Fast Bayesian estimation of disaggregated demographic rates by borrowing strength from international databases

John Bryant[1] and Junni L. Zhang[2]

[1]Bayesian Demography Limited, Christchurch, New Zealand
[2]National School of Development, Peking University, Beijing, China

1 Introduction

Demographers are analysing increasingly large datasets in which outcomes are disaggregated not only by age, sex, and time, but also by variables such as region, ethnicity, and education. Using statistical models to estimate rates in these big datasets can be slow and difficult.

Demographers have long recognized that the key to modelling demographic rates is to take advantage of regularities in the rates (Keyfitz, 1982). Age-sex profiles for demographic processes such as mortality, fertility, and migration all follow distinctive patterns that are repeated across widely varying populations. Moreover, international databases such as the Human Mortality Database (HMD) provide demographers with the data they need to quantify the regularities.

We have developed new methods and software for estimating demographic rates that take advantage of the regularities found in international databases. Our methods and software are flexible, allowing users to define their own outcomes, classifications of the data, and model specifications. The methods and software are also fast and scalable, as we illustrate with the case of regional mortality rates in Sweden.

2 Using international databases to quantify regularities in rates

One of the most important developments in demography in the 21st century has been the appearance of online international databases of high-quality demographic data. Examples include the Human Mortality Database, the Human

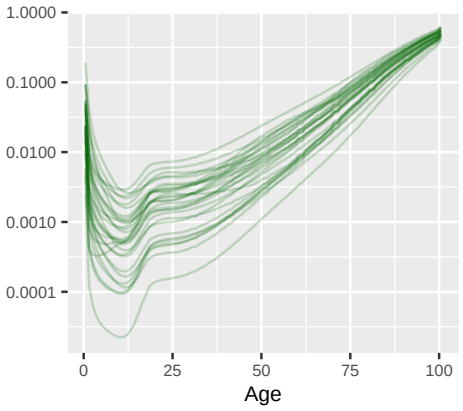

Figure 1: *Thirty randomly-generated age profiles for mortality, generated using SVD components from the Human Mortality Database. Output from an SVD analysis of the HMD is scaled to produce a matrix A and vector b. Generating 30 vectors z of normally-distributed variables, and then setting $x = Az + b$ produces 30 random profiles x. Matrix A in this example has only 3 columns.*

Fertility Database, IPUMS, and OECD and UN databases on subjects such as labor force participation, marriage, and household structure.

It turns out that regularities in these databases can be represented parsimoniously, using a mathematical technique called the Singular Value Decomposition (SVD) (Alexander et al., 2017; Clark, 2019). The ability of the SVD to capture important features of the data is illustrated in Figure 1. The profiles were generated from output from an SVD analysis. Despite the simplicity of the procedure, the profiles look like real mortality profiles drawn at random from the Human Mortality Database data.

3 Incorporating regularities into Bayesian statistical models

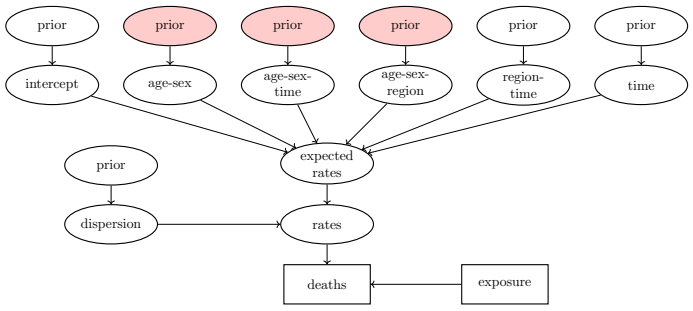

Figure 2: *Components of a statistical model for rates. Rectangles represent data, and ellipses represent quantities to be estimated. The three priors containing information from international databases are shown in red.*

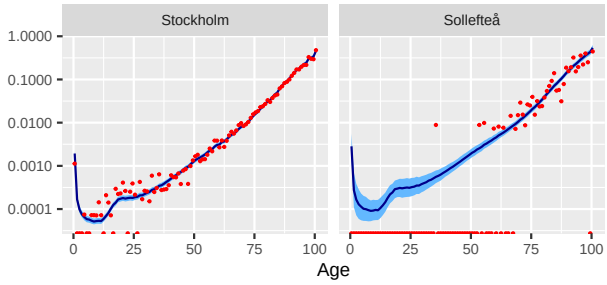

Figure 3: *Estimates of mortality rates (on a log scale) for females in two Swedish regions, 2023.*

We have developed a family of Bayesian hierarchical models for estimating demographic rates. A typical model from the family (in fact, the model we use for Swedish mortality data) is shown in Figure 2. Expected values for demographic rates are determined by main effects and interactions formed from variables such as age, sex, region, and time. Each of these terms has a prior distribution describing its expected behavior. Many of the priors that we use are only "weakly informative", and do nothing more than encode rough orders of magnitude. In contrast, the priors that we use for terms involving age, which are based on output from SVD analyses of the HMD, are strongly informative, pushing the estimates towards age-sex patterns that are consistent with patterns found in the HMD.

The models are implemented in an R package called **bage**, available on the public repository CRAN.

4 Case study: Mortality in Swedish regions

We illustrate our methods with an analysis of mortality rates in Swedish regions. Our dataset is obtained from the Statistics Sweden website[1]. It includes 101 single-year age groups, 2 sexes, 70 regions, and 10 years, for a total of 141,400 rates to be estimated. The median death count for each cell within the age-sex-region-time classification is 1.

We model deaths as draws from Poisson distributions. As shown in Figure 2, our model includes an age-sex interaction, an age-sex-region interaction, an age-sex-time interaction, a region-time interaction, and a time main effect. We use priors based on HMD data for all terms involving age. Model-fitting is fast, taking about 50 seconds on a laptop.

Figure 3 shows estimates of mortality rates (on a log scale) for females in 2023, for the largest and smallest regions. The blue lines and bands represent point estimates and 95% credible intervals from the model. The red dots represent 'direct' estimates, that is, deaths in each cell divided by population in that cell. Most direct estimates for the smaller region are 0, or negative infinity

[1]Tables *Population by region, marital status, age and sex. Year 1968 - 2023* and *Deaths by region, age (during the year) and sex. Year 1968 - 2023* on the Statistics Sweden website, accessed 9 September 2024.

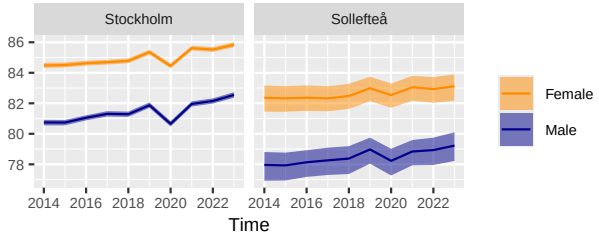

Figure 4: *Estimates of life expectancy at birth in two Swedish regions, 2014–2023.*

on the log scale. The modelled estimates smooth through the dots, drawing strength from across the Swedish dataset, and from the information encoded in the SVD-based priors.

Figure 4 shows life expectancies derived from the rates. Life expectancy is lower, and less precisely estimated, in the smaller region.

5 Discussion

Our models continue a long demographic tradition of identifying empirical regularities in demographic rates, and then incorporating these regularities into analyses of noisy datasets. The contribution of our work is to show how this strategy can be implemented within formal statistical models, at scale.

Although our case study focuses on mortality, the methods can be applied to other demographic processes where international demographic databases are available, and where there are strong cross-national regularities in age-sex profiles. Potential applications include fertility, migration, labor force participation, living arrangements, and marriage.

References

Alexander, M., Zagheni, E., and Barbieri, M. (2017). A flexible bayesian model for estimating subnational mortality. *Demography*, 54:2025–2041.

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.

Keyfitz, N. (1982). Choice of function for mortality analysis: Effective forecasting depends on a minimum parameter representation. *Theoretical Population Biology*, 21(3):329–352.