Jointly Estimating Subnational Mortality for Multiple Populations

Ameer Dharamshi[1,2]

Monica Alexander[1]

Celeste Winant[3]

Magali Barbieri[3,4]

[1] University of Toronto

[2] University of Washington, Seattle

[3] University of California, Berkeley

[4] Institut National d'Études Démographiques

July 27, 2023

Abstract

Understanding patterns in mortality across subpopulations is essential for local health policy decision making. One of the key challenges of subnational mortality rate estimation is the presence of small populations and zero or near zero death counts. When studying differences between subpopulations, this challenge is compounded as the small populations are further divided along socioeconomic or demographic lines. In this paper, we build on principal component-based Bayesian hierarchical approaches for subnational mortality rate estimation to model correlations across subpopulations. The principal components identify structural differences between subpopulations, and coefficient and error models track the correlations between subpopulations over time. We illustrate the use of the model in a simulation study as well as on county-level sex-specific US mortality data. We find that results from the model are reasonable and that it successfully extracts meaningful patterns in US sex-specific mortality. Additionally, we show that ancillary correlation parameters are a useful tool for studying the convergence and divergence of mortality patterns over time.

1 Introduction

Reliable mortality rate estimates are crucial for understanding, formulating, and validating health policy decision making. Historically, developments in both mortality modelling and understanding mortality differences across populations has focused on national-level estimates for cross-country comparisons. However, patterns in more granular subnational level mortality rates are needed to facilitate local level decisions. More recent work has thus focused on disparities in mortality outcomes within countries at the subnational level. Additionally, with evidence of increasing mortality inequality, there is increased necessity in monitoring mortality differences across key population subgroups within subnational areas, such as differences by sex or gender, race/ethnicity, or socioeconomic groups like income or education. For example, Bhutta (2016) argued that subnational determinants explain a greater portion of child mortality than country-level characteristics. In the collective effort to address health inequalities experienced by segments of the population, estimates of subpopulation mortality discrepancies are needed to identify and understand the mortality patterns of vulnerable groups, and track the effects of policy responses.

A challenge with producing reliable estimates for subnational populations is issues associated with small populations. When considering deaths by age at an increasingly granular geographic level, the number of deaths observed gets smaller, and the chance of observing no deaths at all increases. This issue is exacerbated when populations are further disaggregated by other demographic characteristics. This means that inferring underlying mortality risks from raw death rates is challenging, due to the erratic or unclear patterns over age. Although raw death rates can usually be calculated, the natural variation leads to uncertain estimates. In order to obtain more reliable estimates of mortality rates and associated uncertainty, models that take into account the stochasticity in the data are needed.

In this paper, we propose a general model framework to estimate age-specific mortality rates at the subnational level jointly for multiple populations. The model incorporates characteristic shapes of mortality age schedules within a Bayesian hierarchical framework, which allows information on mortality patterns to be shared across populations. The model extends previous approaches by accounting for correlation in mortality experiences across subpopulations, rather than assuming subpopulations are independent. As well as producing estimates and uncertainty for mortality rates, higher-level parameters are also estimated, which summarize trends in different dimensions of mortality over time and how these trends are similar or different across groups. We illustrate the model with an application to estimating sex-specific mortality by county in the United States. While we focus on sex-specific mortality, the modelling framework is generalizable to population groups defined by other characteristics (such as race/ethnicity).

The remainder of the paper is structured as follows. We first give a brief overview of recent developments in subnational mortality estimation. We then begin our methods discussion by illustrating that the selection of principal components captures structural differences in subpopulation specific mortality patterns, followed by a formal statement of the model. Results from the model along with validation exercises using simulated and real sex-specific US county-level mortality data are then provided. Finally, we conclude with a discussion of our findings and identify directions for future research.

2 Background

A large body of research exists on small area estimation issues, and recently demographers have increasingly taken advantage of computational advances which make fitting complex statistical models to small-scale mortality data feasible. In particular, there has been a notable increase in the use of Bayesian methods in demographic estimation, particularly subnational estimation. Bayesian methods are particularly suited to demographic contexts as they provide a useful framework to incorporate different data sources in the same model, account for various types of uncertainty, and allow for information exchange across time and space (Bijak and Bryant 2016).

One area of previous work has focused on estimating aggregate indicators at the subnational level, such as life expectancy and child mortality (Mercer et al. 2015; Ševčíková and Raftery 2021). Models on aggregate indicators generally involve temporal and spatial smoothing, allowing for information in mortality trends to be shared across these dimensions. In some cases, models rely on covariates (such as education or income) to stabilize mortality rate estimates from noisy data (Wang et al. 2013; Arias et al. 2018).

In addition to aggregate mortality indicators, research has focused on producing estimates of age-specific mortality rates. The advantage of modelling age-specific rates is that they can be converted into estimates of life tables. Recent advances in this area build on classical demographic approaches of model life tables and relational models, which identify key patterns in mortality over age across a wide range of populations, and allow patterns to be shifted based on a reduced set of parameters (Coale, Demeny, and Vaughan 1983). For example, TOPALS models consist of a standard age schedule and population-specific deviations away from that standard, which are smoothed using linear splines (de Beer 2012). TOPALS-type models have been used to produce subnational estimates of migration and mortality in varying data quality contexts (Schmertmann and Gonzaga 2016, 2018; Dyrting 2020).

A related modelling approach derives a set of ‘principal components’ from reference mortality curves which are then used as the basis of a regression framework. This allows a large set of plausible mortality curves to be estimated using a reduced set of parameters. This approach is in the spirit of the Lee-Carter model for mortality forecasting and related work (Lee and Carter 1992). For example, Clark (2019) uses this approach to formulate a new set of model age schedules in data-sparse contexts. In particular, this paper extends an earlier paper by Alexander, Zagheni, and Barbieri (2017) which introduces a Bayesian hierarchical framework that builds on principal components derived from national mortality schedules for use at the subnational level. In more recent work, Alexander and Alkema (2021) have used a principal components approach to model mortality in the broader context of subnational population estimation using a cohort component projection framework.

In general, most existing mortality models inherently assume subpopulations are independent. They model each population separately, then combine or interpret after modelling. However, ideally we would model all subpopulations within one framework as mortality is generally correlated over groups. Existing joint models focus on modelling mortality by sex. At the national level, sex-specific life expectancy estimates have been produced by the UN using gap-based approaches such as in Raftery, Lalic, and Gerland (2014). Under this approach, female life expectancy is estimated with standard one-sex methods. Estimates of male life expectancy are then constructed by modelling the gap between female and male life expectancy. In the present case, we are interested in estimates of all age-specific mortality rates, not just life expectancy, and a purely gap-based approach to this problem at smaller scales may not be appropriate due to the small counts in both areas and age groups. Additionally, gap-based approaches require the selection of an anchor population such as females in the case of sex-specific modelling. This works in situations where the composition within a population is roughly balanced, but could be problematic when considering variables such as race or ethnicity that vary dramatically across jurisdictions, particularly if the anchor population is near absent in certain regions.

At the subnational level, Rau and Schmertmann (2020) jointly model age- and sex-specific mortality in regions across Germany using a Bayesian spatial TOPALS approach. With regards to sex differences, they obtain a typical pattern of age-specific sex differences across aggregated regions of Germany, and then use this information as a basis for a prior for differences at smaller scales, penalizing large deviations away from the observed differences by sex. In this paper, we take a different approach, and flip the viewpoint from thinking about sex differences, to thinking about covariation in mortality by sex (or any other population subgroups), and explicitly allow for group mortality rates to move together.

3 Methods

To estimate subnational mortality rates while also extracting patterns across key subpopulations, we propose a Bayesian hierarchical model that builds on a principal component-based approach, and incorporates structures that capture subpopulation interactions. Before defining the model equations, we first discuss why principal components are an attractive modelling strategy in this context.

3.1 Principal components models

The use of principal components is motivated by the fact that age-specific mortality rates tend to display strong regularities across space and over time. This means that systematic variation in mortality rates is well captured by a reduced set of parameters which can be modelled such that information is shared across space and time.

To generate principal components, we compute the singular value decomposition of a collection of regional log-mortality curves. The collection of log-mortality curves, $\mathbf{X}$, is a $N \times A$ matrix where N is the number of region-subpopulation-years under consideration and A is the number of age groups. For US sex-specific mortality, $N = 6066$ (60 years of state-level sex-specific data) and $A = 19$ (ages <1, 1-4, 5-9, ..., 75-79, 80-84, 85+). Note that different age groups, such as one-year age groups, could also be considered. The singular value decomposition of $\mathbf{X}$ is then:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}' \tag{1}$$

where $\mathbf{U}$ is the 6066×19 matrix of left singular values, $\mathbf{\Sigma}$ is a 19×19 diagonal matrix of scaling factors, and $\mathbf{V}$ is the 19×19 matrix of right singular values, which we term 'principal components'. These extract key patterns in mortality over age. The first principal component explains the most variation across mortality curves with each successive component explaining less variation.

A natural question to ask is how many principal components to use in a model. Choosing the number of components is a balance between incorporating components that only pick up on systematic patterns, while still allowing for enough flexibility in the model.

The set of principal components that offer useful information on sex-specific mortality can be identified using the $\mathbf{U}$ matrix. If the ith principal component contains material subpopulation-specific differences, we would expect the distributions of the values in the ith column of $\mathbf{U}$ to differ between subpopulations. We thus separate the rows of $\mathbf{U}$ by sex and examine the resulting distributions. Figure 1 displays the coefficients by sex for the first eight principal components. We choose eight as the singular values suggest there is limited additional information beyond eight.

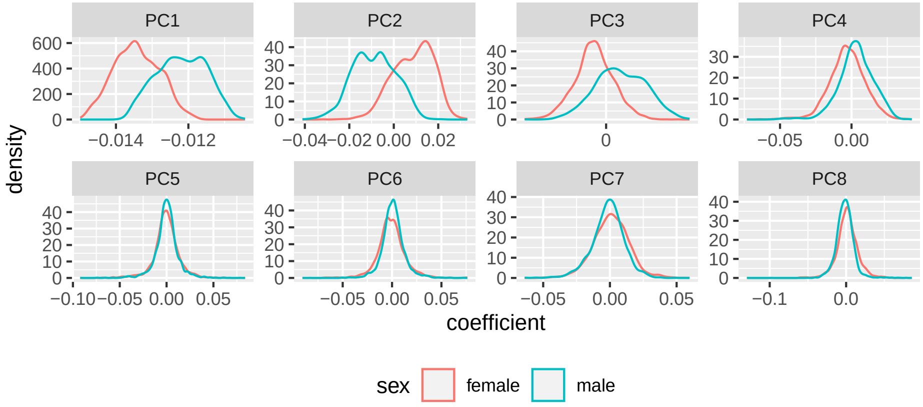


Figure 1: Distribution of observed state-level left singular values by sex and principal component.

The first four principal components demonstrate clear differences in the sex-specific distributions. In particular, the location of the distributions are noticeably different suggesting structural differences between sexes. The distributions of the remaining principal components, including those not presented in Figure 1, are largely similar. Beyond a visual inspection, simple t-tests suggest that aside from the fifth, all of the first eight principal components have location differences. Balancing these findings with the practical consideration that each additional principal component will lead to a significant computational burden during model estimation, we suggest that at least three but ideally four principal components should be used in modelling sex-specific mortality in the US.

The first four principal component curves are plotted in Figure 2. Note that the first principal component is inverted for clarity as its coefficients are exclusively negative.

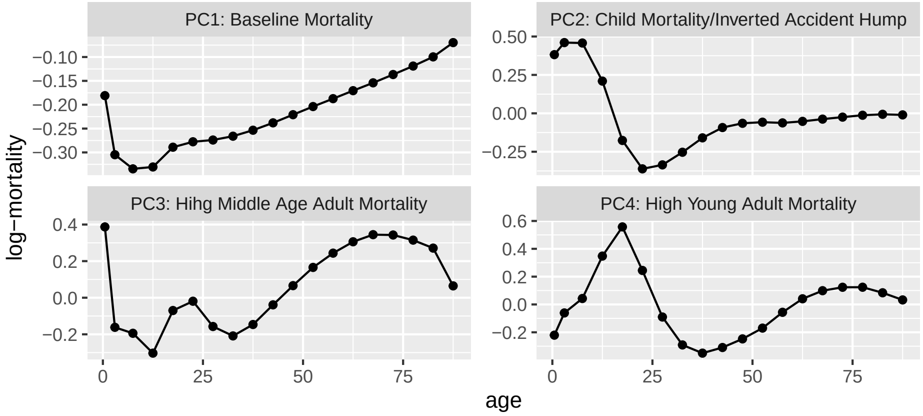


Figure 2: First four state-level log-mortality principal components.

Considering Figure 1 and Figure 2 together offers insight into patterns of interest. The first principal component has the characteristic 'J' shape of log mortality. In addition, the distribution of coefficients for the second and fourth principal components indicate higher young adult mortality in the male population which is consistent with the "accident hump" described in the broader mortality literature, as males are more likely to suffer mortality due to risky behaviour (Heligman and Pollard 1980).

The evidence that the principal components are adequately capturing understood sex-specific differences motivates their use as the basis for a subpopulation mortality model. Using principal components as a basis for a statistical model can be robust in small population settings as the components allow plausible mortality rates to be estimated even in the presence of high variability.

3.2 Model summary

We begin by defining $y_{a,s,c,t}$ as the observed number of deaths in age group a, subpopulation s, county c, and year t. Then, we assume that

$$y_{a,s,c,t}|\lambda_{a,s,c,t} \sim \text{Poisson}\left(P_{a,s,c,t} \cdot \lambda_{a,s,c,t}\right), \tag{2}$$

where $P_{a,s,c,t}$ is the population corresponding to age group a, subpopulation s, county c, and year t, and $\lambda_{a,s,c,t}$ is the mortality rate to be estimated for age group a, subpopulation s, county c, and year t.

The mortality rates $\lambda_{a,s,c,t}$ are estimated on the log scale as follows:

$$\log\left(\lambda_{a,s,c,t}\right) = \sum_{i=1}^{P} \left(\beta_{i,s,c,t} \cdot Y_{i,a}\right) + \gamma_{a,s,c,t}, \tag{3}$$

where Y_i is the ith principal component, P is the number of principal components, β are the estimated coefficients, and γ is an overdispersion term. The number of principal components, P can be selected based on their contributions and differences between groups. In the case of sex-specific mortality in US counties, we use the $P = 4$ principal components plotted in Figure 2.

Equation (3) can be intuitively understood as constructing a log-mortality curve for each age-sex-county-year as a linear combination of the four principal components plus some additional age-specific variation. Different linear combinations of the principal components (that is, different estimated values of the β coefficients) lead to different plausible log-mortality curves. The overdispersion term accounts for the possibility that deaths will be overdispersed relative to the patterns captured by the principal components.

3.2.1 Core model

For many counties, population and death counts are small and highly variable, which would lead to uncertain estimates of β if each county and year were estimated independently. As such, we propose a hierarchical model for β based on the nesting structure of counties within states. Specifically, we will assume that counties within each state are more likely to share similar mortality patterns than counties in different states, and allow high-data counties to share information with low-data counties in the same state. A similar relationship holds across subgroups. Within a county, subgroups of the population may experience similar drivers of mortality, whether they be due to local policy, environmental, or other factors. This structure suggests that β coefficients, and thus by extension log-mortality curves, for all subgroups within a county should be modelled jointly, thereby exploiting the dependence between groups.

To capture geographic and subgroup dependence, we propose to model the vector of β coefficients for all S groups within each county jointly as multivariate normal with a common state-level mean vector and covariance matrix:

$$\begin{pmatrix} \beta_{i,1,c,t} \\ \dots \\ \beta_{i,S,c,t} \end{pmatrix} = \begin{pmatrix} \mu_{\beta_{i,1,t}} \\ \dots \\ \mu_{\beta_{i,S,t}} \end{pmatrix} + \begin{pmatrix} \omega_{i,1,c,t} \\ \dots \\ \omega_{i,S,c,t} \end{pmatrix}, \quad i = 1, \dots, P \tag{4}$$

$$\begin{pmatrix} \omega_{i,1,c,t} \\ \dots \\ \omega_{i,S,c,t} \end{pmatrix} \Bigg| \sigma_{\beta_{i,t}}, L_{i,t}^{(\beta)} \sim \mathcal{N}\left(\mathbf{0}_S, \sigma_{\beta_{i,t}} \mathbf{1}_S L_{i,t}^{(\beta)} L_{i,t}^{(\beta)\top} \mathbf{1}_S \sigma_{\beta_{i,t}}\right) \tag{5}$$

$$L_{i,t}^{(\beta)} L_{i,t}^{(\beta)\top} \sim \mathrm{LKJ}(1) \tag{6}$$

$$\sigma_{\beta_{i,t}} \sim \mathcal{N}(0,1) \tag{7}$$

where $\mu_{\beta_{i,s,t}}$ is the state-level coefficient for the ith principal component, subpopulation s, and year t, and $\omega_{i,s,c,t}$ is the county deviation for the ith principal component, subpopulation s, county c, and year t.

β is written as the sum of two terms, a state-level mean vector, μ_β, and a vector of deviations from the mean, ω. Information-rich observations from county-subgroups with large populations are the primary contributors to estimates of μ_β for the corresponding subpopulation. Estimates of β for small county-subgroups with less informative observed death counts are then partially informed by the high-data counties through the shared state-level means. This pooling effect stabilizes estimates of β for small populations by pulling the estimates towards the state mean. It is important to note that this effect occurs at the county-subgroup level and not uniformly within a county: a large population county with an uneven subgroup composition can experience stronger pooling effects in its smaller subgroups. This is not particularly important for the application to sex-specific mortality as the populations are roughly balanced, but may be critical in other applications involving subgroups defined by other demographic variables.

The ω vector models the specific deviations in the β vector for each county from the state-level mean vector. By jointly modelling the deviations for all subpopulations as in (4), the model captures patterns in the dispersion of β by principal component in $L_{i,t}^{(\beta)} L_{i,t}^{(\beta)\top}$. This enables information sharing across subpopulations. In counties where there is an imbalance in the number of non-zero death observations across subpopulations, jointly generating β coefficients allows higher data groups to support lower data groups. An example of such a county is given in Figure 12 in Appendix B.

We assign the uninformative LKJ(1) prior to the subpopulation correlation Cholesky factors to allow the observed data to fully determine the correlation structure (Lewandowski, Kurowicka, and Joe 2009).

By modelling correlations between principal components, it is possible for joint movement in certain parts of the mortality curve but not in others. For example, baseline mortality could experience strong correlation but the accident hump may not. An illustration of the correlated principal component coefficients is provided in Figure 3. Each plot captures the patterns of one of the four principal component coefficients for California in 2017. The red point represents the state-level value, the black points represent each county, and the blue contours are the distribution of ω centered at the state means. High correlation in the first principal component is expected since when the general conditions in a region improve, both sexes are positively impacted, leading to highly correlated changes in baseline mortality. Similarly, the low correlation in the second principal component is consistent with males experiencing a relatively elevated level of mortality in young adult ages (i.e. the accident hump) whereas females tend not to.

The β specification is intended to be flexible. If there is only a small number of years of data available or there is prior information suggesting that correlation structures are time-invariant, one could reasonably omit the time index or share correlations across a short time interval. Similarly, if certain correlations are known in advance, the correlation matrices themselves could be constrained accordingly.

3.2.2 Temporal smoothing

State-level means are smoothed over time by penalizing the second-order differences to produce gradual changes. Smoothing occurs at the state-level as opposed to the county-level to allow the needed flexibility for

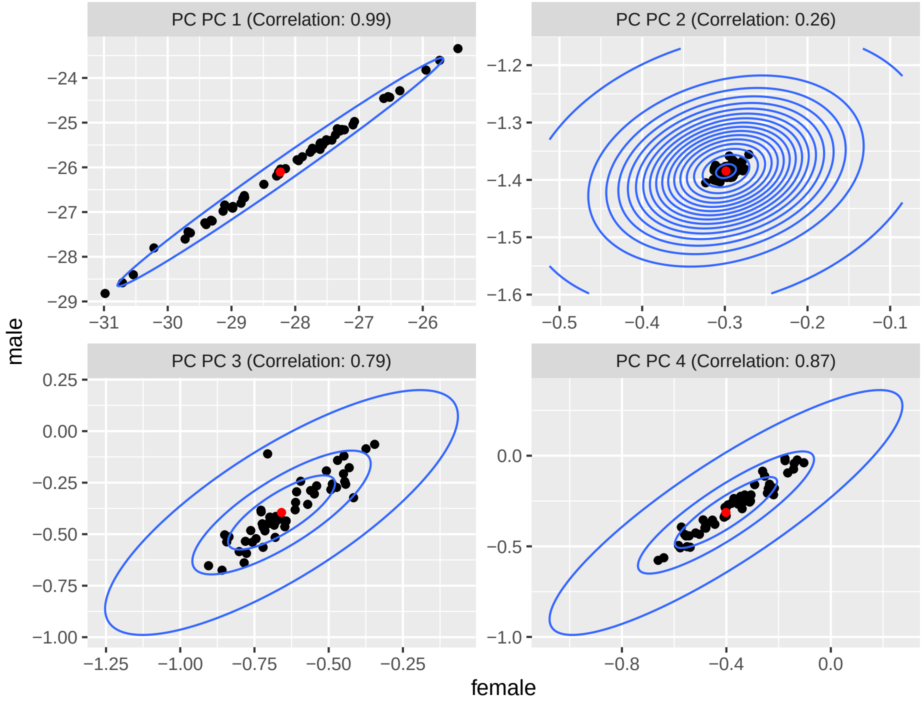


Figure 3: Median posterior county-level principal component coefficients (β) and the corresponding state-level mean (μ_β) for California in 2017. The blue contour plots indicate the recovered correlation patters between male and female population.

counties to experience irregular mortality patterns driven by localized events such as a natural disaster, or a pandemic.

$$\mu_{\beta_{i,s,t}} | \mu_{\beta_{i,s,t-1}}, \mu_{\beta_{i,s,t-2}}, \sigma_{\mu_{\beta_i}} \sim \mathcal{N}\left(2 \cdot \mu_{\beta_{i,s,t-1}} - \mu_{\beta_{i,s,t-2}}, \sigma_{\mu_{\beta_i}}\right) \tag{8}$$

$$\sigma_{\mu_{\beta_i}} \sim \mathcal{LN}\left(-1.5, 0.5\right). \tag{9}$$

3.2.3 Overdispersion term

Finally, the $\gamma_{a,s,c,t}$ term that allows for overdispersion of the log-mortality rate is modelled similarly to $\omega_{i,s,c,t}$ in a S-dimensional multivariate normal setup:

$$\left.\begin{pmatrix}\gamma_{a,1,c,t}\\ \dots\\ \gamma_{a,S,c,t}\end{pmatrix}\right|\sigma_a, L^{(\gamma)}_{a,t} \sim \left(\mathbf{0}_S, \sigma_a \mathbf{1}_S L^{(\gamma)}_{a,t} L^{(\gamma)\top}_{a,t} \mathbf{1}_S \sigma_a\right) \tag{10}$$

$$\sigma_a \sim \mathcal{N}\left(0, 0.25\right) \tag{11}$$

$$L^{(\gamma)}_{a,t} L^{(\gamma)\top}_{a,t} \sim \text{LKJ}(1). \tag{12}$$

Relationships between subpopulation γ's are captured using the age-year correlation matrix $L^{(\gamma)}_{a,t} L^{(\gamma)\top}_{a,t}$. An intuitive way of understanding this component of the model is that the principal components produce the expected mortality derived from aggregate and local patterns while γ captures additional age-specific deviations. These deviations from the expectation are often correlated, though this correlation may differ across ages and over time. Given that the log-mortality values for different age groups occur in substantially different parts of the log curve, we estimate a separate scaling factor σ for each age group.

3.3 Computation

The model described here was fit using a Bayesian framework. Posterior samples are drawn using the No-U-turn sampling (NUTS) Hamiltonian Monte Carlo algorithm (Hoffman and Gelman 2014; Neal 2011) implemented in the Stan R package (Stan Development Team 2021). We execute 4 chains with 500 iterations of burn-in and 2 500 iterations of samples. Convergence was diagnosed using trace plots, effective sample sizes, and the Gelman and Rubin diagnostic (Gelman and Rubin 1992).

4 Results

4.1 Simulation study

We conducted a simulation study to test the ability of the model to simultaneously estimate mortality rates and extract patterns in subgroup mortality. We generated 10 years of population data for 25 simulated counties of various sizes composed of 5 population subgroups ranging in size from 10% to 50% of the total population. We then generated deaths for these subgroups jointly under a variety of correlation patterns. Specific details on the data generating process are provided in Appendix A.1.

After generating the data, we ran the model and compared posterior estimates of log-mortality rates and correlation values against the corresponding simulation study parameters. We found that the model successfully estimated both sets of parameters. Detailed results are provided in Appendix A.2.

4.2 US county mortality

We applied the model to estimate sex-specific mortality at the county-level in the US over the period 1982-2018.

In the US, relatively high quality data are collected through vital registration systems. The National Center for Health Statistics (NCHS) publishes detailed mortality records for each year included in our series from which we can compile the mortality tabulations. All required variables are available in the public data for years 1982 to 2003. As all geographic identifiers have been suppressed in the public data for years 2004 onward, for purposes of the present analysis, we obtained the geographic identifiers for this period from the NCHS through a restricted-access data agreement with the National Association for Public Health Statistics and Information Systems (NAPHSIS), giving us access to records from 1989 to 2018. The United States Census Bureau publishes mid-year population estimates by county of residence, year, sex, age. These data are available to the public through the Census Bureau. To comply with our data agreement, counties are anonymized in all figures.

The model is applied to each state individually using the first four principal components plotted in Figure 2. Note that estimates based on this model are published as part of the United States Mortality DataBase (USMDB) (https://usa.mortality.org/).

As an illustration of the model outputs, in Figure 4 we plot the estimated sex-specific log-mortality curves in 1982 and 2018 for a subset of US counties along with associated uncertainty intervals. The observed log-mortality rates for age groups with non-zero deaths are presented as points, and the female and male county-level estimates and 95% credible intervals are presented in red and blue respectively. The selected counties represent settings with small, medium, and large populations. As expected, smaller counties such as County 1 have greater uncertainty as compared to larger ones such as County 3 due to the higher levels of noise caused by low or zero death counts. For County 3, modelled estimates follow the data exactly.

4.2.1 State-level patterns in sex-specific mortality

In addition to producing mortality rates for subnational areas, the proposed model specification contains a number of parameters that are useful in observing broad patterns in mortatity at the state-level. Specifically, trends in μ_β, the state-level principal component coefficients, describe structural themes in mortality patterns over time and the correlation matrices offer insight into county-level trends.

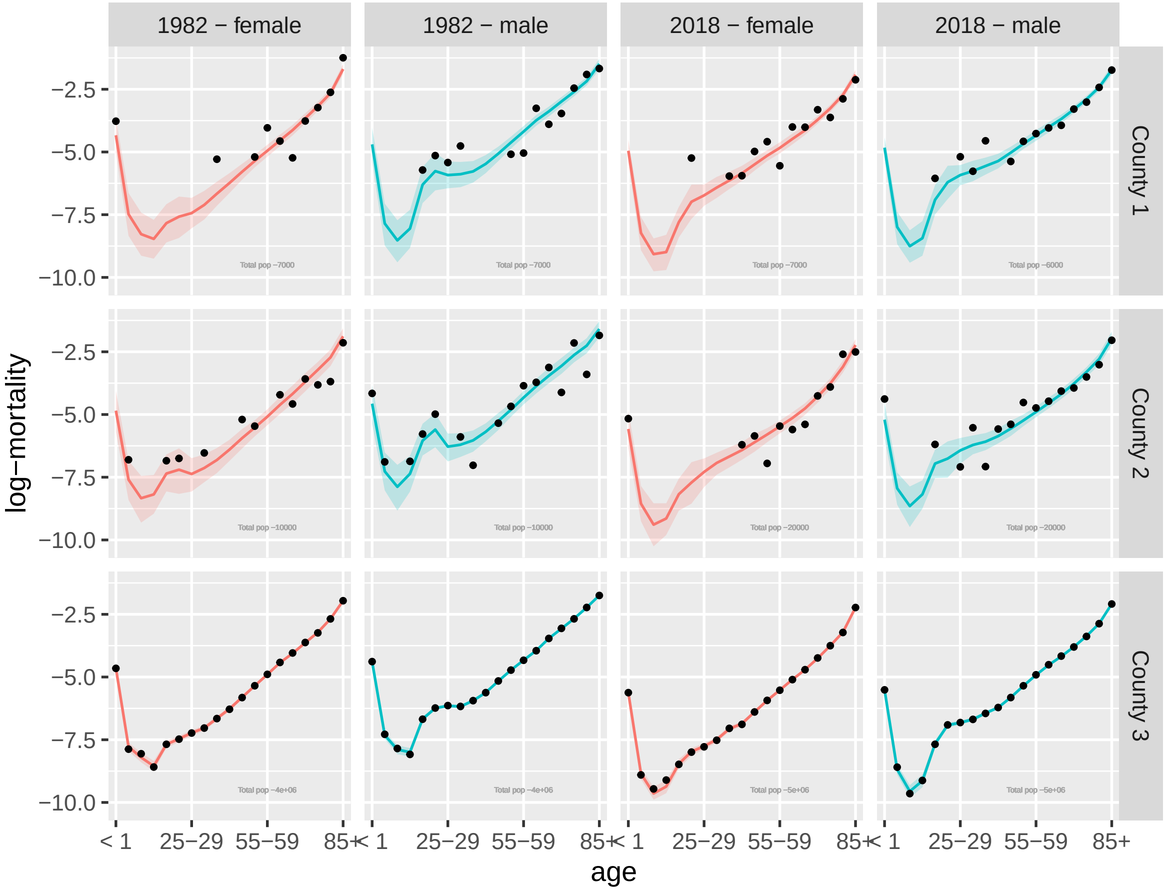


Figure 4: Examples of US county-level sex-specific log-mortality plots by county-year-sex for three counties of varying sizes. Black dots represent observed values and coloured curves and regions indicate posterior medians and 95% credible intervals respectively.

In Figures 5 and 6, we plot posterior medians and 95% credible intervals of the estimated state-level coefficients for the first and second principal components for male and female populations in all states and years. The most notable finding is that in the plot for the first principal component, the gap between baseline mortality coefficients is shrinking over time. This suggests a convergence in mortality across the sexes (Seligman, Greenberg, and Tuljapurkar 2016).

Geographically, we see that baseline mortality patterns differ substantially by state, suggesting evidence for a divergence in mortality across US states due to stagnation or regression in some states (Fenelon 2013). For example, states such as California and New Jersey experienced significant declines in baseline mortality in both sexes. Others such as Louisiana experienced limited declines and have stagnated in the 2010s. Some states such as Ohio saw regression in baseline mortality in the 2010s. This trend reversal occurs across the eastern half of the US.

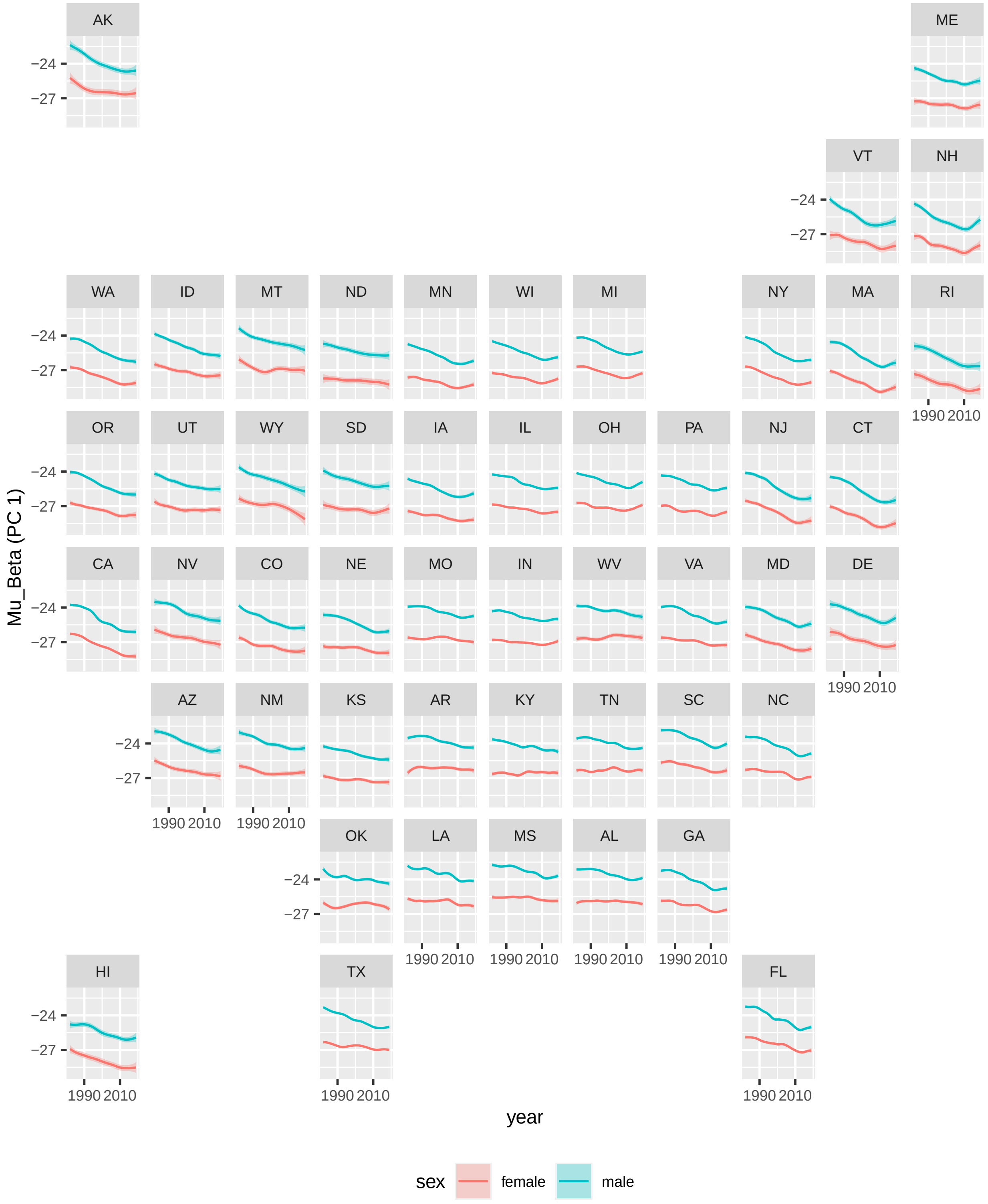


Figure 5: Posterior medians and 95% credible intervals for the state-level coefficients for the first principal component (μ_β).

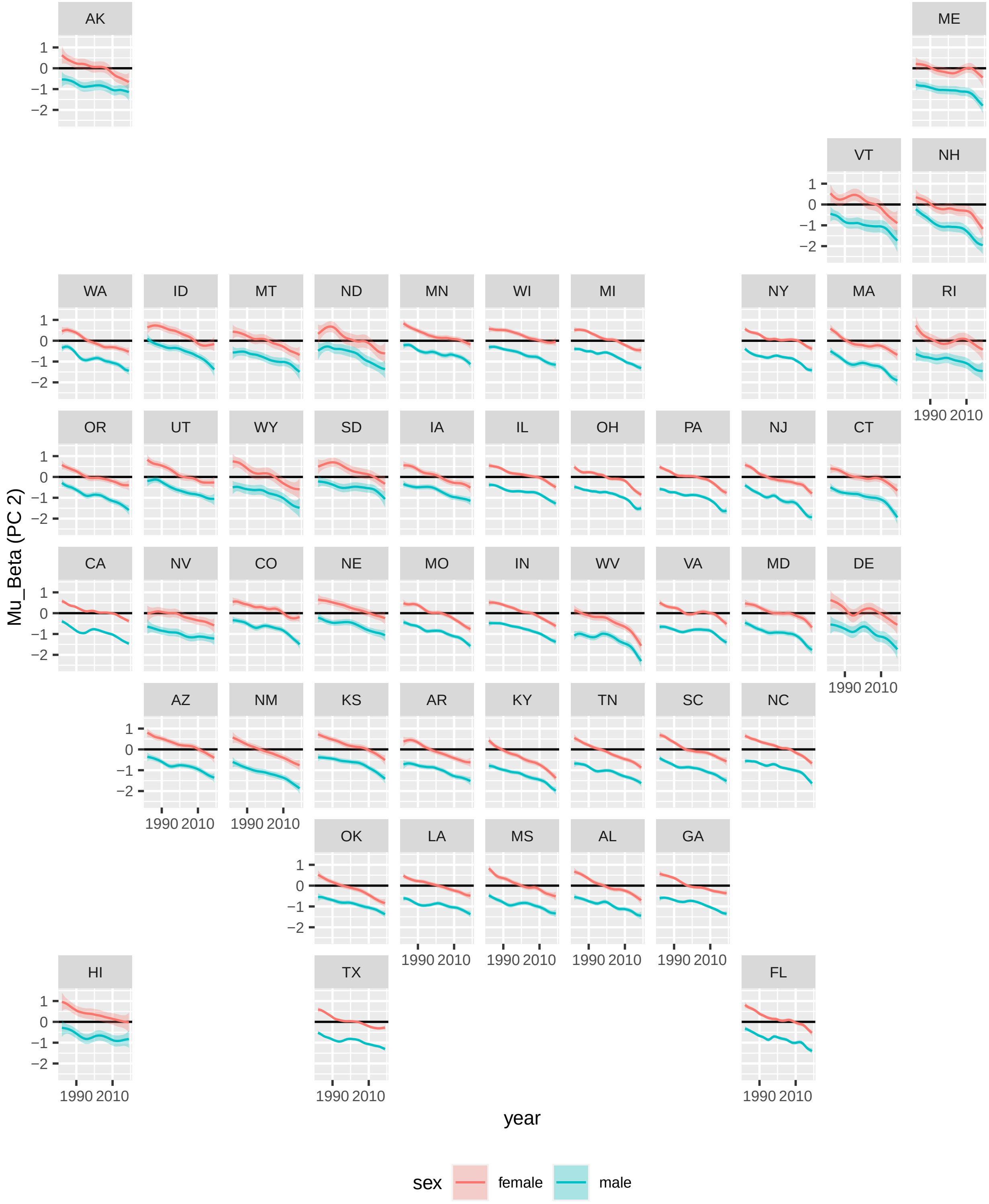


Figure 6: Posterior medians and 95% credible intervals for the second principal component (μ_β).

With the second principal component, declines are experienced across the US in both sexes with female values transitioning from positive to negative. As the second principal component contains an inverted accident hump shape, this suggests increasing young adult mortality relative to the baseline. Our findings suggest that this trend accelerated in the 2010s for males in New Jersey, West Virginia, Ohio, and more broadly across the Northeast and Midwest. This is consistent with the sharp increase in opioid overdose deaths in these regions of the US (Alexander, Kiang, and Barbieri 2018).

For the final two principal components (shown in Appendix B) there are again clear geographic clusters of patterns over time. For the third principal component, which has relatively higher mortality in mid adult ages, trends are generally declining, but stagnating more in the South, and there is evidence of a trend reversal in Kentucky and West Virginia. Alaska specifically also offers interesting results. Unlike the rest of the states, Alaska historically had no significant gap in the third principal component suggesting that it contributed similarly to the mortality patterns of both sexes. However, it is important to recognize that the population in Alaska is small, leading to large uncertainty in the extracted trends.

Figure 7 plots the posterior values of the between sex principal component correlations extracted from $L_{i,t}^{(\beta)}L_{i,t}^{(\beta)\top}$ for a subset of states over time. Plots for all states are given in Appendix B. Each series includes the associated uncertainty intervals along with a horizontal line at zero. Across states, the first principal component has high correlation reinforcing the notion that baseline mortality for both sexes move together. In contrast, there is a limited relationship in the second principal component. This is consistent with the idea that the accident hump is a predominantly male phenomenon, which would imply weak correlation between sexes. The third and fourth principal component correlations deviate by state. For Alaska, the relationships are weak at best, for California and Texas, there are strong or growing correlations, and New Jersey has declining correlations.

In Figure 8, we plot the posterior medians and 95% credible intervals for the correlations in the overdispersion terms captured by $L_{a,t}^{(\gamma)}L_{a,t}^{(\gamma)\top}$ for California. We find that patterns vary substantially by age group. For the 85+ category, consistent high correlation is found. However, in the youth age groups, there does not appear to be any meaningful correlation. The most interesting patterns are those in the mid adult age groups where transitions from limited correlation to strong positive correlation are found.

4.2.2 Validation

To formally evaluate the model, we perform out-of-sample validation exercises comparing the present model with the model proposed by Alexander, Zagheni, and Barbieri (2017). By comparing these two models, we can focus on the contributions of the between subpopulation correlation matrices introduced. The validation

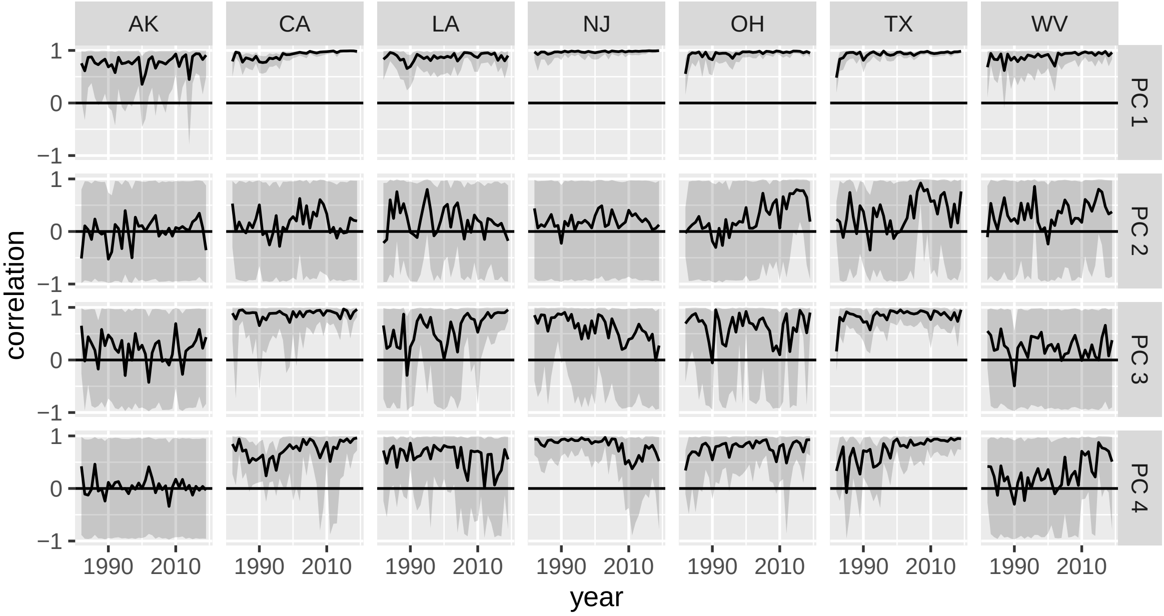


Figure 7: Time-series of posterior principal component correlations for seven states.

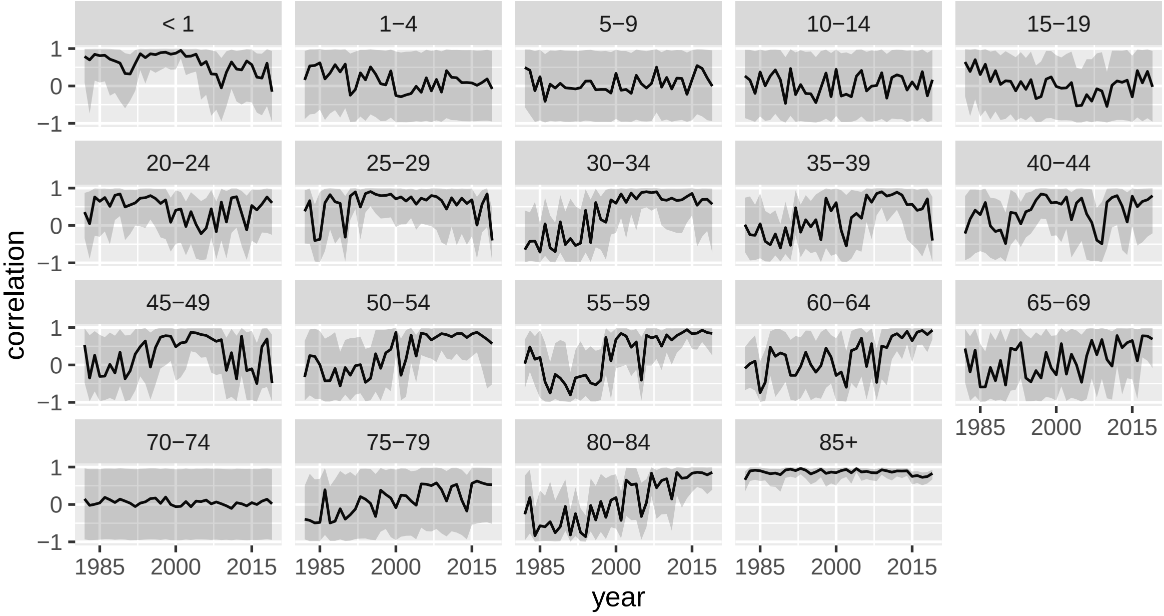


Figure 8: Posterior medians and 95% credible intervals for γ correlation over time by age group for California.

exercises focus on five states (Alaska, California, Louisiana, New Jersey, and Texas) that capture the diversity in population size and number of counties across the US.

For each state, we leave out 20% of the observed data for each county in the years 1982-2018 as an out-of-sample dataset, then execute the model on the remaining in-sample dataset. We then generate a distribution of deaths for all left-out observations using the posterior log-mortality rates for the corresponding county-age-sex-years. Finally, we calculate coverage at the 80%, 90%, and 95% nominal levels for the death uncertainty intervals as well as mean squared errors (MSE) and mean absolute deviations (MAD) between the median death estimates and the observed deaths. In Table 1, the results of this exercise by state are presented. Note that the model introduced in this paper is labeled as the “joint” model and the version in Alexander, Zagheni, and Barbieri (2017) with independent sex modelling is denoted “independent”.

Table 1: Out-of-Sample Coverage and Errors

State	Model	Cov80	Cov90	Cov95	MAD	MSE
Alaska	independent	0.884	0.943	0.973	1.931	14.447
Alaska	joint	0.887	0.946	0.974	1.899	12.614
California	independent	0.843	0.920	0.961	9.747	2004.650
California	joint	0.846	0.923	0.963	8.568	1194.914
Louisiana	independent	0.873	0.939	0.970	3.054	40.795
Louisiana	joint	0.874	0.940	0.972	2.921	30.257
New Jersey	independent	0.843	0.925	0.963	8.388	298.082
New Jersey	joint	0.850	0.930	0.964	7.838	238.458
Texas	independent	0.878	0.940	0.970	3.232	92.449
Texas	joint	0.879	0.941	0.970	3.030	61.618

The joint model consistently outperforms the independent model on the out-of-sample set across metrics. The level of outperformance is most pronounced in the larger states, notably California. The outperformance of the joint model on the out-of-sample set suggests that it may have uses beyond the US county context for estimating subnational mortality by subpopulation in jurisdictions without complete data.

5 Discussion

In this paper we extended principal component-based methods to jointly estimate subnational mortality across subpopulations. This approach leverages the inherent structural mortality patterns associated with the individual principal components and extracts correlations between groups to offer insight into the joint movements of mortality trends across groups. The model centers on a regression-based framework with four principal components that allow core patterns in age-specific mortality to be captured. The principal component coefficients are modelled hierarchically to allow for information exchange within states. County-specific effects on each component are assumed to be correlated across subgroups within a county.

Our approach is validated using both a simulation study and with out-of-sample exercises using real subnational mortality data. We find that the proposed model is well calibrated and that its errors are smaller in the out-of-sample exercises.

We illustrate the model through estimating US county-level sex-specific mortality rates. An investigation into the parameters of the model highlights general trends in US mortality as well as state specific patterns. Notably, while movements in baseline mortality tend to manifest in both sexes simultaneously, specific features of the mortality curve such as elevated young adult mortality appear independent by sex. These results can offer direction to those working to reduce mortality pressures. For example, when correlations are high, aggregate policies may be effective. However, in jurisdictions where correlations are low or zero, more targeted policies may be necessary. State-level parameter estimates also highlighted clear geographic clustering of mortality patterns over time. Notably, there is evidence of stagnation in parts of the country, with increasing early adult mortality particularly in the eastern states, and a stagnating improvement in middle-age mortality in the southern states.

We identify two directions for future extensions to this work. The first is to study how the model performs in real-world circumstances similar to the simulation study where there are more than two subpopulations and the composition of the population is not approximately balanced across groups. For example, race/ethnicity-based mortality data in US counties. Many counties are dominated by one subgroup, and thus the frequency of low or zero death data is significantly higher than with aggregate or sex-specific data. The second direction is to apply the model in contexts where there are differences in mortality data collection or death registration coverage. In many countries, complete mortality data such as the US county data used here are not available and death registration coverage can vary substantially by sex, geographical area, or time (Peralta et al. 2019; Basu and Adair 2021). Typically, male coverage is higher than female coverage. By studying relationships between sexes in the years and regions with higher quality data, one may be able to support estimation of female mortality rates by using the male data and the modelled correlations.

6 References

Alexander, Monica, and Leontine Alkema. 2021. "A Bayesian Cohort Component Projection Model to Estimate Adult Populations at the Subnational Level in Data-Sparse Settings." https://doi.org/10.48550/arXiv.2102.06121.

Alexander, Monica, Mathew V Kiang, and Magali Barbieri. 2018. "Trends in Black and White Opioid Mortality in the United States, 1979–2015." *Epidemiology (Cambridge, Mass.)* 29 (5): 707.

Alexander, Monica, Emilio Zagheni, and Magali Barbieri. 2017. "A Flexible Bayesian Model for Estimating Subnational Mortality." *Demography* 54. https://doi.org/10.1007/s13524-017-0618-7.

Arias, Elizabeth, Loraine A. Escobedo, Jocelyn Kennedy, Chunxia Fu, and Jodi Cisewki. 2018. "U.s. Small-Area Life Expectancy Estimates Project: Methodology and Results Summary." *Vital and Health Statistics Series 2* 181. https://www.cdc.gov/nchs/data/series/sr_02/sr02_181.pdf.

Basu, J. K., and T. Adair. 2021. "Have inequalities in completeness of death registration between states in India narrowed during two decades of civil registration system strengthening?" *Int J Equity Health* 20 (1): 195.

Bhutta, Zulfiqar A. 2016. "Mapping the geography of child mortality: a key step in addressing disparities." *The Lancet Global Health* 4 (12): E877–78. https://doi.org/10.1016/S2214-109X(16)30264-9.

Bijak, J., and J. Bryant. 2016. "Bayesian demography 250 years after Bayes." *Population Studies* 70 (1): 1–19. https://doi.org/10.1080/00324728.2015.1122826.

Clark, Samuel J. 2019. "A General Age-Specific Mortality Model with an Example Indexed by Child Mortality or Both Child and Adult Mortality." *Demography* 56 (3). https://doi.org/10.1007/s13524-019-00785-3.

Coale, Ansley J., Paul Demeny, and Barbara Vaughan, eds. 1983. Second Edition. Academic Press. https://doi.org/10.1016/C2013-0-07295-7.

de Beer, Joop. 2012. "Smoothing and Projecting Age-Specific Probabilities of Death by TOPALS." *Demographic Research* 27: 543–92. http://www.jstor.org/stable/26349934.

Dyrting, Sigurd. 2020. "Smoothing migration intensities with P-TOPALS." *Demographic Research* 43 (55): 1607–50. https://doi.org/10.4054/DemRes.2020.43.55.

Fenelon, Andrew. 2013. "Geographic Divergence in Mortality in the United States." *Population and Development Review* 39 (4): 611–34. https://doi.org/10.1111/j.1728-4457.2013.00630.x.

Gelman, Andrew, and Donald B. Rubin. 1992. "Inference from Iterative Simulation Using Multiple Sequences." *Statistical Science* 7 (4): 457–72. http://www.jstor.org/stable/2246093.

Heligman, L., and J. H. Pollard. 1980. "The Age Pattern of Mortality." *Journal of the Institute of Actuaries* 107 (1): 49–80. https://doi.org/10.1017/S0020268100040257.

Hoffman, Matthew D., and Andrew Gelman. 2014. "The No-u-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research* 15 (47): 1593–623. http://jmlr.org/papers/v15/hoffman14a.html.

Lee, Ronald D., and Lawrence R. Carter. 1992. "Modeling and Forecasting u. S. Mortality." *Journal of the American Statistical Association* 87 (419): 659–71. http://www.jstor.org/stable/2290201.

Lewandowski, Daniel, Dorota Kurowicka, and Harry Joe. 2009. "Generating Random Correlation Matrices Based on Vines and Extended Onion Method." *Journal of Multivariate Analysis* 100 (9): 1989–2001. https://doi.org/https://doi.org/10.1016/j.jmva.2009.04.008.

Mercer, Laina D., Jon Wakefield, Athena Pantazis, Angelina M. Lutambi, Honorati Masanja, and Samuel Clark. 2015. "Space-Time Smoothing of Complex Survey Data: Small Area Estimation for Child Mortality." *The Annals of Applied Statistics* 9 (December): 1889–1905. https://doi.org/10.1214/15-AOAS872.

Neal, Radford M. 2011. "MCMC Using Hamiltonian Dynamics." In *Handbook of Markov Chain Monte Carlo*. Chapman; Hall/CRC. https://doi.org/10.1201/b10905.

Peralta, A., J. Benach, C. Borrell, V. Espinel-Flores, L. Cash-Gibson, B. L. Queiroz, and M. Marí-Dell'Olmo. 2019. "Evaluation of the mortality registry in Ecuador (2001-2013) - social and geographical inequalities in completeness and quality." *Popul Health Metr* 17 (1): 3.

Raftery, Adrian E., Nevena Lalic, and Patrick Gerland. 2014. "Joint probabilistic projection of female and male life expectancy." *Demographic Research* 30 (27): 795–822. https://doi.org/10.4054/DemRes.2014.30.27.

Rau, Roland, and Carl P. Schmertmann. 2020. "Lebenserwartung auf Kreisebene in Deutschland." *Dtsch Arztebl International* 117 (29-30): 493–99. https://doi.org/10.3238/arztebl.2020.0493.

Schmertmann, Carl P., and Marcos R. Gonzaga. 2016. "Estimation of Mortality Rates by Age and Sex for Small Areas with TOPALS Regression: An Application for Brazil in 2010." *Revista Brasileira De Estudos De População* 33 (3): 629–52. https://doi.org/10.20947/S0102-30982016c0009.

———. 2018. "Bayesian Estimation of Age-Specific Mortality and Life Expectancy for Small Areas With Defective Vital Records." *Demography* 55 (4): 1363–88. https://doi.org/10.1007/s13524-018-0695-2.

Seligman, Benjamin, Gabi Greenberg, and Shripad Tuljapurkar. 2016. "Convergence in male and female life expectancy: Direction, age pattern, and causes." *Demographic Research* 34 (38): 1063–74. https://doi.org/10.4054/DemRes.2016.34.38.

Ševčíková, Hana, and Adrian E. Raftery. 2021. "Probabilistic Projection of Subnational Life Expectancy." *Journal of Official Statistics* 37 (3): 591–610. https://doi.org/doi:10.2478/jos-2021-0027.

Stan Development Team. 2021. "RStan: The R Interface to Stan." https://mc-stan.org/.

Wang, Haidong, Austin E. Schumacher, Carly E. Levitz, Ali H. Mokdad, and Christopher JL Murray. 2013. "Left Behind: Widening Disparities for Males and Females in US County Life Expectancy, 1985–2010."

Population Health Metrics 11 (8). https://doi.org/10.1186/1478-7954-11-8.

A Simulation study details

A.1 Simulation setup

To illustrate the model's ability to estimate mortality rates and extract mortality patterns across subgroups, we simulate population and death data for populations composed of five subgroups of varying sizes. The simulation study data includes population and mortality data for 10 years, 25 subnational areas, and 5 population subgroups. For each year-area-age-subgroup, we construct the population data as follows:

1. The total population in year 1 in each area is set as 100,000 times the area index, which ranges from 1-25.
2. The total population in each area in subsequent years is computed by increasing the total population in each area by 1% annually.
3. Total populations in each year-area are allocated to the 19 age groups using a jittered version of the age distribution of Los Angeles County in California.
4. Each year-area-age population is then disaggregated into the 5 subgroups using the following proportions: 50% in Group A, 20% in Group B, and 10% in Groups C, D, E.

To generate corresponding mortality data for the populations in each year-area-age-subgroup, we first construct true log-mortality curves for each year-area-subgroup. The mortality curves are constructed using a linear combination of two standard curves: the first representing a baseline mortality curve, and the second containing an accident hump. Plots of the standard curves are given in Figure 9.

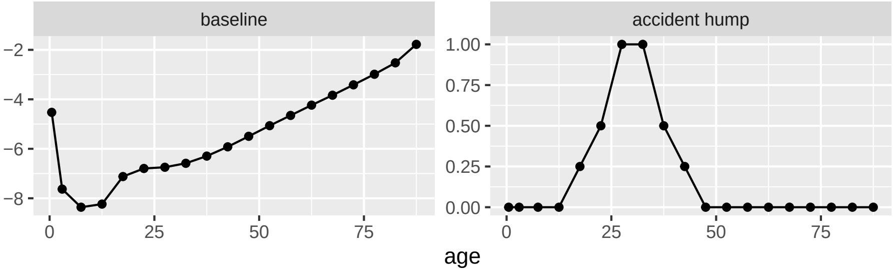


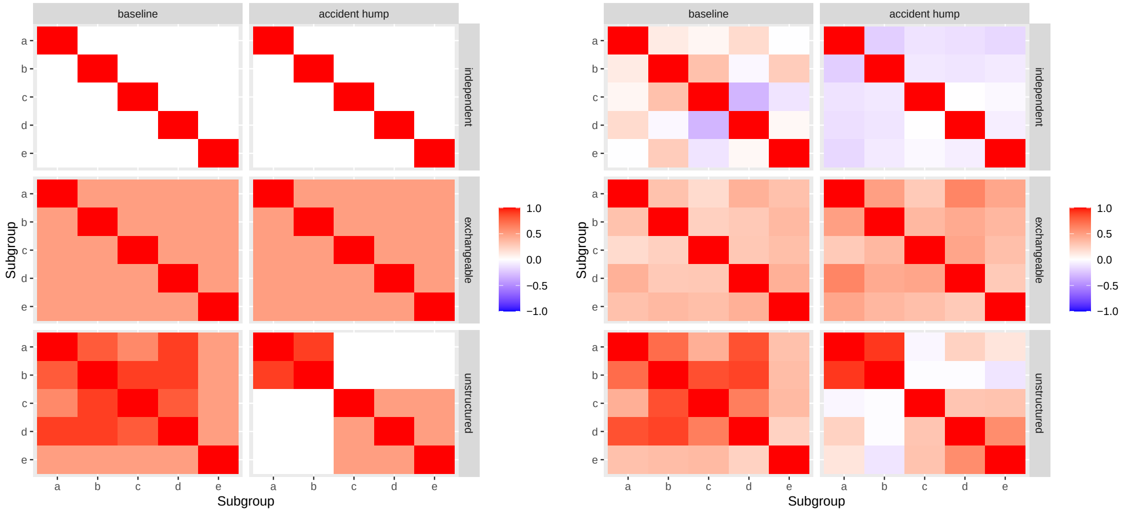
Figure 9: Simulation study mortality curve components

The log-mortality curves are then constructed with the following steps:

1. The baseline mortality curve is multiplied by a random coefficients sampled from a 5-dimensional multivariate normal distribution with mean $\mathbf{1}_5$ and a year-dependent covariance matrix. The covariance matrix describes the dependence between the 5 subgroups. It is assigned a constant variance of 0.1^2 and one of the correlation structures plotted in Figure 10a. In years 1-3, the independent correlation matrix

is used. In years 4-6, the exchangeable correlation matrix (ie. constant off-diagonal values) is used. In years 7-10, the unstructured correlation matrix based on state-level observed race/ethnicity-based US mortality is used.

2. An accident hump is added to the mortality curve by adding the accident hump standard curve multiplied by random coefficients sampled from a 5-dimensional multivariate normal distribution with mean $\mathbf{0}_5$, and covariance matrix with a constant variance of 0.5^2 and one of the correlation matrices plotted in Figure 10. The matrices used in each year mirror those used for the baseline mortality standard.



(a) True correlation matrices

(b) Estimated correlation matrices

Figure 10: Simulation study true and estimated posterior median correlation matrices.

Finally, observed deaths for each year-area-age-subgroup are generated from the log-mortality curves using a Poisson likelihood with rate equal to the corresponding population multiplied by the exponential of the corresponding log-mortality rate.

Examples of simulated data in two subnational areas are given in Figure 11. Points correspond to log-mortality observations (ie. the log of observed deaths divided by population). Note that zero death observations are encoded as -10 on the log-scale.

A.2 Results

The model is run using the standard curves as the principal components. To validate the results, we compare the estimated correlation matrices and log-mortality rates to the corresponding true values.

Figure 10b plots estimated correlation matrices extracted from the model. Comparing corresponding facets

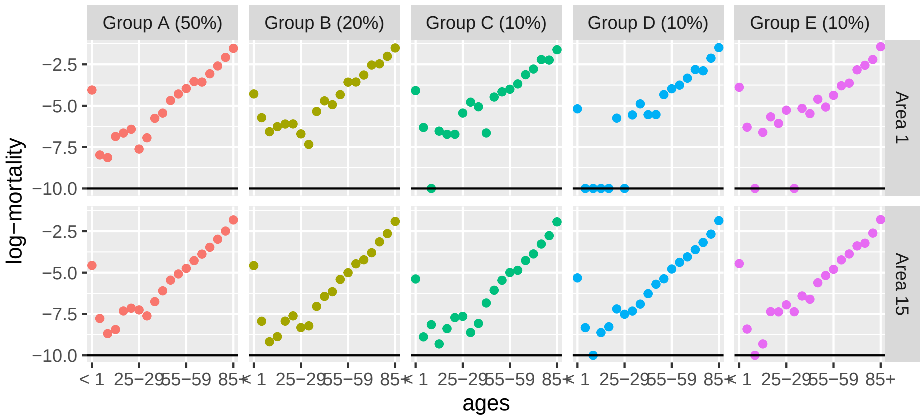


Figure 11: Examples of simulated log-mortality observations.

Table 2: Simulation study coverage values for correlation and log-mortality rates.

	Coverage (80%)	Coverage (90%)	Coverage (95%)
Correlations	0.78	0.90	0.94
Log-mortality rates	0.83	0.92	0.96

against Figure 10a, it appears that the model has successfully identified and extracted the patterns observed in the data.

Table 2 presents the coverage values for correlation and log-mortality rate parameters at the 80%, 90%, and 95% nominal levels. Coverage is defined for correlation and log-mortality rates as $\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{l_i\leq\theta_i\leq u_i}$ where n is the number of parameters, θ_i refers to an individual parameter, and l_i and u_i are respectively the lower and upper bounds of the posterior intervals. For the correlation matrices, we compute entrywise coverage values for the off-diagonals of the lower triangles of the correlation matrices.

For both sets of parameters, the model's coverage values are in line with the nominal values at all levels, suggesting that the model is well calibrated.

B Additional results

Figure 12 plots two county-years where the amount of non-zero death data varies by sex. The observed log-mortality rates for age groups with non-zero deaths are presented as points, the estimated state means for each sex are denoted by the grey lines, and the male and female county-level estimates and 95% credible intervals are presented in blue and red respectively. In both county-years, the majority of age groups had zero female deaths in the year whereas the male data is nearly complete. As opposed to defaulting to the female state-level means, estimates for female mortality are informed by the behaviour of male mortality relative to the male state-level means.

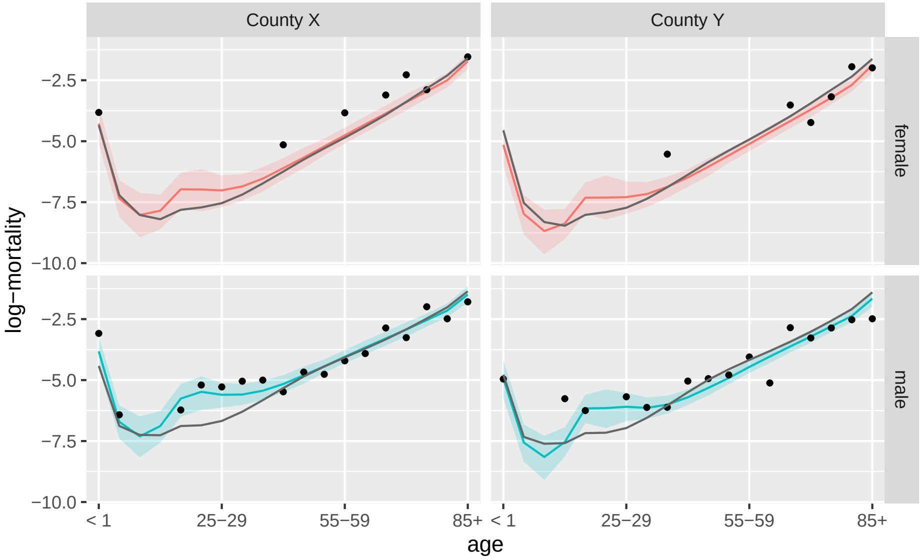


Figure 12: Examples of US counties with limited female but sufficient male data.

Figures 13 and 14 plot maps of the posterior medians and 95% credible intervals of the estimated state-level coefficients for the third and fourth principal components for male and female populations for all years.

Figures 15, 16, 17, and 18 plot the posterior medians and 95% credible intervals for principal component coefficient correlations between sexes over time. As expected, strong correlation in the first principal component is found in most states whereas there is negligible correlation in the second principal component. Patterns in the third and fourth principal component coefficient correlations vary regionally with some evidence for stronger correlations in the fourth principal component coefficients in the Great Lakes region.

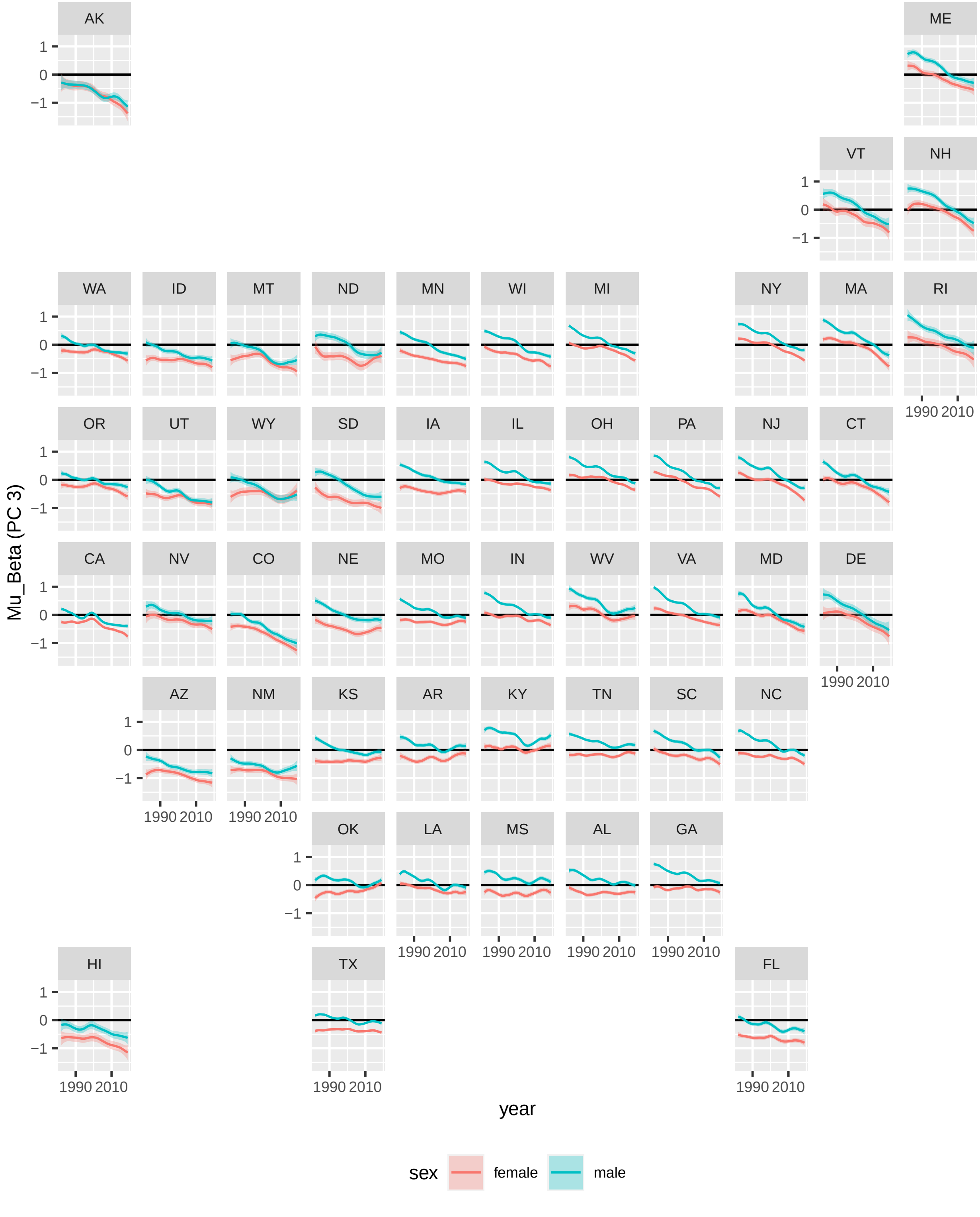


Figure 13: Posterior medians and 95% credible intervals for the third principal component (μ_β).

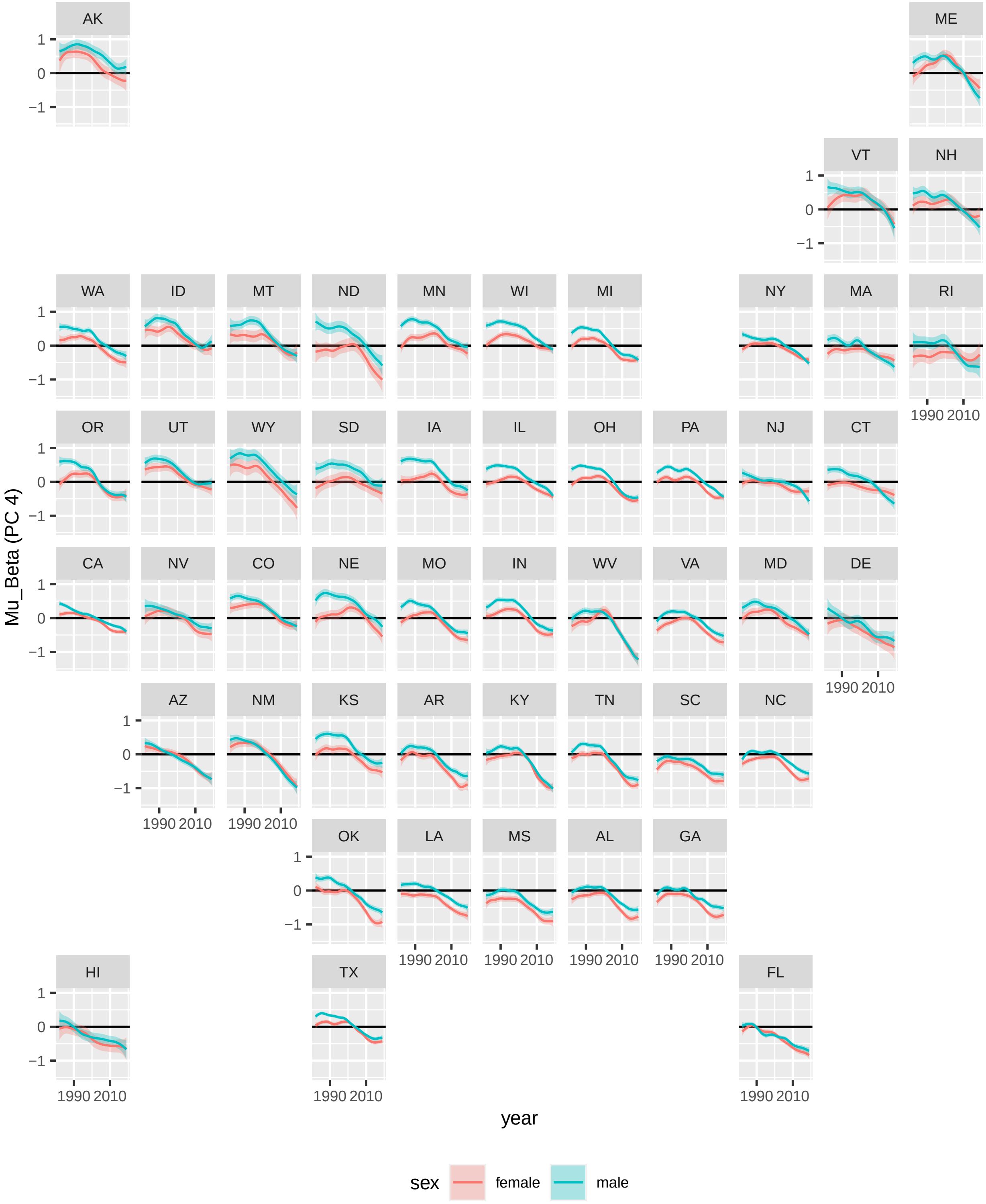


Figure 14: Posterior medians and 95% credible intervals for the fourth principal component (μ_β).

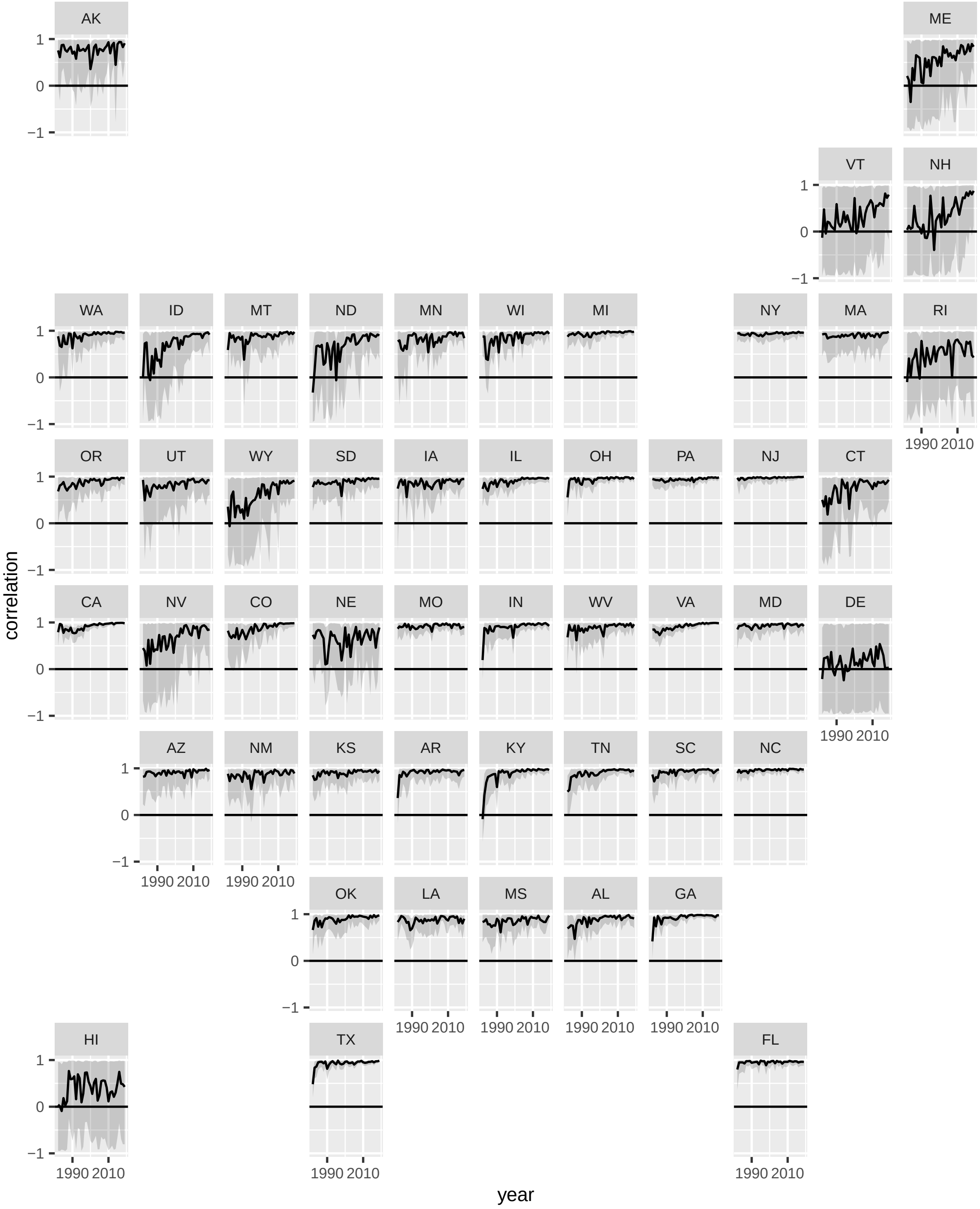


Figure 15: Time-series of posterior medians and 95% credible intervals for the first principal component correlations.

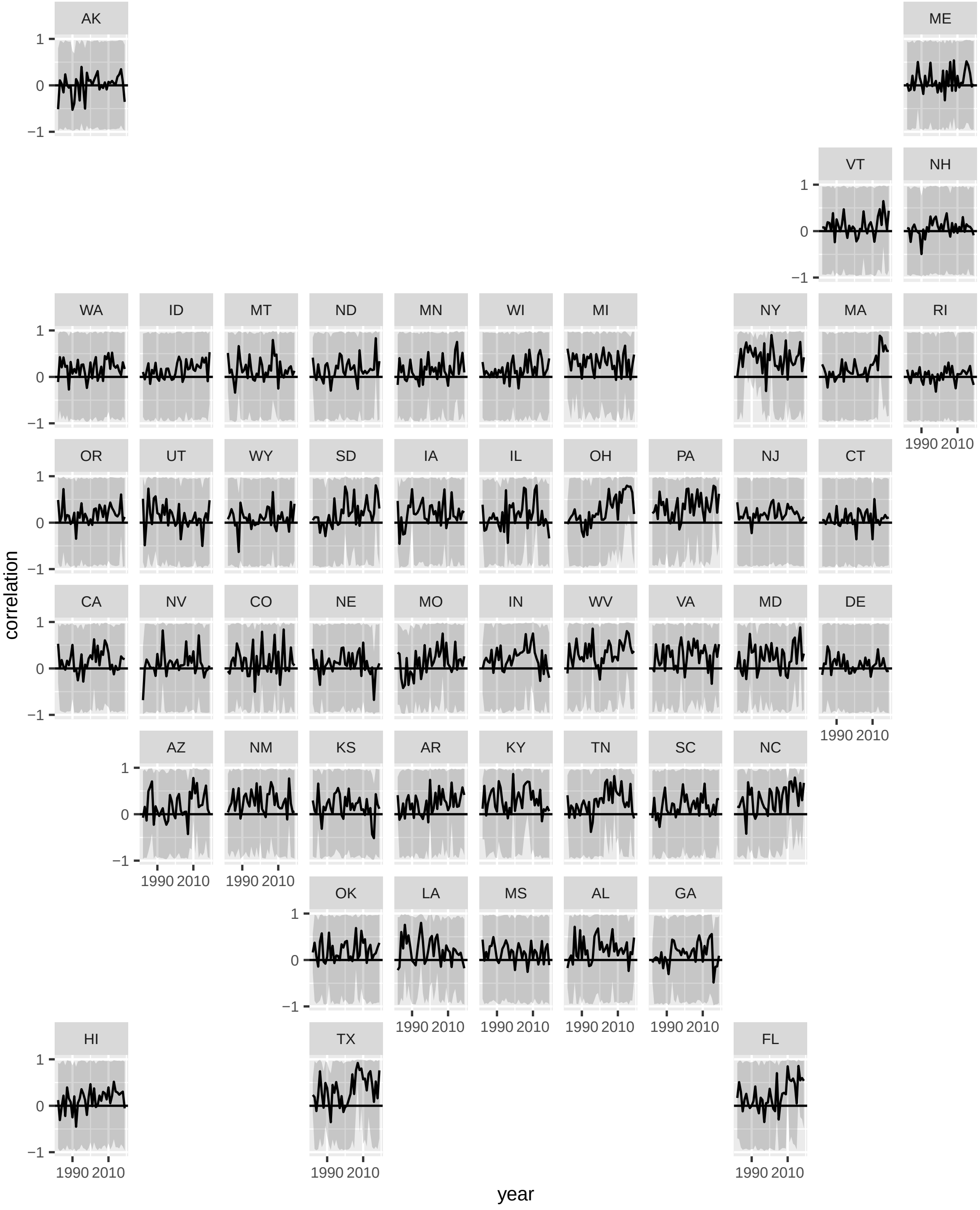


Figure 16: Time-series of posterior medians and 95% credible intervals for the second principal component correlations.

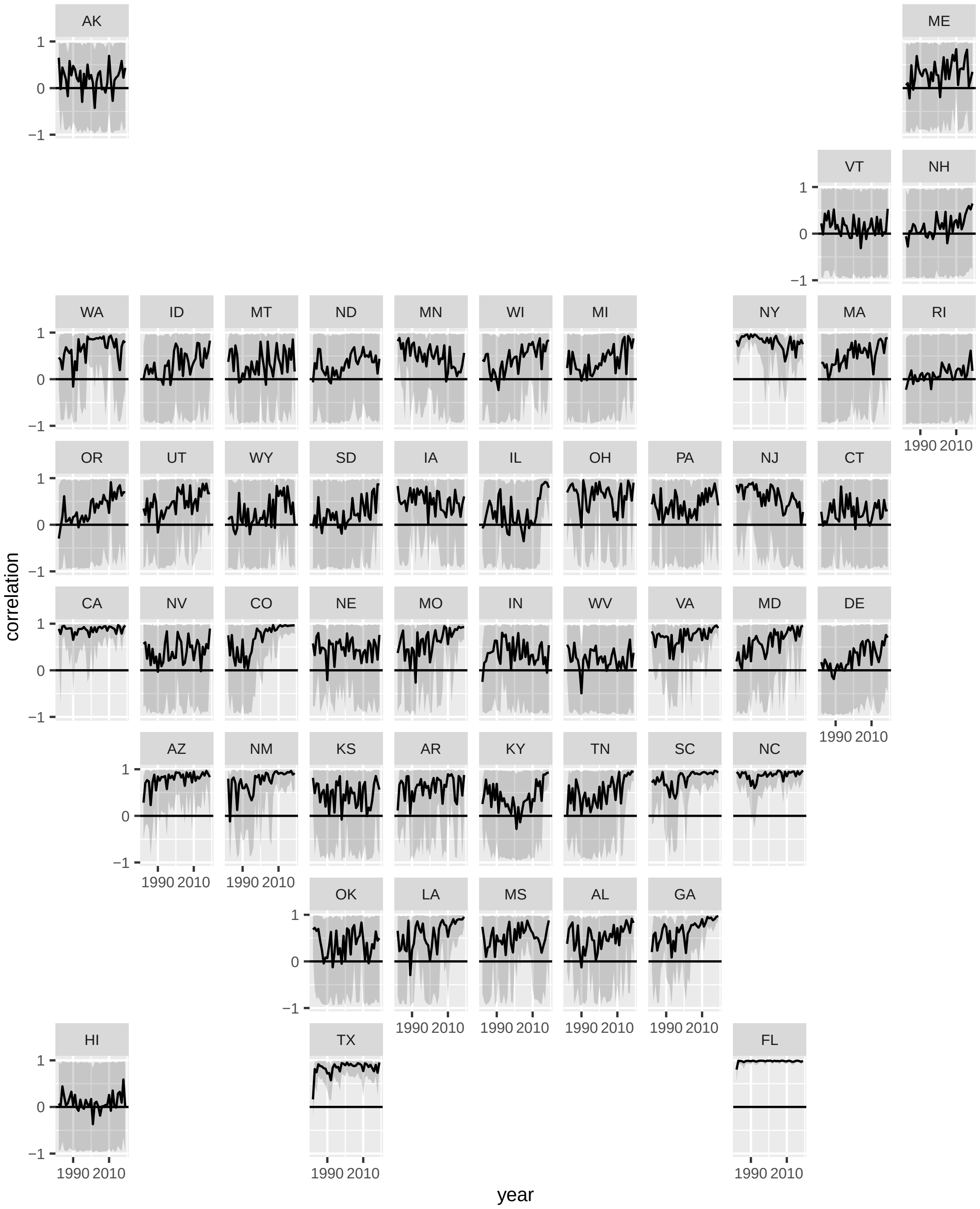


Figure 17: Time-series of posterior medians and 95% credible intervals for the third principal component correlations.

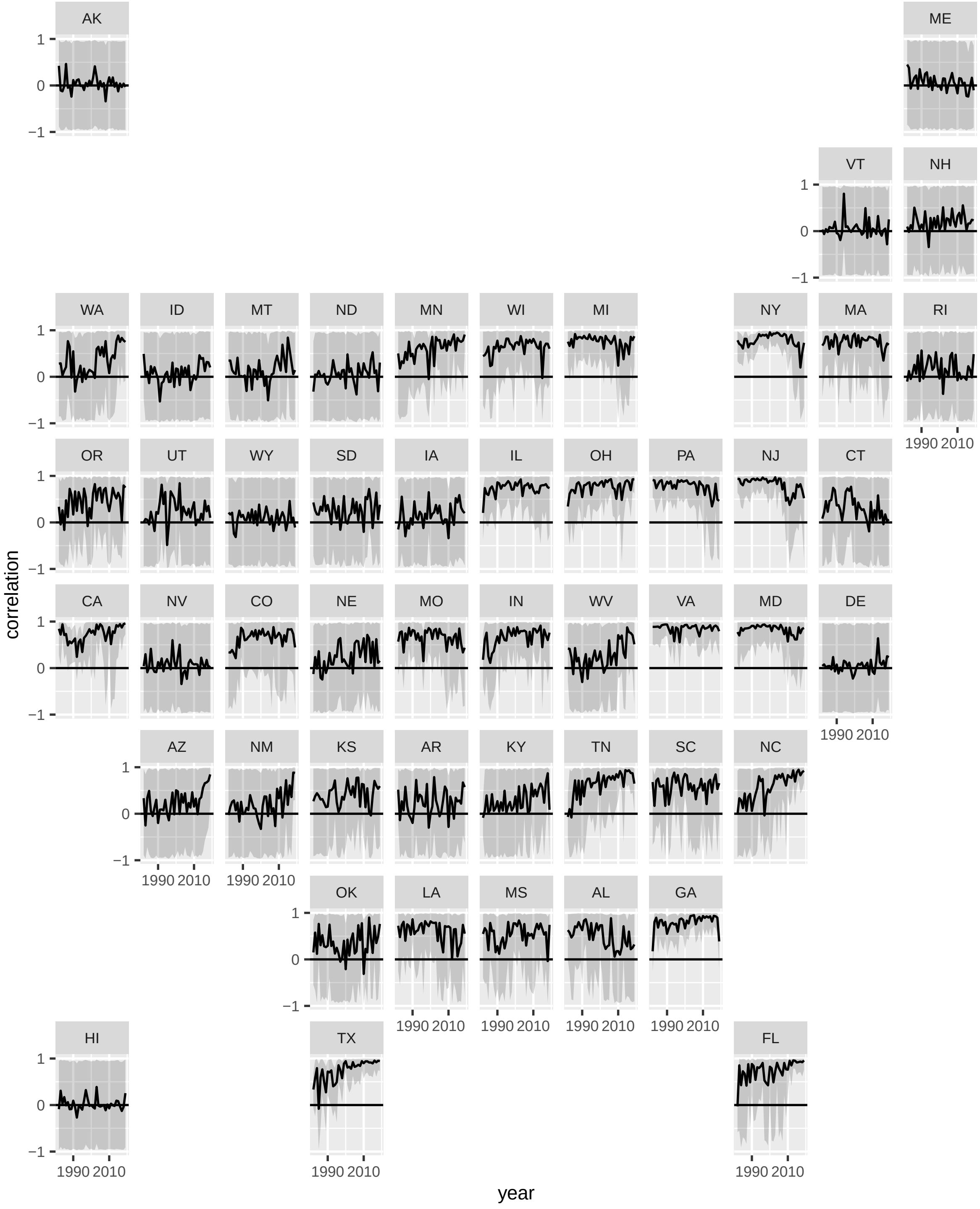


Figure 18: Time-series of posterior medians and 95% credible intervals for the fourth principal component correlations.