Estimating demographic and global health indicators for multiple countries and periods in the context of missing data and data quality issues

Introducing a class of temporal models for multiple populations to facilitate model comparison

Herb Susmann[a], Monica Alexander[b], Leontine Alkema[a]
[a]University of Massachusetts Amherst, [b]University of Toronto

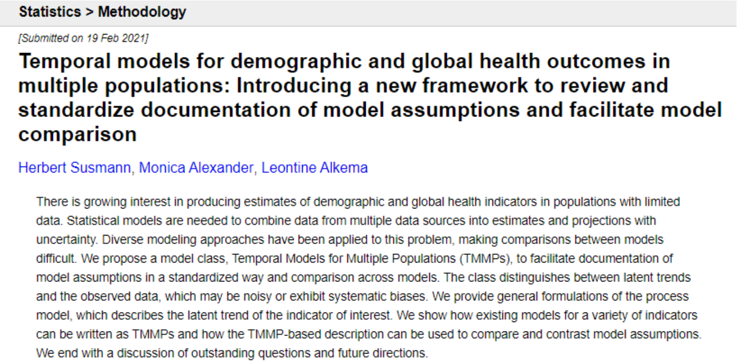
PAA 2021 Annual Meeting
May 6, 2021

This work was supported by the Bill & Melinda Gates Foundation.

Resources

UMass Amherst

- Annotated slides: http://herbsusmann.com/paa2021
- Preprint: https://arxiv.org/abs/2102.10020



Statistics > Methodology

[Submitted on 19 Feb 2021]

Temporal models for demographic and global health outcomes in multiple populations: Introducing a new framework to review and standardize documentation of model assumptions and facilitate model comparison

Herbert Susmann, Monica Alexander, Leontine Alkema

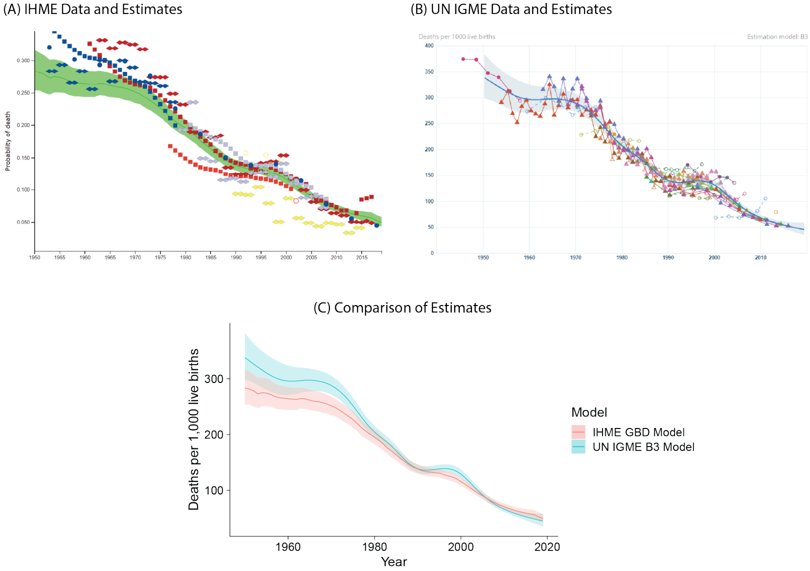
There is growing interest in producing estimates of demographic and global health indicators in populations with limited data. Statistical models are needed to combine data from multiple data sources into estimates and projections with uncertainty. Diverse modeling approaches have been applied to this problem, making comparisons between models difficult. We propose a model class, Temporal Models for Multiple Populations (TMMPs), to facilitate documentation of model assumptions in a standardized way and comparison across models. The class distinguishes between latent trends and the observed data, which may be noisy or exhibit systematic biases. We provide general formulations of the process model, which describes the latent trend of the indicator of interest. We show how existing models for a variety of indicators can be written as TMMPs and how the TMMP-based description can be used to compare and contrast model assumptions. We end with a discussion of outstanding questions and future directions.

Background

- There is interest in modeling demographic and health indicators in order to measure progress towards international goals.
- Data availability and quality are varied.
- Many statistical models have been created to provide estimates and projections.
- Comparing across models can be difficult.
- Proposed overarching model class: Temporal Models for Multiple Populations (TMMPs).

Case Study: U5MR

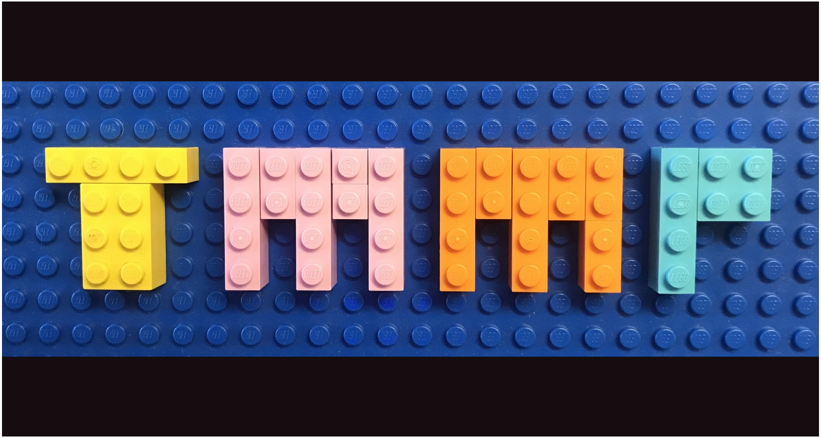
Under-five Mortality Rate Estimates in Senegal, 1950-2019



(A) IHME Data and Estimates

(B) UN IGME Data and Estimates

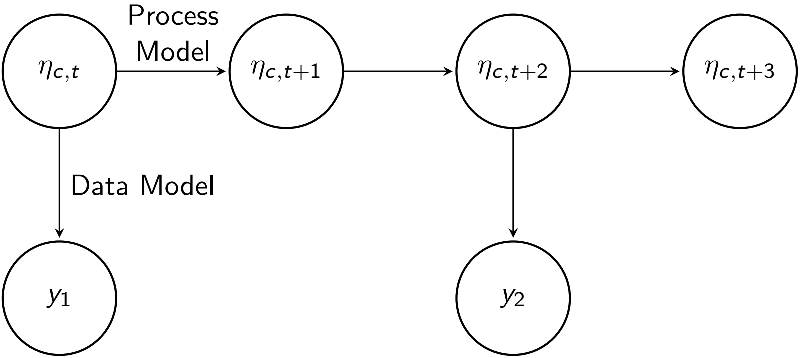
(C) Comparison of Estimates



Modeling Framework

- True value of indicator: $\eta_{c,t}$ for $c = 1, \ldots, C$, $t = 1, \ldots T$.
- *Process model* describes evolution of $\eta_{c,t}$.
  - Covariates
  - Systematic trends
- Observed data y_i, with associated properties $c[i]$, $t[i]$, $s[i]$, ...
- *Data model* describes relationship between y_i and $\eta_{c[i],t[i]}$.

Modeling Framework



Process Model

$$g_1(\eta_{c,t}) = \underbrace{g_2(X_{c,t}, \beta_c)}_{\text{covariate}} + \underbrace{g_3(t, \eta_{c,s\neq t}, \alpha_c)}_{\text{systematic}} + \underbrace{a_{c,t}}_{\text{offset}} + \underbrace{\epsilon_{c,t}}_{\text{smoothing}}$$

Covariate component

$$g_1(\eta_{c,t}) = \underbrace{g_2(\boldsymbol{X}_{c,t}, \boldsymbol{\beta}_c)}_{\text{covariate}} + \underbrace{g_3(t, \eta_{c,s\neq t}, \boldsymbol{\alpha}_c)}_{\text{systematic}} + \underbrace{a_{c,t}}_{\text{offset}} + \underbrace{\epsilon_{c,t}}_{\text{smoothing}}$$

- Regression function for incorporating covariates.
- Example: IHME U5MR

$$g_2(\boldsymbol{X}_{c,t}, \boldsymbol{\beta}_c) = \exp\left[\beta_{c,1} \cdot \log(X_{c,t}^{LDI}) + \beta_{2,c} \cdot X_{c,t}^{EDU} + \beta_{3,c}\right] + \beta_{4,c} x_{c,t}^{HIV}$$

Systematic component

$$g_1(\eta_{c,t}) = \underbrace{g_2(X_{c,t}, \beta_c)}_{\text{covariate}} + \underbrace{g_3(t, \eta_{c,s\neq t}, \boldsymbol{\alpha}_c)}_{\text{systematic}} + \underbrace{a_{c,t}}_{\text{offset}} + \underbrace{\epsilon_{c,t}}_{\text{smoothing}}$$

- Parametric function for modeling systematic temporal trends.
- Example: modeling the rate of change in adoption of modern family planning as following logistic growth [Cahill et al., 2018].

Offset

$$g_1(\eta_{c,t}) = \underbrace{g_2(X_{c,t}, \beta_c)}_{\text{covariate}} + \underbrace{g_3(t, \eta_{c,s\neq t}, \alpha_c)}_{\text{systematic}} + \underbrace{a_{c,t}}_{\text{offset}} + \underbrace{\epsilon_{c,t}}_{\text{smoothing}}$$

- The offset term incorporates external information, for example from a separate modeling step.
- Example: IHME U5MR model uses an offset derived from the smoothed residuals of a separate mixed-effects model.

Smoothing component

$$g_1(\eta_{c,t}) = \underbrace{g_2(X_{c,t}, \beta_c)}_{\text{covariate}} + \underbrace{g_3(t, \eta_{c,s\neq t}, \alpha_c)}_{\text{systematic}} + \underbrace{a_{c,t}}_{\text{offset}} + \underbrace{\epsilon_{c,t}}_{\text{smoothing}}$$

- The smoothing component allows data-driven deviations from the other components, while still enforcing smoothness.
- Many choices B-splines, Gaussian processes, AR(p), RW(p), spatial smoothing (ICAR), ...
- Model class introduces some additional structure to help understand the smoothing component.

Smoothing component

- Define $\boldsymbol{\epsilon}_c = [\epsilon_{c,1}, \cdots, \epsilon_{c,T}]$.
- Smoothing model defined as

$$\boldsymbol{\epsilon}_c = \boldsymbol{B}_c \boldsymbol{\delta}_c,$$

where $\boldsymbol{B}_c$ is a full rank matrix, and

$$\triangle_r \boldsymbol{\delta}_c \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_c\right)$$

with $\boldsymbol{\Sigma}_c$ defined via an autocovariance function $s(t_1, t_2)$.

Smoothing component

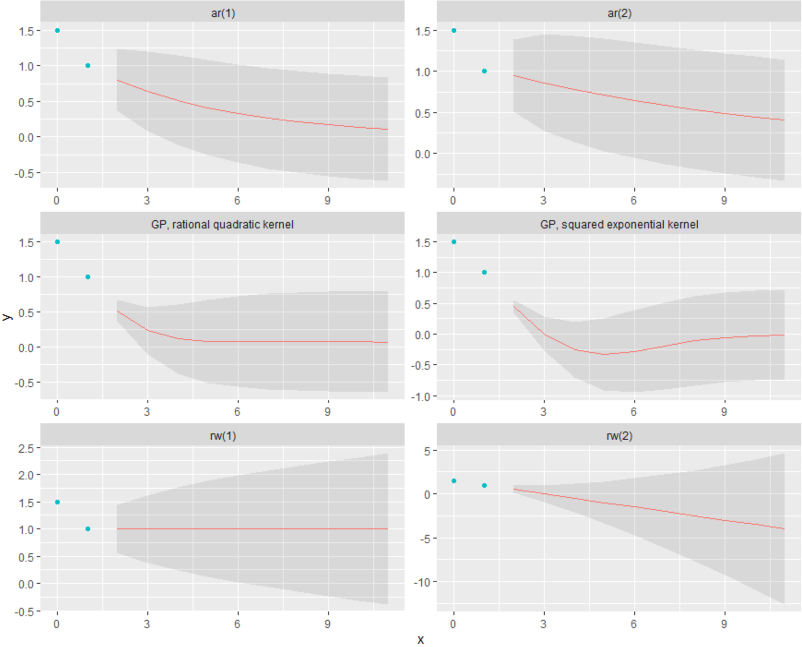
- UN-IGME: RW(2), $r = 2$, cubic B-splines with $\boldsymbol{B}_{c,t,k} = b_{c,k}(t)$, and

$$s(t_1, t_2) = \kappa^2 I(t_1 = t_2)$$

- IHME: Matérn Gaussian Process, $r = 0$, $\boldsymbol{B}_c = \boldsymbol{I}$, and

$$s_{\text{Matérn}}(t_1, t_2) = \kappa^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{|t_1 - t_2|}{\ell} \right)^{\nu} K_{\nu} \left(\sqrt{2\nu} \frac{|t_1 - t_2|}{\ell} \right)$$

Smoothing component



Estimation

- Each component introduces many unit specific parameters that need to be estimated.
- Hierarchical modeling is a way to share information between units.
- UN-IGME: hierarchical model on RW(2) variance
- IHME: fixed smoother hyperparameters based on data availability

UN-IGME U5MR Model

- Process model [Alkema and New, 2014]

$$\log(\eta_{c,t}) = \underbrace{g_3(t, \eta_{c,s\neq t}, \boldsymbol{\alpha}_c)}_{\text{systematic}} + \underbrace{\epsilon_{c,t}}_{\text{smoothing}}$$

- Systematic component during estimation period:

$$g_3(t, \eta_{c,s\neq t}, \boldsymbol{\alpha}_c) = \alpha_{c,0} + \alpha_{c,1}(t - t_c^*)$$

- Smoothing component during estimation period: cubic B-splines with K_c knots per country and RW(2) process on spline coefficients

$$\boldsymbol{\epsilon}_c = \boldsymbol{B}_c \boldsymbol{\delta}_c,$$

where after two levels of differencing ($r = 2$), $\boldsymbol{\delta}_c$ is normally distributed with mean zero:

$$\Delta_2 \boldsymbol{\delta}_c \sim N(\boldsymbol{0}, \sigma^2_{\delta,c}\boldsymbol{I}).$$

IHME U5MR Model

- Process model [Dicker et al., 2018]

$$\log_{10}(\eta_{c,t}) = \underbrace{g_2(\boldsymbol{X}_{c,t}, \boldsymbol{\beta}_c)}_{\text{covariate}} + \underbrace{a_{c,t}}_{\text{offset}} + \underbrace{\epsilon_{c,t}}_{\text{smoothing}}$$

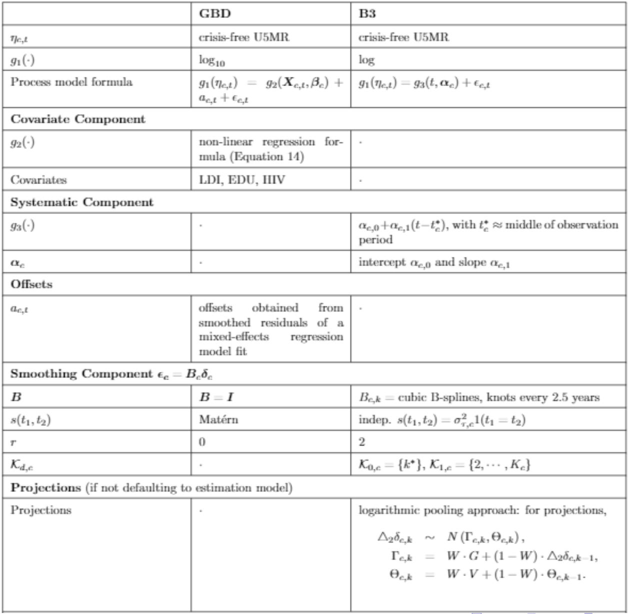
- Covariate component:

$$g_2(\boldsymbol{X}_{c,t}, \boldsymbol{\beta}_c) = \exp\left[\beta_{c,1} \cdot \log(X_{c,t}^{LDI}) + \beta_{2,c} \cdot X_{c,t}^{EDU} + \beta_{3,c}\right] + \beta_{4,c} X_{c,t}^{HIV}$$

- Offset: adjusts covariate component using smoothed residuals from separate mixed-effects model.
- Smoother: Gaussian process with no transformation or differencing ($\boldsymbol{B} = \boldsymbol{I}, \epsilon_c = \boldsymbol{\delta}_c,\ r = 0$) and Matérn covariance function

$$\boldsymbol{\delta}_c \sim N(\mathbf{0}, \boldsymbol{\Sigma}_c).$$

U5MR Model Comparison



	GBD	**B3**
$\eta_{c,t}$	crisis-free U5MR	crisis-free U5MR
$g_1(\cdot)$	$\log_{10}$	log
Process model formula	$g_1(\eta_{c,t}) = g_2(\boldsymbol{X}_{c,t}, \boldsymbol{\beta}_c) + a_{c,t} + \epsilon_{c,t}$	$g_1(\eta_{c,t}) = g_3(t, \boldsymbol{\alpha}_c) + \epsilon_{c,t}$
Covariate Component		
$g_2(\cdot)$	non-linear regression formula (Equation 14)	$\cdot$
Covariates	LDI, EDU, HIV	$\cdot$
Systematic Component		
$g_3(\cdot)$	$\cdot$	$\alpha_{c,0} + \alpha_{c,1}(t - t_c^*)$, with $t_c^* \approx$ middle of observation period
$\boldsymbol{\alpha}_c$	$\cdot$	intercept $\alpha_{c,0}$ and slope $\alpha_{c,1}$
Offsets		
$a_{c,t}$	offsets obtained from smoothed residuals of a mixed-effects regression model fit	$\cdot$
Smoothing Component $\boldsymbol{\epsilon}_c = \boldsymbol{B}_c \boldsymbol{\delta}_c$		
$\boldsymbol{B}$	$\boldsymbol{B} = \boldsymbol{I}$	$B_{c,k}$ = cubic B-splines, knots every 2.5 years
$s(t_1, t_2)$	Matérn	indep. $s(t_1, t_2) = \sigma^2_{\tau,c} 1(t_1 = t_2)$
r	0	2
$\mathcal{K}_{d,c}$	$\cdot$	$\mathcal{K}_{0,c} = \{k^*\}$, $\mathcal{K}_{1,c} = \{2, \cdots, K_c\}$
Projections (if not defaulting to estimation model)		
Projections	$\cdot$	logarithmic pooling approach: for projections, $\triangle_2 \delta_{c,k} \sim N(\Gamma_{c,k}, \Theta_{c,k})$, $\Gamma_{c,k} = W \cdot G + (1 - W) \cdot \triangle_2 \delta_{c,k-1}$, $\Theta_{c,k} = W \cdot V + (1 - W) \cdot \Theta_{c,k-1}$.

Additional Examples

Our preprint includes additional examples of existing models written using the TMMP notation:

- Family Planning [Cahill et al., 2018]
- Neonatal Mortality [Alexander and Alkema, 2018]
- Maternal Mortality [Alkema et al., 2017]
- Subnational Mortality [Alexander et al., 2017]

We also include a table template for documenting models in TMMP form.

Discussion

- Problem: many existing models using different approaches and notations, hard to compare across them
- We introduce Temporal Models for Multiple Populations (TMMPs)
  - Model class makes a clear distinction between the *process model* and the *data model*.
  - Process model is split into building blocks: covariates, systematic trends, offsets, and smoothing components.
- TMMP framework useful for:
  - Systematizing model documentation,
  - Facilitating comparisons between existing models,
  - Developing new models.
- To improve model comparison, we propose that standardized documentation be considered for GATHER reporting guidelines
- Contact: Herb Susmann (hsusmann@umass.edu, @herbps10)

References

Alexander, M. and Alkema, L. (2018).
Global estimation of neonatal mortality using a Bayesian hierarchical splines regression model.
Demographic Research, 38:335–372.

Alexander, M., Zagheni, E., and Barbieri, M. (2017).
A Flexible Bayesian Model for Estimating Subnational Mortality.
Demography, 54(6):2025–2041.

Alkema, L. and New, J. R. (2014).
Global estimation of child mortality using a Bayesian B-spline Bias-reduction model.
The Annals of Applied Statistics, 8(4):2122–2149.

Alkema, L., Zhang, S., Chou, D., Gemmill, A., Moller, A.-B., Fat, D. M., Say, L., Mathers, C., and Hogan, D. (2017).
A Bayesian approach to the global estimation of maternal mortality.
The Annals of Applied Statistics, 11(3):1245–1274.

Cahill, N., Sonneveldt, E., Stover, J., Weinberger, M., Williamson, J., Wei, C., Brown, W., and Alkema, L. (2018).
Modern contraceptive use, unmet need, and demand satisfied among women of reproductive age who are married or in a union in the focus countries of the Family Planning 2020 initiative: a systematic analysis using the Family Planning Estimation Tool.
The Lancet, 391(10123):870–882.

Dicker, D., Nguyen, G., Abate, D., and other GBD 2017 Mortality Collaborators (2018).
Global, regional, and national age-sex-specific mortality and life expectancy, 1950–2017: a systematic analysis for the Global Burden of Disease Study 2017.
The Lancet, 392(10159):1684–1735.