carlo methods that we are using to update the accounts.

Kantas, N., Doucet, A., Singh, S. S., Maciejowski, J., & Chopin, N. (2015). On particle methods for parameter estimation in state-space models. *Statistical science*, *30*(3), 328-351.

van de Schoot, R., Depaoli, S., King, R. *et al.* Bayesian statistics and modelling. *Nat Rev Methods Primers* **1,** 1 (2021). https://doi.org/10.1038/s43586-020-00001-2

Appendix – Presentation from MARP23 (6th December 2021)

Demographic Accounts

Prototype model to inform the 2023 National Statistician's recommendation on the traditional census

Presentation to MARP December 6th 2021

Louisa Blackwell, John Bryant, Duncan Elliott.