# Bayesian Modeling of Demographic Data in Many Subpopulations*

GREGOR ZENS

*International Institute for Applied Systems Analysis (IIASA)*

*Wittgenstein Centre for Demography and Global Human Capital (WIC)*

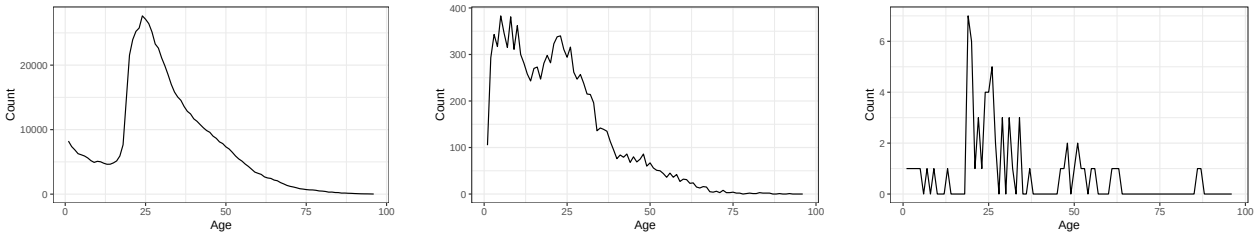
January 17, 2024

**Abstract**

Analyzing age-specific mortality, fertility, and migration in subpopulations is a crucial task in demography, with significant policy relevance. In practice, such analysis is challenging when studying numerous subpopulations, due to small sample sizes and demographic heterogeneity. To address this issue, a Bayesian model for the joint analysis of many, potentially small, demographic subgroups is proposed. The model combines three common assumptions about demographic processes in a unified probabilistic framework. The approach provides robust estimates of the demographic process in each subpopulation, allows testing for heterogeneity between subpopulations, and can be used to assess the impact of covariates on the demographic process. This makes the model suitable for probabilistic projection exercises and scenario analysis. An in-depth analysis of age-specific immigration flows to Austria, disaggregated by sex and 155 countries of origin, is used to illustrate the framework. Comparative analysis shows that the model outperforms commonly used benchmark frameworks in both in-sample imputation and out-of-sample prediction exercises.

**Keywords:** Multidimensional Demography, Multipopulation Model, Hierarchical Model, Probabilistic PCA, Functional Data Analysis.

*Correspondence to Gregor Zens, International Institute for Applied Systems Analysis. Schlossplatz 1, 2361 Laxenburg, Austria. E-mail: zens@iias.ac.at.




**(a)** Aggregate Counts. **(b)** Males from Syria. **(c)** Females from Guatemala.

**Figure 1:** Age-specific immigration flows to Austria, 2016-2020. Panel (a) shows the aggregate counts. Panel (b) shows counts of male immigrants from Syria, characterized by patterns that differ substantially from the aggregate. Panel (c) shows counts of female immigrants from Guatemala, with considerable sampling noise.

## 1 Introduction

In demography, analyzing data recorded by age and additional individual characteristics, such as sex and education, plays a prominent role in empirical research. Modeling such multidimensional data is crucial for informing policymakers and population projections, for exploring the drivers of heterogeneity among subpopulations, and for predicting the demographic characteristics of populations for which actual data are not available. The latter is a fundamental challenge, particularly in the context of developing countries.

In practice, modeling multidimensional demographic data can be difficult. This is particularly true when considering many, potentially small subpopulations. To illustrate this, consider the data shown in Fig. 1. Each of the three panels shows immigrant flows as a function of age. These migrant counts are based on register data from Austria and include all international immigrants who arrived in the country between 2016 and 2020.[1] Panel (a) shows the aggregate age-specific immigrant counts, revealing typical regularities of migration intensity across the life course. Panel (b) shows immigration flows by age for male immigrants from Syria. This subpopulation deviates from the aggregate pattern, with a much larger proportion of immigrants falling into the younger age groups. This reflects the distinct refugee dynamics in this subpopulation. Panel (c) shows immigrant counts by age for female immigrants from Guatemala. Only a small number of migrants fall into this subpopulation, resulting in significant sampling noise. This makes it difficult to make reliable statements about the underlying demographic process based on the raw data alone.

Model-based analysis of demographic phenomena in such multipopulation settings is challenging, precisely due to the interplay of small sample sizes and demographic heterogeneity. Models need to be flexible enough to accurately represent systematic heterogeneity between subpopulations. At the same time, models must be robust to avoid misinterpreting sampling noise in small subpopulations as systematic demographic patterns. Additionally, handling the complex interdependencies and correlations among subpopulations in high-dimensional settings demands parsimonious, yet comprehensive modeling approaches.

[1] More details on the data are provided in Sec. 4.

To address these multifaceted challenges, this article presents a Bayesian framework for demographic analysis of numerous, potentially small, subpopulations. The framework is based on three widely recognized assumptions about demographic processes, which are formalized and combined into a single hierarchical model. Its key applications include testing for heterogeneity across subpopulations, exploring systematic drivers of this heterogeneity, and producing predictions as well as probabilistic projections. These are illustrated through an in-depth analysis of data on international migration flows to Austria, disaggregated by age, sex, and country of origin, resulting in a total of 300 subpopulations. The results indicate pronounced heterogeneity in age patterns across subpopulations, with conflict in origin countries identified as one of the key modulating factors. Systematic simulation studies and real data exercises further demonstrate the relative advantage of the model over competing models for in-sample smoothing, imputation of partially missing data, and out-of-sample forecasting tasks.

The rest of this article is structured as follows. Sec. 2 summarises related literature in empirical demography and statistics. Sec. 3 discusses the details of the proposed framework. Sec. 4 applies the framework to data on international immigration to Austria. Sec. 5 provides insights into the comparative performance of the model based on simulated and real data. Sec. 6 concludes and provides directions for future research.

## 2 Related Literature: Three Key Assumptions

Due to the importance of empirical demographic analysis, numerous statistical models for demographic data have been developed in recent decades. Examples include Rogers et al. (1978), Lee and Carter (1992), Hyndman and Ullah (2007), Camarda (2012), Alexander et al. (2017), Durowaa-Boateng et al. (2023) and Dharamshi et al. (2023). Susmann et al. (2022) provide an overview and a classification of many of these frameworks. This section summarises three key ideas and underlying assumptions about demographic processes that are regularly put forward in this literature. These are later formalized into a single, unified statistical framework.

*Smoothing*

First, a common strategy is to assume that demographic processes are smooth in the age dimension. This translates into the assumption that a priori, the demographic characteristics and behaviors of 5-year-olds and 6-year-olds are expected to be relatively more similar than the characteristics of 5-year-olds and 60-year-olds. This assumption motivates methods that make direct use of tools such as spline smoothing, kernel smoothing, or functional data analysis. Such approaches have been extensively studied and applied in

demographic literature; see for example McNeil et al. (1977), Hyndman and Ullah (2007), Camarda (2012), Hyndman et al. (2013), or Pavone et al. (2022).

*Latent Commonalities*

Second, a frequent assumption is the existence of common underlying patterns shared by many or all subpopulations in the data. This idea arises naturally from the regularity of demographic processes in different contexts. Examples include a higher-than-average probability of dying in old age or a higher-than-average probability of migrating between the ages of 20 and 40. This concept is reflected in methods based on principal components analysis (PCA) and singular value decompositions (SVDs). These methods are highly popular in empirical demography, see for instance Shang et al. (2011), Alexander et al. (2017), Clark (2019) or the famous Lee-Carter framework (Lee and Carter, 1992; Wiśniowski et al., 2015).

Similarly, traditional demographic methods based on fitting pre-specified functions to demographic data can also be seen as exploiting the concept of common underlying patterns. For example, the Rogers-Castro migration model (Rogers et al., 1978) or the Heligman-Pollard mortality model (Heligman and Pollard, 1980) postulate that age-specific demographic patterns in different populations can typically be reconstructed by appropriately weighting and combining common underlying components. Finally, similar ideas about latent commonalities underlie parametric mixture models for demographic data (Mazzuco et al., 2018).

*Information Sharing*

A third key idea is based on the assumption that some notion of 'similarity' can be formulated across subpopulations and that 'similar' subgroups will be characterized by 'similar' demographic processes. In terms of the motivating application, this assumption might imply, for example, that male migrants from Kenya and Ethiopia are a priori expected to be characterized by a relatively more similar age pattern than male migrants from Kenya and female migrants from Germany. Different notions of similarity have been proposed, based on spatial proximity (Alexander et al., 2017), temporal proximity (Lee and Carter, 1992; Hyndman and Ullah, 2007), or based on externally available covariates (Clark, 2019). Statistically, the idea of sharing information across subpopulations based on similarity is most often formalized using hierarchical modeling frameworks (Susmann et al., 2022). Hierarchical models allow demographic processes to be reconstructed, even in subpopulations with sparse data, by using information from 'similar' subpopulations. In addition, hierarchical structures facilitate forecasting, imputation of missing data, and projection exercises.[2]

[2] See Sec. 4 for an illustration.

*A Unified Perspective*

While these three assumptions are ubiquitous in empirical demography, their application in isolation suffers from several shortcomings. For example, in noisy multipopulation data, applying age smoothing methods separately to each subpopulation can easily lead to overfitting or underfitting. Similarly, extracting latent commonalities using SVD or PCA in isolation has serious potential for error propagation when facing noisy data. That is, if the data contains small and noisy subpopulations, the extracted principal components will contain a certain amount of sampling noise. This noise is then multiplicatively reintroduced into the PCA-based predictions for each subpopulation. Finally, traditional parametric models often lack the flexibility needed to capture the demographic heterogeneity in granular multipopulation data.[3]

This paper proposes a modeling framework for multidimensional demographic analysis that superimposes all three assumptions jointly in a single probabilistic framework. The framework exploits the advantages of the three stated assumptions without the shortcomings of applying them in isolation. In essence, the model represents potentially noisy and heterogeneous multipopulation data using a small number of latent smooth components. The relative importance of these components for a given subpopulation is based on the similarity between subpopulations. Employing a Bayesian framework for the model further allows for valid probabilistic inference.

From a statistical perspective, the proposed framework is a Bayesian functional probabilistic PCA model for count data. The scores are modeled using a hierarchical regression structure, allowing for information sharing across subpopulations. The statistical approach thus integrates ideas from Bayesian factor models (Conti et al., 2014), Bayesian probabilistic PCA for count outcomes (Chiquet et al., 2018), Bayesian demographic modeling (Czado et al., 2005; Alexander et al., 2017; Raftery and Ševčíková, 2023), functional data analysis (Hyndman and Ullah, 2007; Montagna et al., 2012; Kowal and Bourgeois, 2020), and literature on Bayesian smoothing and shrinkage priors (Lang and Brezger, 2004; Piironen and Vehtari, 2017).

## 3 Statistical Framework

Let $y_{ix}$ denote a demographic count outcome observed at a discrete age $x$ in a subpopulation $i$.[4] The counts $y_{ix}$ are modeled as noisy observations from a Poisson distribution[5]

$$y_{ix} \sim \mathcal{P}(e^{z_{ix}}). \tag{3.1}$$

[3] See Sec. 5 for empirical illustrations of these phenomena in various modeling frameworks.
[4] For notational simplicity, a single subscript is used to refer to subpopulations, rather than a longer panel-type notation based on multiple subscripts.
[5] While the focus on a count data framework is motivated by its application in empirical demography, Eq. (3.1) can be easily replaced by alternative likelihood specifications, for example for Gaussian outcomes or for binary and binomial outcomes based on data augmentation techniques (Tanner and Wong, 1987; Polson et al., 2013; Zens et al., 2023).

The log-mean parameter $z_{ix}$ is a noisy realization of a Gaussian model

$$z_{ix} = \alpha_i + z_i(x) + O_{ix} + \varepsilon_{ix} \qquad \varepsilon_{ix} \sim \mathcal{N}(0, \sigma^2), \tag{3.2}$$

where $\alpha_i$ is a subpopulation-specific intercept, $z_i(x)$ is a smooth function of age and $\varepsilon_{ix}$ is a subpopulation and age-specific iid zero mean Gaussian noise term with variance $\sigma^2$. $O_{ix}$ is a known offset term, representing, for example, a log total population count for subpopulation $i$ and age $x$. The Poisson lognormal model (Aitchison and Ho, 1989) defined by Eq. (3.1) - Eq. (3.2) accounts for overdispersion and hence helps to adequately quantify estimation uncertainty. Related empirical approaches often directly impose Gaussian regression models on log-transformed counts or rates. This implicitly approximates a Poisson lognormal model, see Sec. A1 for details.

It is further assumed that the functions $z_i(x)$ can be written as a linear combination of $Q$ ($q = 1, \ldots, Q$) unknown, smooth zero-mean functions $\Phi_q(x)$, weighted by subpopulation specific loadings $\lambda_{iq}$ such that

$$z_i(x) = \sum_{q=1}^{Q} \Phi_q(x)\lambda_{iq}. \tag{3.3}$$

The latent functions $\Phi_q(x)$ represent underlying commonalities across subpopulations, reflecting the assumptions discussed in Sec. 2. The loadings $\lambda_{iq}$ determine how important a particular function $\Phi_q(x)$ is in describing the observed patterns within a given subpopulation $i$. Assuming that $Q << N$ provides an adequate description of the data, this approach is related to probabilistic principal component models for counts (Chiquet et al., 2018), functional principal component frameworks (Ramsay and Silverman, 2005), functional regression models (Kowal and Bourgeois, 2020), and functional factor models (Montagna et al., 2012). In related demographic work, the functions $\Phi_q(x)$ are often implicitly approximated via principle components of the log counts or log rates (Alexander et al., 2017; Clark, 2019; Dharamshi et al., 2023). In this paper, we treat them as parameters that are surrounded by uncertainty and that need to be estimated along with the rest of the parameters. This approach has several advantages, including an appropriate quantification of parameter uncertainty, straightforward estimation from partially incomplete data, and a reduced likelihood of capturing noise in the factors, due to the presence of an error term in the model. To facilitate estimation, we assume that the unknown functions $\Phi(x)$ can be represented as a linear combination of $K$ B-spline basis functions $B_{qk}$

$$\Phi_q(x) = \sum_{k=1}^{K} f_{qk} B_{qk}(x) \tag{3.4}$$

where $f_{qk}$ is the spline base coefficient of knot $k$ and function $q$.

A Bayesian approach to estimation is pursued to allow for probabilistic inference. The Bayesian paradigm requires the choice of appropriate prior distributions for all model parameters. In this article, the prior distributions are chosen in an informative way that reflects the assumptions on information sharing and smoothing outlined in Sec. 2.

### 3.1 Information Sharing

First, it is assumed that 'similar' subpopulations are a priori characterized by 'similar' demographic processes. Formally, this is implemented by assuming a hierarchical structure for the loadings $\lambda_{iq}$ and intercepts $\alpha_i$ which determine the shape and level of the demographic process, respectively. In demographic trend analysis, such hierarchical structures often include time series frameworks (Lee and Carter, 1992; Hyndman and Ullah, 2007; Susmann et al., 2022). In subnational mortality modeling, Alexander et al. (2017) allow for similarity based on spatial proximity of subnational units. Dharamshi et al. (2023) present a more complex approach, based on the geographic nesting structure of subnational units. In this article, we assume that the loadings $\lambda_{iq}$ and intercepts $\alpha_i$ are a function of observed subpopulation-specific covariates:

$$\begin{aligned} \alpha_i &\sim \mathcal{N}(\boldsymbol{w}_i'\boldsymbol{\delta}, \sigma_a^2) \\ \lambda_{iq} &\sim \mathcal{N}(\boldsymbol{w}_i'\boldsymbol{\beta}_q, \sigma_{\lambda,q}^2). \end{aligned} \tag{3.5}$$

where $\boldsymbol{w}_i$ is a $R \times 1$ vector of observed covariates describing subpopulation $i$, $\boldsymbol{\beta}_q$ and $\boldsymbol{\delta}$ are $R \times 1$ vectors of linear regression coefficients, and $\sigma_{\lambda,q}^2$ and $\sigma_a^2$ are variance terms.[6] This specification is related to random effects and multilevel models and allows for heterogeneity around the a priori conditional mean of the loadings $\lambda_{iq}$ and intercepts $\alpha_i$. As will be discussed in Sec. 4.3, covariate effects on the shape and level of demographic processes can be derived from Eq. (3.5) to explore drivers of heterogeneity across subpopulations. In Sec. 4.4, it is demonstrated that Eq. (3.5) is useful for probabilistic projection exercises.

The hierarchical information sharing mechanism Eq. (3.5) is key when working with heterogeneous and sparse multipopulation data. It allows subpopulations with informative data to contribute to demographic process estimates in 'similar' subpopulations, where the data may be too sparse to produce reliable estimates based on the 'local' subpopulation likelihood alone. If the data in a subpopulation are informative, then estimates of $\lambda_{iq}$ will be based to a large extent on the data in that particular subpopulation. The more noisy the data in a subpopulation, the more $\lambda_{iq}$ will be influenced by information from 'similar' subpopulations.

[6] Extending the model to include, e.g., a temporal smoothing component is straightforward. This is not considered in this paper, partially due to the immense challenges of modeling temporal patterns in migration data (Bijak et al., 2019).

In related demographic work, hierarchical modeling set-ups such as Eq. (3.5) are often implicitly approximated using a two-step procedure, where $\lambda_{iq}$ are extracted using an SVD or PCA in a first step and then used as input to time series or regression models in a second step. In this paper, a joint estimation approach is adopted instead. Importantly, this enables leveraging both the 'local' subpopulation information and the information obtained from 'similar' subpopulations during model estimation. Combining both sources of information stabilizes estimates in noisy subpopulations. In addition, the joint hierarchical estimation approach allows for valid uncertainty quantification.

The priors on $\boldsymbol{\delta}$ and $\boldsymbol{\beta}_q$ are specified as zero-centred horseshoe priors (Carvalho et al., 2010). This informative prior has most of its mass at zero and only allows the coefficients in $\boldsymbol{\delta}$ and $\boldsymbol{\beta}_q$ to deviate from zero if the data are sufficiently informative. This provides a variable selection and regularisation mechanism, effectively reducing concerns about overfitting in contexts with noisy data. As will be illustrated in Sec. 4.3, regularising $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ towards zero implies that, a priori, the level of the subpopulation-specific process (governed by $\boldsymbol{\delta}$) and the shape of the subpopulation-specific process (governed by $\boldsymbol{\beta}_q$) are only affected by covariates where the data are clearly informative about these effects. To complete the prior setup, weakly informative $\mathcal{IG}(c_0, d_0)$ priors are chosen on the remaining variance parameters, with $c_0$ and $d_0$ set to small constants, as is standard in the literature.

### 3.2 Smoothing Mechanism

Finally, it is assumed that the underlying latent functions $\Phi_q(x)$ are a priori smooth, in the sense that small changes in $x$ imply only small changes in $\Phi_q(x)$. Formally, a smoothing prior on the spline coefficients $f_{qk}$ is used to introduce this assumption. In particular, we follow Lang and Brezger (2004) and work under a stochastic difference penalty in the form of a second-order random walk on $f_{qk}$, such that, a priori

$$f_{qk} = 2f_{q,k-1} - f_{q,k-2} + u_{qk} \tag{3.6}$$

where $u_{qk}$ is a noise term. Diffuse priors are specified for the initial values $f_{q,1}$ and $f_{q,2}$ for each $q$. To allow for locally adaptive smoothing behavior, we let

$$u_{qk} \sim \mathcal{N}\left(0, \frac{\tau_q}{\kappa_{qk}}\right) \tag{3.7}$$

with $\kappa_{qk} \sim \mathcal{G}(\frac{1}{2}, \frac{1}{2})$. Marginally, this implies a Cauchy distribution for $u_{qk}$. In essence, the prior specification Eq. (3.6) - Eq. (3.7) strongly regularises 'neighboring' spline coefficients to be similar, resulting in smooth estimates of the underlying components $\Phi_q(x)$. In addition, this prior renders the analysis less sensitive to knot placement (Lang and Brezger, 2004).

### 3.3 Summary

In summary, the main mechanisms of the proposed model can be conceptualized in the following way. The $Q$ latent functions $\Phi_q(x)$ summarise information on age-specific patterns that are common to many (or all) subpopulations. The functions $\Phi_q(x)$ are further penalized to ensure smoothness in the age dimension. Given $\Phi_q(x)$, the conditional mean of $z_{ix}$ is fully determined by the intercepts $\alpha_i$ and the loadings $\lambda_{iq}$. For $\lambda_{iq}$ and $\alpha_i$, hierarchical structures are used to share information between 'similar' subpopulations, using the observed covariates $\boldsymbol{w}_i$ to measure similarity. The modeling approach therefore reflects all three key assumptions discussed in Sec. 2.

The estimate of the demographic process in a subpopulation is then based on combining the information about $\alpha_i$ and $\lambda_{iq}$ contained in the 'local' subpopulation likelihood with information on those parameters obtained 'globally' from 'similar' subpopulations. The informativeness of the data determines the extent to which the 'local' subpopulation information influences the final estimates relative to the 'global' information from other subpopulations. In subpopulations with informative data, the effect of information-sharing may be overwritten, whereas, in subpopulations with noisy data, the information-sharing effect will be more pronounced. A discussion of model identification and parameter estimation using tuning-free and fully automatic Markov chain Monte Carlo methods is given in Sec. A1.

## 4 Case Study: International Migration to Austria, 2016-2020

Knowledge of migration processes is key to understanding future population change. In addition to its direct impact on population growth, migration has indirect effects by influencing fertility and mortality patterns in the host population. However, data on human mobility are often imperfectly measured or unavailable (Raymer et al., 2013). In general, modeling migration patterns is a highly challenging empirical task (Bijak et al., 2019).

In this context, register data on international immigration flows to Austria are analyzed to illustrate the merits of the proposed modeling framework. The data are obtained from the Austrian national statistical office (*Statistik Austria*) and consist of counts of incoming international migrants, summed over the period 2016-2020, and reported by individual years of age ($x = 0, \ldots, 95$), sex and country of origin of the immigrants.[7] After excluding subpopulations with fewer than 10 observed immigrants, the data includes immigrants from 155 countries of origin and a total of 300 subpopulations.

As covariates $\boldsymbol{w}_i$, 15 binary indicators for the region of origin are included. These indicators are based on the UN sub-regional classification of countries. Additional indicator variables capture contiguity to Austria

[7] Immigrant status is defined as being officially registered in Austria for more than 90 days.

and EU membership of the country of origin, as well as the sex of migrants in a given subpopulation. Moreover, gross national income per capita, the number of battle deaths, population size, and working-age population share are included, each with respect to the origin country. Finally, the covariates include the number of migrants from a given country residing in Austria as of 2016, and the distance from that country's capital city to Vienna.[8] These variables enter the model after a logarithmic transformation, except for the share of the working-age population. All continuous variables enter the model with an additional quadratic term to allow for non-linear effects. This set of observables is in the spirit of classical gravity models of migration (Beine et al., 2016). The covariates cover many important theoretical channels driving international migration, such as economic opportunity, social networks, and fleeing violence (Czaika and Reinprecht, 2022). In addition, this set of variables allows for a rich information-sharing structure that facilitates borrowing information across different subpopulations based on regional similarities as well as similarities in living standards, conditions in the country of origin, and network ties to Austria.

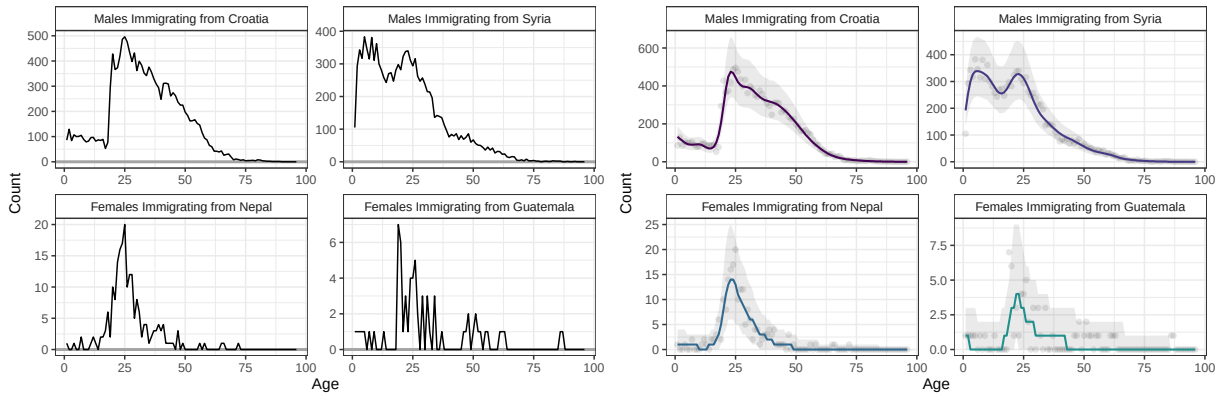
Posterior inference is based on 50,000 posterior samples stored after an initial burn-in period of 25,000 iterations. Every 5th draw is saved to thin the posterior chain. In the absence of a suitable offset term, we set $O_{ix} = 0$ for all $i$ and $x$. Cubic splines are used as base functions, with knots placed every five years from age 6 to 66, and boundary knots at ages 0 and 95. This reflects less dynamics in migration intensity at later ages and results in a reasonable balance between model flexibility and smoothness of the estimated demographic functions.[9] To determine an appropriate number of latent functions $Q$, model runs varying Q in $\{1, \dots, 10\}$ were considered. Several model selection criteria based on in-sample likelihood evaluations (such as BICs and AICs) were computed. In addition, a five-fold cross-validation exercise was performed and several predictive scores (such as RMSEs and MAEs) were compared.[10] There is no consensus between all the scores and criteria.[11] Overall, however, $Q = 6$ seems to be a reasonable choice for the data set at hand. This is slightly higher than the number of factors in related papers on mortality modeling, such as Clark (2019) and Alexander et al. (2017), potentially reflecting a higher degree of heterogeneity in migration data compared to mortality data. Results are largely similar between $Q = 5$ and $Q = 10$. The $Q = 6$ extracted latent functions with uncertainty measures are shown in Fig. A1.

[8] These covariates are sourced from the *CEPII gravity database*, *Statistik Austria*, the *World Development Indicators* (WDI) of the *World Bank*, and the *United Nations Development Programme* (UNDP).

[9] A future avenue of research likely to lead to further improvements in model performance is to investigate an 'optimal' placement of nodes in different demographic modeling contexts.

[10] Details are provided in Sec. 5.3.

[11] Determining an 'optimal' number of latent factors $Q$ in factor modeling or probabilistic PCA is a challenging and unresolved statistical problem. There are no universally valid model selection criteria, and existing criteria are not directly applicable to the model outlined in this paper. Furthermore, because the hold-out data are potentially very noisy realizations of the process of interest, even model selection criteria based on out-of-sample predictive power (such as cross-validation scores) run an increased risk of overfitting.




**(a)** Raw age-specific counts. **(b)** Model fit.

**Figure 2:** Raw (a) and estimated (b) age-specific counts. Shaded areas correspond to 95% credible intervals. Points correspond to observed data. Upper left: Males migrating from Croatia. Upper right: Males immigrating from Syria. Bottom left: Females immigrating from Nepal. Bottom right: Females immigrating from Guatemala. Note that the scales of the $y$-axes differ between panels.

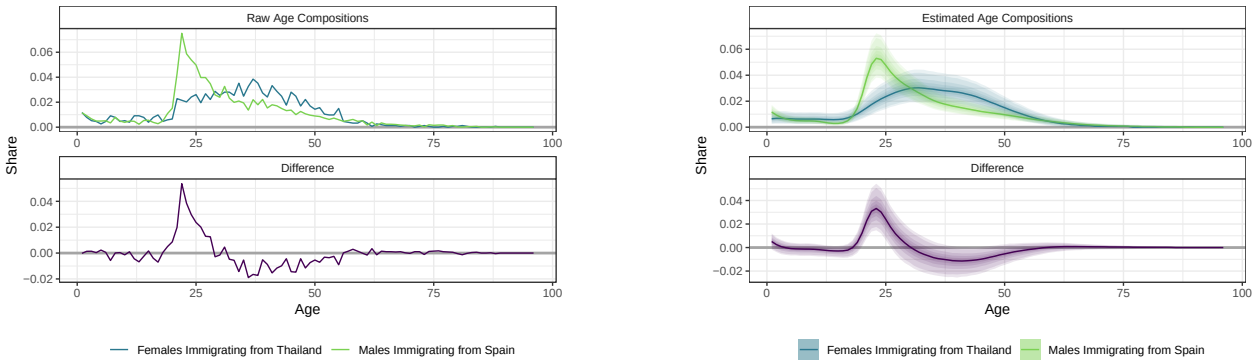
### 4.1 Illustration of Model Fit

Fig. 2 provides a visual representation of the raw data and model-based estimates of age-specific counts in four example subpopulations. These subpopulations include male migrants from Croatia (14,359 observed migrants in total), male migrants from Syria (11,642), female migrants from Nepal (204), and female migrants from Guatemala (68). Panel (a) shows the raw data, illustrating some of the heterogeneity, sparsity, and sampling noise present in the age-specific immigration counts.

In panel (b), the raw counts are overlaid with the estimation results, including probabilistic uncertainty bounds. The estimated mean of the predictive distribution of $y_{ix}$ is shown as a smoothed model fit. In many applications, such model-based smooth estimates of the underlying demographic process are already the final output of interest. For instance, model-based estimates allow tracking of demographic patterns and trends in many subpopulations jointly or can serve as inputs to demographic projection models. By visual inspection of the results in Fig. 2, the model appears flexible enough to capture heterogeneity while remaining robust to the high level of noise in the input data. This will be confirmed more systematically using simulation studies in Sec. 5.

### 4.2 Assessing the Presence of Heterogeneity: Homogeneity Testing

In many cases, researchers may be interested in exploring and making formal statements about homogeneity or heterogeneity of subpopulations. Such an exercise may serve as a purely exploratory tool or may be motivated by a desire to test hypotheses derived from theoretical models. The results of homogeneity tests may, in turn, inform projection models and initiate the development of new theoretical insights. However,




**(a)** Raw age compositions.

**(b)** Estimated age compositions.

**Figure 3:** Raw (a) and estimated (b) age compositions of females immigrating from Thailand (blue), males immigrating from Australia (green) and the difference between the two age compositions (purple). Shaded areas correspond to 95% Bayesian credible intervals.

testing for heterogeneity in multipopulation settings can be a challenging exercise, as it may not be easy to distinguish between systematic demographic patterns and sampling noise when comparing raw data in a given set of subpopulations.

To give an example, consider the problem of estimating whether there is a significant difference in the age composition of two migrant subpopulations.[12] More formally, suppose the goal is to make inferential statements about $(y_{jx}/\sum_x y_{jx}) - (y_{kx}/\sum_x y_{kx})$ for some $j$ and $k$ with $j \neq k$. To illustrate, consider Fig. 3, which compares the age composition of male Spanish immigrants and female Thai immigrants. The age compositions computed from the raw counts $y_{ix}$ (top) as well as the difference between the two subpopulations (bottom) are shown in panel (a). It appears that the proportion of male Spanish migrants in their early 20s is larger than that of female Thai migrants. However, it is difficult to draw statistically valid conclusions from the raw age compositions alone. This is because it is inherently unclear whether the visually observed differences are due to systematic variation or sampling noise. The Bayesian approach is highly useful in this context, as uncertainty bounds for functions of $y_{ix}$ (such as age compositions) can be easily obtained via Monte Carlo simulation.

Panel (b) shows the smoothed estimated age compositions (top) and the difference between the smoothed curves (bottom), including 95% credible intervals. From this model-based analysis, it becomes clear that the difference observed in the raw data is likely to be a systematic phenomenon rather than an artifact of sampling noise. In addition, there appears to be a slightly higher proportion of migrants aged 35-50 in

[12]The age *composition* (as opposed to the raw counts) is a natural object of interest when comparing the characteristics of migrants between subpopulations in settings where the respective levels of the counts are different. In addition, age compositions are key quantities in some demographic frameworks, e.g., for ex-post distribution of known migrant count aggregates in the absence of observed age-specific data.

the Thai female subpopulation. However, this difference is not estimated with fully conclusive precision, as indicated by the size of the uncertainty bands.

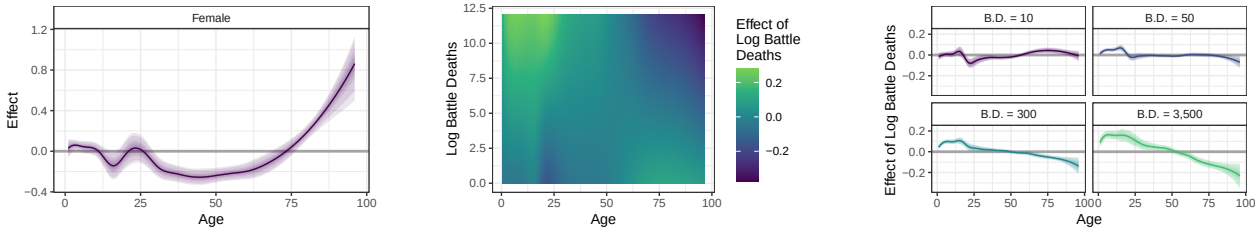
### 4.3 Exploring Drivers of Heterogeneity: Linear and Nonlinear Covariate Effects

In addition to testing for the presence of heterogeneity, there is usually considerable interest in exploring potential drivers of heterogeneity. Eq. (3.5) implies that the model can be used to make statements about the partial effects of specific covariates in $\boldsymbol{w}_i$ on the shape and level of outcomes via their effects on $\lambda_{iq}$ and $\alpha_i$. Since the functions $\Phi_q(x)$ are constrained to have a mean of zero, the level effects are entirely determined by $\boldsymbol{\delta}$. Estimates of these level effects are given in Fig. A2 in the appendix. The indicator for EU membership of the sending country and the linear term of the pre-existing migrant population from a sending country have the clearest positive effects on the level of the immigration process. These reflect the free movement of EU citizens within the EU and the importance of network effects in international migration, in line with theoretical expectations.

To assess the partial effect of covariates on the *shape* of the immigration process, the model can be rewritten as a random effects function-on-scalar regression model, as discussed in Kowal and Bourgeois (2020). The partial effect of a covariate $w_{ij}$ on $z_{ix}$ is then given by the quantity $\sum_q \Phi_q(x)\beta_{qj}$, where $\beta_{qj}$ corresponds to the coefficient of the covariate $j$ in the $q$-th regression equation of the loadings $\lambda_{iq}$. Again, the Bayesian approach allows to easily construct uncertainty bounds for this quantity based on Monte Carlo simulation.

Examples of such shape effect estimates are given in Fig. 4. Panel (a) shows the posterior distribution of the shape effect of the binary indicator for female subpopulations. These results imply that, on average and holding other factors constant, there are fewer female immigrants between the ages of 30 and 60 and more female immigrants at older ages relative to males. A plot of all estimated covariate effects on the shape of the immigration process can be found in Fig. A3 in the appendix. The regularising effect of the horseshoe prior on $\boldsymbol{\beta}$ is visible, as the shape effect of some covariates - e.g., the regional intercepts for Australia/New Zealand and Northern America - are estimated to be very close to zero for all ages. This implies that, on average, age-specific immigration patterns from these regions are not significantly different from the respective baseline estimates. For completeness, Fig. A4 additionally provides the posterior mean estimates of all coefficients $\boldsymbol{\beta_q}$.

For variables that enter the equation of $\lambda_{iq}$ with a linear and a quadratic term, the partial effect of a covariate $w_{ij}$ on $z_{ix}$ is given by $\sum_q \Phi_q(x)\beta_{qj} + 2w_{ij}\sum_q \Phi_q(x)\beta_{qn}$ where $\beta_{qj}$ is the coefficient of the linear




**(a)** **(b)** **(c)**

**Figure 4:** Examples of linear and non-linear covariate effects on the shape of the immigration process. Panel (a) shows the shape effect of the binary indicator for female subpopulations. Panel (b) shows a non-linear effect surface with point estimates of the impact of battle deaths in the country of origin on immigration flows. Panel (c) shows the impact of battle deaths in the country of origin on immigration flows for selected values of battle deaths. The shaded areas correspond to 95% credible intervals.

term $w_{ij}$ in the $q$-th equation and $\beta_{qn}$ is the coefficient of the quadratic term $w_{ij}^2$ in the $q$-th equation.[13] Such non-linear effect estimates can be conveniently summarised in an effect surface. Panel (b) in Fig. 4 shows the effect of (log) battle deaths in the country of origin on the shape of the immigration curve (color) as a function of age ($x$-axis) and the level of battle deaths ($y$-axis). Finally, panel (c) of Fig. 4 shows selected shape effects, evaluated at 10, 50, 300, and 3,500 battle deaths, roughly the 20th, 40th, 60th, and 80th quantiles of the log battle death distribution. The overall pattern is that as battle deaths in the origin country increase, immigration flows from this country to Austria are modulated, on average, towards more younger and fewer older immigrants. This shape effect becomes stronger with increasing levels of battle deaths. The estimates therefore imply a pronounced increase in emigration rates of children and teenagers following conflict events. This reflects the empirical pattern of many young immigrants moving to Austria as refugees from conflict-affected countries.

### 4.4 Probabilistic Projections & Scenarios

The ability to model the effect of covariates on both the level and shape of the immigration process also lends itself to probabilistic projections, a common task in empirical demography. Such exercises aim to project demographic patterns into the future based on given scenario paths for $\boldsymbol{w}_i$. A small projection exercise is implemented to illustrate the capabilities of the model in this context. Suppose one is interested in estimating how immigration flows of male Syrians to Austria might be affected if violent conflict in Syria were to end. Model-based answers to this question can be explored by comparing the baseline fit of the model with a prediction for a given subpopulation, where the observed covariate vector is swapped with a covariate vector corresponding to the scenario narrative.

[13] While the use of quadratic terms to model non-linear effects is a simplistic approach, extensions to more complex non-linear modeling techniques such as spline expansions of covariates can be analyzed on similar grounds.

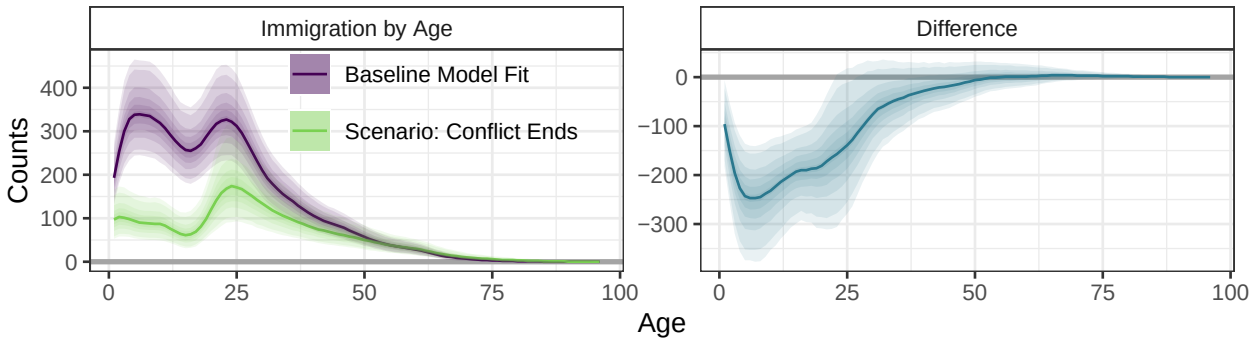



**Figure 5:** Scenario exercise. In-sample baseline fit for the Syrian male subpopulation (purple) versus model predictions for the Syrian male subpopulation with battle death indicators set to zero (green), as well as their difference (blue). Shaded areas correspond to 95% credible intervals.

The results of this exercise are given in Fig. 5. The figure shows in purple the baseline estimate for the subpopulation of males immigrating from Syria to Austria. In green, the figure shows the predictions of the model when the combat-related indicators in the covariate vector of the Syrian male subpopulation are set to zero.[14] The difference between the two scenarios is shown in blue. The estimated difference and associated uncertainty bounds indicate that the number of migrants under the age of 20 is expected to decrease significantly in the end-of-conflict scenario relative to the baseline. This reflects, by construction, the estimates on the relationship of conflict and age-specific migration discussed in Sec. 4.3.

## 5 Predictive Ability & Comparison to Alternative Approaches

To assess the comparative performance of the proposed modeling framework and to highlight its predictive power, this section provides several illustrative comparisons with related modeling frameworks. First, selected in-sample estimates are discussed to outline some general insights. Second, further insights into the in-sample and out-of-sample predictive performance of different modeling frameworks are provided through systematic simulation studies. Finally, a cross-validation exercise is carried out using the Austrian immigration data.

### 5.1 Insights on In-Sample Performance Using Selected Examples

To provide some initial insights, three common alternative modeling frameworks are compared to the Bayesian model. For each model, the goal is to provide subpopulation-specific estimates of age-specific migration, based on the Austrian immigration data. First, an SVD-based reconstruction is computed, based on the first six singular vectors of the $\log(1+x)$ transformed counts. Second, results from a local smoother based on a Poisson

[14] It should be noted that this approach to a scenario capturing the end of a large-scale conflict is rather simplistic, and does not take into account potential relocation effects on migrants already residing in Austria, effects on per capita income, etc. If desired, this type of scenario and projection exercise can be made arbitrarily complex, given suitable scenario paths for the covariates.

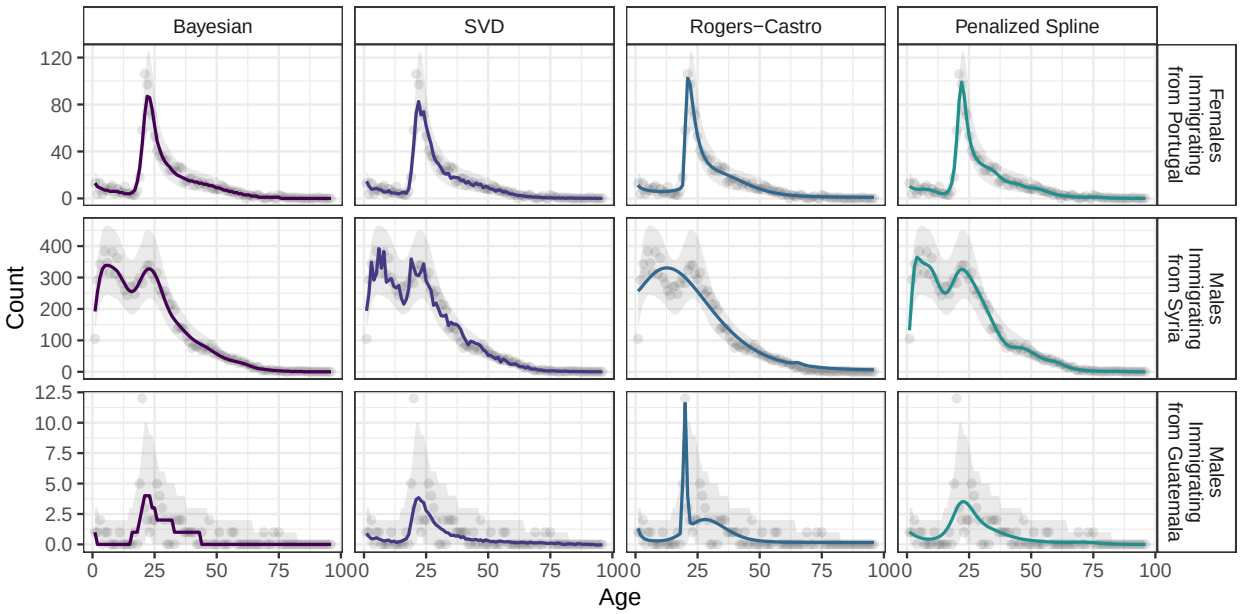



**Figure 6:** Selected in-sample fits obtained by four different modeling frameworks (columns) for three different subpopulations (rows). Shaded areas correspond to 95% credible intervals of the Bayesian model. Points correspond to observed data. Refer to the text for details.

penalized spline model, as in Camarda (2012), are obtained. Third, the Rogers-Castro model (Rogers et al., 1978), a traditional demographic method for reconstructing migration schedules by age, is applied to the data.[15]

Results for four selected subpopulations are shown in Fig. 6. The top row shows the fit for a well-behaved subpopulation in an informative setting with relatively large counts. Roughly similar results are obtained from all four models. The second row shows the fit for a subpopulation with a large sample size but a relatively 'irregular' pattern. Here, the Rogers-Castro model is not flexible enough to account for the heterogeneity of the pattern and underfits. This happens because the parametric function underlying the Rogers-Castro model is based on regularities in *aggregate* migration data. In comparison, the SVD model, spline smoothing, and the Bayesian model give similar results, with the SVD model capturing significantly more sampling noise. The last row shows the fit for a subpopulation that is relatively regular in shape, but with a small sample size. The resulting fits show that the SVD and Bayesian approaches, as well as the spline model, can still recover a reasonable age pattern despite the extremely noisy data. In contrast, the Rogers-Castro model provides a degenerate fit in this setting, clearly overfitting the noisy data.

[15]The penalised spline model is fitted using the function `Mort1Dsmooth` in the defunct R package `MortalitySmooth`. In the absence of an available procedure to fit the Rogers-Castro model directly to count data, the Rogers-Castro model is fitted by first transforming the count data to an age composition, then fitting the compositional data with a Gaussian likelihood model using the R package `rcbayes`, and then transforming back to the count scale based on median modeled estimates of the age composition.

These preliminary investigations broadly illustrate some of the relationships and comparative advantages of the Bayesian approach relative to the alternative modeling frameworks. The Bayesian approach combines the idea of underlying common components from the SVD and Rogers-Castro model, as well as the idea of smoothing from the penalized spline framework. This allows the Bayesian model to overcome the inflexibility of the Rogers Castro framework, while remaining more robust to noise than the SVD framework. Compared to local spline smoothing methods, the Bayesian approach allows for information sharing, further improving the robustness of the model fit and additionally providing a mechanism for projections and predictions.

### 5.2 Systematic Simulation Studies Based On Synthetic Data

To corroborate the selected results from the previous subsection, a systematic simulation study is carried out to evaluate the performance of the proposed model. The purpose of this simulation study is to identify scenarios in which the additional complexity of the Bayesian model pays off relative to simpler modeling frameworks. Three different exercises are considered. First, the accuracy of *in-sample* estimates is assessed, providing insight into the overall quality of the smoothed model fits. Second, prediction quality when *imputing missing data* is assessed to explore the utility of the model in scenarios with partially incomplete data. Third, *out-of-sample predictions* are evaluated in settings where demographic data are missing completely for some subpopulations. In all exercises, the aim is to estimate a known systematic signal $\alpha_i + z_i(x)$ from the noisy observations $y_{ix}$.

The synthetic data sets are simulated as follows. First, the five leading principal components are extracted using a singular value decomposition of the $log(1 + x)$ transformed Austrian immigration counts. These principal components are then smoothed using a B-spline basis expansion with 7 equally spaced interior nodes. These represent the ground truth of the underlying functions $\Phi_q(x)$. Ten covariates and all elements of $\boldsymbol{\beta}_q$ are simulated from $\mathcal{N}(0, 1)$. The intercept term in $\boldsymbol{\delta}$ is set to 15 and all other elements of $\boldsymbol{\delta}$ are simulated from $\mathcal{N}(0, 1)$.[16] Three elements in $\boldsymbol{\delta}$ and three elements in $\boldsymbol{\beta}$ are set to zero to simulate a scenario where some covariates are uninformative predictors. All $\sigma_{\lambda,q}$ are set to one and $\sigma_a^2 = 0.5$. We consider a low noise scenario with $\sigma^2 = 0.1$ and a high noise scenario with $\sigma^2 = 1$. For the partially missing data exercise, one observation per series is randomly selected and dropped from the training data. Each simulation is repeated 25 times. Root mean square errors, mean absolute errors, and mean percentage errors of all competing models relative to the ground truth are recorded, averaged over the 25 replications.

[16] Note that large positive counts $y_{ix}$ result from the intercept value of 15 in $\boldsymbol{\delta}$. This makes $z_{ix}$ essentially equivalent to $\log(y_{ix})$ (see Sec. A1). This is done to minimize the effect of the common $log(1 + x)$ transformation of count data prior to modeling the transformed counts. Therefore, the results of the simulation study can be seen as an upper bound on the performance of competing models that rely on log transformation of count data, free from the detrimental effects of log transformation of small counts.

As competing models for the in-sample simulation, we consider a reconstruction based on the first five principal components of the log counts using SVD, a penalized spline smoothing procedure on the log counts that minimizes the generalized cross-validation criterion[17] and a procedure that first smooths the log counts using penalized splines and then extracts the first five principal components from the smoothed series. For imputing partially missing data, we consider linear interpolation of missing log counts and penalized spline interpolation of the missing log counts as competing frameworks. Finally, for the out-of-sample prediction exercise, we consider as competitors two models that learn linear regression functions in the scores $\lambda_{iq}$ and intercepts $\alpha_i$. Both $\lambda_{iq}$ and $\alpha_i$ are estimated from an SVD of the log counts, once with and once without prior smoothing of the log counts.[18]

The results of these three systematic simulation exercises can be summarised as follows. First, the Bayesian model improves over all competing models in all considered settings (see Fig. A5). This implies that when the three assumptions in Sec. 2 are assumed to hold, incorporating them jointly leads to clear performance gains relative to simpler methods that leverage these assumptions only partially and rely on approximate estimation algorithms. Second, the performance of the various methods is not very different when noise is low, but diverges when noise is high (see Fig. A6 for an example). This suggests that local smoothers and SVD-based methods are useful modeling tools in settings where subpopulations are characterized by large counts and where sampling noise is not an issue. However, in the presence of subpopulations with small sample sizes and noise, the Bayesian approach significantly improves over alternative methods. Third, when imputing missing data, the Bayesian model outperforms basic interpolation rules already in the simplest settings, where noise is low and only single data points are missing. These gains are likely to be even more pronounced in noisy data or when larger sets of data points are missing.

### 5.3 Cross-Validation Exercise Based On Austrian Immigration Data

In addition to the systematic exercises using simulated data, a real data out-of-sample prediction exercise is carried out using the Austrian immigration data. The purpose of this is to gain further insight into the predictive power of the framework and to select an appropriate value of $Q$.[19]

A five-fold cross-validation approach is implemented where, in each of the five runs, the total of 300 observed subpopulations are divided into 240 (80%) training subpopulations and 60 (20%) hold-out subpopulations on which the model predictions are evaluated. Each subpopulation was therefore part of a hold-out

[17] The Poisson penalized spline model is fitted using the function `smooth.Pspline` in the R package `pspline`.

[18] The Rogers-Castro model and the Poisson count smoothing framework of Camarda (2012) are not explicitly considered, as both are difficult to fit in sparse and noisy multipopulation environments, leading to repeated software failures, degenerate fits, and uncompetitive predictive performance.

[19] It is important to note that the hold-out sample is inherently noisy. This encourages a degree of out-of-sample overfitting compared to simulation exercises where the ground truth signal is known. Therefore, cross-validation results should be taken with a grain of salt, especially for model selection purposes.

sample after the five runs. The Bayesian model is estimated with $Q$ varying in $\{1, \ldots, 10\}$ in each of the five runs. Predictions from conceptually simpler two-stage approximation approaches are also collected. The first stage is an SVD on the $\log(1 + x)$-transformed counts, once with and once without prior smoothing of the log counts. In the second stage, linear regression models with SVD-based approximations of $\lambda_{iq}$ and $\alpha_i$ as outcomes are estimated using ordinary least squares. Model predictions are then obtained by predicting $\lambda_{iq}$ and $\alpha_i$ out-of-sample and combining these predictions with the appropriately scaled singular vectors. All models assume $\log(1 + y_{ix})$ as the outcome of interest, in favor of SVD-based methods that rely on a transformed count outcome.

To summarise the results of this exercise, we find that SVD with and without prior smoothing of the counts perform relatively similarly in terms of mean absolute error and root mean square error. The Bayesian model outperforms the SVD approaches for all values of $Q$ considered. A visual summary of the results is given in Fig. A7.

## 6 Concluding Remarks

This paper presents a Bayesian framework for the joint modeling of many, potentially small, demographic subpopulations. The framework combines three common assumptions from empirical demographic literature into a unified probabilistic framework. The model is based on the idea that latent and smooth underlying common patterns can be extracted from the data. These patterns are then recombined in a way that allows for information sharing between similar subpopulations. Importantly, the model is flexible enough to describe systematic patterns of heterogeneity while remaining robust to noise through regularisation mechanisms.

The method is applied to Austrian register data on age-specific immigrant flows in 300 subpopulations, based on sex and country of origin. This case study is used to demonstrate the main applications of the model, including obtaining smoothed estimates of the demographic process, making probabilistic statements about the presence of heterogeneity, exploring systematic drivers of heterogeneity, and forecasting exercises, including probabilistic projections. Further illustrative and systematic exercises are used to compare the modeling framework with various competing models. The generally good performance of the Bayesian model is highlighted.

The results of this study carry two key messages for applied research in empirical demography. First, this study confirms the considerable efficacy of dimensionality reduction techniques, such as PCA or SVD, in quantitative demographic research. Given the structure of demographic data and the great success of frameworks such as the Lee-Carter model, this is somewhat self-evident. However, it is worth emphasizing that of the three model assumptions described, the assumption of latent underlying common patterns

across multiple populations appears to be the most important for providing reasonable model estimates and predictions, based on the case study and systematic explorations of model performance. A combination of log-transformed data and SVD-based modeling is expected to work well in settings with large counts and relatively little noise (e.g. age-specific mortality at the country level). Combined with a hierarchical structure exploiting temporal, spatial, or covariate-based proximity of subpopulations, powerful demographic modeling tools can be obtained. The assumption of smoothness across ages becomes increasingly important when data are noisy. The second takeaway message is that 'local' smoothing methods (such as splines) should be used with caution when dealing with many subpopulations. This is particularly true when subpopulations are sparse, as overfitting or underfitting of some subpopulations is likely to occur. Furthermore, models developed with aggregate patterns in mind, such as the Rogers-Castro framework, should only be used in the simplest and most regular cases when modeling disaggregated data. Results can be misleading if applied to noisy and heterogeneous multipopulation data.

In terms of applied work, several avenues of future research might be interesting to explore. First, although we apply the model in a migration setting, applications to fertility and mortality data and comparisons with other standard modeling frameworks would be informative. Second, applications to partially missing data, as in the context of life tables in developing countries, appear promising. Third, it would be interesting to consider alternative dimensions for smoothing demographic outcomes. For example, the model can be extended to consider smoothing in the dimension of a continuous variable such as population density, to explore demographic phenomena along an urban-rural continuum.

In addition, several methodological extensions are potentially worth investigating. First, an extension of the hierarchical part of the model to jointly accommodate covariate effects, a smooth temporal component, and spatial random effects is a further generalization. Moreover, a more complex nonlinear regression approach for the priors on $\alpha_i$ and $\lambda_{iq}$ (possibly based on another hierarchical layer including P-splines or regression trees) could further improve predictive performance. Finally, only a moderate number of subpopulations (i.e. a few hundred) are considered in the case study in this paper. Some interesting settings will potentially have thousands of very sparse subpopulations. It would be informative and potentially lead to further model refinement to explore the performance of the model in such settings.

## REFERENCES


Aitchison, J., & Ho, C. (1989). The multivariate Poisson-log normal distribution. *Biometrika*, *76*(4), 643–653.

Alexander, M., Zagheni, E., & Barbieri, M. (2017). A flexible Bayesian model for estimating subnational mortality. *Demography*, *54*(6), 2025–2041.

Beine, M., Bertoli, S., & Fernández-Huertas Moraga, J. (2016). A practitioners' guide to gravity models of international migration. *The World Economy*, *39*(4), 496–512.

Bijak, J., Disney, G., Findlay, A. M., Forster, J. J., Smith, P. W., & Wiśniowski, A. (2019). Assessing time series models for forecasting international migration: Lessons from the United Kingdom. *Journal of Forecasting*, *38*(5), 470–487.

Camarda, C. G. (2012). MortalitySmooth: An R package for smoothing Poisson counts with P-splines. *Journal of Statistical Software*, *50*, 1–24.

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*, 465–480.

Chan, A. B., & Vasconcelos, N. (2009). Bayesian poisson regression for crowd counting. *2009 IEEE 12th international conference on computer vision*, 545–551.

Chiquet, J., Mariadassou, M., & Robin, S. (2018). Variational inference for probabilistic Poisson PCA. *The Annals of Applied Statistics*, *12*(4), 2674–2698.

Clark, S. J. (2019). A general age-specific mortality model with an example indexed by child mortality or both child and adult mortality. *Demography*, *56*(3), 1131–1159.

Conti, G., Frühwirth-Schnatter, S., Heckman, J. J., & Piatek, R. (2014). Bayesian exploratory factor analysis. *Journal of econometrics*, *183*(1), 31–57.

Czado, C., Delwarde, A., & Denuit, M. (2005). Bayesian Poisson log-bilinear mortality projections. *Insurance: Mathematics and Economics*, *36*(3), 260–284.

Czaika, M., & Reinprecht, C. (2022). Migration drivers: Why do people migrate. *Introduction to Migration Studies: An Interactive Guide to the Literatures on Migration and Diversity*, 49–82.

Dharamshi, A., Alexander, M., Winant, C., & Barbieri, M. (2023). Jointly estimating subnational mortality for multiple populations. *arXiv preprint arXiv:2310.03113*.

Durowaa-Boateng, A., Goujon, A., & Yildiz, D. (2023). A Bayesian model for the reconstruction of education- and age-specific fertility rates: An application to African and Latin American countries. *Demographic Research*, *49*(31), 809–848.

El-Sayyad, G. (1973). Bayesian and classical analysis of Poisson regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, *35*(3), 445–451.

Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing*, *7*, 57–68.

Heligman, L., & Pollard, J. H. (1980). The age pattern of mortality. *Journal of the Institute of Actuaries*, *107*(1), 49–80.

Hyndman, R. J., Booth, H., & Yasmeen, F. (2013). Coherent mortality forecasting: The product-ratio method with functional time series models. *Demography*, *50*(1), 261–283.

Hyndman, R. J., & Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, *51*(10), 4942–4956.

Kastner, G., Frühwirth-Schnatter, S., & Lopes, H. F. (2017). Efficient Bayesian inference for multivariate factor stochastic volatility models. *Journal of Computational and Graphical Statistics*, *26*(4), 905–917.

Kowal, D. R., & Bourgeois, D. C. (2020). Bayesian function-on-scalars regression for high-dimensional data. *Journal of Computational and Graphical Statistics*, *29*(3), 629–638.

Lang, S., & Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, *13*(1), 183–212.

Lee, R. D., & Carter, L. R. (1992). Modeling and forecasting US mortality. *Journal of the American statistical association*, *87*(419), 659–671.

Mazzuco, S., Scarpa, B., & Zanotto, L. (2018). A mortality model based on a mixture distribution function. *Population Studies*, *72*(2), 191–200.

McNeil, D. R., Trussell, T. J., & Turner, J. C. (1977). Spline interpolation of demographic data. *Demography*, *14*(2), 245–252.

Montagna, S., Tokdar, S. T., Neelon, B., & Dunson, D. B. (2012). Bayesian latent factor regression for functional and longitudinal data. *Biometrics*, *68*(4), 1064–1073.

Pavone, F., Legramanti, S., & Durante, D. (2022). Learning and forecasting of age-specific period mortality via B-spline processes with locally-adaptive dynamic coefficients. *arXiv preprint arXiv:2209.12047*.

Piironen, J., & Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Stasitistics*, *11*, 5018–5051.

Polson, N. G., Scott, J. G., & Windle, J. (2013). Bayesian inference for logistic models using Pólya-Gamma latent variables. *Journal of the American Statistical Association*, *108*, 1339–49.

Raftery, A. E., & Ševčíková, H. (2023). Probabilistic population forecasting: Short to very long-term. *International Journal of Forecasting*, *39*(1), 73–97.

Ramsay, J., & Silverman, B. (2005). Principal components analysis for functional data. *Functional data analysis*, 147–172.

Raymer, J., Wiśniowski, A., Forster, J. J., Smith, P. W., & Bijak, J. (2013). Integrated modeling of European migration. *Journal of the American Statistical Association*, *108*(503), 801–819.

Rogers, A., Raquillet, R., & Castro, L. J. (1978). Model migration schedules and their applications. *Environment and Planning A*, *10*(5), 475–502.

Shang, H. L., Booth, H., & Hyndman, R. J. (2011). Point and interval forecasts of mortality rates and life expectancy: A comparison of ten principal component methods. *Demographic Research*, *25*, 173–214.

Susmann, H., Alexander, M., & Alkema, L. (2022). Temporal models for demographic and global health outcomes in multiple populations: Introducing a new framework to review and standardise documentation of model assumptions and facilitate model comparison. *International Statistical Review*, *90*(3), 437–467.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*, 528–540.

Wiśniowski, A., Smith, P. W., Bijak, J., Raymer, J., & Forster, J. J. (2015). Bayesian population forecasting: extending the Lee-Carter method. *Demography*, *52*(3), 1035–1059.

Zens, G., Frühwirth-Schnatter, S., & Wagner, H. (2023). Ultimate Pólya Gamma Samplers–Efficient MCMC for possibly imbalanced binary and categorical data. *Journal of the American Statistical Association*.

# APPENDIX

## A1 Identification and Bayesian Estimation Using MCMC

Eq. (3.3) implies that the model is only identified up to rotating, scaling, and changing the sign of the latent functions $\Phi_q(x)$. To solve the scaling and rotational problem, the constraints of Kowal and Bourgeois (2020) are implemented. This effectively fixes the system to an orthonormal rotation by conditioning the estimation process on $\Phi_j(x)'\Phi_j(x) = 1$ and $\Phi_j(x)'\Phi_k(x) = 0$ for $j \neq k$. This orthonormality constraint makes the model similar to a probabilistic principal components model under a Poisson likelihood, see for example Chiquet et al. (2018). An order constraint on $\sigma^2_{\lambda,q}$ can then be used to fully resolve the rotational invariance problem. The signs of $\Phi_q(x)$ are left unidentified. This is not an issue for MCMC sampling in the applications considered in this article. In case sign switching becomes an issue, simple and efficient ex-post reordering schemes are available (Kastner et al., 2017).

Given the model specification and the identification scheme, the goal is then to obtain estimates of the unknown parameters $\alpha_i$, $\boldsymbol{\delta}$, $\lambda_{iq}$, $\boldsymbol{\beta}_q$, $\sigma^2$, $\sigma^2_a$ and $\sigma^2_{\lambda,q}$, as well as the spline coefficients $f_{qk}$ jointly. From these estimates we can then construct the predictive posterior distribution, and hence estimates and probabilistic uncertainty quantification, for $\alpha_i + z_i(x)$, the demographic process of interest.

To make model estimation feasible when confronted with count-valued $y_{ix}$, we will use the idea of data augmentation (Tanner and Wong, 1987) and work with explicit imputation of the latent $z_{ix}$ within an MCMC scheme. Sampling of $z_{ix}$ is based on adaptive Metropolis-Hastings steps within a Gibbs sampler. Conditional on $z_{ix}$, the model is then essentially a Gaussian regression model, which facilitates sampling of the remaining parameters. In fact, the model for $z_{ix}$ coincides almost perfectly with the functional regression model of Kowal and Bourgeois (2020), allowing us to make use of the highly efficient posterior simulation algorithm developed there. In addition, handling missing data becomes straightforward based on this approach.

In case exact updating of $z_{ix}$ is too computationally expensive, several approximations to Poisson and Poisson lognormal models (El-Sayyad, 1973; Chan and Vasconcelos, 2009) are available. These work well when the outcomes are large counts. In addition, in large count settings, a simple approximation to the Poisson lognormal model is to fix $z_{ix} = log(y_{ix})$. To see this, observe that as $y_{ix} \to \infty$, $z_{ix} \to \log(y_{ix})$ and a Gaussian model for the log counts becomes an excellent approximation to the Poisson lognormal model, see also Pavone et al. (2022). In small count settings and in the presence of zero counts, however, relying on logarithmic transforms becomes increasingly problematic. As an in-between solution that balances computational tractability and approximation quality, the ideas of Gamerman (1997) could be applied. Specifically, an iteratively reweighted least squares algorithm can be used to obtain a Gaussian approximation of the conditional posterior of $z_{ix}$ in each MCMC iteration, which can then be used to efficiently produce approximate posterior samples. Again, the quality of this approximation will increase with the size of the counts $y_{ix}$.

## A2 Additional Results

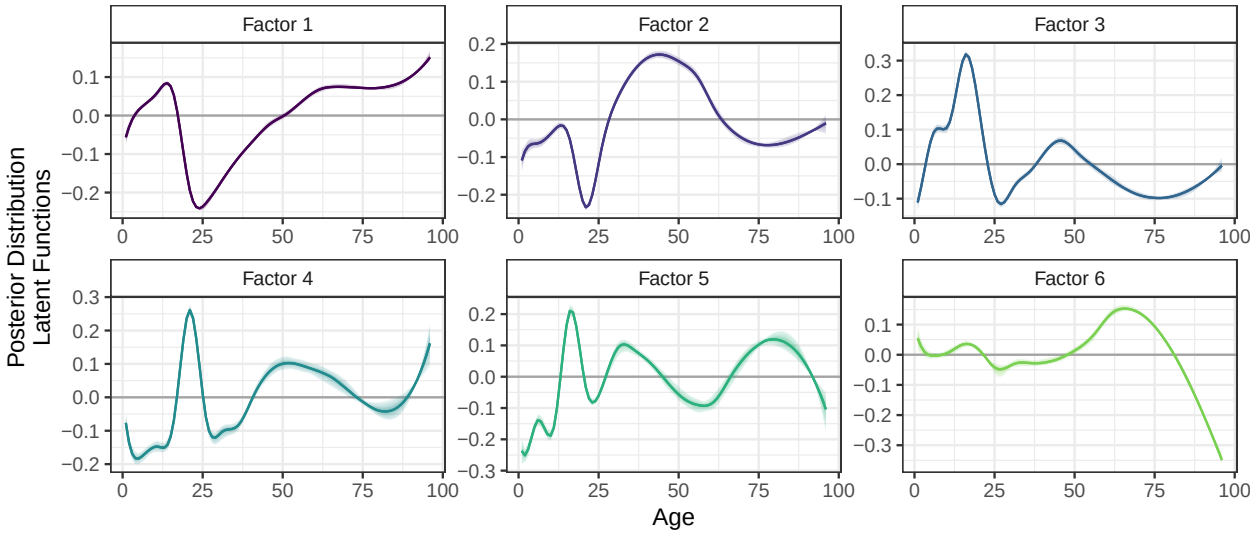



**Figure A1:** Estimated latent functions $\Phi_q(x)$ as well as 95% credible intervals.

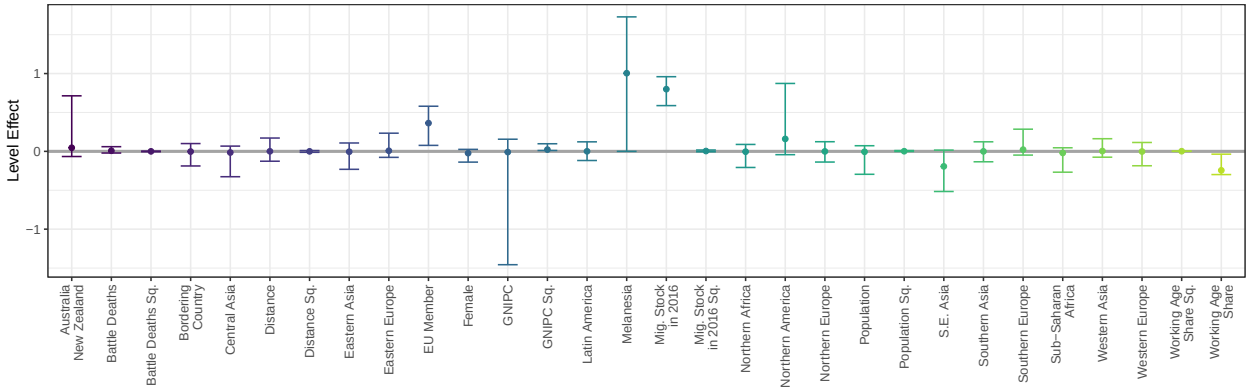



**Figure A2:** Posterior mean estimates of $\boldsymbol{\delta}$ as well as 95% credible intervals. These estimates describe the average level shift of the demographic process of interest after a unit increase in a given covariate.

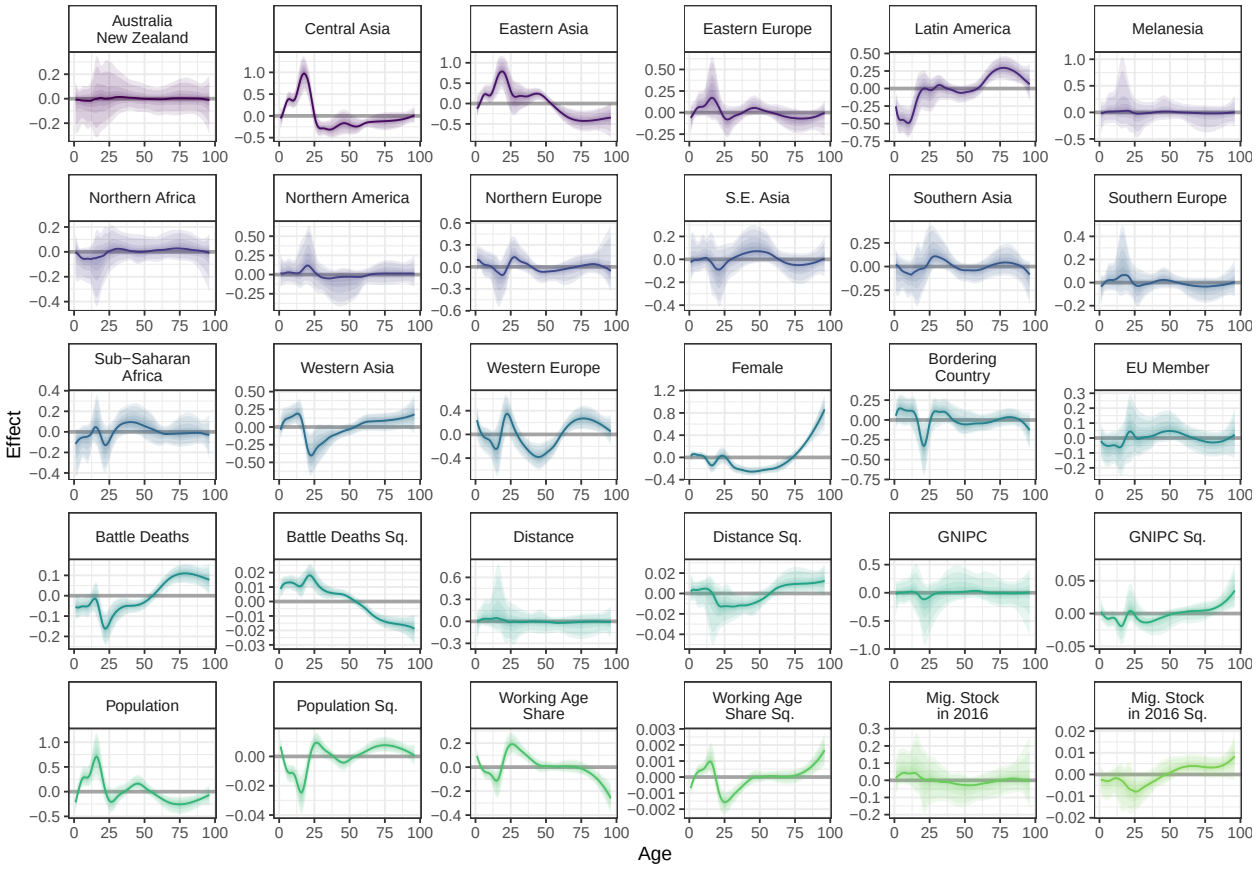



**Figure A3:** Posterior mean estimates of the shape effects of the included covariates as well as 95% credible intervals. These estimates describe the average modulation of the demographic process of interest after a unit increase in a given covariate.

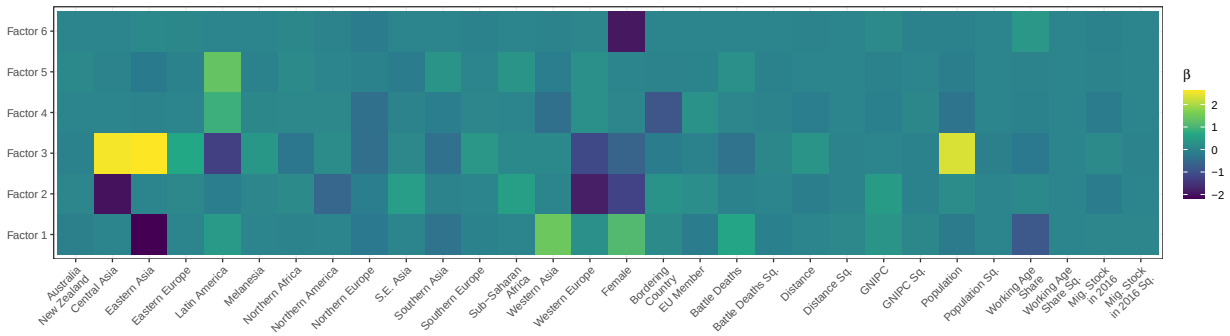
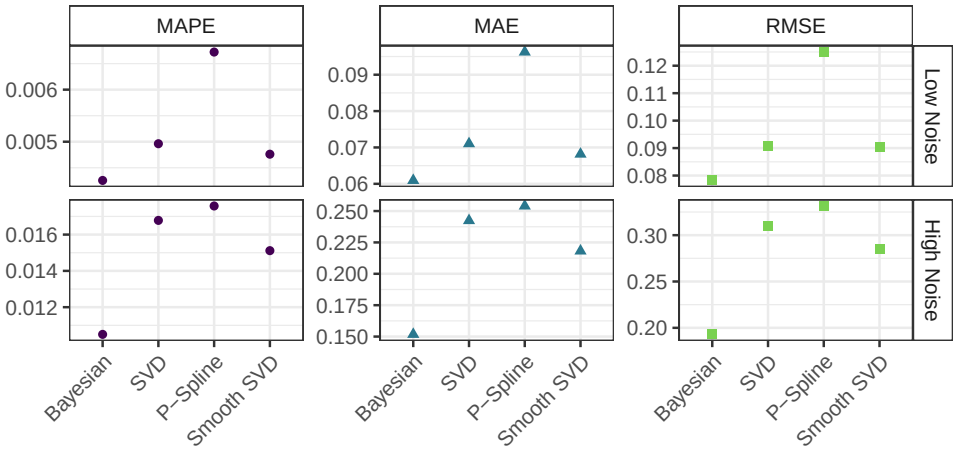
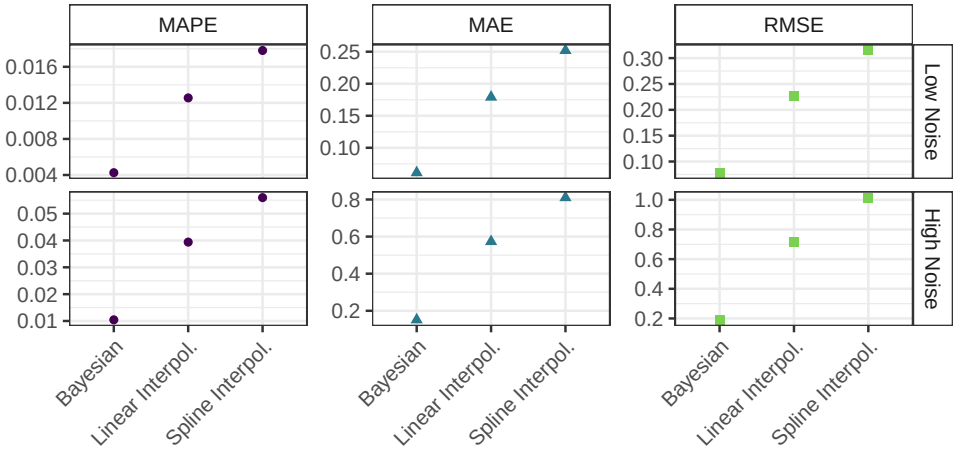


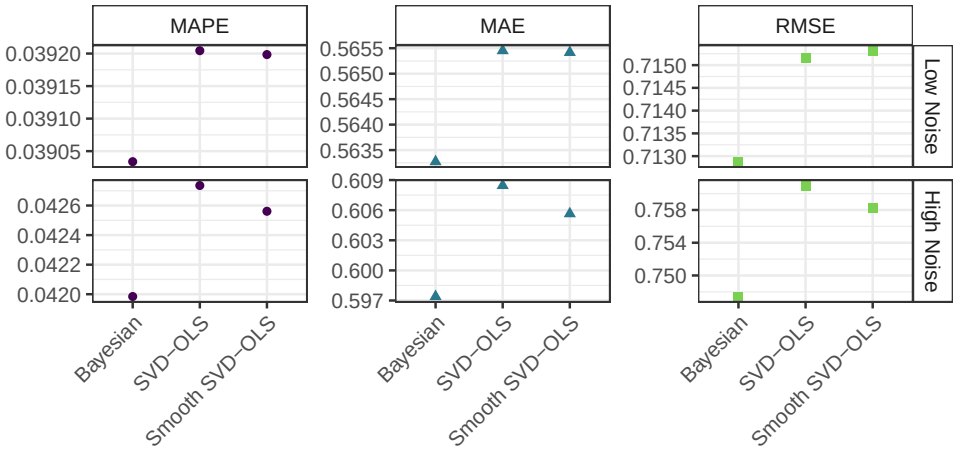

**Figure A4:** Estimated posterior means for coefficients $\boldsymbol{\beta}$ for each covariate / factor combination.




**(a)** Simulated In-Sample Exercise.




**(b)** Simulated Missing Data Exercise.




**(c)** Simulated Out-Of-Sample Exercise.

**Figure A5:** Visual summary of the results of the systematic simulation studies. Competing models are on the x-axis. Panels refer to combinations of high/low noise settings and the three different evaluation scores computed. MAPE = mean absolute percentage error. MAE = mean absolute error. RMSE = root mean square error.

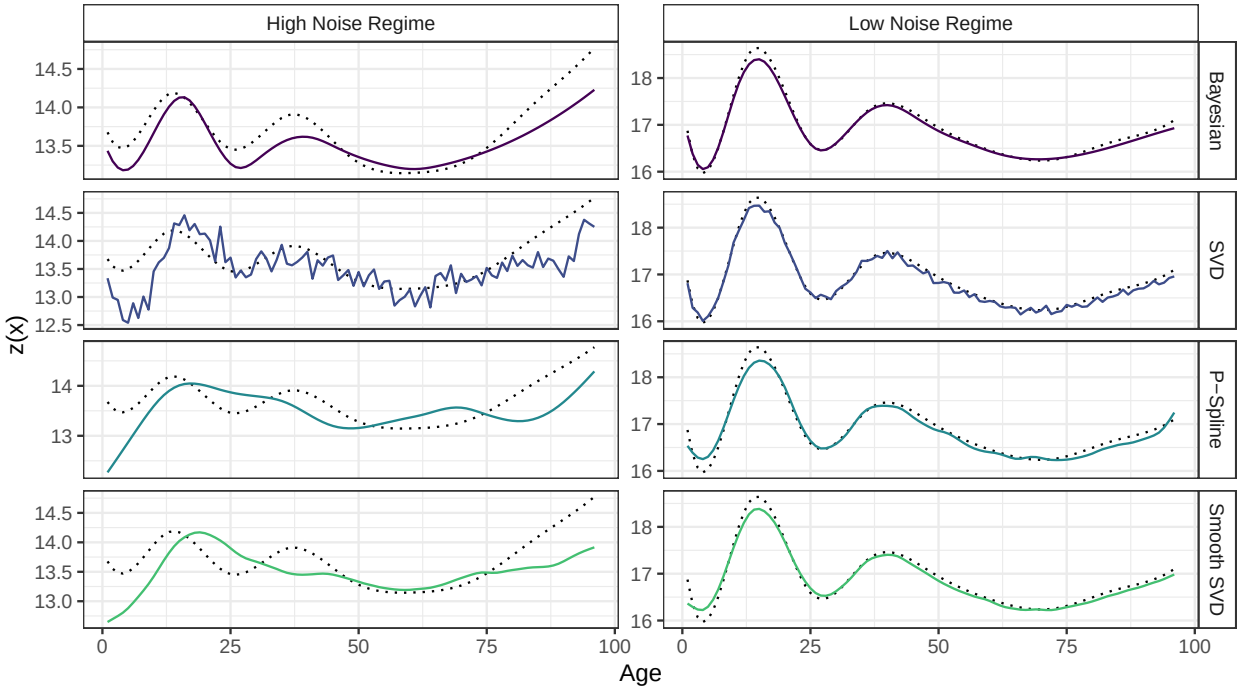



**Figure A6:** Two selected examples for in-sample fit based on simulated data. Columns correspond to high/low noise regimes. Rows correspond to four different models considered. Refer to the text for details.

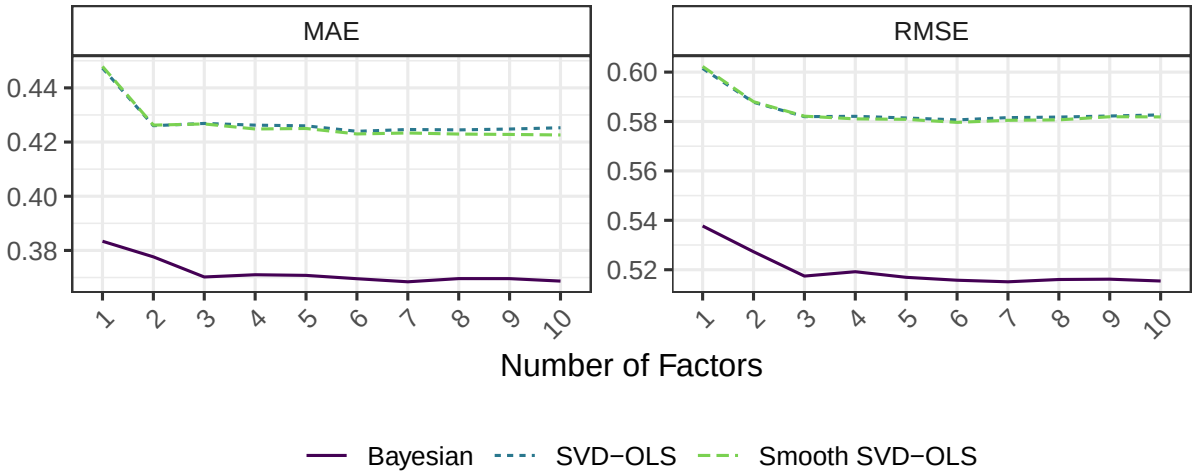



**Figure A7:** Visual summary of the results of the cross-validation exercise using the Austrian immigration data. Number of estimated factors is on the x-axis. Models correspond to colored lines. Panels refer to the two different evaluation scores computed. MAE = mean absolute error. RMSE = root mean square error.