Hierarchical models and aggregate uncertainty in the Dynamic Population Model

Duncan Elliott and Salah Merad

2025-02-20

Summary

This paper presents further developments to the computational framework of the Dynamic Population Model that is used for estimating Admin-Based Population Estimates. The developments presented in the paper address the problem of estimating uncertainty for aggregate estimates such as Local Authority or England and Wales population totals. We would like to thank all members of the team that have been involved in the development and testing of models as well as the critical feedback from collaborators John Bryant (Bayesian Demography Ltd.) and the University of Southampton (Professor Peter Smith, Professor Jakub Bijak, Dr Jason Hilton, Professor Erengul Dodd, Dr Joanne Ellison, Andrew Hind).

Questions for MARP

1. Is the proposed modification to the computational framework of the Dynamic Population Model sensible?
2. Do the panel have suggestions for improvements to the coverage adjustment models that we are fitting in step 1 that might improve the estimates of uncertainty for aggregates?
3. Do panel members have any thoughts on the need for regular reviews of the step 1 models and how these should be conducted.

1 Introduction

The Dynamic Population Model (DPM) combines estimates of births, deaths, migration and different estimates of the population at mid-year points from 2011 to 2023 by Local Authority, single year of age and sex to estimate a coherent demographic account (population, events and demographic rates) that we refer to as Admin Based Population Estimates (ABPEs). The development of the DPM has followed a software-engineering project-management approach, first establishing a basic working model, and then successively refining it. We have provided updates on developments at previous MARP meetings:

- outlining plans for a Bayesian demographic accounts (Blackwell et al, 2021),
- the initial computational breakthrough, demonstrating the feasibility of the DPM (Blackwell et al, 2022),
- R package developments and model refinements, publictions and engagement activities with LAs (Elliott and Blackwell, 2023),

- faster estimation methods (Merad et al, 2024).

In this paper we provide a summary of the work to date on modifications to the computational framework of the DPM to improve estimates of uncertainty, in particular for aggregates such as Local Authority or England and Wales population totals.

In the papers listed above we have described the justification of using different steps in the DPM to make the computation feasible. Here we focus only on the first (initial adjustment for bias in observed data and estimation of demographic rates) and second (cohort estimation) steps. In Section 2 we describe the objective in producing ABPEs and explain the limitations in the approach we have been using in publications to date in calculating uncertainty in our estimates.

To obtain improved uncertainty we have proposed a multiple draws approach which involves fitting the cohort estimation step of the DPM multiple times using draws of certain hyperparameters estimated in step 1. Using the multiple draws approach, estimates of aggregate uncertainty (aggregating cells after step 2) are sensitive to the models specified in step 1.

In Section 3 we summarise the models that have been used in publications to date and briefly describe the models we have been exploring that aim to improve estimates of aggregate uncertainty (technical specifications of these models are provided for information in section Section 7). In the context of usual models in statistical demography we are exploring unusually large models (hundreds of thousands of parameters). These are being reviewed internally for quality assurance and user acceptance testing, as well as through reviews and discussions with external academics supporting the project. This is also proposed as a topic for discussion at the MARP subgroup that is being formed for the DPM.

Section 4 briefly outlines the multiple draw approach that is proposed for step 2 to approximate the posterior distribution of interest. Section 5 presents results for some example step 1 models and the impact on aggregate uncertainty after step 2, while Section 6 summarises the discussion.

2 DPM

The main objective is to estimate the components of an unobserved demographic account (x), including demographic rates (γ) and parameters that address bias (ρ) in the observed data (y). Hence we wish to produce draws from the target distribution

$$p(x,\rho,\gamma|y) = p(\rho,\gamma|x,y)p(x|y) \qquad (1)$$

Let ϕ and ψ be hyperparameters of initial models for γ and ρ respectively (see section Section 3 for further details). We can write

$$p(x|y) = \int p(x|y,\phi,\psi)p(\phi,\psi|y)d\phi d\psi \qquad (2)$$

In the current implementation of the model we approximate the true distribution by conditioning on point estimates $\hat{\phi}$ and $\hat{\psi}$, that is we use

$$p(x|y) = p(x|y, \hat{\phi}, \hat{\psi}) \qquad (3)$$

This approach provides good annual point estimates by single year of age, sex, and local authority, as well as aggregates such as LA totals, but credible intervals are underestimated and severely so for aggregates such as LA totals. The reason for this is that the estimation does not account for correlations between cells (defined by year, single year of age, sex, and LA) in both ϕ and ψ.

To address this we propose to approximate the true distribution as

$$\int p(x|y, \phi, \psi) d\phi d\psi \approx \frac{1}{M} \sum_{m=1}^{M} p\left(x|y, \phi^{(m)}, \psi^{(m)}\right) \qquad (4)$$

where $\phi^{(m)}$ and $\psi^{(m)}$, $m = 1, \ldots, M$, are M draws from the distribution of the hyperparameters ϕ and ψ from step 1. Note that we take a sample of size J/M from $p(x|y, \phi^{(m)}, \psi^{(m)})$ for the estimates of the demographic account, see the diagram in Section 3.

Individual cohorts (c) are independent conditional on the hyperparameters so

$$p(x|y, \phi^{(m)}, \psi^{(m)}) = \prod_c p\left(x_c|y_c, \phi_c^{(m)}, \psi_c^{(m)}\right) \qquad (5)$$

In terms of the computation, we describe step 1 as models used for the initial estimation of ϕ and ψ, whereas as step 2 is the estimation of a consistent demographic account over time for a single cohort.

3 Step 1 models (ϕ and ψ)

In our previous papers listed in Section 1 we have distinguished between data models (accounting for bias and precision of observed data) and system models (estimation of demographic rates). Note that ϕ correspond to hyperparameters for system models and ψ are hyperparameters for data models.

In publications to date the estimation of the expected values, $\hat{\phi}$ and $\hat{\psi}$, has used generalised additive models. However, the way in which they have been estimated does not capture well correlations across all cells within England and Wales. We briefly discuss the main issues with the estimation for these data and system model hyperparameters and describe the models being explored to improve their estimation.

3.1 Data model hyperparameters (ψ)

The current models for ψ are fit to clusters of LAs and use parametric bootstrapping to produce estimates of uncertainty. This approach was developed to produce approximate uncertainty intervals for admin-based population estimates by Local Authority, single year of age and sex (ONS, 2020). The main problem with this approach is that it was not designed for aggregate estimates and unfortunately leads to very high correlations for all pairs of cells. Moreover, there is no accounting for changes over time as the model is fit on data from a single year. For example, the Patient Register estimates coverage ratios using 2011 data only, while the Statistical Population Dataset uses 2021 data only.

To address these limitations we have been exploring hierarchical models that are fit to all data in England and Wales from 2011 to the present. These models include covariates that involve age, time, sex and region to obtain draws $\psi^{(m)}$ that can better capture correlations, leading to improved estimates of uncertainty from step 2. Note that the data to fit the models is restricted because benchmark estimates for these models are only available in 2011 and 2021, where we make use of mid-year estimates from census years, and we make the assumption that they are unbiased.

3.2 System model hyperparameters (ϕ)

The current models for ϕ are fit independently by LA and sex and hence do not capture correlation between LAs or sexes. Similarly to the models being explored for ψ we are exploring hierarchical model that are fit to all data in England and Wales from 2012 to present, including covariates that involve age, time, sex and region.

Note that the exposure terms in the system models for birth, deaths and outflows are calculated using expected coverage adjusted estimates. Potentially we could use $\psi^{(m)}$, and run the models M times, but this becomes very computationally expensive for negligible effects on the resulting $\phi^{(m)}$ and therefore we use $\hat{\psi}$.

4 Step 2 cohort estimation

In step 2 we calculate estimates for each of the M draws of the hyperparameters from Step 1. Step 2 involves estimation by individual annual birth cohort (for each Local Authority and sex combination).

In Step 2, for each $m = 1, \cdots, M$, we obtain J/M draws from

$$p\left(x^{\text{stk}}, x^{\text{mig}}, \gamma, \rho | \phi^{(m)}, \psi^{(m)}, y\right) \tag{6}$$

We concatenate the resulting M sets of draws, to obtain a total of J approximate draws from $p(x, \rho, \gamma | y)$ with structure

$$
\begin{array}{ccc}
x^{(1,1)} & \gamma^{(1,1)} & \rho^{(1,1)} \\
\vdots & \vdots & \vdots \\
x^{(J/M,1)} & \gamma^{(J/M,1)} & \rho^{(J/M,1)} \\
x^{(1,2)} & \gamma^{(1,2)} & \rho^{(1,2)} \\
\vdots & \vdots & \vdots \\
x^{(J/M,M)} & \gamma^{(J/M,M)} & \rho^{(J/M,M)}.
\end{array}
$$

We are testing alternative values of M and J to ensure that we get at good representation of the posterior distribution while ensuring that the computation time is reasonable for research and production purposes.

These concatenated sets of draws for every cell (with cells being defined by the unique combinations of single year of age, sex, LA, time and cohort) can be aggregated across some dimensions for each draw and from these we then calculate point estimates and credible intervals.

Let D be a sub-population group that comprises cells $i_1, \cdots, i_K$. To obtain the 95% credible interval of the population estimate for domain D we compute the 2.5 and 97.5 percentiles of $\sum_{k=1}^{k=K} x_{i_k}^{\text{stk}(1,1)}, \cdots, \sum_{k=1}^{k=K} x_{i_k}^{\text{stk}(J/M,M)}$.

The ABPEs in the Population estimates for England and Wales: mid-2023 publication included point estimates and 95% credible intervals for cells defined by single year of age, sex, LA and time. Point estimates and 95% credible intervals for LA population totals were also derived using a limited multiple draws approach.

The reason for describing this as a limited multiple draws approach is that, as discussed in Section 3, the step 1 models for ϕ and ψ may not appropriately capture correlations for calculating uncertainty for estimates at any level of aggregation.

5 Results

In this section we present some results that demonstrate the impact of the models we have been testing to estimate coverage ratios (ie ρ which are a component of ψ) for the stock data in step 1 and the results of using the multiple draws approach at step 2. The full step 1 coverage ratio model specifications, including priors, are provided in Section 7. Below we focus on the differences between the models in terms of variables included (ie model 2 is a nested version of model 1 and model 3 is a nested version of model 2). Note that the priors for terms included in models do not change.

Let cell i denote a combination of age a by sex s by LA r by time t. We observe y_i (SPD population estimate) for all cells, but only observe Census year mid-year estimates (y_i^{MYE}) at $t = 2011{,}2021$).

Model 1 is

$$y_i \sim \text{Poisson}\left(\rho_i y_i^{\text{MYE}}\right) \tag{7}$$

$$\rho_i \sim \text{Gamma}(\xi^{-1}, (\xi\mu_i)^{-1}) \quad (8)$$

$$\log\mu_i = \beta_i^{\text{Intercept}} + \beta_{a_i}^{\text{age}} + \beta_{r_i}^{\text{region}} + \beta_{s_i}^{\text{sex}} + \beta_{(a,s)_i}^{\text{age:sex}} + \beta_{(a,r)_i}^{\text{age:region}} + \beta_{(s,r)_i}^{\text{sex:region}} + \beta_{(t,r)_i}^{\text{time:region}} + \beta_{(t,a)_i}^{\text{time:age}} + \beta_{(t,s)_i}^{\text{time:sex}} \quad (9)$$

where β and ξ are parameters to be estimated in the model. We require the posterior distribution of ρ for the multiple draws approach in step 2.

Model 2 excludes the terms $\beta_{(t,a)_i}^{\text{time:age}}, \beta_{(t,s)_i}^{\text{time:sex}}$ and then model 3 further excludes $\beta_{a_i}^{\text{age}}, \beta_{r_i}^{\text{region}}, \beta_{s_i}^{\text{sex}}$.

We have evaluated these models: one of the validation results has been to compute in-sample and out-of-sample coverage. These are calculated by randomly setting 10% of observed SPD estimates to missing for model fitting. We then predict those missing values based on the model and report the percentage of observed values that fall within the 95% credible intervals of the predicted counts. As can be seen in Table 1, there is little difference between these models. The in-sample percentage is about right in 2011 but too high in 2021. For the out-of-sample the 2021 coverage is reasonable but slightly low in 2011.

Table 1: In- and out-of-sample coverage for selected coverage ratio models for SPD.

Model	Year	In_sample	Out_sample
Model 1	2011	0.953	0.891
Model 1	2021	0.991	0.970
Model 2	2011	0.954	0.887
Model 2	2021	0.987	0.966
Model 3	2011	0.953	0.892
Model 3	2021	0.988	0.966

A second validation of these models involved computing the coefficient of variation of the coverage adjusted estimates at the England and Wales level obtained using draws from the posterior distribution of the coverage ratios ρ. It is clear from Table 2 that model 1, that includes terms for age by time $\beta_{(t,a)_i}^{\text{time:age}}$ and sex by time $\beta_{(t,s)_i}^{\text{time:sex}}$ covariates, has much lower precision, with coefficients of variation about 20 times higher than those of models 2 and 3.

Table 2: Estimated coefficients of variation for England & Wales level coverage adjusted SPD estimates

Year	Model1	Model2	Model3
2016	3.141	0.164	0.180
2017	2.967	0.155	0.143
2018	2.899	0.141	0.134
2019	2.537	0.131	0.115
2020	2.044	0.097	0.103
2022	1.878	0.099	0.093
2023	2.602	0.138	0.133

Given the poor precision from model 1, we have tested step 2 estimation with draws from model 3, on the basis that it is more precise than model 1 and while similar to model 2 it can be fitted faster.

We now briefly present some results from running step 2, focussing on Local Authority, and England and Wales population totals. In order to demonstrate the importance of using the multiple draws approach for estimating aggregate uncertainty, we show the 95% credible intervals derived from using a single run and multiple runs for two Local Authorities. As can be seen in Figure 1 the credible intervals are wider for the multiple draws approach, that is a result of capturing correlations between cohorts, that are not accounted for in the single run approach.

Figure 1: Population estimates for Blackpool (left) and Cambridge (right) with 95% credible intervals

[Figure 2](#) shows boxplots of the the estimated coeffificents of variation for Local Authority level estimates of population over time for both the coverage adjusted inputs (ie the results of step 1 that are passed to step 2) and the estimates after running step 2 using the multiple draws approach. As can be seen, the precision is slightly increased in 2011 and 2021 after step 2. We expect the precision to only improve slightly as the mid-year estimates from census years are assumed to be precise relative to the other stock data. The precision of the coverage adjusted inputs increases the further away from a Census year, and Patient Register data (2012 to 2015) is slightly less precise than SPD (2016 to 2020 and 2022 to 2023). As would be expected, the more precise population estimates after step 2 are obtained for the years 2011 and 2021, while in those years further from census years the precision reduces.

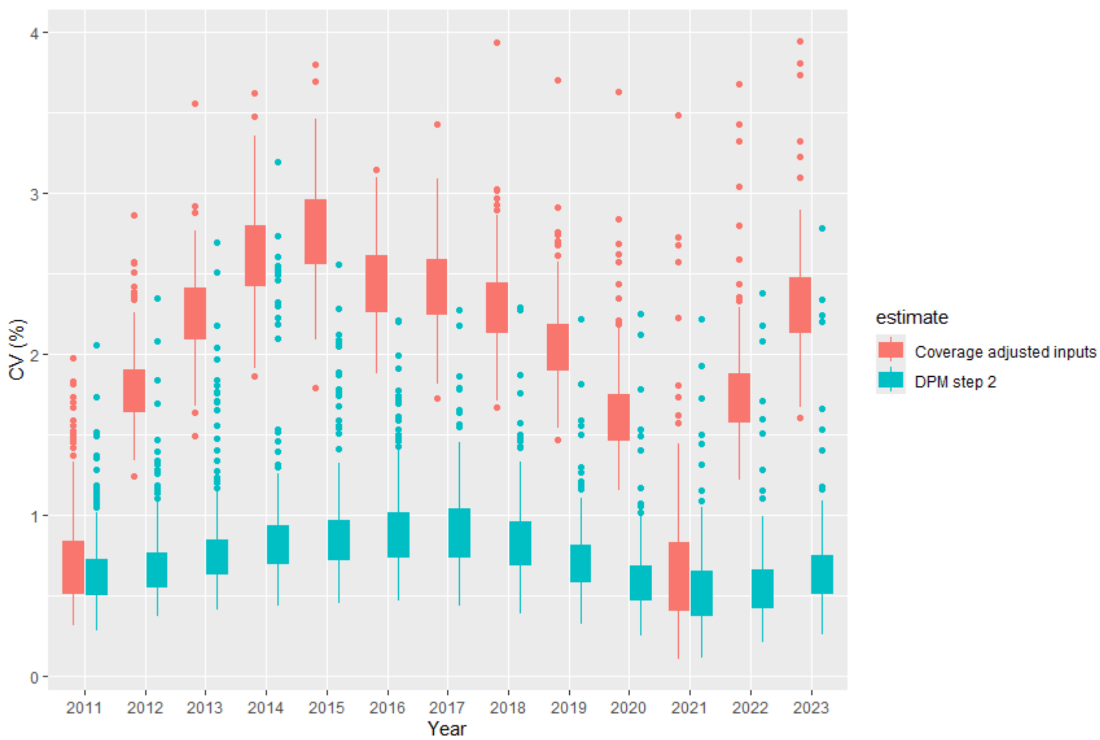

Figure 2: Estimated coefficients of variation (%) of Local Authority level estimates of population

At the England and Wales level we observe a similar pattern, as shown in Figure 3, except that the coefficient of variation for the DPM estimate after step 2 (multiple runs) for 2021 is greater than those for a number of other years, which is not what we would expect. We believe that this is a sign the coverage ratio models may not be sufficiently accounting for correlations between regions.

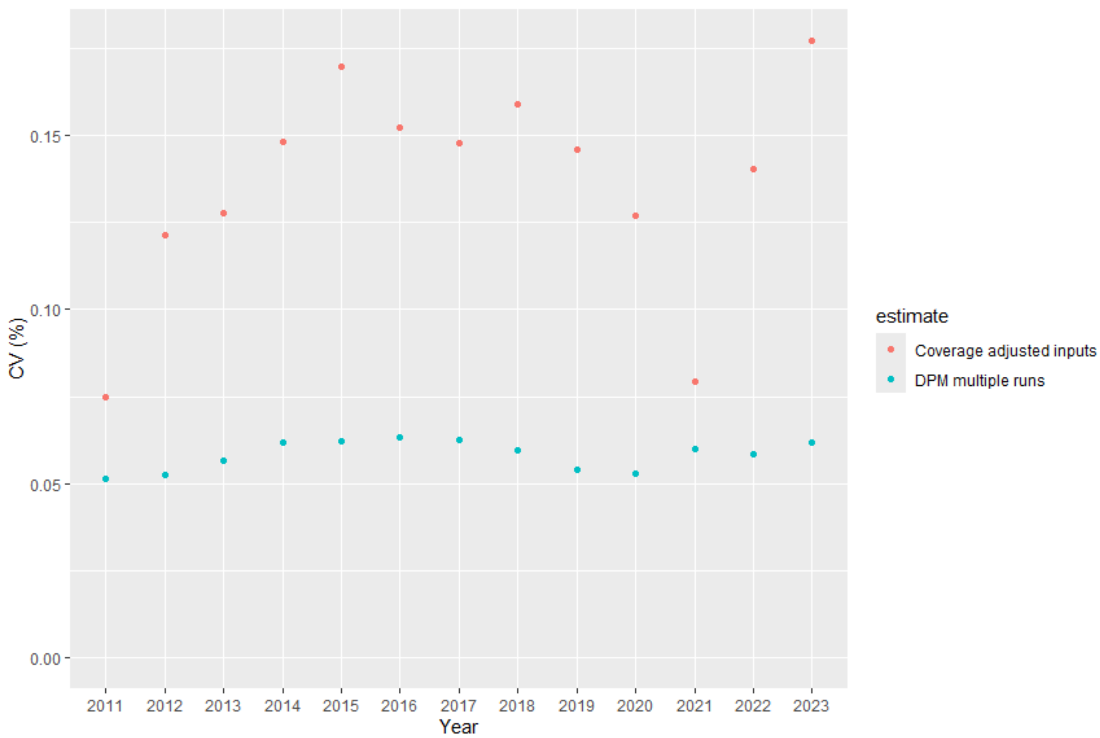

Figure 3: Estimated coefficients of variation (%) of England and Wales level estimates of population

6 Discussion

We believe that the multiple draws approach is an important development of the DPM that allows us to produce reasonably accurate estimates of uncertainty at aggregate levels. The results that we have presented show plausible estimates of uncertainty for Local Authority population totals. However, it appears that the coverage ratio models are perhaps not appropriately capturing regional correlation, which leads to England and Wales level estimates of uncertainty that are less plausible.

An important practical consideration for a regular production cycle is the issue of revisions. Clearly as new data arrives or historical data is revised, potentially causing structural changes, it will be important to review and potentially modify the step 1 hierarchical model specifications. We would be interested to hear panel members opinions on the necessity of this and how such reviews of the models might be conducted.

The hierarchical models that have been presented to use for coverage adjustment of the population stock data are a necessary part of the estimation as we do not have an alternative coverage adjustment method for these periods. These models will need to be modified appropriately when new coverage adjustment methods are developed.

References

- Blackwell. L., Elliott, D. and Bryant, J. (2021), Integrated statistical design for the transformed population and social statistics system- Bayesian methods for demographic estimation https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2022/03/EAP174-Bayesian-Methods-for-Demographic-Estimation.pdf
- Blackwell. L., Elliott, D. and Bryant, J. (2022), Demographic Accounts Project Update https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fuksa.statisticsauthority.gov.uk%2Fwp-content%2Fuploads%2F2022%2F08%2FEAP179-Demographic-Accounts-MARP-Update.pptx&wdOrigin=BROWSELINK
- Elliott, D. and Blackwell, L. (2023), Progress on an integrated statistical design for the transformed population and social statistics system: Bayesian methods for demographic estimation https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2023/02/EAP193-DPM-progress-report-and-technical-paper.pdf
- Merad, S., Elliott, D. and Metcalfe, A. (2024), Evaluating the DPM estimation method using a simulation study https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2024/10/EAP209-Evaluation-of-TMB-using-a-simulation-study.pdf
- ONS (2020) Indicative uncertainty intervals for the admin-based population estimates: July 2020. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/indicativeuncertaintyintervalsfortheadminbasedpopulationestimatesjuly2020

7 Technical appendix

- Cell i is defined as a combination of age a by sex s by LA r by time t.
- y_i: SPD population estimate for cell i.
- y_i^{MYE}: MYE population for cell i (observed only at $t = 2011{,}2021$).
- β, τ, and ξ are parameters to be estimated in the model.
- We require the posterior distribution of ρ for the multiple draws approach in step 2.

The coverage ratio Model 1 is specified as

$$y_i \sim \text{Poisson}\left(\rho_i y_i^{\text{MYE}}\right)$$

$$\rho_i \sim \text{Gamma}(\xi^{-1}, (\xi\mu_i)^{-1})$$

$$\log\mu_i = \beta_i^{\text{Intercept}} + \beta_{a_i}^{\text{age}} + \beta_{r_i}^{\text{region}} + \beta_{s_i}^{\text{sex}} +$$

$$\beta_{(a,s)_i}^{\text{age:sex}} + \beta_{(a,r)_i}^{\text{age:region}} + \beta_{(s,r)_i}^{\text{sex:region}} +$$

$$\beta^{\text{time:region}}_{(t,r)_i} + \beta^{\text{time:age}}_{(t,a)_i} + \beta^{\text{time:sex}}_{(t,s)_i}$$

With priors:

$\beta^{\text{Intercept}} \sim \mathrm{N}(0,10^2)$

$\beta^{\text{age}}_{a} \sim \mathrm{N}\left(\beta^{\text{age}}_{a-1}, \tau^2_{\text{age}}\right), \quad a = 1, \dots 105$

$$\frac{1}{106}\sum_{a=0}^{a=105} \beta^{\text{age}}_{a} \sim \mathrm{N}(0,1)$$

$\tau^2_{\text{age}} \sim \mathrm{N}^{+}(0,1)$

$\beta^{\text{sex}}_{d} \sim \mathrm{N}(0,1), \quad s = \text{Female,Male}$

$\beta^{\text{region}}_{r} \sim \mathrm{N}\left(0, \tau^2_{\text{region}}\right), \quad r = LA_1, \dots, LA_{318}$

$\tau^2_{\text{region}} \sim \mathrm{N}^{+}(0,1)$

$\beta^{\text{age:sex}}_{a,s} \sim \mathrm{N}\left(\beta^{\text{age:sex}}_{a-1,s}, \tau^2_{\text{age:sex}}\right), \quad s = \text{Female,Male}, \quad a = 1, \dots 105$

$$\frac{1}{106}\sum_{a=0}^{a=105} \beta^{\text{age:sex}}_{a,s} \sim \mathrm{N}(0,1), \quad s = \text{Female,Male}$$

$\tau^2_{\text{age:sex}} \sim \mathrm{N}^{+}(0,1)$

$\beta^{\text{age:region}}_{a,r} \sim \mathrm{N}\left(\beta^{\text{age:region}}_{a-1,r}, \tau^2_{\text{age:region}}\right), \quad r = LA_1, \dots, LA_{318}, \quad a = 1, \dots 105$

$$\frac{1}{106}\sum_{a=0}^{a=105} \beta^{\text{age:region}}_{a,r} \sim \mathrm{N}(0,1), \quad r = LA_1, \dots, LA_{318}, \quad a = 1, \dots 105$$

$\tau^2_{\text{age:region}} \sim \mathrm{N}^{+}(0,1)$

$\beta^{\text{sex:region}}_{s,r} \sim \mathrm{N}\left(0, \tau^2_{\text{sex:region}}\right), \quad s = \text{Female,Male}, \quad r = LA_1, \dots, LA_{318}$

$\tau^2_{\text{sex:region}} \sim \mathrm{N}^{+}(0,1)$

$\beta^{\text{time:region}}_{t,r} \sim \mathrm{N}\left(\beta^{\text{time:region}}_{t-1,r}, \tau^2_{\text{time:region}}\right), \quad r = LA_1, \dots, LA_{318}, \quad t = 2012, \dots, 2023$

$$\frac{1}{13}\sum_{t=2011}^{t=2023} \beta^{\text{time:region}}_{t,r} \sim \mathrm{N}(0,1), \quad r = LA_1, \dots, LA_{318}$$

$\tau^2_{\text{time:region}} \sim \mathrm{N}^{+}(0,1)$

$$\beta_{t,a}^{\text{time:age}} \sim \mathrm{N}\left(\beta_{t-1,a}^{\text{time:age}}, \tau_{\text{time:age}}^2\right), \quad a = 0, \ldots, 105, \quad t = 2012, \ldots, 2023$$

$$\frac{1}{13} \sum_{t=2011}^{t=2023} \beta_{t,a}^{\text{time:age}} \sim \mathrm{N}(0,1), \quad a = 0, \ldots, 105$$

$$\tau_{\text{time:age}}^2 \sim \mathrm{N}^+(0,1)$$

$$\beta_{t,s}^{\text{time:sex}} \sim \mathrm{N}\left(\beta_{t-1,s}^{\text{time:sex}}, \tau_{\text{time:sex}}^2\right), \quad s = \text{Female,Male}, \quad t = 2012, \ldots, 2023$$

$$\frac{1}{13} \sum_{t=2011}^{t=2023} \beta_{t,s}^{\text{time:sex}} \sim \mathrm{N}(0,1), \quad s = \text{Female,Male}$$

$$\tau_{\text{time:sex}}^2 \sim \mathrm{N}^+(0,1)$$

$$p(\xi) = \frac{1}{2\sqrt{\xi}} e^{-\sqrt{\xi}}$$