Progress on an integrated statistical design for the transformed population and social statistics system: Bayesian methods for demographic estimation

Duncan Elliott (Methodology and Quality) and Louisa Blackwell (Social Statistics Transformation and Research)

1. Purpose

This paper is a very short progress report on the research to develop a prototype demographic accounting model for the population of England and Wales and is presented to the Methodological Assurance Review Panel (MARP) for information only. A detailed technical paper describing current and proposed developments is also appended for interest.

2. Background

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

At the December 2021 meeting of MARP we presented plans for transforming the population and social statistics system using Bayesian methods for demographic estimation. At the May 2022 meeting of MARP we provided a short presentation on progress to date and shared with the panel our proposed publication schedule and an early draft of a technical paper for those panel members interested in further details of the approach.

The main aims of the research are to provide a proof-of-concept for incorporating survey, and administrative data on population stocks and flows within a statistical framework to produce,

i. annual estimates of population stocks and flows for England and Wales by Local Authority (LA) single year of age and sex from mid-2011 onward, including a nowcast of population for the current year,
ii. monthly estimates of population stocks and flows for England and Wales by single year of age and sex from mid-2011 onward, including a nowcast of population for the current month,
iii. monthly estimates of population stocks and flows for England and Wales by LA, single year of age and sex from mid-2011 onward, including a nowcast of population for the current month.

3. Progress

The main progress points since we last presented our work to MARP include

- Reprioritization to focus on developing annual estimates of population stocks and flows from the DPM as National Statistics. Currently paused development work on monthly estimates
- Extension of method to incorporate uncertainty in population flow rates in the model
- R packages developed for estimating dynamic population model using particle filter method

- R package for splitting combined migration estimates into full origin destination matrices developed
- Publication of model results for a synthetic Local Authority (https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/dynamicpopulationmodelforenglandandwales/2022-07-14)
- Engagement exercise with 14 Local Authorities and subsequent publication of model results for those Local Authorities (https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/dynamicpopulationmodelforlocalauthoritycasestudiesinenglandandwales/2011to2022#data-sources-in-the-dpm)
- Improvements to models for expected population flow rates that are used as inputs
- Testing new faster estimation method for the model using R package Template Model Builder that uses Laplace's method to generate a fast approximation of the full posterior distribution

4. Documentation and peer review

The first version of the technical documentation described the development of the model and particle filtering algorithms used for estimation. We have received useful comments and suggestions on this from colleagues at the University of Southampton, Professor Peter W. Smith, Professor Jakub Bijak and Doctor Jason Hilton, and also from Professor Jon Forster at the University of Warwick.

The version of the technical paper that we have attached to this brief progress update includes amendments based on the feedback we have received as well as expanding upon the developments that we have since made and including further detail on the way in which the methods have been implemented with data for England and Wales, and some example model outputs.

We are still in a research phase and so will continue to engage with experts on the methods we are using. We have had, and will continue to have, regular engagement with stakeholders. We plan to submit appropriately adapted sections of the technical paper to academic journals for fully independent peer review.

5. Next steps

The next steps for developing the dynamic population model include

- Accuracy, robustness: Improve the accuracy of the estimates, and make the production system more maintainable and reliable.
- Inputs: Expand the range of data sources that the DPM can ingest.
- Outputs: Expand the level of detail the DPM provides, and the topics that it addresses.
- Checking: Build tools to monitor the performance of the DPM.

The following publications are also planned

- February 2023 publication of DPM estimates for 331 LAs to June 2022
- Summer 2023 publication comparing Mid-Year Estimates, Statistical Population Dataset (SPD) Estimates and DPM estimates for 331 LAs to June 2022
- Autumn 2023 publication of DPM estimates for 331 LAs to June 2023

Design and Implementation of the Dynamic Population Model: Version 2.0

DPM Project Team

February 10, 2023

Abstract

This paper is a technical description of the Dynamic Population Model (DPM), a new approach to population estimation being developed at the Office for National Statistics. The core of the DPM is a demographic account, a set of disaggregated, internally-consistent estimates of population, births, deaths, and migration. The information sources used to build the demographic account include multiple imperfect quantitative datasets, and qualitative information on features such as data reliability, which are combined and synthesised within a formal statistical framework.

The principal challenge in building the DPM has been developing methods which are fast enough to produce estimates, by age and sex, for 331 Local Authorities. The DPM breaks the estimation process down into multiple parts, and uses a novel method for the most computationally-demanding part, which is the estimation of values within each Local Authority. This paper describes the work we have done to date, and outlines possible extensions. The paper will be continually revised as the methods and supporting software develop. Comments and suggestions are welcome.

Acknowledgements

We are very grateful to have received detailed comments and suggestions for improvements on earlier versions of this paper from colleagues at the University of Southampton, Professor Peter W. Smith, Professor Jakub Bijak and Doctor Jason Hilton, and also from Professor Jon Forster at the University of Warwick. We are also very grateful for comments and suggestions that we have received from members of the Methodological Assurance Review Panel. Any errors are of course our own.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The Dynamic Population Model (DPM) project team at the Office for National Statistics (ONS) is developing a new set of methods and software for demographic estimation and forecasting. The aim is to produce estimates and forecasts that are more timely, detailed, flexible, and transparent than is possible with existing ONS methods. These improvements all depend on the use of formal statistical modelling. Expert judgment and qualitative information will play a role, but will take the form of explicit modelling assumptions.

This paper describes the state of the model in early 2023. The model is still under active development. However, initial results from the model will help inform the National Statistician's 2023 recommendations to government on the future of population and social statistics [Benton, 2021].

Once it is sufficiently mature, the DPM will be used to produce official statistics as part of routine production processes. A system that is to be used in a production process for official statistics must be reliable and maintainable. It also needs to be fast, and able to accommodate changes to input data or to the specification of outputs.

There are no existing systems for population estimation that meet all the requirements of the DPM. The difficulty of scaling up existing statistical methods for population estimation was, for instance, a recurring theme in discussions at the United Nations expert group meeting on population estimation and forecasting in 2020 [United Nations Population Division, 2020]. In the absence of an existing comprehensive solution, the DPM project has adopted a framework that meets its non-computational requirements, and changed the way the framework is implemented so that it meets the computational requirements as well.

The framework that the DPM has adopted is Bayesian demographic accounts [Bryant and Graham, 2013, Bryant and Zhang, 2018]. Bayesian demographic accounts yield internally-consistent demographic estimates from multiple noisy datasets, in a transparent and reproducible way.

The existing software for estimating Bayesian demographic accounts, the R package **demest** [Bryant et al., 2021], takes approximately 30 hours to produce annual estimates, by age and sex, for England and Wales over the period 2011–

2021. This is far too slow to be scaled up to 331 Local Authority estimates, or even to do thorough testing of the England and Wales model.

The DPM's new approach to estimating Bayesian demographic accounts has two distinctive features aimed at maximising speed, simplicity, and reliability:

Sequential estimation. Rather than estimating all unknown quantities simultaneously, the DPM breaks the estimation into discrete steps, carried out sequentially.

Cohort state space estimation. We implement a novel approach to estimating demographic accounts where we apply state space methods to cohorts within accounts.

This paper sets out our strategy for estimating Bayesian demographic accounts. It provides an overview of the system as a whole, gives details of the current implementation of the system, and outlines further work that will be required to make it ready for the production of official statistics.

Chapter 2

Framework

2.1 Bayesian demographic accounts

A demographic account is a systematic tabulation of demographic stocks and flows over time, disaggregated by dimensions such as age, sex, and geography. It is the demographic equivalent of national accounts, and in fact shares a common intellectual origin with them [Rees, 1979, Stone, 1984, Willekens, 2011]. As with a national account, the detail, standardisation, and consistency of a demographic account mean that it can be used in many ways.

The stocks and flows in a demographic account conform to the accounting identity that change in stock over a period equals inflows during that period minus outflows. Inflows include births and in-migration, and outflows include deaths and out-migration. The accounting identity applies to the account as a whole, and also to sub-populations, such as the population of a particular area.

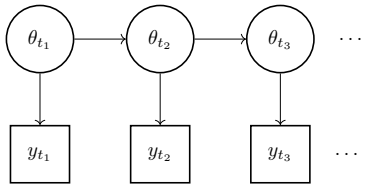


Figure 2.1: Structure of a state space model. The y_t are data and the θ_t are latent quantities that must be inferred. Arrows represent probabilistic relationships.

Bayesian demographic accounts are an example of a Bayesian state space model. Bayesian state space models have become popular, in recent years, among demographers, ecologists, and epidemiologists working on challenging problems in demographic inference [e.g. King et al., 2009, Wheldon et al., 2013, Raymer et al., 2013, Newman et al., 2014, Alkema et al., 2016, Flaxman et al., 2020, Auger-Méthé et al., 2021]. A generic state space framework is depicted in Figure 2.1. Data y_t are assumed to be generated by unobserved processes

described by vector θ_t. The θ_t are assumed to give a sufficiently complete description of the system that (i) data y_t are independent conditional on the θ_t, and (ii) θ_{t-1} is no help in predicting θ_{t+1} if we already know θ_t. The problem of estimating unknown quantities within a state space model is well suited to Bayesian statistical methods, which are extremely flexible, and can accommodate large systems with many unobserved components. The main feature distinguishing Bayesian demographic accounts from other Bayesian state space applications is their size, in that a model such as the DPM requires unusually large amounts of input data and produces unusually detailed outputs [Bryant and Zhang, 2018].

Figure 2.2 shows the contents of a Bayesian demographic account. In contrast to Figure 2.1, which disaggregates by time period, Figure 2.2 disaggregates by component.

The circles marked 'hyper-parameters' and 'rates' at the top of Figure 2.2 together form system models for the births, deaths, and migration. These system models describe demographic regularities, such as the tendency for mortality rates to fall and then rise with age. The circles marked 'model' at the bottom of Figure 2.2 are data models. Each data model depicts the assumed relationship between the (unknown) true counts and the (known) reported counts.

The datasets in Figure 2.2, represented by squares, correspond to the y in Figure 2.1. For simplicity, the diagram shows each demographic series as having a single dataset. In practice, however, a series can have any number of datasets, including zero.

The system models approximate what a skilled analyst would know about plausible values for demographic parameters, while the data models approximate what a skilled analyst would know about the quality of the data sources. Specifying and fitting a Bayesian demographic account is a more formal and statistical way of doing the production tasks of assessing data sources, adjudicating among inconsistent values, smoothing through random variation, reconciling stocks and flows, and checking for demographic plausibility.

The output from a Bayesian demographic account is a comprehensive and internally-consistent set of counts and rates. This consistency takes a strong form. Population, births, deaths, and migration satisfy the demographic accounting identities. Birth rates, death rates, and migration rates are all integrated with the counts. Uncertainties about death rates, or about the population at risk of dying, for instance, are reflected in uncertainties about death rates.

Bayesian demographic accounts can be estimated using the open source R package **demest** [R Core Team, 2020, Bryant et al., 2021]. This package uses Markov chain Monte Carlo methods, customised for demographic accounts. However, despite the computationally-intensive code being translated into C and optimised for speed, the calculations are still prohibitively slow. In addition, the code dealing with demographic accounts is complex and difficult to maintain.

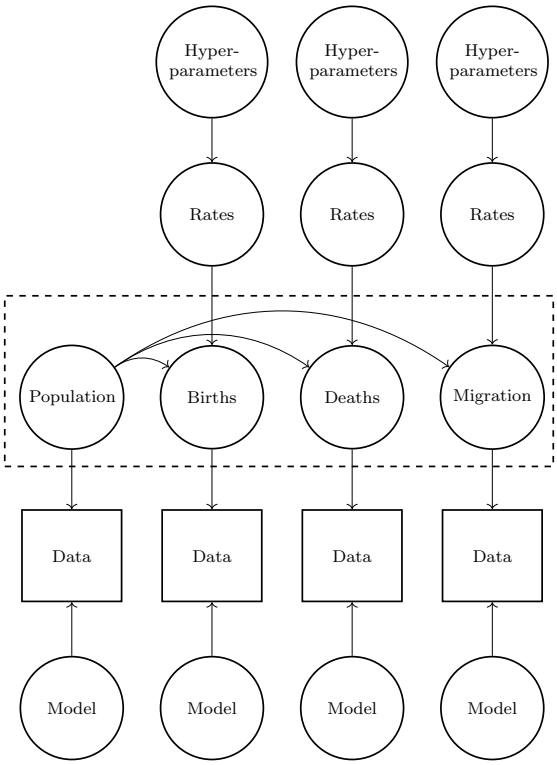
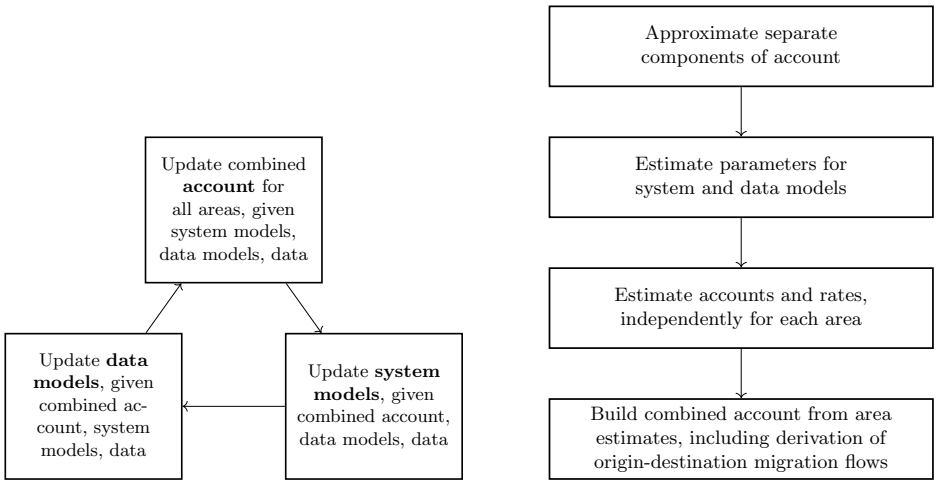


Figure 2.2: Structure of a Bayesian demographic account. Squares represent quantities that are treated as known (corresponding to y in Figure 2.1); circles represent quantities that are treated as unknown (corresponding to θ in Figure 2.1); arrows represent probabilistic relationships; and the dashed line marks the boundaries of the demographic account. Although each demographic series is depicted in the figure as having exactly one dataset and one associated data model, each series can in fact have zero, one, or more datasets, with the corresponding number of data models. Similarly, the "Migration" series can consist of multiple series, each with its own datasets and data models. The arrows from population to births, deaths, and migration reflect the fact that models for births, deaths, and migration include exposure terms.

2.2 Computational strategy

The DPM project team is completely redesigning and rebuilding the methods and software for estimating Bayesian demographic accounts, prioritising simplicity and speed. In Section 2.2, we introduce the two main features of the new approach that support these objectives. Subsequent sections and the Appendix fill in the details.

2.2.1 Sequential calculations



(a) Classic Bayesian approach

(b) DPM sequential approach

Figure 2.3: Contrasting strategies for estimating Bayesian demographic accounts. The classic Bayesian computational strategy is to cycle through components of the model, updating each component conditional on the remaining components until convergence is reached. The DPM strategy is to carry out each stage only once. This strategy requires some use of approximations.

The DPM has achieved major gains in speed, scalability, and simplicity by avoiding the classic Bayesian iterative approach to computation, and instead taking a more sequential approach. Figure 2.3 summarises the differences. Sections 5.1.2 and 5.1.3 describe our plans to address the potential loss of accuracy that may arise in the sequential approach from not accounting for dependencies between areas.

The methods implemented by R package **demest**, for other methods for Bayesian population estimation [Wheldon et al., 2013, Alexander and Alkema, 2022], are based on Markov chain Monte Carlo (MCMC). A large model containing all the unknown quantities is set up, and estimation consists of cycling through the model, updating each component conditional on all the others, until the process converges. This approach is very powerful, and produces compre-

hensive, internally-consistent estimates for all the unknowns. The disadvantage is that the wait for convergence can be long.

Convergence is particularly slow when the unknown quantities within the model are strongly correlated with each other. The unknown quantities within a demographic account, like those of many demographic models [Yackulic et al., 2020], often *are* strongly correlated with each other. Estimates of birth, death, and migration rates, for instance, are strongly correlated with estimates of birth, death, and migration counts. The result is that, even with careful specification of models, convergence can require infeasibly long computation times.

A second practical disadvantage of large MCMC-based models is that they can be difficult to understand and maintain. When every part of the system can affect every other part, tracing the origins of a possible error is challenging. The result is reduced transparency, increased cost, and increased risk.

The DPM avoids extensive MCMC-style iteration. An account disaggregated by age, sex, and Local Authority is constructed in four steps:

1. **Approximate components.** Build approximations of the series for births, deaths, migration, and population. Unlike in the final account, these series do not have to be mutually consistent.

2. **Fit system and data models.** Use the approximate series to fit models for births, deaths, and migration. The models all contain hyper-parameters, which are kept, and rates, which are discarded. Similar calculations, based on the approximate series for births, deaths, migration, and population, are also done for data models.

3. **Estimate individual accounts and rates** Using the hyper-parameters and the raw data, (re-)estimate demographic accounts and rates. Each Local Authority is estimated independently (conditional on the hyper-parameters and data).

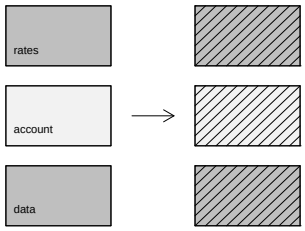
4. **Combine accounts, derive migration** Combine the individual accounts into a unified account for all of England and Wales. As part of this process, derive values for all migration flows between Local Authorities, and between all Local Authorities and the outside world.

The DPM sequential strategy relies on having relatively good data on which to base the initial estimates in Step 1, and on designing steps 2–4 so that they still work even when the initial estimates do have errors.

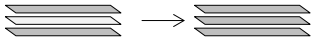
2.2.2 Cohort state space estimation

Of the steps described above, the most challenging is Step 3, the estimation of counts and rates for each Local Authority. The DPM team has developed a novel approach to Step 3, which is summarised in Figure 2.4.

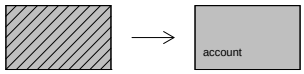
The new approach exploits the fact that, under certain conditions, the posterior distribution for the account as a whole can be expressed as the product



(a) Split system models, account, and data models into cohorts.



(b) Use state space methods to obtain cohort-level rates and counts.



(c) Assemble cohort-level counts into a demographic account.

Figure 2.4: The cohort state space approach to estimating a demographic account.

of posterior distributions for each cohort within the account. A sufficient set of conditions for cohort independence is that

- hyper-parameters in the system and data models are fixed;
- birth counts are fixed; and
- data are disaggregated by age, sex, cohort, and time.

Expressing the posterior distribution for the whole account as the product of cohort-level posterior distributions allows us to build up the posterior distribution for the whole account one cohort at a time. Estimating rates and counts for a single cohort is much easier than estimating rates and counts for the whole account. Each cohort has the generic state space format depicted in Figure 2.1, but with only a few dimensions. This simple, low-dimensional structure permits the use of specialised estimation methods that are many times faster than Markov chain Monte Carlo (MCMC).

Section 3.3.1 describes how cohort rates and counts can be estimated using particle filters [Doucet et al., 2009, Kantas et al., 2015, van de Schoot et al., 2021], which is the method we have been using in our work so far. Section 3.3.2 describes a possible alternative, using R package **TMB** [Kristensen et al., 2016], which may be even faster.

2.3 Demographic account

This section provides a detailed description of a demographic account, including notation. We describe an account for a single area. As described in Chapter 3,

we assemble an account for all of England and Wales by combining individual accounts for each Local Authority.

An account for a single area contains five generic series: population, births, deaths, in-migration, and out-migration. In-migration is defined as movement into the geographical unit, and out-migration is defined as movement out.

Counts within a demographic series are classified along four dimensions: age, sex, time, and cohort. The dimensions are defined as follows.

age Age is measured in completed years, and takes values $0, 1, \cdots, A$, where A is the maximum age that we are considering in the account. Age group A is not open ended, as this would entail mixing together cohorts.

sex The account includes a dimension called sex. One of the sexes is chosen to be used as the denominator when calculating fertility rates (we have chosen female, following the practice in other ONS outputs).

cohort "Cohort" means "birth cohort", that is, a group of people all born during the same year.

time We specify points one year apart, and use these to divide time into one-year periods. We refer to periods by the time point at the end, so that the interval between time points $t-1$ and t is called period t.

We use q^{pop}_{asct} to denote the count of people in age group a, sex s, and cohort c at time point t. Quantities describing births, deaths, in-migration and out-migration are defined similarly, and are summarised in the upper panel of Table 2.1. The table also includes a quantity q^{acc}_{asct}, called accession [Preston and Coale, 1982; Moultrie et al., 2013, p. 258]. Accession is the number of people ascending from age a to age $a+1$ during a period t, such as the number of people attaining age 65 during the year 2020. It is useful in demographic accounting, where it functions as a type of stock measure.

When working with demographic accounting identities, or when estimating counts and rates within cohorts, we have found it helpful to have a second system of notation for describing counts and rates. We refer to this notation as 'cohort-oriented' notation, as opposed to the standard 'age-oriented' notation.

Figure 2.5 illustrates age-oriented and cohort-oriented notation for stocks. The estimation period starts at time t_0. The shaded area depicts a cohort. The left panel shows stock measures for this cohort using age-oriented notation, and the right panel shows stock measures using cohort-oriented notation. The index k starts at 0 with the initial population, and then increments by 1 with each successive value for population or accession. Cohort-oriented notation uses the same superscript (stk) to denote population and accession.

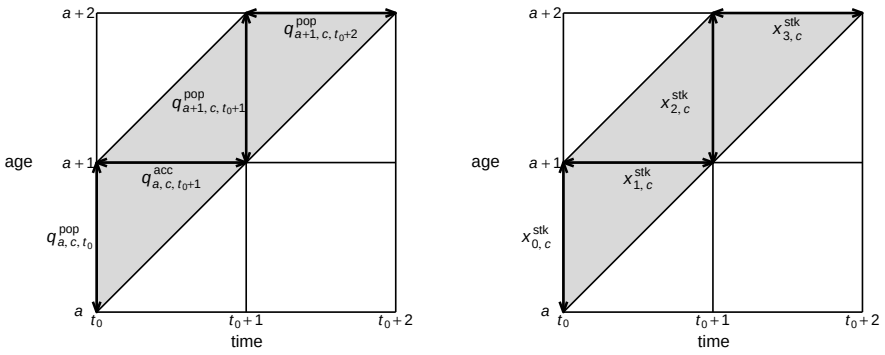
If a cohort is born during the estimation period, rather than before it, then the initial stock is the number of births occurring during the year, summed over the age and cohort of parents,

$$x^{\text{stk}}_{0,s,t} = \sum_a \sum_c q^{\text{bth}}_{asct}. \tag{2.1}$$

Table 2.1: Notation for demographic accounts

Quantity	Definition
Age-oriented notation	
q^{pop}_{asct}	Count of people belonging to age group a, sex s, and cohort c at time point t
q^{acc}_{asct}	Count of people in sex s and cohort c attaining age $a+1$ during period t (accession)
q^{bth}_{asct}	Count of births of sex s to members of the sex chosen for calculating the exposure term in fertility rates in age group a and cohort c during period t.
q^{dth}_{asct}	Count of deaths of people in age group a, sex s, and cohort c during period t.
q^{in}_{asct}	Count of in-migrations by people in age group a, sex s, and cohort c during period t.
q^{out}_{asct}	Count of out-migrations by people in age group a, sex s, and cohort c during period t.
Cohort-oriented notation	
$x^{\text{stk}}_{ksc}, k=0, c \leq t_0$	Count of population in sex s and cohort c at time point t_0.
$x^{\text{stk}}_{ksc}, k=0, c > t_0$	Count of births of sex s during period c.
$x^{\text{stk}}_{ksc}, k>0$	Count of population or accession for people in sex s and cohort c at the right or upper boundary of Lexis triangle k.
x^{bth}_{ksc}	Count of births of sex s to people in Lexis triangle k and cohort c.
x^{dth}_{ksc}	Count of deaths of people in Lexis triangle k, sex s, and cohort c.
x^{in}_{ksc}	Count of in-migrations by people in Lexis triangle k, sex s, and cohort c.
x^{out}_{ksc}	Count of out-migrations by people in Lexis triangle k, sex s, and cohort c.

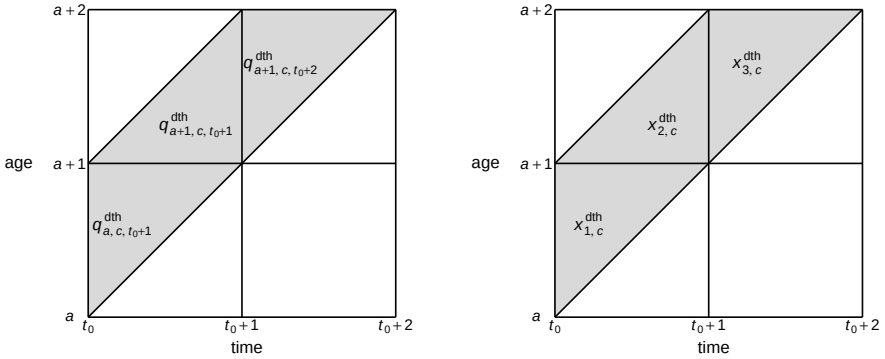
Period t is the interval between time points $t-1$ and t. Estimation starts at point t_0, implying that the first period for which flows are estimated is period t_0+1.



(a) Age-oriented (b) Cohort-oriented

Figure 2.5: Age-oriented versus cohort-oriented notation for stocks in a cohort born before the start of the estimation period. Sex subscripts s have been omitted.

If we define births as accession to age 0, then for a cohort born during the estimation period, stock at $k = 0$ is accession, stock at $k = 1$ is population, stock at $k = 2$ is accession, and so on.



(a) Age-oriented (b) Cohort-oriented

Figure 2.6: Age-oriented versus cohort-oriented notation for deaths in a cohort born before the start of the estimation period. Sex subscripts s have been omitted.

Figure 2.6 illustrates notation for flows, using the example of deaths. Flows are classified according to the 'Lexis triangle' that they belong to. Lexis triangles are defined in diagrams such as Figure 2.6 by the vertical lines marking out time, the horizontal lines marking out age, and the diagonal lines marking out cohort. The cohort in Figure 2.6a, for instance, traverses three Lexis triangles, from the bottom left to the top right. The first triangle is an "Upper" Lexis triangle, the second is a "Lower" Lexis triangle, and the third is an "Upper" Lexis triangle.

The notation for flows starts with $k = 1$, and then increments with each Lexis triangle.

Using cohort-oriented notation, the basic accounting identity for the demographic account in the base model can be stated very simply:

$$x^{\text{stk}}_{ksc} = x^{\text{stk}}_{k-1,s,c} - x^{\text{dth}}_{ksc} + x^{\text{in}}_{ksc} - x^{\text{out}}_{ksc}. \tag{2.2}$$

2.4 System models

We model deaths using

$$q^{\text{dth}}_{asct} \mid \gamma^{\text{dth}}_{asct}, e_{asct} \overset{\text{ind}}{\sim} \text{pois}(\gamma^{\text{dth}}_{asct} e_{asct}), \tag{2.3}$$

which states that, conditional on mortality rate $\gamma^{\text{dth}}_{asct}$ and exposure e_{asct}, deaths q^{dth}_{asct} are drawn independently from a Poisson distribution with mean $\gamma^{\text{dth}}_{asct} e_{asct}$.

Exposure e_{asct} is calculated using

$$e_{asct} = \begin{cases} (q^{\text{acc}}_{a-1,s,c,t} + q^{\text{pop}}_{a,s,c,t+1})/4 & \text{if } a, c, t \text{ refer to a lower Lexis triangle} \\ (q^{\text{pop}}_{a,s,c,t-1} + q^{\text{acc}}_{asct})/4 & \text{if } a, c, t \text{ refer to an upper Lexis triangle,} \end{cases} \tag{2.4}$$

or, in cohort-oriented notation,

$$e_{ksc} = (x^{\text{stk}}_{k-1,s,c} + x^{\text{stk}}_{ksc})/4. \tag{2.5}$$

The expressions for exposure are derived by multiplying the average number of people in a Lexis triangle by the average number of years that a person spends in a triangle, which is $\frac{1}{2}$.

The model for out-migration has an identical structure to the model for deaths,

$$q^{\text{out}}_{asct} \mid \gamma^{\text{out}}_{asct}, e_{asct} \overset{\text{ind}}{\sim} \text{pois}(\gamma^{\text{out}}_{asct} e_{asct}). \tag{2.6}$$

Let s_{E} denote the sex chosen for calculating exposure. Our model for fertility is

$$q^{\text{bth}}_{asct} \mid \gamma^{\text{bth}}_{asct}, e_{a,s_{\text{E}},c,t} \overset{\text{ind}}{\sim} \text{pois}(\gamma^{\text{bth}}_{asct} e_{a,s_{\text{E}},c,t}), \tag{2.7}$$

where the a, c, and s_{E} subscripts describe the parent, and the s subscript describes the child.

Finally, the model for in-migration is

$$q^{\text{in}}_{asct} \mid \gamma^{\text{in}}_{asct} \overset{\text{ind}}{\sim} \text{pois}(\gamma^{\text{in}}_{asct}). \tag{2.8}$$

Unlike the models for deaths, out-migration, and births, the model for in-migration does not include an exposure term. The reason for omitting the exposure term is that in-migration originates from outside the system and does not have a well-defined population at risk.

In these models the assumption that counts conditional on the rates and exposure follow a Poisson is not as restrictive as it might first appear. This is

because the partly hierarchical models that we use for estimating expected rates allow for over-dispersion. Section 5.1.3 outlines our plans for investigating the further developments of these hierarchical models.

Birth, death, and migration rates $\gamma^{\text{bth}}_{asct}$, $\gamma^{\text{dth}}_{asct}$, $\gamma^{\text{in}}_{asct}$, and $\gamma^{\text{out}}_{asct}$ are in turn modelled, with these higher-level models having their own hyper-parameters. The specifications that we have used in our modelling to date are described in Chapter 4.

2.5 Data models

The DPM project team is developing a suite of data models, ranging from simple to complex, including models that are general-purpose and models that are customised to particular datasets. Here we illustrate the design of data models using a simple example. Details on the implementation of the model are given in Section 4, and plans for extensions are described in Section 5.2.1.

Our simple model employs a normal distribution with means and variances that are treated as known. We assume that $y^{(d)}_{asct}$, the reported population value from population dataset d, follows a normal distribution centered on the true population value, q^{pop}_{asct}, multiplied by net coverage ratio $\rho^{(d)}_{asct}$. When we allow for the fact that q^{pop}_{asct}, unlike the normal distribution, is discrete, we obtain

$$\begin{aligned} p\left(y^{(d)}_{asct}\right) = \Phi\left(y^{(d)}_{asct} + 0.5 \mid \rho^{(d)}_{asct} q^{\text{pop}}_{asct}, (\sigma^{(d)}_{asct})^2\right) \\ - \Phi\left(y^{(d)}_{asct} - 0.5 \mid \rho^{(d)}_{asct} q^{\text{pop}}_{asct}, (\sigma^{(d)}_{asct})^2\right), \end{aligned} \tag{2.9}$$

where Φ denotes the cumulative distribution function for the normal distribution.

Values for $\rho^{(d)}_{asct}$ and $\sigma^{(d)}_{asct}$ are supplied by the user and are treated as fixed. Information on values for $\rho^{(d)}_{asct}$ and $\sigma^{(d)}_{asct}$ can come, for instance, from coverage surveys, metadata, or previous studies of data quality. A value of 0.9 for $\rho^{(d)}_{asct}$, for instance, implies that any over-coverage is outweighed by under-coverage so that the reported value $y^{(d)}_{asct}$ is expected to understate the true value q^{pop}_{asct} by 10%.

Under our normal-distribution data model, the $y^{(d)}_{asct}$ are independent, conditional on the model parameters. Conditional independence is something we build into all our data models.

2.6 Back series, extending series, and forecasting

The DPM needs to provide three types of estimates:

Back series Historical back series, such as accounts and associated rates for the period 2011–2022.

Extending series Adding new values to an existing series, such as adding values for 2023 to an existing series for 2011–2022.

Forecasting Producing values for future years, such as doing a forecast for the period 2024–2028 in the year 2023.

The distinction between extending series and forecasting can become blurred, since values for existing series sometimes have to be produced before all the main data sources have yielded numbers of the period in question.

The question of how long to wait before producing estimates can be difficult. Some data users require the most up-to-date values possible, but the longer the reporting lag, the more input data becomes available, and the more accurate the estimates can be. National Accounts face this same trade-off, which they resolve by producing provisional and final estimates. The same solution may make sense for demographic accounts. The DPM team is currently discussing options with our stakeholders.

Chapter 3

Computation

Dependencies between rates and counts creates an awkward circularity in the estimation of demographic systems. To infer birth, death, and migration rates, we need counts of births, deaths, migration, and population. But to estimate counts of births, deaths, migration, and population, we would like to make use of regularities in birth, death, and migration rates. MCMC-based methods deal with the circularity by updating counts conditional on rates, and then updating rates conditional on counts, repeating the process thousands of times.

In the DPM, we make do with only two conditional updates. In Step 1, we construct a high-quality approximation to the true account; in Step 2, we use this approximation to estimate system and data models (the first conditional update); and in Step 3 we use system and data models to estimate Local Authority accounts and rates (the second conditional update). The Local Authority accounts are combined into an England and Wales account, and origin-destination migration flows are derived, in Step 4.

3.1 Step 1: Initial approximation of components of account

We start by constructing an initial approximation of the birth, death, migration, and population series that make up Local Authority accounts. The likely accuray of this approximation varies by demographic series. Births and deaths data are extremely high quality, as is population data in census years. Population data in non-census years, data on migration flows within the England and Wales, and data on migration between England and Wales and Scotland and Northern Ireland are somewhat less reliable, but still good. Data on international migration are currently the least reliable. Chapter 4 provides more detail on the specific data sources that we have been using, and on the processes we follow to clean the data.

When we approximate the components of the demographic account, we do not require that the components satisfy the demographic accounting identities.

Satisfaction of the identities is not required for the next step in the process, the estimation of system and data models.

3.2 Step 2: Estimation of system and data models

As shown in Figure 2.2, the system models for births, deaths, and migration are all hierarchical, with parameters for rates governed in turn by sets of hyper-parameters. The hyper-parameters capture patterns in the rates. A model for births, for instance, might contain a rate parameter for every combination of age, sex, cohort, area, and time, but also a smaller number of hyper-parameters that capture overall levels in each Local Authority, or age-profiles for all England and Wales.

Inclusion of hyper-parameters can enable system models to pool strength across ages, sexes, areas, and times, leading to more stable estimates. Hyper-parameters also provide a way of modelling shared trends across all of England and Wales, such as the long-term downward drift in mortality rates, interrupted by the COVID pandemic. As described in Chapter 4, in our work to date, we have not fully exploited the potential benefits of hyper-parameters, but this is something we will pursue in future.

Estimates for rate parameters reflect local variation, while estimates for hyper-parameters reflect broader trends. Although estimates for hyper-parameters, like those for rate parameters, are vulnerable to systematic errors in the input data, they are at least partly protected against idiosyncratic errors affecting small numbers of cells. Our estimation strategy reflects these differences, in that we retain the hyper-parameters that we estimates in Step 2, but discard the rates parameters. The aim of Step 2 is to capture the main features of fertility, mortality, and migration rates, and not the fine details.

As we discuss in Section 2.5 and Chapter 4, none of the data models that we are currently using have the same rates versus hyper-parameters structure as the system models. In fact, in our work to date, we have, for simplicity, used data models in which all parameters are estimated from the initial approximation of the account. This is likely to change, however, as the DPM develops. It is likely that in future we will estimate the parameters of at least some data models, like the parameters of system models, in two stages.

3.3 Step 3: Estimation of demographic counts and rates

Step 3, the estimation of the demographic account for each Local Authority, and the associated birth, death, and migration rates, is the most challenging of the four steps. We describe two alternative estimation methods, the first based on particle filters and the second based on Laplace's Method, as implemented

R package **TMB**. We have used the first method to produce full scale accounts, and are still experimenting with the second method.

3.3.1 Estimation via particle filters

We describe the particle filters that we have been using for estimating a back series, and then describe the slightly different particle filters that we use for adding values to existing series. The particle filters are applied separately to every combination of sex s and cohort c. To reduce clutter, we omit references to s and c. We suppress dependence on hyper-parameters and data model parameters, which we are treating as fixed. In the models that we have implemented to date, we have also treated births and deaths, which are measured extremely accurately in the UK, as known and fixed.

Let

$$\boldsymbol{x}_0 = x_0^{\text{stk}} \tag{3.1}$$

$$\boldsymbol{x}_k = (x_k^{\text{stk}}, x_k^{\text{bth}}, x_k^{\text{dth}}, x_k^{\text{in}}, x_k^{\text{out}}), \quad k = 1, \cdots, K \tag{3.2}$$

$$\boldsymbol{\gamma}_k = (\gamma_k^{\text{bth}}, \gamma_k^{\text{dth}}, \gamma_k^{\text{in}}, \gamma_k^{\text{out}}), \quad k = 1, \cdots, K. \tag{3.3}$$

$$\tag{3.4}$$

Similarly, let

$$\boldsymbol{y}_k = \{y_k^{(d)}\}_{d=1,\cdots,D}, \quad k = 0, \cdots, K, \tag{3.5}$$

with the understanding that, for any given k, some or all values for $y_k^{(d)}$ may be missing. For any variable u, we use $u_{t_1:t_2}$ as shorthand for $u_{t_1}, \cdots, u_{t_2}$.

The structure of system models (2.3), (2.6), (2.7), and (2.8), and expressions for exposure (2.4) and (2.5) imply that, conditional on demographic rates,

$$p(\boldsymbol{x}_k, \boldsymbol{\gamma}_k \mid \boldsymbol{x}_{0:k-1}, \boldsymbol{\gamma}_{1:k-1}) = p(\boldsymbol{x}_k, \boldsymbol{\gamma}_k \mid \boldsymbol{x}_{k-1}). \tag{3.6}$$

In addition, the structure of data models such as (2.9) implies that

$$p(\boldsymbol{y}_k \mid \boldsymbol{y}_{0:k-1}, \boldsymbol{x}_{0:K}) = p(\boldsymbol{y}_k \mid \boldsymbol{x}_k). \tag{3.7}$$

We will make sure that any data models we introduce in future will also satisfy (3.7).

Deriving a back series for an individual cohort entails sampling from

$$p(\boldsymbol{x}_{0:K}, \boldsymbol{\gamma}_{1:K} \mid \boldsymbol{y}_{0:K}) \propto p(\boldsymbol{x}_{0:K}, \boldsymbol{\gamma}_{1:K}) p(\boldsymbol{y}_{0:K} \mid \boldsymbol{x}_{0:K}), \tag{3.8}$$

which can be decomposed into

$$p(\boldsymbol{x}_0) p(\boldsymbol{y}_0 \mid \boldsymbol{x}_{0:K}) \prod_{k=1}^{K} p(\boldsymbol{x}_k, \boldsymbol{\gamma}_k | \boldsymbol{x}_{0:k-1} \boldsymbol{\gamma}_{1:k-1}) p(\boldsymbol{y}_k \mid \boldsymbol{y}_{0:k-1}, \boldsymbol{x}_{0:K}). \tag{3.9}$$

Equations (3.6) and (3.7) allow us to simplify (3.9) to

$$p(\boldsymbol{x}_0) p(\boldsymbol{y}_0 \mid \boldsymbol{x}_0) \prod_{k=1}^{K} p(\boldsymbol{x}_k, \boldsymbol{\gamma}_k \mid \boldsymbol{x}_{k-1}) p(\boldsymbol{y}_k \mid \boldsymbol{x}_k). \tag{3.10}$$

Our algorithm for drawing from (3.10) is set out in Figure 3.1. Following standard terminology for particle filters, Figure 3.1 refers to $f(\boldsymbol{x}_k, \boldsymbol{\gamma}_k \mid \boldsymbol{x}_{k-1}) = p(\boldsymbol{x}_k, \boldsymbol{\gamma}_k \mid \boldsymbol{x}_{k-1})$ from (3.10) as the transition function, and $g(\boldsymbol{y}_k \mid \boldsymbol{x}_k) = p(\boldsymbol{y}_k \mid \boldsymbol{x}_k)$ as the likelihood. It also refers to importance function $q(\boldsymbol{x}_k, \boldsymbol{\gamma}_k \mid \boldsymbol{y}_k, \boldsymbol{x}_{k-1})$, which is an approximation of our (unnormalised) target distribution, $f(\boldsymbol{x}_k, \boldsymbol{\gamma}_k \mid \boldsymbol{x}_{k-1})g(\boldsymbol{y}_k \mid \boldsymbol{x}_k)$. The importance function is designed so that, in contrast to $f(\boldsymbol{x}_k, \boldsymbol{\gamma}_k \mid \boldsymbol{x}_{k-1})g(\boldsymbol{y}_k \mid \boldsymbol{x}_k)$, we are able to draw from it directly. The algorithm in Figure 3.1 corrects for the difference between the importance function and the target distribution through a combination of weighting and resampling.

The efficiency of a particle filter depends heavily on the ability of the importance function $q(\boldsymbol{x}_k \mid \boldsymbol{y}_k, \boldsymbol{x}_{k-1})$ to successfully generate values that are proportional to $f(\boldsymbol{x}_k \mid \boldsymbol{x}_{k-1})g(\boldsymbol{y}_k \mid \boldsymbol{x}_k)$. Our importance function is described in Section A.1. The transition function is described in Section A.2.

Input

$p_0(\boldsymbol{x}_0)$	Prior distribution for $\boldsymbol{x}_0$
$q_0(\boldsymbol{x}_0\|\boldsymbol{y}_0)$	Importance function for $\boldsymbol{x}_0$
$f(\boldsymbol{x}_k, \boldsymbol{\gamma}_k\|\boldsymbol{x}_{k-1})$	Transition function
$g(\boldsymbol{y}_k\|\boldsymbol{x}_k)$	Likelihood
$q(\boldsymbol{x}_k, \boldsymbol{\gamma}_k\|\boldsymbol{y}_k, \boldsymbol{x}_{k-1})$	Importance function for $\boldsymbol{x}_k$
a	Resampling threshold, $0 \leq a \leq 1$

Algorithm

- Generate initial values and weights
  1. For $i = 1, \cdots, N$
     (a) Draw $\tilde{\boldsymbol{x}}_0^{(i)} \sim q_0(\boldsymbol{x}_0|\boldsymbol{y}_0)$
     (b) Calculate unnormalised weights

$$\tilde{w}_0^{(i)} = \frac{g(\boldsymbol{y}_0|\tilde{\boldsymbol{x}}_0^{(i)}) p_0(\tilde{\boldsymbol{x}}_0^{(i)})}{q(\tilde{\boldsymbol{x}}_0^{(i)}|\boldsymbol{y}_0)}$$

  2. Calculate normalised weights $\tilde{W}_0^{(i)} = \tilde{w}_0^{(i)} / \sum_{i'=1}^{N} \tilde{w}_0^{(i')}$
  3. Calculate effective sample size $\hat{N}_0 = 1/\left(\sum_{i=1}^{N} (\tilde{W}_0^{(i)})^2\right)$
  4. If $\hat{N}_0 < aN$ then resample, obtaining N particles $\boldsymbol{x}_0^{(i)}$ with weights $W_0^{(i)} = 1/N$. Otherwise set $\boldsymbol{x}_0^{(i)} = \tilde{\boldsymbol{x}}_0^{(i)}$ with weights $W_0^{(i)} = \tilde{W}_0^{(i)}$.

Figure 3.1: Particle filter for estimating a back series for one cohort. The description is modified from Doucet et al. [2009]. We use a standard algorithm for resampling Carpenter et al. [1999].

(continued on next page)

(continued from previous page)

- Generate values and weights for remaining intervals
  - For $k = 1, \cdots, K$
    1. For $i = 1, \cdots, N$
       (a) Draw $(\tilde{\boldsymbol{x}}_k^{(i)}, \tilde{\boldsymbol{\gamma}}_k^{(i)}) \sim q(\boldsymbol{x}_k, \boldsymbol{\gamma}_k | \boldsymbol{y}_k, \boldsymbol{x}_{k-1}^{(i)})$
       (b) Set $(\tilde{\boldsymbol{x}}_{0:k}^{(i)}, \tilde{\boldsymbol{\gamma}}_{1:k}^{(i)}) = (\boldsymbol{x}_{0:k-1}^{(i)}, \tilde{\boldsymbol{x}}_k^{(i)}, \boldsymbol{\gamma}_{1:k-1}^{(i)}, \tilde{\boldsymbol{\gamma}}_k^{(i)})$
       (c) Calculate unnormalised weights

$$\tilde{w}_k^{(i)} = \frac{g(\boldsymbol{y}_k | \tilde{\boldsymbol{x}}_k^{(i)}) f(\tilde{\boldsymbol{x}}_k^{(i)}, \tilde{\boldsymbol{\gamma}}_k^{(i)} | \boldsymbol{x}_{k-1}^{(i)})}{q(\tilde{\boldsymbol{x}}_k^{(i)}, \tilde{\boldsymbol{\gamma}}_k^{(i)} | \boldsymbol{y}_k, \boldsymbol{x}_{k-1}^{(i)})} W_{k-1}^{(i)}$$

    2. Calculate normalised weights $\tilde{W}_k^{(i)} = \tilde{w}_k^{(i)} / \sum_{i'=1}^{N} \tilde{w}_k^{(i')}$
    3. Calculate effective sample size $\hat{N}_k = 1 / \left(\sum_{i=1}^{N} (\tilde{W}_k^{(i)})^2 \right)$
    4. If $\hat{N}_k < aN$ or if $k = K$ then resample, obtaining N particles $\boldsymbol{x}_{0:k}^{(i)}$ with weights $W_k^{(i)} = 1/N$. Otherwise set $\boldsymbol{x}_{0:k}^{(i)} = \tilde{\boldsymbol{x}}_{0:k}^{(i)}$ with weights $W_k^{(i)} = \tilde{W}_k^{(i)}$.

Output

$\{\boldsymbol{x}_{0:K}^{(i)}, \boldsymbol{\gamma}_{1:K}^{(i)}\}$ N draws from the posterior distribution $p(\boldsymbol{x}_{0:K}, \boldsymbol{\gamma}_{1:K} | \boldsymbol{y}_{0:K})$

3.3.2 Estimation via TMB

TMB, short for Template Model Builder, is an R package that has become increasingly popular for large-scale demographic estimation [Kristensen et al., 2016, Eaton et al., 2021, Dwyer-Lindgren et al., 2022, Osgood-Zimmerman and Wakefield, 2022]. TMB is often used for non-Bayesian inference, but can be used for a large class of Bayesian models, including ones with complicated prior structures. TMB uses Laplace's Method [MacKay, 2003, ch. 27] to generate a fast approximation of the full posterior distribution. The user supplies TMB with a description of the posterior distribution, on the log scale, via a template written in the C++ language. (Appendix B gives an example of a template for estimating cohort counts and rates.) TMB finds the maximum, and then uses information on curvature to approximate the distribution around this point. Laplace's Method requires first and second derivatives of the log posterior, but TMB calculates these itself, using automatic differentiation [Bell, 2006, Kristensen et al., 2016].

When using TMB to estimate rates and counts for a cohort, we work with the log posterior

$$\log p(\boldsymbol{x}_{0:K}, \boldsymbol{\gamma}_{1:K} \mid \boldsymbol{y}_{0:K}, \boldsymbol{\phi}), \tag{3.11}$$

where $\boldsymbol{\phi}$ denotes hyper-parameters. TMB requires that all unknowns in the log posterior be continuous, so we allow counts of stocks and flows to be non-integers. The use of non-integers does not prevent us from using the Poisson distribution in our model for events, since TMB uses gamma functions (which allow non-integer values) in its definition of Poisson densities. The use of non-integers within the model is not apparent to end-users, who only see summary statistics, such as means or quantiles, which already take non-integer values.

Laplace's Method produces an approximation of the posterior distribution, and, depending on the particular application, approximation errors can be non-negligible [Kristensen et al., 2016, Osgood-Zimmerman and Wakefield, 2022]. We have conducted simulation experiments to see how large approximation errors are likely to be when estimating cohorts. We find that the errors are on the order of a few percent at most, and no worse than those of particle filters, provided the number of people in a cohort is around 20 or more. When cohorts are smaller, errors in estimates of events can be large in relative terms, though small in absolute terms.

The simulations also suggested that TMB is an order of magnitude faster than particle filters. TMB requires much less computer code than hand-written particle filters. The form of the output – means and covariance matrices for the joint posterior distribution of the unknowns – is also convenient, in that it requires relatively little storage space, but can be used to generate arbitrarily large samples from the posterior distribution. As discussed in Section 5.1.2, we plan to develop a full-scale system for estimation based on TMB, as a possible replacement for particle filters.

3.4 Step 4: Combining accounts, splitting migration

To produce an account for all England and Wales, we concatenate the individual Local Authority accounts. The resulting England and Wales account contains a "Local Authority" dimension alongside the age, sex, cohort, and time. At this point, the estimation of population, births, and deaths is complete, but migration requires further work.

Concatenating Local Authority accounts allows us to derive net external migration for all of England and Wales. The output from Step 3 is vectors

$$(x_r^{\text{stk}}, x_r^{\text{dth}}, x_r^{\text{in}}, x_r^{\text{out}})$$

for every combination of age, cohort, sex, and time. (In Section 3.4, to reduce clutter, we omit age, cohort, sex, and time subscripts, retaining only subscripts for geographical areas.) Let x_{r_1,r_2}^{int} denote migration between LAs r_1 and r_2, with the convention that $x_{r_1,r_2}^{\text{int}} \equiv 0$ when $r_1 = r_2$. Let x_{wr}^{ext} denote immigration into LA r from external region w, and let x_{rw}^{ext} denote emigration from LA r to external region w. By definition,

$$x_r^{\text{in}} = \sum_{r'=1}^{R} x_{r',r}^{\text{int}} + \sum_{w=1}^{W} x_{wr}^{\text{ext}} \tag{3.12}$$

and

$$x_r^{\text{out}} = \sum_{r'=1}^{R} x_{r,r'}^{\text{int}} + \sum_{w=1}^{W} x_{rw}^{\text{ext}}. \tag{3.13}$$

Summing (3.12) and (3.13) over r, and using the fact that

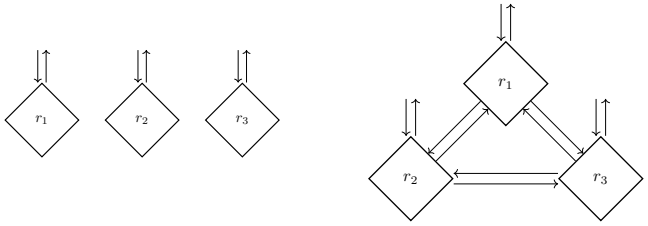
$$\sum_{r=1}^{R}\sum_{r'=1}^{R} x_{r',r}^{\text{int}} = \sum_{r=1}^{R}\sum_{r'=1}^{R} x_{r,r'}^{\text{int}}, \tag{3.14}$$

we obtain

$$\sum_{r=1}^{R}\sum_{w=1}^{W} x_{wr}^{\text{ext}} - \sum_{r=1}^{R}\sum_{w=1}^{W} x_{rw}^{\text{ext}} = \sum_{r=1}^{R} x_r^{\text{in}} - \sum_{r=1}^{R} x_r^{\text{out}}. \tag{3.15}$$

The quantity $\sum_{r=1}^{R}\sum_{w=1}^{W} x_{wr}^{\text{ext}}$ on the left hand side of Equation (3.15) is total external in-migration, and $\sum_{r=1}^{R}\sum_{w=1}^{W} x_{rw}^{\text{ext}}$ is total external out-migration. Total external net migration is the difference between external in-migration and external out-migration.

Figure 3.2 depicts the extra detail on migration that is added during Step 4. Before Step 4, all R Local Authorities have migration flows, but the origins and destinations are unspecified. After Step 4, migration flows between every pair of areas have been estimated, along with flows to and from the outside world. Though not apparent from the diagram, the sum of all flows into and out of each Local Authority is unchanged.



(a) Before disaggregation (b) After disaggregation

Figure 3.2: Disaggregating migration flows. Diamonds represent local authorities and arrows represent migration flows. Step 3 yields estimates of total in-migration into each area and total out-migration out of each area. In Step 4, we split these totals, to obtain in-migration into and out-migration out of the system as a whole, and flows between each pair of areas.

The migration flows we would like to estimate are set out in Figure 3.3a, and the inputs to the estimation are set out in Figure 3.3b. We have data $z^{\text{int}}_{r1,r2}$ for internal flows and data z^{ext}_{wr} and z^{ext}_{rw} for external flows. At present, the data on internal flows come from the Patient Register, and the data on external flows come from a mix of administrative sources, survey data, and modelling.

The problem depicted in Figure 3.3 is close to the type of problem that is solved by Iterative Proportional Fitting (IPF) [Fienberg, 1970, Willekens, 1999]. The difference is that we are missing $2 \times W$ marginal totals: the totals for immigration and emigration. However, a variant of IPF can be implemented without these totals.

We rewrite the $\boldsymbol{X}$ matrix in Figure 3.3a as

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}^{\text{int}} & \boldsymbol{X}^{\text{em}} \\ \boldsymbol{X}^{\text{im}} & \boldsymbol{0} \end{bmatrix} \tag{3.16}$$

where $\boldsymbol{X}^{\text{int}}$ is an $R \times R$ matrix, $\boldsymbol{X}^{\text{em}}$ is an $R \times W$ matrix, and $\boldsymbol{X}^{\text{im}}$ is a $W \times R$ matrix. We rewrite the Z matrix in Figure 3.3b as

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}^{\text{int}} & \boldsymbol{Z}^{\text{em}} \\ \boldsymbol{Z}^{\text{im}} & \boldsymbol{0} \end{bmatrix}, \tag{3.17}$$

with row constraints $\boldsymbol{x}^{\text{out}}$ and column constraints $\boldsymbol{x}^{\text{in}}$. We apply iterative proportional fitting, but adjusting only the first R rows and R columns at each iteration. The system converges to

$$\boldsymbol{X}^{\text{int}} = \boldsymbol{P}\boldsymbol{Z}^{\text{int}}\boldsymbol{Q} \tag{3.18}$$

$$\boldsymbol{X}^{\text{em}} = \boldsymbol{P}\boldsymbol{Z}^{\text{em}} \tag{3.19}$$

$$\boldsymbol{X}^{\text{im}} = \boldsymbol{Z}^{\text{im}}\boldsymbol{Q} \tag{3.20}$$

where $\boldsymbol{P}$ and $\boldsymbol{Q}$ are diagonal matrices with elements

$$P_r = \frac{x_r^{\text{out}}}{\sum_{r'} Q_{r'} z_{r,r'}^{\text{int}} + \sum_{w=1}^{W} z_{rw}^{\text{em}}} \tag{3.21}$$

$$Q_r = \frac{x_r^{\text{in}}}{\sum_{r'} P_{r'} z_{r',r}^{\text{int}} + \sum_{w=1}^{W} z_{wr}^{\text{im}}}. \tag{3.22}$$

The resulting solution for $\boldsymbol{X}$ satisfies the constraint that out-migration to all destinations sums to the $x^{\text{out}(i)}$s and in-migration sums to the $x^{\text{in}(i)}$s, as required. Moreover, zeros on the diagonal in $\boldsymbol{Z}$ are preserved in $\boldsymbol{X}$.

IPF is also fast, has low memory requirements, and works in more than two dimensions should this be required in future.

$$
\begin{array}{cccc:ccc}
0 & x_{1,2}^{\text{int}(i)} & \dots & x_{1,R}^{\text{int}(i)} & x_{1,1}^{\text{ext}(i)} & \dots & x_{1,\text{W}}^{\text{ext}(i)} \\
x_{2,1}^{\text{int}(i)} & 0 & \dots & x_{2,R}^{\text{int}(i)} & x_{1,1}^{\text{ext}(i)} & \dots & x_{2,\text{W}}^{\text{ext}(i)} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
x_{R,1}^{\text{int}(i)} & x_{R,2}^{\text{int}(i)} & \dots & 0 & x_{1,1}^{\text{ext}(i)} & \dots & x_{R,\text{W}}^{\text{ext}(i)} \\
\hdashline
x_{1,1}^{\text{ext}(i)} & x_{1,2}^{\text{ext}(i)} & \dots & x_{1,R}^{\text{ext}(i)} & 0 & \dots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
x_{\text{W},1}^{\text{ext}(i)} & x_{\text{W},2}^{\text{ext}(i)} & \dots & x_{\text{W},R}^{\text{ext}(i)} & 0 & \dots & 0
\end{array}
$$

(a) Flows to estimate

$$
\begin{array}{cccc:ccc:c}
0 & z_{1,2}^{\text{int}} & \dots & z_{1,R}^{\text{int}} & z_{1,1}^{\text{ext}} & \dots & z_{1,\text{W}}^{\text{ext}} & x_1^{\text{out}(i)} \\
z_{2,1}^{\text{int}} & 0 & \dots & z_{2,R}^{\text{int}} & z_{2,1}^{\text{ext}} & \dots & z_{2,\text{W}}^{\text{ext}} & x_2^{\text{out}(i)} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\
z_{R,1}^{\text{int}} & z_{R,2}^{\text{int}} & \dots & 0 & z_{R,1}^{\text{ext}} & \dots & z_{R,\text{W}}^{\text{ext}} & x_R^{\text{out}(i)} \\
\hdashline
z_{1,1}^{\text{ext}} & z_{1,2}^{\text{ext}} & \dots & z_{1,R}^{\text{ext}} & 0 & \dots & 0 & \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \\
z_{\text{W},1}^{\text{ext}} & z_{\text{W},2}^{\text{ext}} & \dots & z_{\text{W},R}^{\text{ext}} & 0 & \dots & 0 & \\
\hline
x_1^{\text{in}(i)} & x_2^{\text{in}(i)} & \dots & x_R^{\text{in}(i)} & & & &
\end{array}
$$

(b) Inputs

Figure 3.3: Outputs and inputs for disaggregating subnational migration flows. The x^{int} and x^{ext} terms are the unknown flows we wish to estimate. The z terms are reported migration flows. The x^{in} and x^{out} terms are outputs from running the base model independently on R subnational areas.

Chapter 4

Dynamic Population Model for England and Wales

In this chapter, we provide a detailed description of the methods used to generate our first full set of estimates by single year of age $(0, 1, \ldots, 105)$ and sex (Female and Male) for all 331 Local Authorities in England and Wales. We illustrate our results using the example of the Local Authority of Cambridge.

4.1 Data sources

In this section we describe the data currently available to the DPM, and process of reformatting the data for inclusion in the DPM. The input data and reformatting are likely to change in future versions of the DPM.

4.1.1 Population data

Mid-year estimates (MYE) Mid-year estimates for 2011 consist of the 2011 Census counts rolled forward from Census Day (27 March) to mid-year (30 June). Because of the proximity to the census, we assume that these data are of high quality. MYE are available by sex, single year of age (0–90+) and Local Authority. Although MYE, plus associated measures of uncertainty, are available for the whole period 2011–2021, we only use estimates for 2011. The methods used for splitting the 90+ age group into single years of age are discussed in Section 4.2.

NHS General Practice Patient Register (PR) The NHS GP Patient Register provides population estimates by sex, single year of age (with no upper bound), and Local Authority for the years 2011–2020. PR data give the number of individuals registered with a GP surgery in England and Wales.

Statistical Population Dataset, version 3 (SPD3) Version 3 of SPD provides estimates by sex, single year of age (0–90+), and Local Authority, at June 30, in 2011, and in 2016–2020. SPD are derived from administrative

data by combining various sources of "activity", and applying rules that predict whether an individual is likely to belong to the usually-resident population.

4.1.2 Births, deaths, and migration data

Births The administrative system from which we derive aggregate counts of births provides us with counts of births by sex, Local Authority, Lexis triangle, and single year of age of mothers (at the time of birth), for years ending June 2012 to June 2022. We assume the data to be of very high quality. The 2022 data is provisional as some births may be missed in the most recent months because of lags in registrations, which can legally take place up to 42 days after the birth.

Deaths The administrative system from which we derive aggregate counts of deaths provides us with counts of deaths by sex, Local Authority, Lexis triangle, and single year of age, for years ending June 2012 to June 2022. Deaths data, like births data, is assumed to be of very high quality. The 2022 data is provisional as some deaths may be missed for the most recent months because of lags in registrations and coroner related delays.

Internal migration We have estimates, constructed with ONS, of internal migration counts by sex, Local Authority of origin, Local Authority of destination, and single year of age, for years ending June 2012 to June 2022. Data for the years ending June 2012 to June 2020 use address changes in GP registrations from the NHS Patient Register (PR) and Personal Demographic Service (PDS), and higher education registrations from Higher Education Statistics Agency (HESA) data, to measure moves between Local Authorities within England and Wales. HESA data are used to move students to university addresses and to adjust for lags in students' post-study moves, ahead of a recorded GP re-registration. Age refers to age at 30 June rather than age at the date of the recorded move.

Data for years ending June 2021 and June 2022 use scaling between MYE-based and PDS-based internal migration estimates in 2018 and 2019. This is then used to impute MYE-based internal migration estimates for 2021 and 2022. These provisional imputed estimates are not comparable with previously-published estimates, and there is no adjustment for students' post-study moves. We are researching ways to produce more timely estimates of internal migration.

Cross-border flows Cross-border flows are migrations to and from (a) England and Wales and (b) Scotland and Northern Ireland. Estimates of cross-border flows are available by sex, Local Authority, and single year of age, for the years ending June 2012 to June 2022. Between 2011 and 2020, we use estimates of cross-border moves from the MYE. The total flows to and from constituent countries of the UK are agreed between the Office for National Statistics (ONS), National Records of Scotland (NRS), and the Northern Ireland Statistics and Research Agency (NISRA), based on records of in-migration to the relevant country.

Data for the period 2021 to 2022 use scaling between MYE-based and PDS-based cross-border migration estimates for 2018 to 2019. This is then used to

impute MYE-based cross-border migration estimates for 2021 to 2022. These provisional imputed estimates are not comparable with previously published estimates and have not been agreed between the constituent countries of the UK. We are researching ways to produce more timely estimates of cross-border moves.

International migration International migration estimates are available by sex, Local Authority, and single year of age (0–90+), for the years ending June 2012 to June 2022. Between 2011 and 2020, we use Long-Term International Migration (LTIM) estimates [ONS, 2022a]. They are predominantly based on the International Passenger Survey (IPS), which was suspended in March 2020 because of the COVID-19 pandemic.

Experimental estimates for March 2020 to June 2020 and for the year ending 30 June 2021 are produced using Home Office Exit Checks data and the Department for Work and Pensions Registration and Population Interaction Database (RAPID). This new method makes greater use of administrative data than previous ONS international migraiton estimates, which relied on IPS data and statistical modelling. Data from March 2020 onward are not comparable with previous estimates and may be subject to revisions.

International migration estimates for 2022 are based on forecasts. Year ending June 2022 migration estimates published on 24 November 2022 will be incorporated in future iterations. There is also active research on methods to improve international migration estimates that we will need to address when any changes are implemented.

4.2 Model specification

System models, data models, and demographic counts and rates are currently estimated separately for each Local Authority. Our notation in this section omits references to Local Authority.

4.2.1 Data models for England and Wales

In our current estimates, for simplicity, we treat the births and deaths data as error-free. This assumption could be relaxed in future versions of the model. Currently our model specification does not include data or data models for migration counts, so that all information on migration enters through the hyperparameters for migration rates. Future versions of the model will incorporate migration data and data models. Our data models for population stocks, with the stocks ordered from most reliable to least reliable, are as follows.

Mid-year estimates (MYE) We assume that mid-year estimates follow the normal-distribution data model of Equation 2.9. The mid-year estimates in 2011 are assumed to be unbiased, and the coverage ratio, ρ_{asct}^{MYE} is set to 1 in all cells. ONS publishes estimates of the uncertainty in the MYE, in the form of standard errors $\hat{\sigma}_{ast}^{\text{MYE}}$ disaggregated by single year of age, sex, and Local Authority [ONS, 2022b].

We split the published population counts by assuming that each LA's population has the same age-structure as England and Wales as a whole. (ONS publishes single year of age estimates beyond age 90 for England and Wales, but not for individual LAs.) We derive disaggregated standard errors by assuming that the coefficient of variation for disaggregated cell is equal to the coefficient of variation for the aggregated cell to which it belongs.

$$\hat{\sigma}^{\text{MYE}}_{ast} = \begin{cases} \frac{\hat{\sigma}^{\text{MYE}}_{90+,s,t,}}{y^{\text{MYE}}_{90+,s,t,}} y^{\text{MYE}}_{ast}, & a \geq 90 \\ \hat{\sigma}^{\text{MYE}}_{ast}, & \text{otherwise} \end{cases} \tag{4.1}$$

Statistical Population Dataset Version 3 (SPD3) Although we have SPD3 counts for 2011, we do not include these directly in the model as they are indirectly used through their role in estimating coverage ratios. The SPD3 counts for 2016-2020 are assumed to follow the normal-distribution data model. As with MYE, we use national-level counts by single-year ages 90–105 to split out LA-level counts in age group 90+. We do not assume that SPD3 data are unbiased and set coverage ratios using

$$\rho^{\text{SPD3}}_{asct} = y^{\text{SPD3}}_{asct} / q^{\text{pop}}_{asct} \tag{4.2}$$

The coverage ratios are estimated by smoothing the ratio of SPD3 to MYE estimates for 2011. The smoothing is performed by fitting the generalized additive model

$$\log \mathbb{E}(y^{\text{SPD3}}_{a,2011}) = \log y^{\text{MYE}}_{a,2011} + f(a_{2011}), \tag{4.3}$$

where $y^{\text{SPD3}}_{a,2011} \sim \text{Poisson}(\mathbb{E}(y^{\text{SPD3}}_{a,2011}))$, and f is a smooth function of age. Separate models are fitted for each Local Authority by sex combination. We use an adaptive smoothers from the **mgcv** R package [Wood, 2011], with the default setting of P-splines for the smoothing and penalty bases. We assume that coverage ratios remain constant at their 2011 levels over the period 2016–2020.

Figure 4.1 compares the raw 2011 SPD coverage ratios $y^{\text{SPD3}}_{a,2011}/y^{\text{MYE}}_{a,2011}$ with the smoothed ratios $\hat{\rho}^{\text{SPD3}}_{a,2011}$ in the Cambridge Local Authority. We calculate standard deviations for these coverage ratios from the confidence intervals for SPD3-based population estimates published in ONS [2020].

Patient Register (PR) We use the normal-distribution data model for the PR-based population estimates for 2012–2015, with coverage ratios calculated using the same method as for the SPD3.

Census 2021 One notable data source that has not been included is Census 2021 data. We withhold Census 2021 data from the model so that we can use it as a yardstick to measure the accuracy of the DPM 2021 estimates. Future versions of the DPM will include Census 2021 data.

4.2.2 System models

We describe how we estimate the hyper-parameters for the system models for births, deaths, in-migration, and out-migration. Since the methods are almost

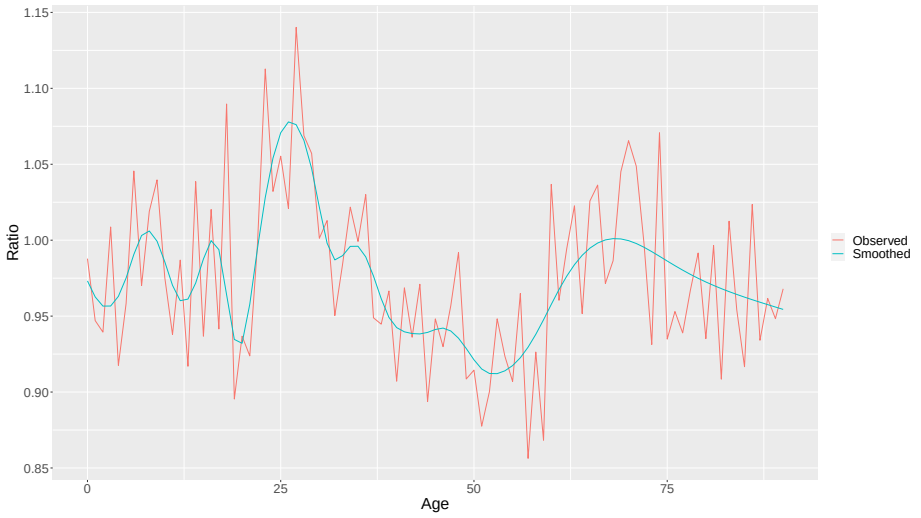


Figure 4.1: Observed ($y^{\text{SPD3}}_{a,2011}/y^{\text{MYE}}_{a,2011}$) and smoothed ($\hat{\rho}^{\text{SPD3}}_{a,2011}$) coverage ratios for females in the Cambridge Local Authority.

identical for all demographic series, most notation in this section omits superscripts denoting series.

We assume that, independently for each Local Authority, rates are drawn from gamma distribution

$$\gamma_{asct} \sim \text{gamma}\left(\frac{\mu_{asct}}{\delta}, \frac{1}{\delta}\right), \tag{4.4}$$

which has mean μ_{asct} and variance $\mu_{asct}\delta$.

Values for μ_{asct} are estimated by smoothing observed rates independently within each sex s, using generalized additive models of the form

$$\log \mathbb{E}(y_{at}) = \log \hat{e}_{at} + x_{at}\theta + f(a,t) \tag{4.5}$$

where y_{at} is observed counts; $\hat{e}_{astr}$ is exposure calculated from coverage-adjusted SPD3 data, or, in years where there are no SPD3 data, an imputed SPD3; x_{astr} is a row vector from a model matrix containing indicator variables for individual ages 0–30; θ is a vector of coefficients to be estimated; $f(a,t)$ is a random factor smooth of an interaction between age and time; and $y_{at} \sim \text{Poisson}(\mathbb{E}(y_{at}))$. There is no manual specification of knots.

For the migration models there is an L1-penalty on the linear terms x_{at}. The linear terms capture the peaks observed in the age profile of migration that are associated, for example, with moves to university or boarding school. The birth models omit the linear term, as the observed age profiles do not exhibit sharp peaks. The death model includes one indicator variable for age 0 without an L1-penalty. The offset term is used in all models except those for in-migration, where exposure is undefined.

The exposure term for rates is calculated as

$$\hat{e}_{at} = \frac{\hat{y}_{at}^{\text{SPD3}} + \hat{y}_{a,t-1}^{\text{SPD3}}}{2} \tag{4.6}$$

where

$$\hat{y}_{at}^{\text{SPD3}} = \begin{cases} \hat{\rho}_{at}^{\text{SPD3}}(\hat{\beta}_0 + \hat{\beta}_1 t_{at}) y_{at}^{\text{PR}}, & t < 2016 \\ \hat{\rho}_{at}^{\text{SPD3}} y_{at}^{\text{SPD3}}, & 2016 \geq t < 2021 \\ \hat{\rho}_{at}^{\text{SPD3}} \tilde{y}_{at}^{\text{SPD3}_{2020}}, & t \geq 2021. \end{cases} \tag{4.7}$$

Values $\hat{\beta}_0$ and $\hat{\beta}_1$ in (4.7) are estimated using

$$y_{at}^{\text{SPD3}} / y_{at}^{\text{PR}} = \beta_0 + \beta_1 t_{at} + \epsilon_{at}, \tag{4.8}$$

where $\epsilon_{at} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. The $\tilde{y}_{at}^{\text{SPD3}_{2020}}$ are SPD3 estimates calculated from the demographic accounting identity starting from the SPD3 stock estimate in 2020, $y_{a,2020}^{\text{SPD3}}$ using observed births, deaths and combined migration flows to estimate for the years 2021 and 2022. The $\tilde{y}_{at}^{\text{SPD3}_{2020}}$ in (4.7) are SPD3 estimates calculated from the demographic accounting identity starting from the SPD3 stock estimate in 2020, $y_{a,2020}^{\text{SPD3}}$, using observed births, deaths and combined migration flows to estimate for the years 2021 and 2022.

The generalized additive models used for estimation in our initial publications ONS [2022c] and ONS [2022d] only uses a smooth function of age as a predictor, with no linear term, and is fitted independently for each year.

Figure 4.2 compares the observed combined in-migration rates for females in Cambridge aged 0 to 50 for the year ending 30 June 2021 ("unsmoothed"), with our initial estimates ("GAM") and the improved estimates ("GAM-LASSO"). The new approach has been effective in dealing with the differing and sometimes extreme spikes in migration observed in different Local Authorities. The initial GAM model oversmooths in-migration, especially around ages 17-23.

Figure 4.3 shows the expected rates used as inputs for the base model for females in Cambridge.

We calculate separate values for the dispersion parameter δ for each combination of Local Authority and sex. The calculations are based on published ONS estimates of standard errors. For each combination of LA, sex, and demographic series, we calculate dispersions by age and time, δ_{at}, and then take the maximum of these values.

In the case of birth and death rates, we assume that all uncertainty arises from the population denominator. Using Taylor linearisation, and assuming $\text{Var}(\hat{e}_{a,s,t,r}) = \text{Var}(y_{a,s,t,r}^{\text{SPD3}})$, the approximate variances of birth and death rates are

$$\text{Var}(\gamma_{at}^{\text{bth}}) = (\frac{x_{at}^{\text{bth}}}{\hat{e}_{at}})^2 \frac{\text{Var}(\hat{e}_{at})}{\hat{e}_{at}^2} \tag{4.9}$$

$$\text{Var}(\gamma_{at}^{\text{dth}}) = (\frac{x_{at}^{\text{dth}}}{\hat{e}_{at}})^2 \frac{\text{Var}(\hat{e}_{at})}{\hat{e}_{at}^2}, \tag{4.10}$$

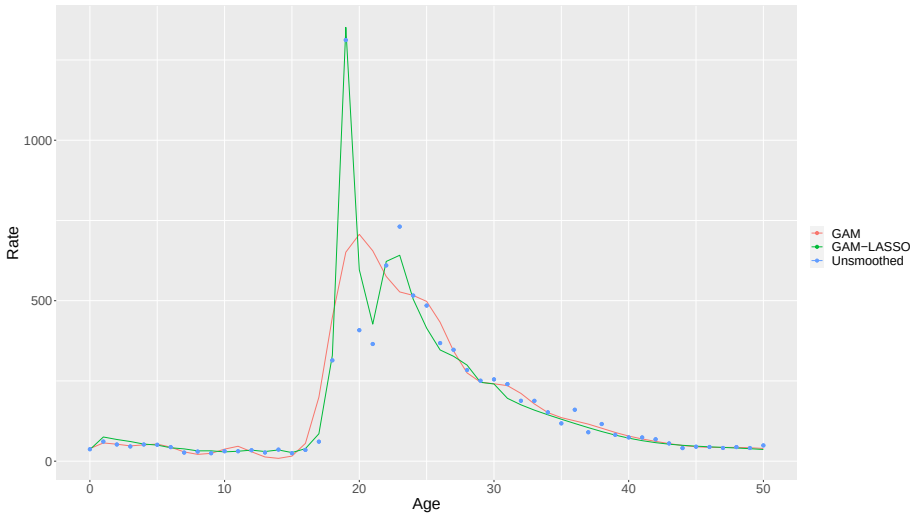


Figure 4.2: Combined in-migration for females in Cambridge ages 0-50 for the year ending 30 June 2021, observed counts ("unsmoothed"), estimates used in [ONS, 2022d] ("GAM") estimates from equation 4.5 ("GAM-LASSO")

which yields dispersion terms

$$\delta_{at}^{\text{bth}} = \frac{\hat{e}_{at}}{x_{at}^{\text{bth}}} \text{Var}(\gamma_{at}^{\text{bth}}) \tag{4.11}$$

$$\delta_{at}^{\text{dth}} = \frac{\hat{e}_{at}}{x_{at}^{\text{dth}}} \text{Var}(\gamma_{at}^{\text{dth}}). \tag{4.12}$$

Uncertainty for in-migration y_{at}^{in} is calculated assuming that the counts of immigration, $y_{at}^{\text{in,ext}}$ and internal (including cross-border) migration, $y_{at}^{\text{in,int}}$, are independent. Values for $\text{Var}(y_{at}^{\text{in,ext}})$ and $\text{Var}(y_{at}^{\text{in,int}})$ are calculated as part of the process for estimating uncertainty in mid-year estimates. We calculate terms for combined in-migration using

$$\delta_{at}^{\text{in}} = \frac{\text{Var}(y_{at}^{\text{in}})}{y_{at}^{\text{in}}}. \tag{4.13}$$

We calculate dispersion for out-migration in the same way, but with the additional assumption that all uncertainty comes from the numerator, and none from the population denominator.

4.3 Results for Cambridge Local Authority

To illustrate our methods, we present results for the Local Authority of Cambridge. We estimated counts and rates for Cambridge using the particle filtering

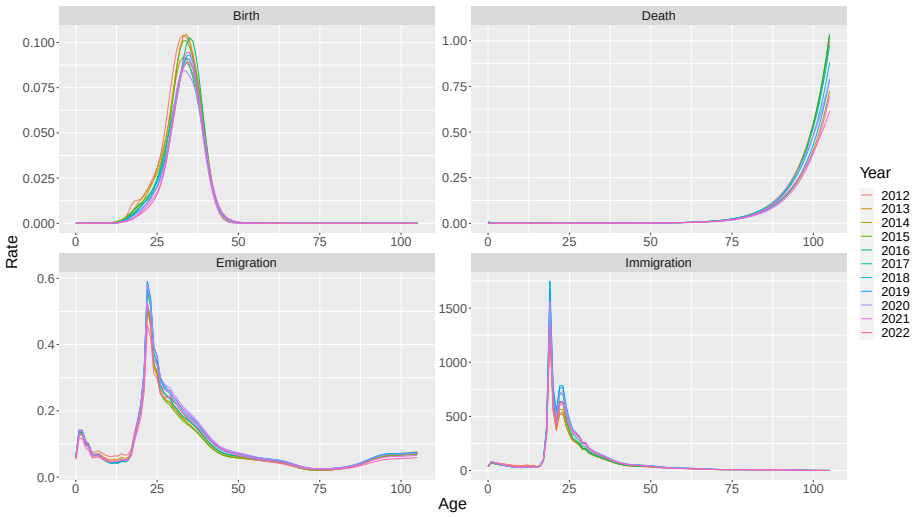


Figure 4.3: Estimated expected flow rates for females in Cambridge based on variations of the model in equation 4.5

method described in Section 3.3.1. Model fitting with 10,000 particles took 15:27 minutes on a laptop with 12th Gen Intel(R) Core(TM) i5-1250P, 1700 Mhz, 12 Cores, 16 Logical Processors, with no parallel processing. This is a dramatic speed-up compared with the many hours it would take to fit an equivalent model in **demest**.

Figures 4.4 and 4.5 show population estimates for 2020 and 2021 for females in Cambridge. Figure 4.4 compares DPM estimates with population counts from SPD3 and MYE. Posterior means from the DPM are shown as triangles while the thin vertical line shows the 95% credible interval and the thicker vertical lines show 65% credible intervals. The DPM estimates are close to SPD3 and are noticeably different from MYE, especially for the ages around 30.

As is apparent in Figure 4.5, while the DPM estimates differ slightly from the Census 2021 data, they are likely to be closer to the census results than the values that would be obtained from rolling forward MYE 20202 estimates. The most noticeable difference between the Census 2021 and DPM estimates occurs at the ages 18–21 and 26–27. The differences at ages 18 and 19 may be caused by the age definitions used for migration data in the DPM, where age is defined as age at the end of 30 June, which appears to shift the age profile to the right.

Figures 4.6 and 4.7 compare observed counts to DPM estimates of combined in-migration and out-migration for females in Cambridge in 2021. DPM estimates of out-migration are noticeably higher than the observed counts for ages 20 to 22, and also noticeably higher for in-migration for age 20. The apparent upward bias in the DPM estimates is probably a result of the smoothing methods used for the rates, which do not currently have a specific intervention for the effect of the COVID-19 pandemic.

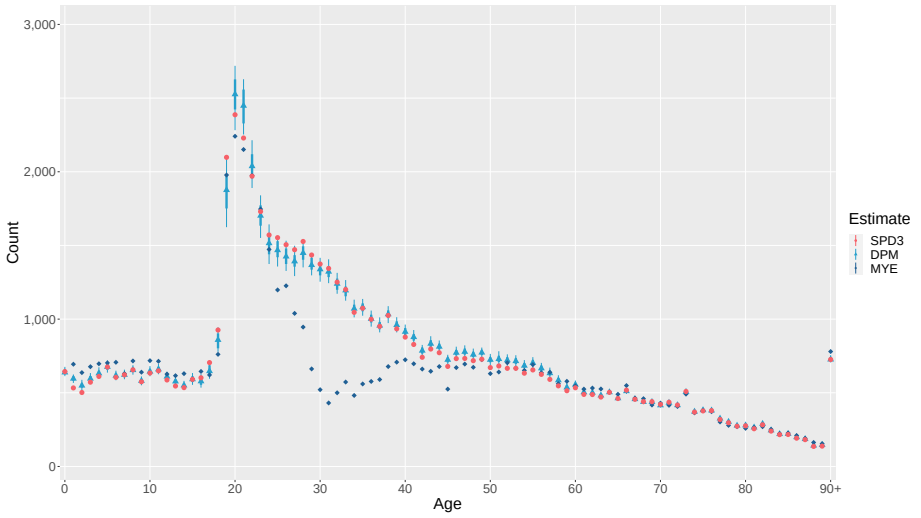


Figure 4.4: Population estimates for females in Cambridge, 2020. MYE and SPD3 estimates are not coverage adjusted. DPM estimates include a thin vertical line representing a 95% credible interval and a thick vertical line for a 65% credible interval.

Figure 4.8 shows the split of combined in-migration to Cambridge in 2020. Combined in-migration is split into international immigration, cross-border flows from Scotland and Northern Ireland, and internal migration from other Local Authorities. In this example, we do not apply the full iterative proportional fitting described in Section 3.4, as we have not run all Local Authorities. Instead we split draws from the posterior for combined in-migration based on the proportion of total observed flows for international immigration, cross-border flows and internal inward migration.

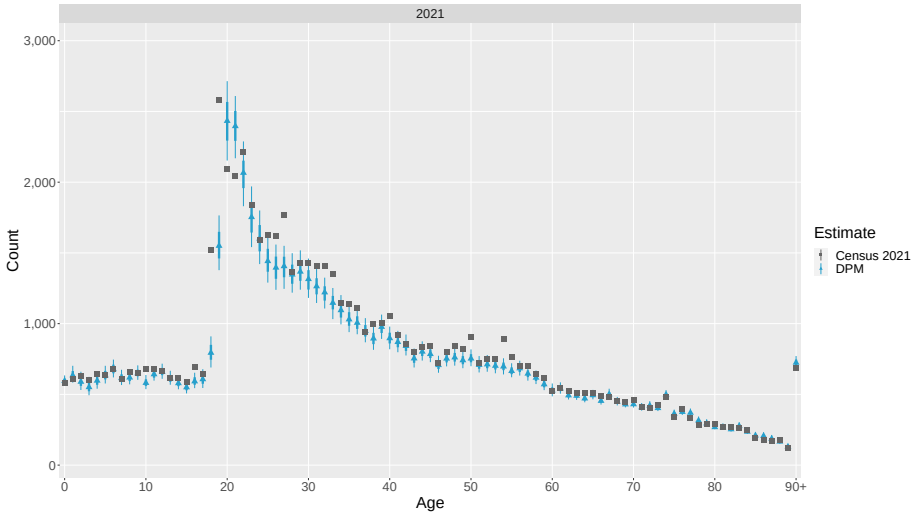


Figure 4.5: Population estimates for females in Cambridge, 2021. DPM estimates include a thin vertical line representing a 95% credible interval and a thick vertical line for a 65% credible interval.

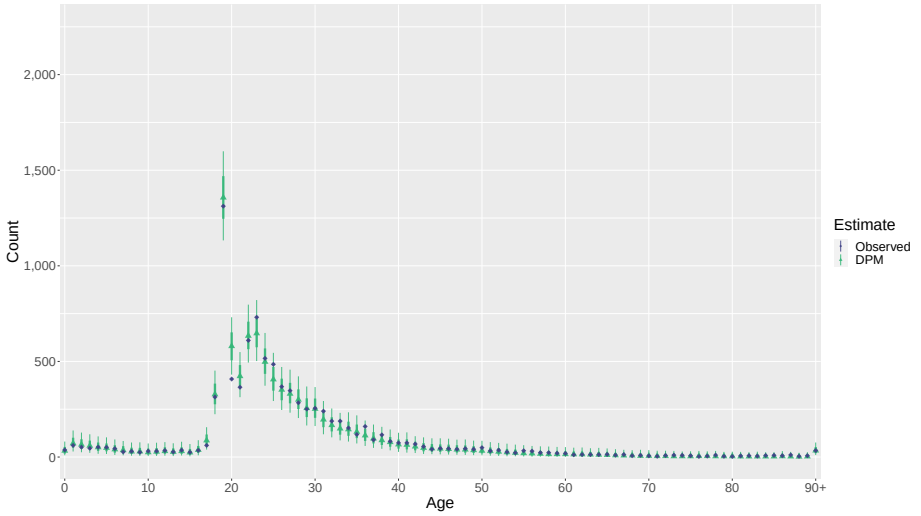


Figure 4.6: Combined in-migration for females in Cambridge, 2021. DPM estimates include a thin vertical line representing a 95% credible interval and a thick vertical line for a 65% credible interval.

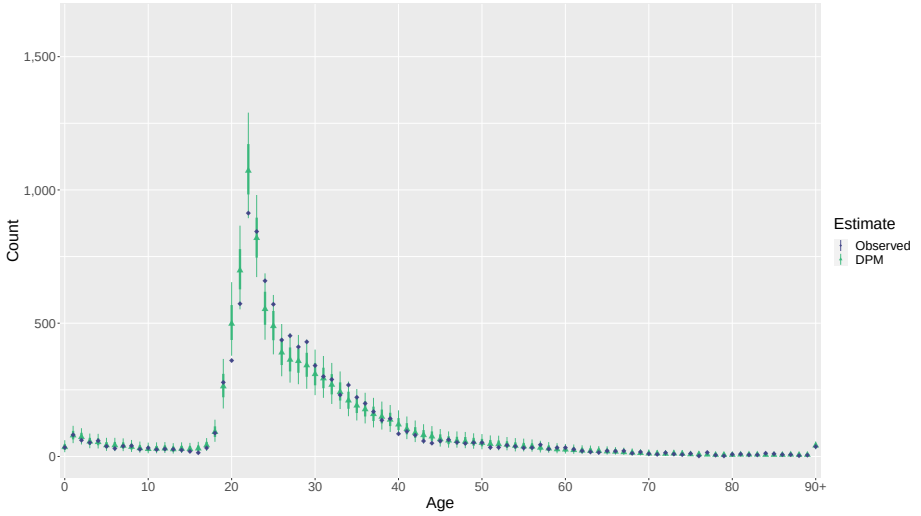


Figure 4.7: Combined out-migration for females in Cambridge, 2021. DPM estimates include a thin vertical line representing a 95% credible interval and a thick vertical line for a 65% credible interval.

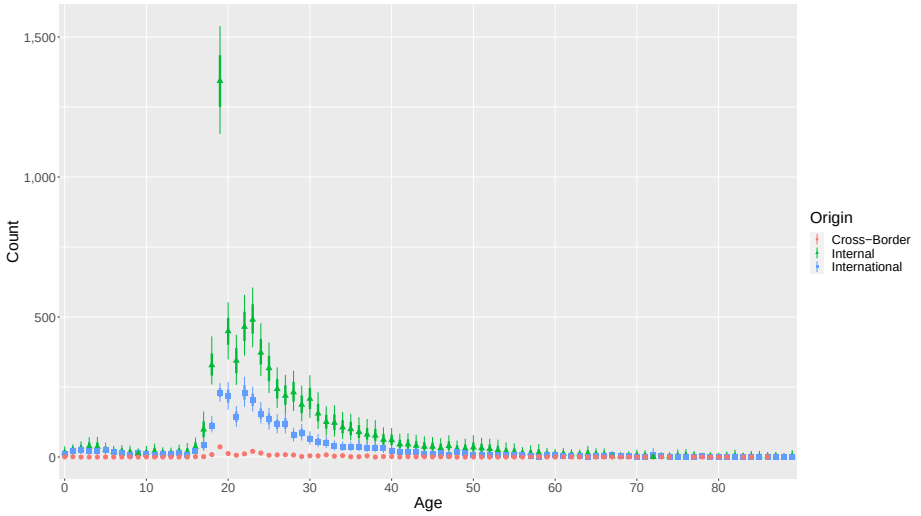


Figure 4.8: Components of combined in-migration for females in Cambridge, 2020. Estimates include a thin vertical line representing a 95% credible interval and a thick vertical line for a 65% credible interval.

Chapter 5

Extensions

The development strategy of the DPM project has been to start with the simplest possible system that can produce Local Authority estimates, and then progressively extend the system, based on performance and user needs. In this chapter, we outline a list of possible extensions. The extensions fall into four groups:

Accuracy, robustness Improve the accuracy of the estimates, and make the production system more maintainable and reliable.

Inputs Expand the range of data sources that the DPM can ingest.

Outputs Expand the level of detail the DPM provides, and the topics that it addresses.

Checking Build mechanisms to monitor the performance of the DPM.

5.1 Increasing accuracy, robustness

5.1.1 Harmonisation of national and subnational estimates

Running the DPM on Local Authority-level data and then summing across LAs will not in general produce the same values as running the DPM on England and Wales-level data. The amount of divergence depends on factors such as the geographical distribution of errors in the data, and the amount of flexibility in system and data models. It is difficult to predict, from first principles, whether adding up disaggregated results or working entirely with aggregate data is likely to give more accurate estimates.

At the time of writing (January 2023), we have only just begun the process of constructing full scale demographic accounts for all 331 Local Authorities in England and Wales. We therefore do not currently have a good sense of the consistency between local-level and national estimates. If it turns out that there are substantial differences, then we will need to understand the source of these differences, and revise the DPM accordingly.

5.1.2 Full-scale estimation using TMB

Experiments with TMB suggest that it may offer substantial advantages over particle filters for estimating counts and rates within cohorts (Section 3.3.) The extra speed offered by TMB is more than a convenience. Faster computation times make it easier to cope with unexpected problems, which is important for meeting production deadlines. Some proposed extensions, such as the use of multiple draws (Section 5.1.4) are only feasible if computation times can be improved. The ability to fit models more quickly makes it easier to explore alternative specifications, or to carry out simulation studies assessing model performance. We will revise the R packages, replacing particle filters with TMB, and then test whether TMB performs well on full scale accounts.

5.1.3 Hierarchical models for birth, death, and migration rates

Our long-term goal is to build hierarchical models for birth, death, and migration rates, with prior structures that can properly account for regularities such as common age-sex profiles or time trends. Our current models for rates have a much simpler structure, and are estimated independently for each Local Authority. Any future models will build on our current methods for dealing with migration peaks, which is one of the most challenging aspects of small-area demographic estimation. We will also build on other ONS work programs on estimating and forecasting demographic rates. We will investigate the use of TMB, which is specifically designed for large-scale hierachical models.

5.1.4 Use multiple draws of demographic rates

Experiments with the current version of the DPM suggest that, although it produces realistic credible intervals at the lowest level of detail – cells defined by age, cohort, sex, Local Authority, and time – it produces spuriously narrow credible intervals for aggregates of these cells, such as the total population of a Local Authority. These narrow credible intervals reflect the fact that the DPM currently has no way of accounting for correlations across age, cohort, sex, Local Authority, and time, with each cell being treated as independent.

Any change to the model to account for these correlations needs to preserve the conditions, described in Section 2.2.2, that allow for the use of cohort state space methods. One way of doing so is to enhance the way that we represent hyper-parameters, moving from point estimates to multiple draws. Taking a multiple-draws approach would require changes to steps two and three of our four-step estimation process, as summarised in Table 5.1.

Preliminary experiments of the effect of using multiple draws for parameters in the data models suggest that a multiple-draws approach does lead to more realistic credible intervals for aggregates. The larger the number of draws, M, the better we would be able to approximate the full multivariate distribution of

Table 5.1: Current estimation approach versus proposed multiple-draws approach to steps 2 and 3 of the estimation process

	Current approach	Proposed approach
Step 2	Construct point estimate $\hat{\phi}$ of hyper-parameters ϕ	Construct draws $\{\phi^{(m)}\}_{m=1}^{M}$, from the posterior distribution for hyper-parameters ϕ
Step 3	Obtain draws of counts and rates $\{x^{(s)}, \gamma^{(s)}\}_{s=1}^{S}$ from $p(x, \gamma \mid y, \hat{\phi})$	Obtain draws of counts and rates $\{x^{(s_m)}, \gamma^{(s_m)}\}_{s=1}^{S/M}$ from $p(x, \gamma \mid y, \phi^{(m)})$, for $m = 1, \ldots, M$

x and γ, but the longer the computations would take. We will experiment with trade-offs.

5.1.5 Importance sampling to reduce approximation errors

The current DPM strategy of using an initial approximation of the demographic account to estimate hyper-parameters is fast and simple. It does, however, introduce approximation errors. Adding a final importance sampling-resampling [Smith and Gelfand, 1992] step to the estimation process is a potential method for reducing these errors.

For each draw $(x^{(s)}, \gamma^{(s)}, \phi^{(s)})$ from the joint posterior distribution of all the unknowns, we would calculate importance weight

$$w^{(s)} = \frac{f(x^{(s)}, \gamma^{(s)}, \phi^{(s)})}{g(x^{(s)}, \gamma^{(s)}, \phi^{(s)})}, \tag{5.1}$$

where ϕ collects holds the hyper-parameters for the system and data models, $f(x, \gamma, \phi)$ is the (unnormalised) density for the target posterior distribution, and $g(x, \gamma, \phi)$ is the density under the approximate model used in the calculations. We would rescale the weights to form selection probabilities, and use the probabilities to resample draws from the posterior distribution. The resampled posterior distribution would hopefully be a better approximation of the true distribution than the original draws.

If we are able to draw accurately from $p(x, \gamma \mid y, \phi)$, then the weights reduce

to[1]

$$w^{(s)} = \frac{1}{p(\hat{x}|\phi^{(s)})}, \tag{5.2}$$

where $\hat{x}$ denotes our initial approximation of the account. Applying these weights during resampling would reduce the influence of $\hat{x}$ on the posterior distribution.

If we were to pursue the option of importance sampling, we would experiment with smoothed versions of the importance weights, based on Vehtari et al. [2015].

5.1.6 Special populations

The current ONS system for generating mid-year population estimates contains dedicated procedures for 'special populations' that are difficult to deal with through standard procedures. Examples include the armed forces and prisoners, both of which have 'static' age structures that remain constant over time.

The DPM may have less trouble than the current ONS population estimation system in dealing with static age structures, because, unlike the current system, it makes use of ongoing information on population stocks. Where data on the size and age-sex structure of special populations is available, we can use it alongside data on births, deaths, and migration. Moreover, if we have evidence that data on special populations contain over-coverage or under-coverage, we can capture this in our data models.

If these measures are not sufficient, it may be necessary to split special populations off from the general population. We could, for instance, divide a local authority with a large special population into two, and use one set of parameters and data for the special population and another set for the remainder.

ONS demographers have extensive experience dealing with special populations. They will be involved in identifying special populations, designing models, and evaluating results.

5.2 Expanding inputs

5.2.1 Expanding the suite of data models

Results from the DPM are sensitive to the specification of the data models, with different models leading to different point estimates and credible intervals. The performance of any one data model depends on which other data models are in use. A complicated and flexible data model can perform well, for instance,

[1]

$$\begin{aligned} g(x,\gamma,\phi) &= p(x,\gamma|y,\phi)p(\phi|\hat{x}) \\ &\propto p(y|x,\gamma,\phi)p(x,\gamma|\phi)p(\hat{x}|\phi)p(\phi) \\ &= p(y|x,\phi)p(x|\gamma)p(\gamma|\phi)p(\phi)p(\hat{x}|\phi) \\ &= f(x,\gamma,\phi)p(\hat{x}|\phi) \end{aligned}$$

when other data models are simple and constrained, but create computational problems when they are not. It is therefore helpful to have a suite of data models to choose from, particularly when tackling new applications, or when adding or omitting data sources.

Our current data models are relatively simple and generic. Our main focus will be on developing models that are optimised for specific tasks or datasets. Examples include the following:

Combined migration counts The in-migration and out-migration data supplied to the base model aggregate over multiple datasets, such as datasets dealing with internal migration and datasets dealing with international migration, with the relative share of each dataset varying by age, sex, and area. Expanding the data models so that they incorporate information about the composition of the data might improve performance.

Components of SPDs Statistical Population Datasets (SPDs) are constructed by linking individual-level data from multiple administrative datasets and then applying a set of business rules to minimise under-coverage and over-coverage. Some of the datasets are available much more quickly than others, but the final SPD cannot be created until all datasets are available. There may be value in constructing an approximation of the SPD using only the most timely components, to use for recent periods where the final SPD is not available. If such a dataset was available, we would ideally use it with a data model that took its special features into account.

Reporting lags There is often a delay between the time when people change residence and the time when they update their administrative data: for instance, people may not update their residential address on the patient register until they visit a doctor. In cases where there is good-quality information on characteristics of reporting lags, exploiting this information may lead to more accurate estimates of the timing of events across the year.

Splitting migration Our procedures for splitting migration streams, described in Section 3.4, assume constant levels of accuracy across subnational areas and datasets. This is equivalent to assuming a very simple data model. If experience suggests that this simple data model is inadequate, we could extend it by, for instance, incorporating weights into the iterative proportional fitting process that reflected relative reliability [Stone et al., 1942, Deville and Särndal, 1992, Stone, 1961, Lahr and De Mesnard, 2004].

5.2.2 Incorporating aggregated data sources

ONS is investigating a variety of novel data sources that are more timely than existing data sources, or that provide information on physical location rather than usual residence (see Section 5.3.4.) Examples include anonymised data on mobile phone use, or data on services such as gas, electricity, or waste water treatment.

Many of these data sources contain little or no information on age, cohort, or sex. Data that does not distinguish cohorts and sexes is difficult to use within cohort state space methods because, as discussed in Section 2.2.2, cohort particle filters require cohorts to be conditionally independent, and conditional independence is lost if cohorts or sexes share the same data values.

One possible solution is to pre-process the data, splitting out any shared values before the data is entered into the base model. The disadvantage of this approach is that it disguises uncertainties about the age-sex distribution. If these uncertainties are substantial, as is likely with data on recent trends or on the population physically present, then the results could be misleading.

An alternative approach that, if feasible, could deal more satisfactorily with uncertainty would be to use an importance sampling scheme, similar to the one that we are considering for reducing approximation errors (Section 5.1.5). Let y^{dis} denote disaggregated data and y^{ag} aggregated data. We would like to estimate

$$p(x|y^{\text{ag}}, y^{\text{dis}}) \propto p(y^{\text{ag}}, y^{\text{dis}}|x)p(x), \tag{5.3}$$

which, if we assume conditional independence of the y variables, we can write as

$$p(y^{\text{ag}}|x)p(y^{\text{dis}}|x)p(x) \tag{5.4}$$

Running the base model only on disaggregated data would yield draws from $p(y^{\text{dis}}|x)p(x)$. Resampling using weights proportional to $p(y^{\text{ag}}|x)$ would yield draws from (5.4). The resampling would increase the relative share of draws from the base model that aligned with the aggregated data.

5.2.3 Coverage surveys

ONS is investigating the potential role of coverage surveys in future systems for population and social statistics. Coverage surveys can play two potential roles within the DPM: as an input to performance monitoring, and as an input to data models.

A coverage survey for an administrative data source could play the same role in monitoring the performance of the DPM as the Census Coverage Survey has traditionally done in monitoring the performance of ONS's current system of population estimation. Estimates of under-coverage and over-coverage from the coverage survey could be used to adjust counts from the administrative data source, and produce population estimates that were, to at least some extent, independent of those of the DPM. Comparing these population estimates with those of the DPM would provide evidence on the accuracy of the DPM, just as comparisons with the census-year population estimates provide evidence on the accuracy of ONS's current system of population estimation.

Performance monitoring based on coverage surveys, although valuable, does have limits. The sample sizes needed to detect small errors in population estimates are extremely large, particularly for disaggregated estimates. Non-response to the coverage survey can compound the problem, by reducing sample sizes and complicating the adjustment process.

A second potential role of coverage surveys is to help with the estimation of parameters in data models. Coverage surveys can, for instance, be used to estimate parameters such as the net coverage ratio ρ_{asct} and variance σ_{asct} in the normal-distribution data model (Equation (2.9) on page 16). In doing so, they can help anchor estimates across all series.

The use of coverage surveys for performance monitoring and the use within data models are partly in conflict with each other. If a coverage survey is an input to a data model, for instance, then the survey cannot provide an independent check on the performance of the DPM. There might, nevertheless, be ways of at least partly achieving both goals. The DPM team might, for instance, compare estimates with and without a coverage survey as a way of gaining insights into the performance of the DPM, but use the full model, with all available coverage surveys, whenever it was producing official population estimates.

5.3 Expanding outputs

5.3.1 Monthly estimates

Monthly estimates of population stocks and flows could be derived either by temporally disaggregating the annual demographic account from the base model or by direct estimation using monthly data as inputs for the modelling.

Temporal disaggregation

Monthly estimation could be carried out once the annual demographic account have been constructed. We obtain monthly estimates for births and deaths straight from the monthly registration data, which we treat as error-free. We obtain monthly estimates for in-migration and out-migration using temporal disaggregation techniques called benchmarking, which are widely used in national accounts [Dagum and Cholette, 2006, Eurostat, 2018, IMF, 2018]. We derive monthly population counts by applying demographic accounting to annual population stocks and monthly births, deaths, in-migration, and out-migration.

We describe here the procedures for benchmarking in-migration; the procedures for out-migration are identical. Let $w_{tcs}^{\text{in}(i)}$ be the total number of in-migrations belonging to cohort c and sex s during year t, according to the ith draw from the posterior distribution for the base model. A value for $w_{tcs}^{\text{in}(i)}$ is obtained by adding together the $x_{kcs}^{\text{in}(i)}$ that refer to year t. (Cohorts that are born or extinguished during year t have one $x_{kcs}^{\text{in}(i)}$, and other cohorts have two: one upper Lexis triangle and one for the lower Lexis triangle.) We do not distinguish between the case when the model has been applied to the whole country and the case when it has been applied to a local area, since the calculations are the same in both. We need to estimate monthly in-migration $u_{mcs}^{\text{in}(i)}$, where m indexes month. Let δ_t denote months that fall within year t. We require that monthly values for the cohort sum to the annual value, $\sum_{m\in\delta_t} u_{mcs}^{\text{in}(i)} = w_{tcs}^{\text{in}(i)}$.

We have data z^{in}_{mcs} that is an imperfect measure of $u^{\text{in}(i)}_{mcs}$ and that does not, in general, add up to $w^{\text{in}(i)}_{tcs}$.

Many different benchmarking procedures have been developed. We are experimenting with the following procedures.

Pro-rata Estimated values are proportional to monthly data,

$$u^{\text{in}(i)}_{mcs} = \frac{w^{\text{in}(i)}_{mcs}}{\sum_{m\in\delta_t} z^{\text{in}}_{mcs}} z^{\text{in}}_{mcs}. \tag{5.5}$$

The pro-rata method is simple, can cope with zeros (unless $\sum_{m\in\delta_t} z^{\text{in}}_{mcs} = 0$), and always leads to non-negative values. However, while it preserves monthly growth rates of z^{in}_{mcs} it can produce sharp jumps between adjacent years.

Proportional Denton Proportional Denton is applied over multiple years, and promotes smoothness, including between adjacent years, by minimising changes in the ratio between estimates and data. The modified proportional first difference Denton estimates that we are exploring are the set of $u^{\text{in}(i)}_{mcs}$ that minimise

$$\sum_m \left(\frac{u^{\text{in}(i)}_{mcs}}{z^{\text{in}}_{mcs}} - \frac{u^{\text{in}(i)}_{m-1,c,s}}{z^{\text{in}}_{m-1,c,s}} \right), \tag{5.6}$$

subject to the constraint that monthly estimates add up to annual estimates.

Additive Denton The modified additive first difference Denton is equivalent to proportional Denton, except that it minimises the quantity

$$\sum_m \left((u^{\text{in}(i)}_{mcs} - z^{\text{in}}_{mcs}) - (u^{\text{in}(i)}_{m-1,c,s} - z^{\text{in}}_{m-1,c,s}) \right)^2. \tag{5.7}$$

Wavelet-based approach [Sayal et al., 2017] describe a novel wavelet-based approach to benchmarking. The approach may be useful for cohorts with many small counts and a few large counts, as it can isolate extreme movements, rather than distributing them across many months, which, among other things, can cause negative values in Denton methods. The basic benchmarking part of the method works by replacing the low frequency wavelets in a wavelet decomposition of z^{in}_{mcs} with those from $w^{\text{in}(i)}_{tcs}$ which, when back transformed to the time domain, meets the benchmarking constraint.

Sometimes monthly estimates are required before annual estimates for the year in question have been constructed. In these situations, benchmarking involves an element of extrapolation. There is an extensive discussion of these issues within the benchmarking literature [Eurostat, 2018].

Having calculated monthly estimates for cohorts, we need to convert them into monthly estimates for single year of age groups. If a cohort spans the interval between exact age a and exact age $a + 1$ on 1 July, then it will span the interval between exact age $a + \frac{1}{12}$ and exact age $a + 1 + \frac{1}{12}$ on 1 August, the interval between exact age $a + \frac{2}{12}$ and exact age $a + 1 + \frac{2}{12}$ on 1 September, and so on. We convert these values into conventional single-year age groups by taking weighted averages, with the weights proportional to the amount of overlap between cohorts and age groups.

Direct monthly estimates

There is no theoretical obstacle to deriving monthly estimates directly, using one-month rather than one-year widths for periods, age groups, and cohorts. The obstacles are instead practical. The computer systems that we have been using for the 2022 prototype cannot cope with the volume of data or calculations, and the confidentiality protocols do not permit the required level of disaggregation.

The practical obstacles to direct calculation of monthly estimates are likely to be reduced as we move into a full production system. Direct calculation would almost certainly be possible when adding to an existing series, as opposed to estimating an entire back series (Section 2.6). As the production system matures, we will compare the two approaches to adding to existing series.

5.3.2 Satellite accounts

> Satellite accounts are one way in which the System of National Accounts may be adapted to meet differing circumstances and needs. They are closely linked to the main system but are not bound to employ exactly the same concepts ... They may also be used to explore new methodologies and to work out new accounting procedures that, when fully developed and accepted, might become absorbed into the main system over time [Eurostat, 2022].

Economic statisticians have used satellite accounts as a way of expanding the scope of national accounts without overloading the main account. Social statisticians could use the same strategy with demographic accounts. Stone [1984, p. 26] argues for this approach, suggesting that statisticians assemble a basic set of relatively simple data, supplemented by "subsidiary sets" of other data, all employing the same conceptual framework.

Some possible candidates for satellite accounts include:

Family and household Counts of families and households, disaggregated by type. Estimates of the distribution of individuals across families and households would need to be consistent with estimates of total numbers of individuals in the main account. (See Section 5.3.5.)

Population present Estimates of numbers of people physically present at a given point in time, possibly in combination with estimates of the usually-resident population. (See Section 5.3.4.)

Labour force Population disaggregated by labour force status (employed, unemployed, not in the labour force), and flows between these statuses.

Education Population disaggregated by current enrolment status or by educational attainment, and flows between these statuses.

Ethnicity Population disaggregated by ethnicity, possibly with flows between statuses reflecting changes in ethnic identification.

Long-term population projections Nowcasts and very short-term forecasts can be handled easily with existing system models. However, longer-term forecasts would typically need longer historical series and specialised models for hyper-parameters.

5.3.3 Additional dimensions

Demographic accounts can contain other dimensions besides age, sex, and subnational area. Possibilities include, for instance, country of birth, and enrolment in school or university. Adding extra dimensions can make the account more useful. There is, for instance, substantial policy interest in the location and characteristics of students. Adding extra dimensions can help stabilise estimates, if groups identified by the new dimension behave differently from each other. Explicitly distinguishing between students and non-students, for instance, might help predict differences among local authorities.

To add a new dimension to the DPM, we need to disaggregate all counts and rates along this new dimension, and apply cohort state space estimation for each of the resulting combinations of cohort, sex, and the new dimension. Under the current framework, all input data needs to include the new dimension, though Section 5.2.2 discusses one way this requirement might be relaxed.

If the characteristic that is measured by the new dimension can change, then we need to include these changes in status in the modelling. We can do so by expanding our definition of migration. When estimating counts and rates for 20-year-old female university students, for instance, we would add university enrolments to the generic "in-migration" component and add graduations to the generic "out-migration" component. The splitting of migration flows described in Section 3.4 would then include the separation of enrolments and graduations from other types of inflows and outflows. The splitting of flows across multiple dimensions could be accommodated through a straightforward extension of our current methods, in which the $\boldsymbol{Z}$ and $\boldsymbol{X}$ of Equations (3.16)–(3.22) changed from matrices to multiway arrays.

A candidate for an additional dimension could be trialled in a satellite account. If the satellite account proved to be feasible and useful, then the dimension could be considered for inclusion in the main account.

5.3.4 Usual residence versus population present

Official population estimates almost always refer to the "usually resident" population of a country or area, that is, to people who have their home, or spend most of their time, in the country or area. The usually-resident population of a country or area is distinct from the number of people physically present in that country or area. The two types of population are related as follows:

$$\begin{array}{ccccccc} \text{Population present} & = & \text{usually-resident population} & + & \text{non-residents temporarily present} & - & \text{residents temporarily absent.} \end{array} \tag{5.8}$$

Estimates of the usually-resident population and population present serve different purposes. The usually-resident population is more relevant, for instance, when deciding how big a school should be, since children typically attend a school near where they live. In contrast, population present is more relevant when deciding where to place carparks, since areas with few residents, such as central business districts, can still have many cars.

Most administrative data on populations are based on registration for a service, often requiring a residential address, and hence measure the usually-resident population. Many of the novel data sources that are currently under investigation at ONS, such as wastewater data, serve as proxies for the population present. Mobile phone data are unusual in that, in addition to measuring population present, they can also measure the usually-resident population, using the night-time population as a proxy for the usually-resident population (ideally with some minimum number of nights.)

When designing a data model for a new data source, it is important to be clear whether the data source is to be treated as a measure of the usually-resident population or as a measure of the population present. Ignoring the distinction can lead to large errors where the two populations diverge, such as in commercial districts or holiday destinations.

Integrated estimates of the usually-resident population and population present would be a good candidate for a satellite account. Data sources for population present often provide much less information on background characteristics, or omit a larger fraction of the total population, than data sources for the usually-resident population. Combining the two sorts of estimates, while allowing for the differences in population definition, could help fill some of the gaps in the information on population present. Publishing estimates of usually-resident population alongside estimates of population present would also help clarify the distinction between the two measures, and highlight areas or subpopulations where the distinction is particularly important.

In principle, an account could be constructed that included stocks and flows for the usually-resident population and population present. However, flows per unit of time for population present can be orders of magnitude higher than those

for usually-resident populations. The population of a commercial district, for instance, can increase and then decrease by a factor of 10 or 20 during a single day. Harmonising flows associated with population present with flows associated with usual residence would be very difficult, without necessarily yielding many new insights. It might be more productive to create a stocks-only account based on the accounting identity of Equation (5.8) rather than the usual accounting identity linking stocks to flows.

5.3.5 Families, households, and dwellings

Many data users need estimates, not just of individuals, but also of families, households, and dwellings. Published estimates of individuals, families, households, and dwellings must be consistent. The minimal type of consistency is that counts of individuals by family type or household type add up to totals in the demographic account. A stronger form of consistency is that changes make demographic sense, so that, for instance, a decline in fertility rates is accompanied by a decline in the proportion of families containing children.

Traditionally, production of household and family statistics has relied heavily on census data. Administrative data typically lack detailed information on relations between family members. Using administrative data to place people in households requires address data to be highly accurate and timely, which is not always the case. A big potential advantage of administrative data, however, is that it is updated far more regularly than the census. Two possible new products at the ONS have substantial potential for household and family statistics. The first is the Census Data Asset, which will roll forward the 2021 Census population. The second is the Statistical Co-resident datasets, which group co-residents based on administrative-based addresses. If these products do go ahead, the DPM project team will investigate how they can be integrated with demographic accounts.

5.4 Model checking

All statistical models are based on simplifications and approximations. Checking that these simplifications and approximations are not having a material effect on outcomes of interest is an essential part of a modelling workflow [Gelman et al., 2020]. If a model is to be deployed within a production process, then model checking needs to be carried out continuously, to detect changes in performance due to, for instance, changes in the input data or changes in the real-world system being modelled.

5.4.1 Priorities for checking

Some aspects of the DPM that are priorities for checking are as follows.

Robustness to violations of assumptions about rates and inputs

We need to assess the performance of the DPM when assumptions about demographic rates and about the input data are violated, including when rates or inputs change but the model does not. Robustness is likely to vary across data sources, data models, and components of the demographic account. We also need to assess the extent to which standard model diagnostics such as leave-one-out cross-validation [Vehtari et al., 2017] are able to detect these sorts of changes. Analysis of robustness are well suited to simulation studies where inputs and model specifications are systematically varied.

Reporting lags and sub-annual estimates

ONS has not previously produced sub-annual population estimates on an ongoing basis, and there is limited evidence about levels of accuracy and possible biases. A potentially important source of error is lags between the time when people change residence and the time when an address change is recorded in an administrative system. Administrative data with lags could potentially give a misleading impression of actual seasonal patterns. The strength of any such effects is, however, unclear. Simulation studies could provide insights into the plausible range. Empirical evidence on the size and prevalence of lags, such as evidence derived from linked census and administrative data, would be helpful in setting up simulation studies. Comparing administrative data against other data sources, such as mobile phone data, that do not have reporting lags could also be useful.

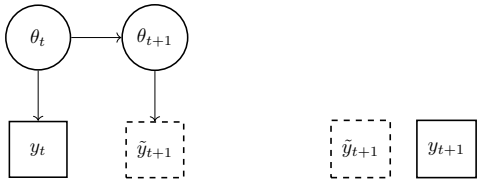
Sensitivity to specification of data models

Choosing appropriate specifications, including prior distributions, for data models can be challenging. Data models are therefore a high priority for sensitivity tests and prior predictive checks [Gelman et al., 2020, Section 2.4]. If important features of the DPM results vary under different, equally defensible, specifications, this should be reported in technical results.

5.4.2 Monitoring performance via forecasts

The DPM offers a novel possibility for monitoring performance. The basic idea is illustrated in Figure 5.1. Each period, the DPM would be used to forecast values for the input data in the next period. Once the actual values for the input data become available, these would be compared with the forecasted values. Analysts would then try to diagnose the reasons for any discrepancies that could not be explained by chance alone. Possible reasons would include errors in the input data, problems with the data model, and problems with the system model. Any such diagnoses would ideally be informed by discussions with data providers and with experts on demographic trends. Comparisons of the forecast and actual data could also provide evidence on whether uncertainty

measures from the DPM were appropriately calibrated—whether the DPM was understating or overstating uncertainty.



(a) Time t (b) Time $t+1$

Figure 5.1: Forecasting next period's data to monitor performance. At time t, forecast next period's data $\tilde{y}_{t+1}$. At time $t+1$, compare the forecasted data $\tilde{y}_{t+1}$ with the actual data y_{t+1}.

The use of forecasts to monitor performance depends on having explicit system and data models. It would not be possible in a conventional population estimation system, where the analysis of demographic trends and data is done in a less formal way.

Using forecasts to monitor performance has potential blind spots. It may, for instance, be poorly suited to identifying persistent flaws, such as consistently incorrect coverage rates, since these sorts of flaws do not necessarily reduce the model's ability to forecast the next period's data. Moreover, even when a discrepancy is discovered, diagnosing the reason for that discrepancy would often be difficult, given the complexity of the DPM. It would be important to have other methods for monitoring performance, such as coverage surveys (discussed in Section 5.2.3) that have complementary strengths and weaknesses.

However, the forecasting approach does have two important strengths. The first is that it is cheap to implement. Forecasting and the detection of discrepancies can be done entirely with existing data, and can easily be automated. The second is that it is hard to manipulate, assuming that the forecasts are carried out and recorded before the data becomes available, so that retrospective adjustment is ruled out. ONS might even consider making the forecasts, and subsequent evaluations, public.

Appendix A

Additional detail on particle filters

A.1 Importance function

Our importance function is described in Figure A.1. The function assumes that data sources can be ranked by overall reliability. We use $y_k^{(*)}$ to denote the value from the most reliable data source available for index k, and use $g_k^{\text{stk}}(\cdot \mid \cdot)$ to denote the likelihood for the data associated with $y_k^{(*)}$. In k where a value for $y_k^{(*)}$ is available, we use $g_k^{\text{stk}}(\cdot \mid \cdot)$ to perturb $y_k^{(*)}$ and use this perturbed value as our estimate of x_k^{stk}. In k where no value for $y_k^{(*)}$ is available, we base the proposal entirely on migration rates.

The algorithm makes use of the Skellam distribution. If U_1 and U_2 are Poisson variates with parameters μ_1 and μ_2, then $V = U_1 - U_2$ is a Skellam variate with parameters μ_1, μ_2. We in fact use a left-truncated Skellam distribution, which we denote $\text{skeltr}(\mu_1, \mu_2, n)$, where n is the lowest value that the draws from the distribution can take. Similarly, we use $\text{poistr}(\lambda, n)$ to denote a left-truncated Poisson distribution.

Towards the end of the algorithm, when we split net migration into in-migration and out-migration, we use a procedure that treats all components symmetrically. We do this to avoid concentrating all variability into one 'residual' component.

Input

$x_{k-1}^{\text{stk}(i)}$	Stock at end of triangle $k-1$
$y_k^{\text{stk}*}$	Reported stock at end of triangle k (optional)
x_k^{bth}	Count of births during triangle k
x_k^{dth}	Count of deaths during triangle k
$\mu_k^{\text{bth}}, \delta_k^{\text{bth}}$	Expected value and dispersion for birth rate γ_k^{bth}
$\mu_k^{\text{dth}}, \delta_k^{\text{dth}}$	Expected value and dispersion for death rate γ_k^{dth}
$\mu_k^{\text{in}}, \delta_k^{\text{in}}$	Expected value and dispersion for in-migration rate γ_k^{in}
$\mu_k^{\text{out}}, \delta_k^{\text{out}}$	Expected value and dispersion for out-migration rate γ_k^{out}
$g_k^{\text{stk}}(\cdot \mid \cdot)$	Data model for stock data $y_k^{\text{stk}*}$

Algorithm

1. Generate rates $\tilde{\gamma}_k^{\text{in}}$, $\tilde{\gamma}_k^{\text{in}}$, $\tilde{\gamma}_k^{\text{in}}$ and $\tilde{\gamma}_k^{\text{out}}$ by drawing from gamma distributions with specified expected values and dispersions.
2. Generate stock at end of triangle $\tilde{x}_k^{\text{stk}(i)}$ and net migration $n_k^{(i)}$
   - If $y_k^{\text{stk}*}$ is available:
     (a) Draw $\tilde{x}_k^{\text{stk}(i)} \sim g_k^{\text{stk}}(\cdot \mid y_k^{\text{stk}*})$
     (b) Set $n_k^{(i)} = \tilde{x}_k^{\text{stk}(i)} - x_{k-1}^{\text{stk}(i)} + x_k^{\text{dth}(i)}$
   - Else:
     (a) Set $e_k'^{(i)} = \max(x_{k-1}^{\text{stk}(i)}/2, \tilde{\gamma}_k^{\text{in}}/4)$
     (b) Set lower bound $n_k^{*(i)} = x_k^{\text{dth}} - x_{k-1}^{\text{stk}(i)}$
     (c) Draw $n_k^{(i)} \sim \text{skeltr}(\tilde{\gamma}_k^{\text{in}}, \tilde{\gamma}_k^{\text{out}} e_k'^{(i)}, n_k^{*(i)})$
     (d) Set $\tilde{x}_k^{\text{stk}(i)} = x_{k-1}^{\text{stk}(i)} - x_k^{\text{dth}(i)} + n_k^{(i)}$
3. Calculate exposure $e_k = (x_{k-1}^{\text{stk}} + x_k^{\text{stk}})/4$.

Figure A.1: Algorithm for drawing from importance function $q(\boldsymbol{x}_k^{(i)}, \boldsymbol{\gamma}_k^{(i)}, \mid \boldsymbol{y}_k, \boldsymbol{x}_{k-1})$, $k = 1, \cdots, K$.

(continued on next page)

(continued from previous page)

4. Split net migration $n_k^{(i)}$ into in-migration $\tilde{x}_k^{\text{in}(i)}$ and out-migration $\tilde{x}_k^{\text{out}(i)}$
   (a) Set lower bound $m_k^{*(i)} = \text{abs}(n_k^{(i)})$
   (b) Draw gross migration $m_k^{(i)} \sim \text{poistr}(\tilde{\gamma}_k^{\text{in}} + \tilde{\gamma}_k^{\text{out}} e_k^{(i)}, m_k^{*(i)})$
   (c) If $m_k^{(i)}$ is odd and $n_k^{(i)}$ is even, or vice versa, set $m_k^{(i)} = m_k^{(i)} + 1$
   (d) Set $\tilde{x}_k^{\text{in}(i)} = (m_k^{(i)} + n_k^{(i)})/2$
   (e) Set $\tilde{x}_k^{\text{out}(i)} = (m_k^{(i)} - n_k^{(i)})/2$

Output

$\tilde{\boldsymbol{x}}_k^{(i)}$ Vector with elements $\tilde{x}_k^{\text{stk}(i)}$, x_k^{bth}, x_k^{dth}, $\tilde{x}_k^{\text{in}(i)}$, $\tilde{x}_k^{\text{out}(i)}$

$\tilde{\boldsymbol{\gamma}}_k^{(i)}$ Vector with elements $\tilde{\gamma}_k^{\text{bth}}$, $\tilde{\gamma}_k^{\text{dth}}$, $\tilde{\gamma}_k^{\text{in}(i)}$, $\tilde{\gamma}_k^{\text{out}(i)}$

A.2 Transition function

We derive the transition function $f(\boldsymbol{x}_k, \boldsymbol{\gamma}_k \mid \boldsymbol{x}_{k-1})$ from the system models described in Section 2.4. We draw the elements of $\boldsymbol{\gamma}_k$ straight from their respective gamma distributions. Conditional on demographic rates, $\boldsymbol{x}_k$ depends on $\boldsymbol{x}_{k-1}$ via x_k^{stk}, so we have

$$p(\boldsymbol{x}_k \mid \boldsymbol{x}_{k-1}) = p(x_k^{\text{stk}}, x_k^{\text{bth}}, x_k^{\text{dth}}, x_k^{\text{in}}, x_k^{\text{out}} \mid x_{k-1}^{\text{stk}}). \tag{A.1}$$

We decompose the right hand side into four conditional probabilities,

$$\begin{aligned} p(x_k^{\text{stk}}, x_k^{\text{bth}}, x_k^{\text{dth}}, x_k^{\text{in}}, x_k^{\text{out}} \mid x_{k-1}^{\text{stk}}) = p(&x_k^{\text{in}} \mid x_{k-1}^{\text{stk}}) \\ &\times p(x_k^{\text{dth}}, x_k^{\text{out}} \mid x_k^{\text{in}}, x_{k-1}^{\text{stk}}) \\ &\times p(x_k^{\text{stk}} \mid x_k^{\text{dth}}, x_k^{\text{out}}, x_k^{\text{in}}, x_{k-1}^{\text{stk}}) \\ &\times p(x_k^{\text{bth}} \mid x_k^{\text{stk}}, x_k^{\text{dth}}, x_k^{\text{out}}, x_k^{\text{in}}, x_{k-1}^{\text{stk}}) \end{aligned} \tag{A.2}$$

The first of these conditional probabilities is easily calculated, using the fact that our system model for in-migration does not include exposure,

$$p(x_k^{\text{in}}) = \text{pois}(x_k^{\text{in}} \mid \gamma_k^{\text{in}}). \tag{A.3}$$

The second conditional probability is the most complicated. We first develop an expression for the special case where γ_k^{in}, and hence x_k^{in}, is 0, leaving only $p(x_k^{\text{dth}}, x_k^{\text{out}} \mid x_{k-1}^{\text{stk}})$. In this special case, stock x_{k-1}^{stk} is subject to decrements via two competing and independent risks: death and out-migration. For simplicity, we treat everyone in x_{k-1}^{stk} as being born exactly halfway through the year, and therefore experiencing exactly 0.5 person-years within each Lexis triangle if they neither die nor emigrate. Using standard methods for multiple decrement processes (see for example Preston et al. [2001, ch. 4]), we can calculate probabilities for the two possible ways of exiting, and the one possible way of remaining in, the Lexis triangle,

Outcome	Probability
Exit via death	$\frac{\gamma_k^{\text{dth}}}{\gamma_k^{\text{dth}}+\gamma_k^{\text{out}}} \left(1 - e^{-\frac{1}{2}(\gamma_k^{\text{dth}}+\gamma_k^{\text{out}})}\right)$
Exit via out-migration	$\frac{\gamma_k^{\text{out}}}{\gamma_k^{\text{dth}}+\gamma_k^{\text{out}}} \left(1 - e^{-\frac{1}{2}(\gamma_k^{\text{dth}}+\gamma_k^{\text{out}})}\right)$
Remain to end	$e^{-\frac{1}{2}(\gamma_k^{\text{dth}}+\gamma_k^{\text{out}})}$

Let $\boldsymbol{\pi}$ denote a vector holding the three probabilities. Applying these probabilities to everyone in x_{k-1}^{stk} leads to

$$p(x_k^{\text{dth}}, x_k^{\text{out}} \mid x_{k-1}^{\text{stk}}) = \text{multinom}((x_k^{\text{dth}}, x_k^{\text{out}}, x_{k-1}^{\text{stk}} - x_k^{\text{dth}} - x_k^{\text{out}}) \mid x_{k-1}^{\text{stk}}; \boldsymbol{\pi}). \tag{A.4}$$

Incorporating in-migration into the calculations is potentially complicated, given that immigrants arrive at different times and are therefore subject to

different probabilities of exiting and remaining. We keep the calculations simple by using a computational shortcut commonly used in population projections [Preston et al., 2001, pp. 125-126]. We assume that half of all immigrants arrive at the start, and are thus subject to the same risks as the initial population, and that the remaining immigrants arrive at the end, and are thus subject to zero risks. If the number of immigrants x_k^{in} is even, we set

$$\begin{aligned} &p(x_k^{\text{dth}}, x_k^{\text{out}} \mid x_k^{\text{in}}, x_{k-1}^{\text{stk}}) = \\ &\quad \text{multinom}((x_k^{\text{dth}}, x_k^{\text{out}}, \tfrac{1}{2}x_k^{\text{in}} + x_{k-1}^{\text{stk}} - x_k^{\text{dth}} - x_k^{\text{out}}) \mid \tfrac{1}{2}x_k^{\text{in}} + x_{k-1}^{\text{stk}}; \boldsymbol{\pi}). \end{aligned} \quad \text{(A.5)}$$

If the number of immigrants is odd, we average over the cases where the x_k^{in}th immigrant arrives at the start and end,

$$\begin{aligned} &p(x_k^{\text{dth}}, x_k^{\text{out}} \mid x_k^{\text{in}}, x_{k-1}^{\text{stk}}) = \\ &\tfrac{1}{2}\text{multinom}((x_k^{\text{dth}}, x_k^{\text{out}}, \tfrac{1}{2}x_k^{\text{in}} + \tfrac{1}{2} + x_{k-1}^{\text{stk}} - x_k^{\text{dth}} - x_k^{\text{out}}) \mid \tfrac{1}{2}x_k^{\text{in}} + \tfrac{1}{2} + x_{k-1}^{\text{stk}}; \boldsymbol{\pi}) \\ &+ \tfrac{1}{2}\text{multinom}((x_k^{\text{dth}}, x_k^{\text{out}}, \tfrac{1}{2}x_k^{\text{in}} - \tfrac{1}{2} + x_{k-1}^{\text{stk}} - x_k^{\text{dth}} - x_k^{\text{out}}) \mid \tfrac{1}{2}x_k^{\text{in}} - \tfrac{1}{2} + x_{k-1}^{\text{stk}}; \boldsymbol{\pi}). \end{aligned} \quad \text{(A.6)}$$

An alternative way to divide out immigrants would be to allocate them probabilistically to the start and end. However, this would complicate the calculation of $p(x_k^{\text{dth}}, x_k^{\text{out}} \mid x_k^{\text{in}}, x_{k-1}^{\text{stk}})$, since we would need to sum over all possible configurations for x_k^{in}.

The third term on the right hand side of Equation (A.2) is equal to 1,

$$p(x_k^{\text{stk}} \mid x_k^{\text{dth}}, x_k^{\text{out}}, x_k^{\text{in}}, x_{k-1}^{\text{stk}}) = 1, \quad \text{(A.7)}$$

since the quantities involved are related deterministically,

$$x_k^{\text{stk}} = x_{k-1}^{\text{stk}} - x_k^{\text{dth}} + x_k^{\text{in}} - x_k^{\text{out}}. \quad \text{(A.8)}$$

Births to a cohort have no effect on the size of that cohort, so the fourth term on the right hand side of Equation (A.2) reduces to

$$p(x_k^{\text{bth}} \mid x_k^{\text{stk}}, x_{k-1}^{\text{stk}}) = \text{pois}(x_k^{\text{bth}} \mid \gamma_k^{\text{bth}} e_k). \quad \text{(A.9)}$$

A.3 Algorithm for extending a series

Figure A.2 presents an algorithm for extending an existing series by making draws from the filtering distribution.

Input

$\{\boldsymbol{x}_K^{(i)}\}$	N draws from the posterior distribution $p(\boldsymbol{x}_K \mid \boldsymbol{y}_{0:K})$
$f(\boldsymbol{x}_k \mid \boldsymbol{x}_{k-1})$	Transition function
$g(\boldsymbol{y}_k \mid \boldsymbol{x}_k)$	Likelihood
$q(\boldsymbol{x}_k \mid \boldsymbol{y}_k, \boldsymbol{x}_{k-1})$	Importance function
a	Resampling threshold, $0 \leq a \leq 1$

Algorithm

- For $k = K+1, K+2$
  1. For $i = 1, \cdots, N$
     (a) Draw $\tilde{\boldsymbol{x}}_k^{(i)} \sim q(\boldsymbol{x}_k \mid \boldsymbol{y}_k, \boldsymbol{x}_{k-1}^{(i)})$
     (b) Set $\tilde{\boldsymbol{x}}_{K+1:k}^{(i)} = (\boldsymbol{x}_{K+1:k-1}^{(i)}, \tilde{\boldsymbol{x}}_k^{(i)})$
     (c) Calculate unnormalised weights

$$\tilde{w}_k^{(i)} = \frac{g(\boldsymbol{y}_k \mid \tilde{\boldsymbol{x}}_k^{(i)}) f(\tilde{\boldsymbol{x}}_k^{(i)} \mid \boldsymbol{x}_{k-1}^{(i)})}{q(\tilde{\boldsymbol{x}}_k^{(i)} \mid \boldsymbol{y}_k, \boldsymbol{x}_{k-1}^{(i)})} W_{k-1}^{(i)}$$

  2. Calculate normalised weights $\tilde{W}_k^{(i)} = \tilde{w}_k^{(i)} / \sum_{j=1}^N \tilde{w}_k^{(j)}$
  3. Calculate effective sample size $\hat{N}_k = 1 / \left(\sum_{i=1}^N (\tilde{W}_k^{(i)})^2 \right)$
  4. If $\hat{N}_k < aN$ or if $k = K+2$ then resample, obtaining N particles $\boldsymbol{x}_{K+1:k}^{(i)}$ with weights $W_k^{(i)} = 1/N$. Otherwise set $\boldsymbol{x}_{K+1:k}^{(i)} = \tilde{\boldsymbol{x}}_{K+1:k}^{(i)}$ with weights $W_k^{(i)} = \tilde{W}_k^{(i)}$.

Output

$\{\boldsymbol{x}_{K+1:K+2}^{(i)}\}$ N draws from distribution $p(\boldsymbol{x}_{K+1:K+2} \mid \boldsymbol{x}_K^{(i)}, \boldsymbol{y}_{K+1:K+2})$

Figure A.2: Particle filter for adding one extra period to an existing cohort.

Appendix B

Example of TMB C++ template for estimating one cohort

```
#include <TMB.hpp>

template<class Type>
Type objective_function<Type>::operator() ()
{
  // input values
  DATA_SCALAR(val_stk_init);
  DATA_VECTOR(val_dth);
  DATA_VECTOR(mean_dth);
  DATA_VECTOR(mean_im);
  DATA_VECTOR(mean_em);
  DATA_SCALAR(sd_dth);
  DATA_SCALAR(sd_im);
  DATA_SCALAR(sd_em);
  DATA_VECTOR(data_stk);
  DATA_VECTOR(data_im);
  DATA_VECTOR(data_em);

  // parameters returned to R
  PARAMETER_VECTOR(log_rate_dth);
  PARAMETER_VECTOR(log_expect_im);
  PARAMETER_VECTOR(log_rate_em);
  PARAMETER_VECTOR(log_val_im);
  PARAMETER_VECTOR(log_val_em);
```

```
  // quantities used in calculations
  int K = val_dth.size();
  vector<Type> val_im = exp(log_val_im);
  vector<Type> val_em = exp(log_val_em);
  vector<Type> val_stk(K);

  // population accounting equation
  val_stk[0] = val_stk_init - val_dth[0] + val_im[0] - val_em[0];
  for (int k = 1; k < K; k++)
    val_stk[k] = val_stk[k-1] - val_dth[k] + val_im[k] - val_em[k];

  // exposure
  vector<Type> exposure(K);
  exposure[0] = 0.5 * (val_stk_init + val_stk[0]);
  for (int k = 1; k < K; k++)
    exposure[k] = 0.5 * (val_stk[k-1] + val_stk[k]);

  // negative log posterior (= negative log likelihood + negative log prior)
  Type ans = 0;

  // contribution from rate_dth, expect_im, rate_em, including Jacobians
  ans -= dnorm(log_rate_dth, mean_dth, sd_dth, true).sum() - log_rate_dth.sum();
  ans -= dnorm(log_expect_im, mean_im, sd_im, true).sum() - log_expect_im.sum();
  ans -= dnorm(log_rate_em, mean_em, sd_em, true).sum() - log_rate_em.sum();

  // contribution from val_dth, val_im, val_em, including Jacobians
  vector<Type> expect_dth = exp(log_rate_dth) * exposure;
  vector<Type> expect_im = exp(log_expect_im);
  vector<Type> expect_em = exp(log_rate_em) * exposure;
  ans -= dpois(val_dth, expect_dth, true).sum();
  ans -= dpois(val_im, expect_im, true).sum() + log_val_im.sum();
  ans -= dpois(val_em, expect_em, true).sum() + log_val_em.sum();

  // contribution from data
  ans -= dpois(data_stk, val_stk, true).sum();
  ans -= dpois(data_im, val_im, true).sum();
  ans -= dpois(data_em, val_em, true).sum();

  return ans;
}
```

Bibliography

Pete Benton, 2021. URL https://blog.ons.gov.uk/2021/07/13/population-and-social-statistics-in-a-rapidly-changing-world/.

United Nations Population Division. Expert group meeting on methods for the world population prospects 2021 and beyond, 2020. URL https://www.un.org/development/desa/pd/events/expert-group-meeting-methods-world-population-prospects-2021-and-beyond.

John Bryant and Patrick Graham. Bayesian demographic accounts: Subnational population estimation using multiple data sources. *Bayesian Analysis*, 8(3): 591–622, 2013.

John Bryant and Junni L Zhang. *Bayesian Demographic Estimation and Forecasting*. CRC Press, 2018.

John Bryant, Jenny Harlow, Junni L. Zhang, Charlotte Taglioni, and Feifei Wang. *demest: Bayesian Demographic Estimation and Forecasting*, 2021. R package version 0.0.0.5.4.

P.H. Rees. Regional Population Project Models and Accounting Methods. *Journal of the Royal Statistical Society. Series A (General)*, 142(2):223–255, 1979. ISSN 0035-9238.

R. Stone. The accounts of society. In *Nobel Prize in Economics Documents*. Nobel Prize Committee, 1984.

Frans Willekens. *Population Accounts*, pages 29–40. Springer, 2011.

Ruth King, Byron Morgan, Olivier Gimenez, and Steve Brooks. *Bayesian analysis for population ecology*. CRC Press, 2009.

Mark C Wheldon, Adrian E Raftery, Samuel J Clark, and Patrick Gerland. Reconstructing past populations with uncertainty from fragmentary data. *Journal of the American Statistical Association*, 108(501):96–110, 2013.

James Raymer, Arkadiusz Wiśniowski, Jonathan J Forster, Peter WF Smith, and Jakub Bijak. Integrated modeling of European migration. *Journal of the American Statistical Association*, 108(503):801–819, 2013.

KB Newman, ST Buckland, BJT Morgan, R King, DL Borchers, DJ Cole, et al. *Modelling population dynamics: model formulation, fitting and assessment using state-space methods*. Springer, 2014.

Leontine Alkema, Doris Chou, Daniel Hogan, Sanqian Zhang, Ann-Beth Moller, Alison Gemmill, Doris Ma Fat, Ties Boerma, Marleen Temmerman, Colin Mathers, et al. Global, regional, and national levels and trends in maternal mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the UN Maternal Mortality Estimation Inter-Agency Group. *The Lancet*, 387(10017):462–474, 2016.

Seth Flaxman, Swapnil Mishra, Axel Gandy, H Juliette T Unwin, Thomas A Mellan, Helen Coupland, Charles Whittaker, Harrison Zhu, Tresnia Berah, Jeffrey W Eaton, et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, 584(7820):257–261, 2020.

Marie Auger-Méthé, Ken Newman, Diana Cole, Fanny Empacher, Rowenna Gryba, Aaron A King, Vianey Leos-Barajas, Joanna Mills Flemming, Anders Nielsen, Giovanni Petris, et al. A guide to state–space modeling of ecological time series. *Ecological Monographs*, 91(4):e01470, 2021.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

Monica Alexander and Leontine Alkema. A bayesian cohort component projection model to estimate women of reproductive age at the subnational level in data-sparse settings. *Demography*, 59(5):1713–1737, 2022.

Charles B Yackulic, Michael Dodrill, Maria Dzul, Jamie S Sanderlin, and Janice A Reid. A need for speed in bayesian population models: a practical guide to marginalizing and recovering discrete latent states. *Ecological Applications*, 30(5):e02112, 2020.

Arnaud Doucet, Adam M Johansen, et al. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, 12(656-704): 3, 2009.

Nikolas Kantas, Arnaud Doucet, Sumeetpal S Singh, Jan Maciejowski, and Nicolas Chopin. On particle methods for parameter estimation in state-space models. *Statistical science*, 30(3):328–351, 2015.

Rens van de Schoot, Sarah Depaoli, Ruth King, Bianca Kramer, Kaspar Märtens, Mahlet G Tadesse, Marina Vannucci, Andrew Gelman, Duco Veen, Joukje Willemsen, et al. Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1):1–26, 2021.

Kasper Kristensen, Anders Nielsen, Casper W Berg, Hans Skaug, and Bradley M Bell. TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, 70:1–21, 2016.

Samuel H Preston and Ansley J Coale. Age structure, growth, attrition, and accession: A new synthesis. *Population Index*, pages 217–259, 1982.

TA Moultrie, RE Dorrington, AG Hill, K Hill, IM Timæus, and B Zaba. *Tools for Demographic Estimation*. International Union for the Scientific Study of Population, Paris, 2013.

J Carpenter, Peter Clifford, and Paul Fearnhead. An improved particle filter for non-linear problems. *IEE Proceedings Radar Sonar and Navigation*, 146 (1):2–7, February 1999. ISSN 1350-2395. doi: 10.1049/ip-rsn:19990255.

Jeffrey W Eaton, Laura Dwyer-Lindgren, Steve Gutreuter, Megan O'Driscoll, Oliver Stevens, Sumali Bajaj, Rob Ashton, Alexandra Hill, Emma Russell, Rachel Esra, et al. Naomi: a new modelling tool for estimating hiv epidemic indicators at the district level in sub-saharan africa. *Journal of the International AIDS Society*, 24:e25788, 2021.

Laura Dwyer-Lindgren, Parkes Kendrick, Yekaterina O Kelly, Dillon O Sylte, Chris Schmidt, Brigette F Blacker, Farah Daoud, Amal A Abdi, Mathew Baumann, Farah Mouhanna, et al. Life expectancy by county, race, and ethnicity in the usa, 2000–19: a systematic analysis of health disparities. *The Lancet*, 400(10345):25–38, 2022.

Aaron Osgood-Zimmerman and Jon Wakefield. A statistical review of template model builder: A flexible tool for spatial modelling. *International Statistical Review*, 2022.

David MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Bradley Bell. Cppad: A package for c++ algorithmic differentiation, 2006. *URL http://www.coin-or.org/CppAD*, 2006.

Stephen E Fienberg. An iterative procedure for estimation in contingency tables. *The Annals of Mathematical Statistics*, 41(3):907–917, 1970.

Frans Willekens. Modeling approaches to the indirect estimation of migration flows: From entropy to em. *Mathematical Population Studies*, 7(3):239–278, 1999.

ONS. `https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/bulletins/longterminternationalmigrationprovisional/june2021`, 2022a.

ONS. `https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/measuresofstatisticaluncertaintyinonslocalauthoritymidyearpopulationestimates/latest`, 2022b.

S. N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36, 2011.

ONS. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/adminbasedpopulationestimatesandstatisticaluncertainty/july2020, 2020.

ONS. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/dynamicpopulationmodelforlocalauthoritycasestudiesinenglandandwales/2011to2022, 2022c.

ONS. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/dynamicpopulationmodelforenglandandwales/2022-07-14, 2022d.

Adrian FM Smith and Alan E Gelfand. Bayesian statistics without tears: a sampling–resampling perspective. *The American Statistician*, 46(2):84–88, 1992.

Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*, 2015.

Richard Stone, D. G. Champernowne, and J. E. Meade. The precision of national income estimates. *The Review of Economic Studies*, 9(2):111–125, 1942.

Jean-Claude Deville and Carl-Erik Särndal. Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418):376–382, 1992.

R. Stone. *Input-Output and National Accounts*. Organisation for Economic Co-operation and Development, Paris, 1961.

Michael L. Lahr and Louis De Mesnard. Biproportional techniques in input-output analysis: Table updating and structural analysis. *Economic Systems Research*, 16(2), 2004.

Estela Dagum and Pierre Cholette. *Benchmarking, Temporal Distribution, and Reconciliation Methods for Time Series*, volume 186. Springer-Verlag, New York, 2006. ISBN 978-0-387-31102-9. doi: 10.1007/0-387-35439-5.

Eurostat. *ESS guidelines on temporal disaggregation, benchmarking and reconciliation: 2018 Edition*, 2018. URL https://ec.europa.eu/eurostat/documents/3859598/9441376/KS-06-18-355-EN.pdf/fce32fc9-966f-4c13-9d20-8ce6ccf079b6.

International Monetary Fund IMF. *Quarterly National Accounts Manual: 2017 Edition*, 2018. URL https://www.imf.org/external/pubs/ft/qna/pdf/2017/QNAManual2017text.pdf.

Homesh Sayal, John A. D. Aston, Duncan Elliott, and Hernando Ombao. An introduction to applications of wavelet benchmarking with seasonal adjustment. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(3):863–889, 2017. doi: https://doi.org/10.1111/rssa.12241. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssa.12241.

Eurostat. Glossary:satellite account. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Satellite_account, 2022. Accessed: 25 June 2022.

Andrew Gelman, Aki Vehtari, Daniel Simpson, Charles C Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. Bayesian workflow. *arXiv preprint arXiv:2011.01808*, 2020.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27(5):1413–1432, 2017.

S.H. Preston, P. Heuveline, and M. Guillot. *Demography: Modelling and Measuring Population Processes*. Blackwell, Oxford, 2001.