Estimating Subnational Under-Five Mortality Rates Using a Spatio-Temporal Age-Period-Cohort Model

Connor Gascoigne

Imperial College London

GEOMED, Hasselt, Belgium

9-11 September 2024

MRC
Centre for Environment & Health

Medical Research Council

Imperial College London

Outline

- Introduction
- Methods
- Results
- Conclusion

Introduction

Under-Five Mortality

- SDG 3.2:
  - By 2030, 25 deaths per 1000 live births.
  - In 2022, 37 deaths per 1000 live births.
  - Disproportionate rates in LMICs
    - $14\times$ higher in sub-Saharan Africa than EU and NA.

- How to help achieve this:
  - Subnational estimates and predictions of U5M in LMICs.
  - Identify areas of current and future high mortality.
  - Targeted intervention.
- Challenges:
  - Data source.
  - Data sparsity.
- Previous U5M literature:
  - Survey data for LMIC.
  - Subnational estimates and predictions of U5M.

U5M: Previous Methods

- Weighted (direct) estimates:
  – Horvitz and Thompson, *Journal of the American Statistical Association*, 1952.
  – Weights reflect the survey design.
  – Data sparsity - estimates are unstable and variance is large.

U5M: Previous Methods

- Weighted (direct) estimates:
  - Horvitz and Thompson, *Journal of the American Statistical Association*, 1952.
  - Weights reflect the survey design.
  - Data sparsity - estimates are unstable and variance is large.
- Fay and Herriot estimates:
  - Fay and Herriot, *Journal of the American Statistical Association*, 1979.
  - Models direct estimates and accounts for trends in time (period) and space to decrease variance.
  - Data sparsity - direct estimates still unstable.

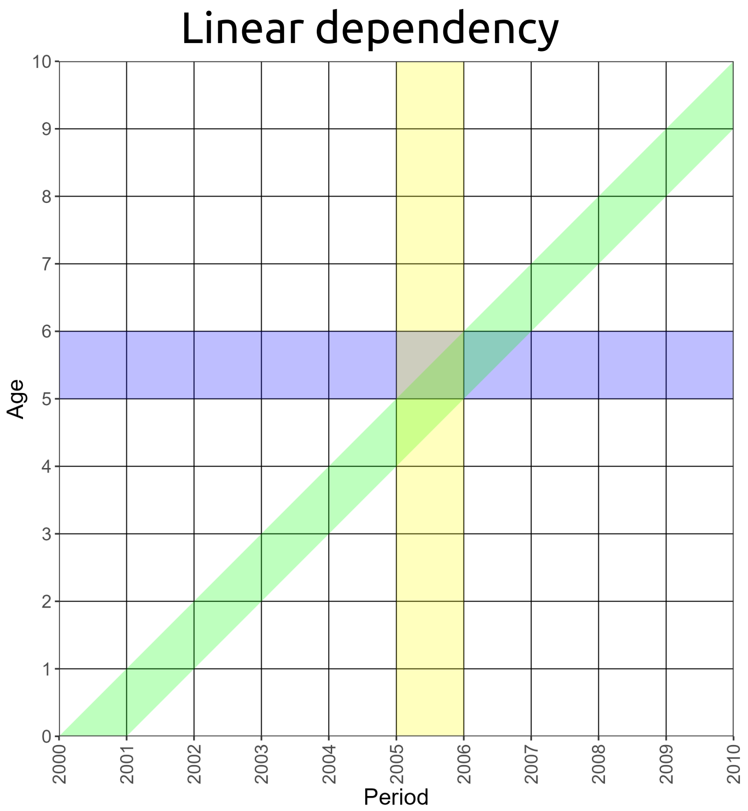
U5M: Previous Methods

- Weighted (direct) estimates:
  - Horvitz and Thompson, *Journal of the American Statistical Association*, 1952.
  - Weights reflect the survey design.
  - Data sparsity - estimates are unstable and variance is large.
- Fay and Herriot estimates:
  - Fay and Herriot, *Journal of the American Statistical Association*, 1979.
  - Models direct estimates and accounts for trends in time (period) and space to decrease variance.
  - Data sparsity - direct estimates still unstable.
- Cluster level model estimates:
  - Wakefield et al., *Statistical Methods in Medical Research*, 2019.
  - Uses random effects to account for survey design and accounts for trends in age, period, and space.
  - Data source - data contains (birth) cohort but not used.

U5M: Previous Methods

- Weighted (direct) estimates:
  - Horvitz and Thompson, *Journal of the American Statistical Association*, 1952.
  - Weights reflect the survey design.
  - Data sparsity - estimates are unstable and variance is large.
- Fay and Herriot estimates:
  - Fay and Herriot, *Journal of the American Statistical Association*, 1979.
  - Models direct estimates and accounts for trends in time (period) and space to decrease variance.
  - Data sparsity - direct estimates still unstable.
- Cluster level model estimates:
  - Wakefield et al., *Statistical Methods in Medical Research*, 2019.
  - Uses random effects to account for survey design and accounts for trends in age, period, and space.
  - Data source - data contains (birth) cohort but not used.
- Why cohort:
  - Facilitates stable estimates and predictions.
  - Produces predcitions less influenced by yearly fluctuations.
  - Include cohort along side age and period for modelling U5M.

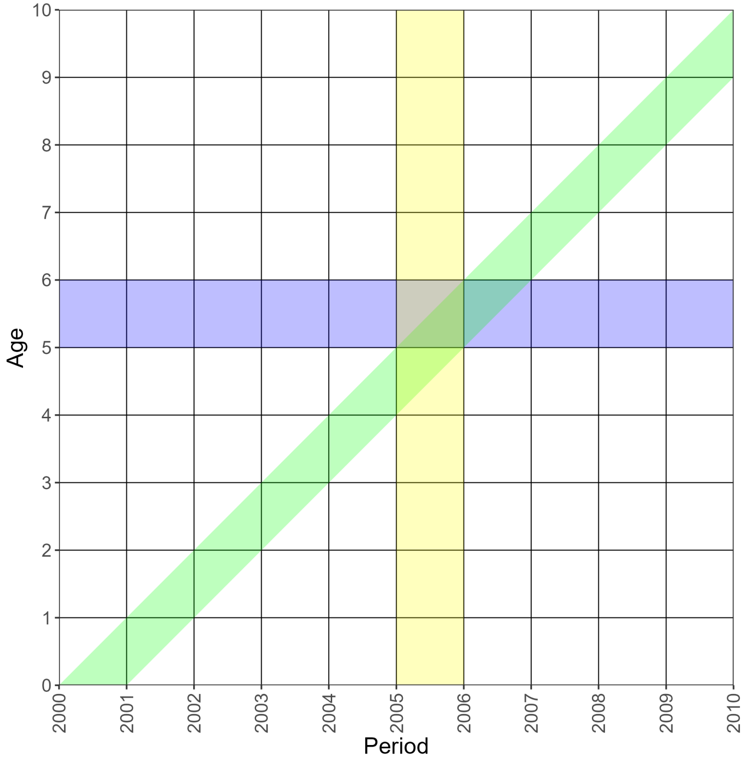
APC: Previous Methods



- Accounted for in numerous ways:
  - Holford, *Biometrics*, 1983.
  - Carstensen, *Statistics in Medicine*, 2006.

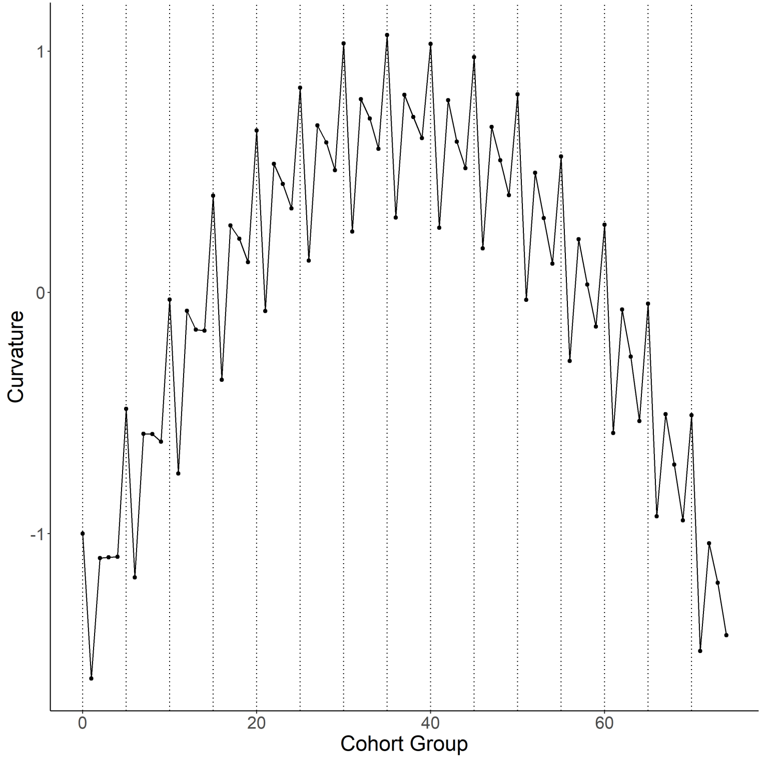
APC: Previous Methods

Linear dependency



- Accounted for in numerous ways:
  - Holford, *Biometrics*, 1983.
  - Carstensen, *Statistics in Medicine*, 2006.

More issues when age, period and cohort are not equally aggregated



- Noted by: Holford, *Statistics in Medicine*, 2006.
- Approach suggested by: Gascoigne and Smith, *Statistics in Medicine*, 2023.

Aims

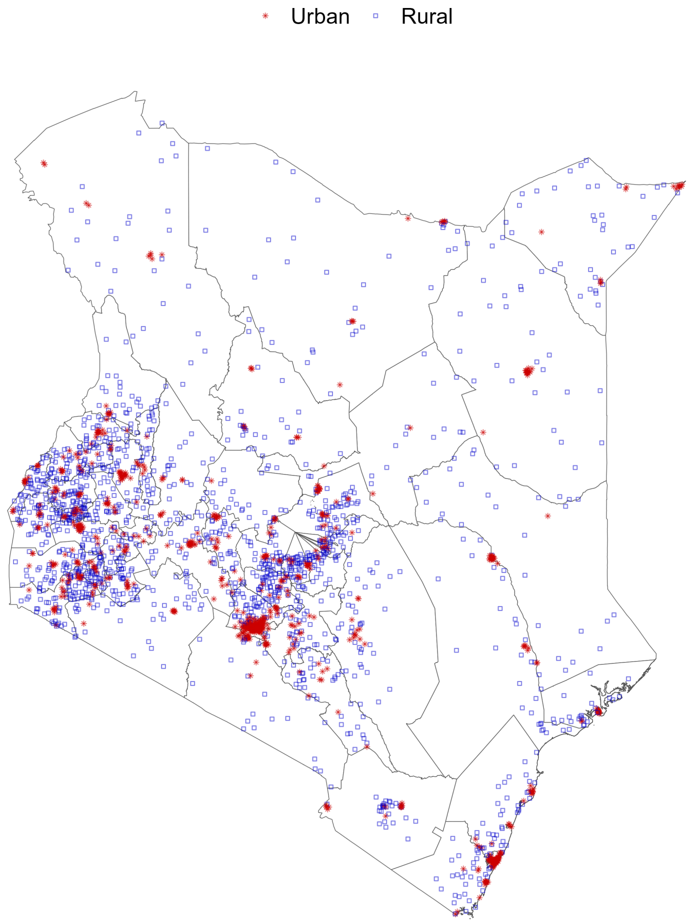
1. Build a flexible model to include cohort alongside the age and period trends to estimate U5M.
   - Account for linear dependency.
   - Account for issues when data are in not equally aggregated.
2. Extend the APC model to account for the complex survey design.
   - Use data from the DHS.

Methods

Data

Demographic Health Survey | Questionnaire and Outcome

- Demographic and Health Surveys (DHS):
  - 400 surveys in 90 countries (incl. sub-Saharan Africa).
- Stratified two-stage cluster design:
  - First stage - Primary Sampling Units (PSUs) or "clusters".
  - Second stage - 25/30 households per cluster.
  - Stratification - geographical region crossed with urban/rural classification.
- We use the 2014 Kenyan DHS (2014 KDHS).
  - 1,612 clusters.
  - 25 households per cluster.
  - 92 stratum.
    - 47 counties (admin-1).
    - 2 all-urban counties (Nairobi and Mombasa).
    - 92 stratum.



Data

Demographic Health Survey | **Questionnaire and Outcome**

- Survey participants:
  - All women aged 15-49 who spent night before in household.
- Survey responses of interest:
  - Cluster Number.
  - Date of birth of child.
  - Child is alive (1 = Yes, 0 = No).
  - Age and death in months.
- Outcome:
  - Mortality counts.
  - Age (monthly), period (yearly), cohort (yearly) and cluster.
- Study period:
  - Data for 2006 - 2013.
  - Predict years 2014 - 2018.

Statistical Model

Let $y_{\tilde{a},p,c,k}$ and $n_{\tilde{a},p,c,k}$ be the number of observed deaths and number of months at risk for age group $\tilde{a}$, period p, cohort c and in cluster k.

$$y_{\tilde{a},p,c,k} | \pi_{\tilde{a},p,c,k}, d \sim \text{BetaBinomial}\left(n_{\tilde{a},p,c,k}, \pi_{\tilde{a},p,c,k}, d\right)$$

$$\text{logit}\left(\pi_{\tilde{a},p,c,k}\right) = \beta_1 + I\left(\boldsymbol{s}_k \in \text{ urban }\right)\beta_2 + t_{1,r}\beta_3 + t_{2,r}\beta_4 + \nu_{\tilde{a}} + \eta_p + \xi_c + S_{r[s_k]} + \delta_{p,r[s_k]}$$

- U5M literature, Wakefield et al., *Statistical Methods in Medical Research*, 2019:
  - Discrete age groups $\tilde{a} = \{0-1, 1-12, 12-24, 24-36, 36-48, 48-60\}$
  - d, overdispersion accounts for clustering.
  - β_1 and $\beta_1 + \beta_2$, intercepts for urban/rural stratification.
  - $S_{r[s_k]}$ accounts for admin-1 region of cluster.
  - $\delta_{p,r[s_k]}$ accounts for space-time interaction of cluster admin-1 region and year (period).

Statistical Model

Let $y_{\tilde{a},p,c,k}$ and $n_{\tilde{a},p,c,k}$ be the number of observed deaths and number of months at risk for age group $\tilde{a}$, period p, cohort c and in cluster k.

$$\begin{aligned} y_{\tilde{a},p,c,k} | \pi_{\tilde{a},p,c,k}, d &\sim \text{BetaBinomial}\left(n_{\tilde{a},p,c,k}, \pi_{\tilde{a},p,c,k}, d\right) \\ \text{logit}\left(\pi_{\tilde{a},p,c,k}\right) &= \beta_1 + I\left(\boldsymbol{s}_k \in \text{ urban }\right)\beta_2 + t_{1,r}\beta_3 + t_{2,r}\beta_4 + \nu_{\tilde{a}} + \eta_p + \xi_c + S_{r[s_k]} + \delta_{p,r[s_k]} \end{aligned}$$

- U5M literature, Wakefield et al., *Statistical Methods in Medical Research*, 2019:
  - Discrete age groups $\tilde{a} = \{0-1, 1-12, 12-24, 24-36, 36-48, 48-60\}$
  - Overdispersion d accounts for clustering.
  - β_1 and $\beta_1 + \beta_2$ account for urban/rural stratification.
  - $S_{r[s_k]}$ accounts for admin-1 region of cluster.
  - $\delta_{p,r[s_k]}$ accounts for space-time interaction of cluster admin-1 region and year (period).
- APC literature, Gascoigne and Smith, *Statistics in Medicine*, 2023:
  - $t_{1,r}\beta_3$ and $t_{2,r}\beta_4$, linear trends of temporal effects.
  - $\nu_{\tilde{a}}$, η_p, and ξ_c, non-linear trends of temporal effects.

Statistical Model

Let $y_{\tilde{a},p,c,k}$ and $n_{\tilde{a},p,c,k}$ be the number of observed deaths and number of months at risk for age group $\tilde{a}$, period p, cohort c and in cluster k.

$$\begin{aligned} y_{\tilde{a},p,c,k} | \pi_{\tilde{a},p,c,k}, d &\sim \text{BetaBinomial}\left(n_{\tilde{a},p,c,k}, \pi_{\tilde{a},p,c,k}, d\right) \\ \text{logit}\left(\pi_{\tilde{a},p,c,k}\right) &= \beta_1 + I\left(\boldsymbol{s}_k \in \text{ urban }\right)\beta_2 + t_{1,r}\beta_3 + t_{2,r}\beta_4 + \nu_{\tilde{a}} + \eta_p + \xi_c + S_{r[s_k]} + \delta_{p,r[s_k]} \end{aligned}$$

- U5M literature, Wakefield et al., *Statistical Methods in Medical Research*, 2019:
  - Discrete age groups $\tilde{a} = \{0-1, 1-12, 12-24, 24-36, 36-48, 48-60\}$
  - Overdispersion d accounts for clustering.
  - β_1 and $\beta_1 + \beta_2$ account for urban/rural stratification.
  - $S_{r[s_k]}$ accounts for admin-1 region of cluster.
  - $\delta_{p,r[s_k]}$ accounts for space-time interaction of cluster admin-1 region and year (period).
- APC literature, Gascoigne and Smith, *Statistics in Medicine*, 2023:
  - 2/3 linear trends of temporal effects $t_{1,r}\beta_3$ and $t_{2,r}\beta_4$.
  - 3/3 non-linear trends of temporal effects $\nu_{\tilde{a}}$, η_p and ξ_c.
- Bayesian paradigm:
  - Priors on all model parameters.
  - Posterior distribution (incl. uncertainty).

Constructing U5MR

General formula:

$$\text{U5MR}_{p,r} = 1 - \prod_{i=1}^{6} \left[1 - \text{expit}\left(\text{logit}\left[\pi_{\tilde{a}(i),p,\bar{c},r}\right]\right)\right]^{z[i]}$$

Constructing U5MR

General formula:

$$\text{U5MR}_{p,r} = 1 - \prod_{i=1}^{6} \left[1 - \text{expit}\left(\text{logit}\left[\pi_{\tilde{a}(i),p,\bar{c},r}\right]\right)\right]^{z[i]}$$

Due to urban and rural intercepts:

$$\text{U5MR}_{p,r,\text{urban}} \quad \text{and} \quad \text{U5MR}_{p,r,\text{rural}}$$

Constructing U5MR

General formula:

$$\text{U5MR}_{p,r} = 1 - \prod_{i=1}^{6} \left[1 - \text{expit}\left(\text{logit}\left[\pi_{\tilde{a}(i),p,\bar{c},r}\right]\right)\right]^{z[i]}$$

Due to urban and rural intercepts:

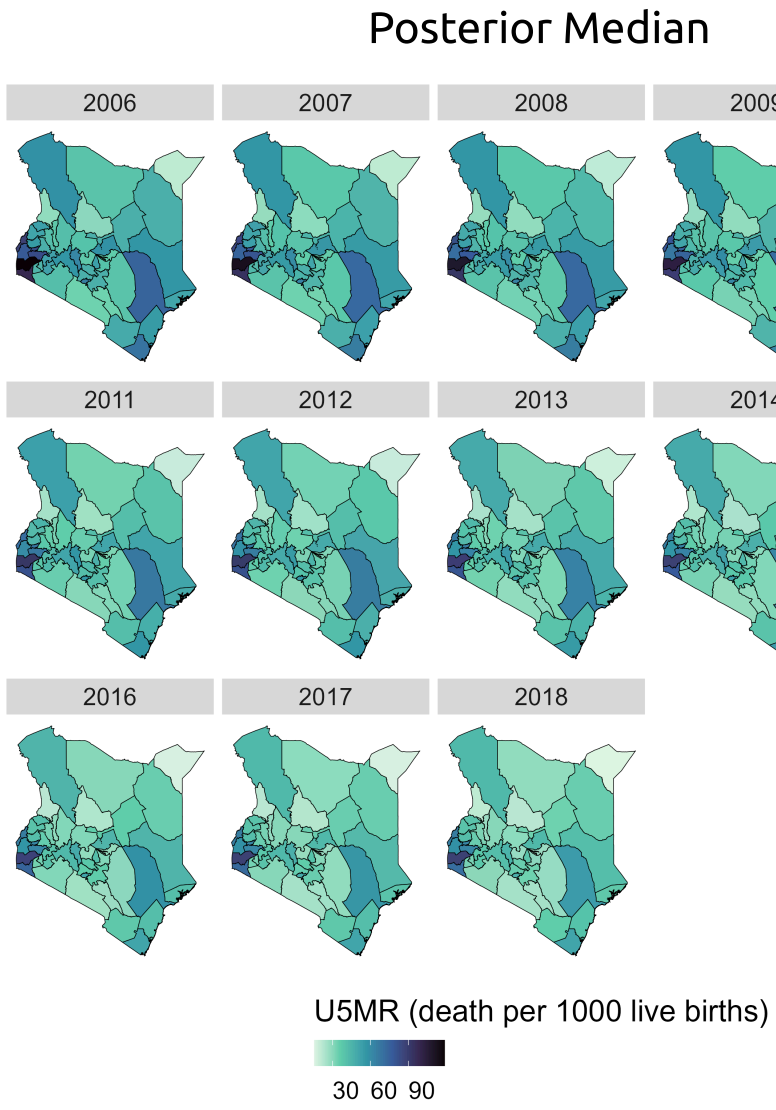
$$\text{U5MR}_{p,r,\text{urban}} \quad \text{and} \quad \text{U5MR}_{p,r,\text{rural}}$$

Using $q_{p,r}$, the proportion of population in period p and admin-1 region r that are from a rural setting:

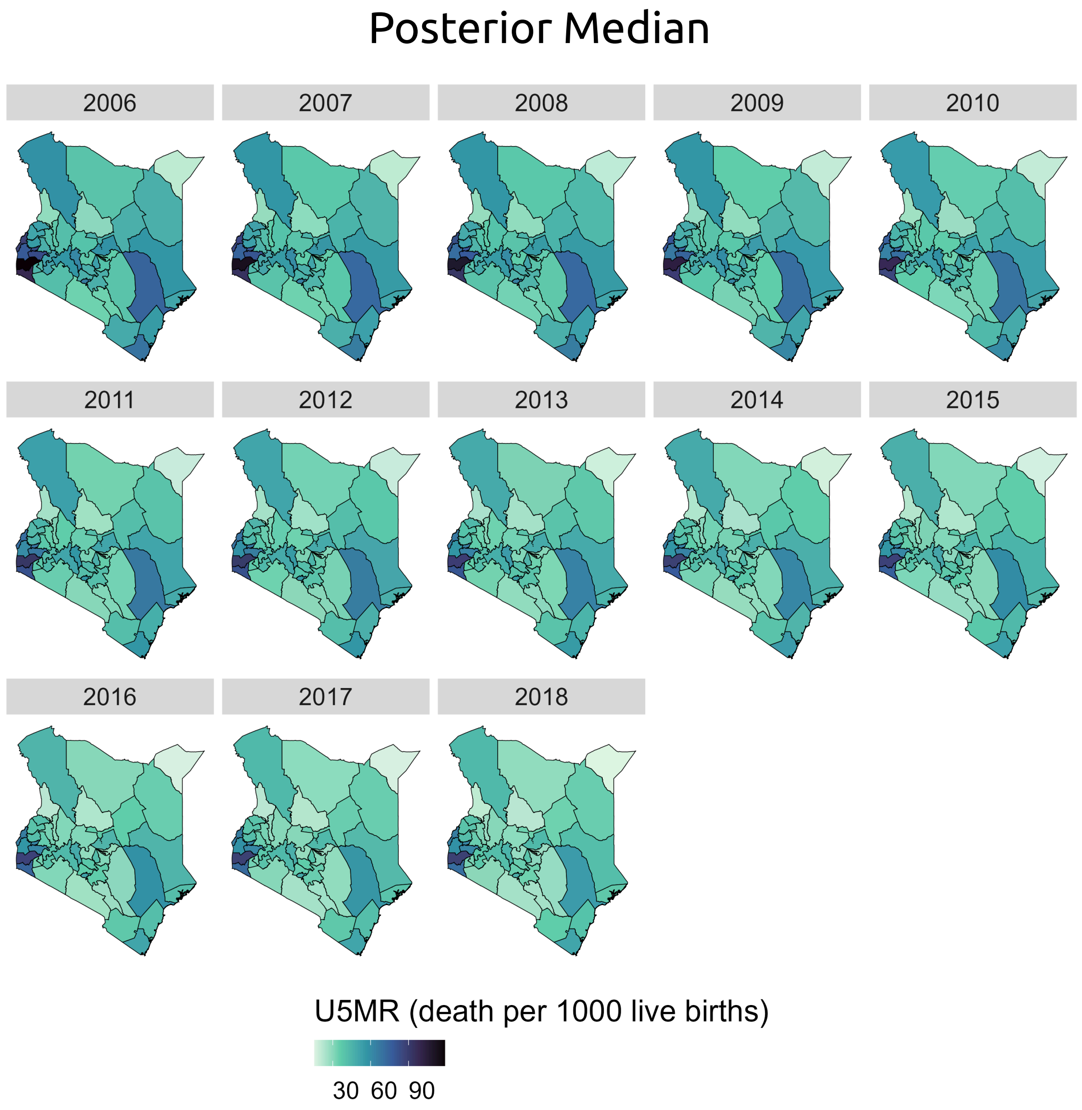
$$\text{U5MR}_{p,r} = \left[q_{p,r} \times \text{U5MR}_{p,r,\text{rural}}\right] + \left[(1 - q_{p,r}) \times \text{U5MR}_{p,r,\text{urban}}\right]$$

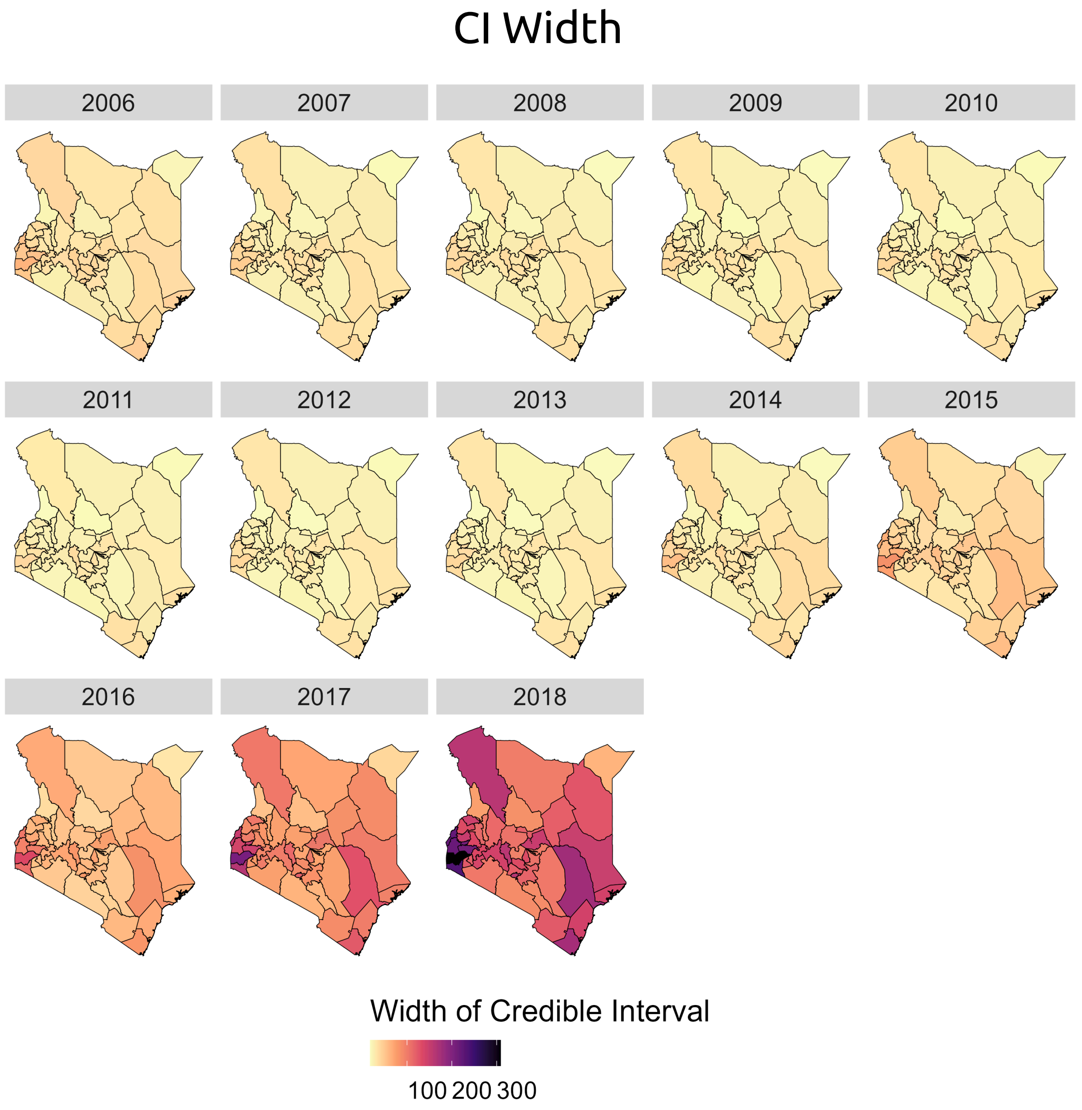
Results

Spatio-temporal

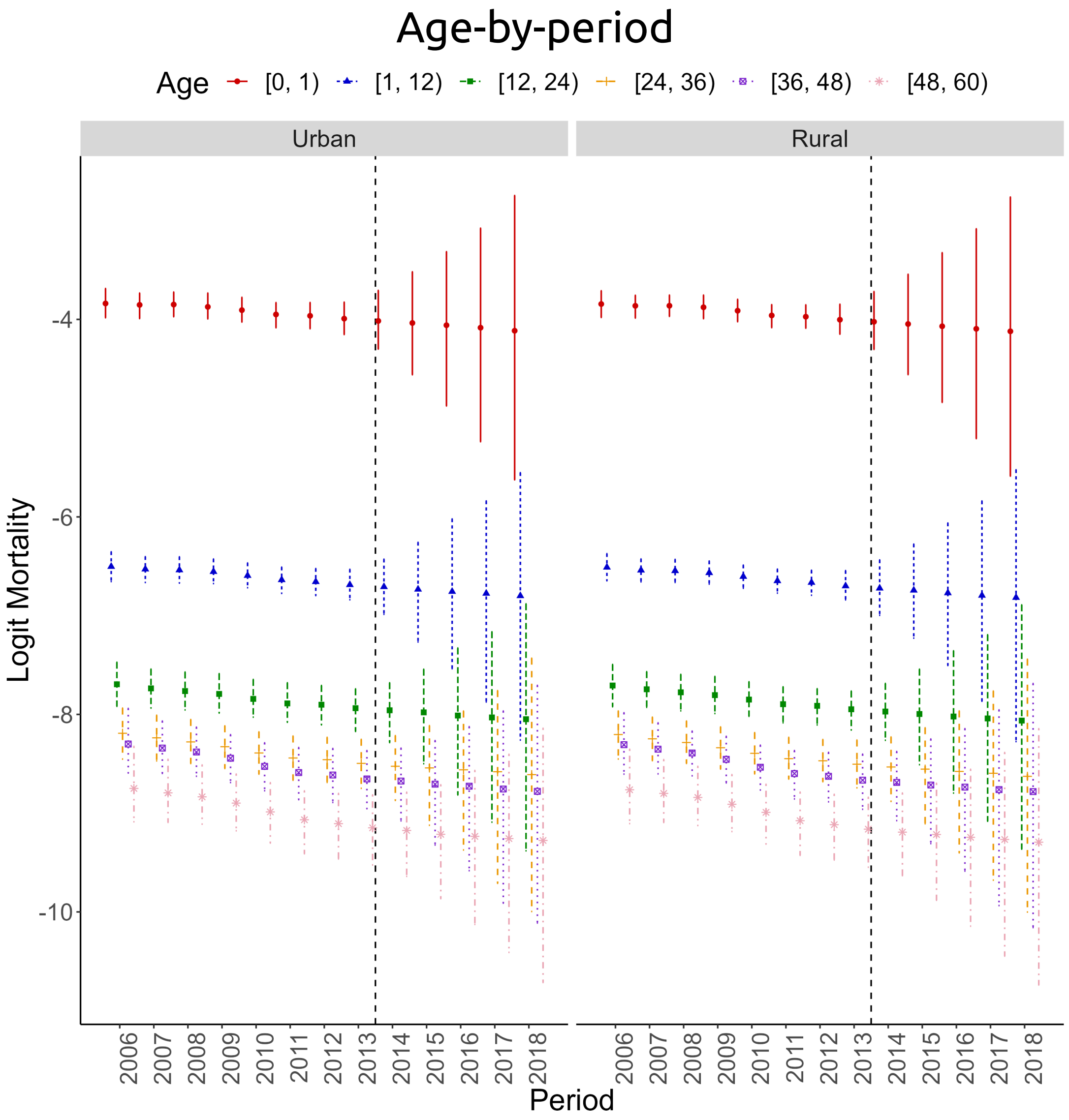


Spatio-temporal

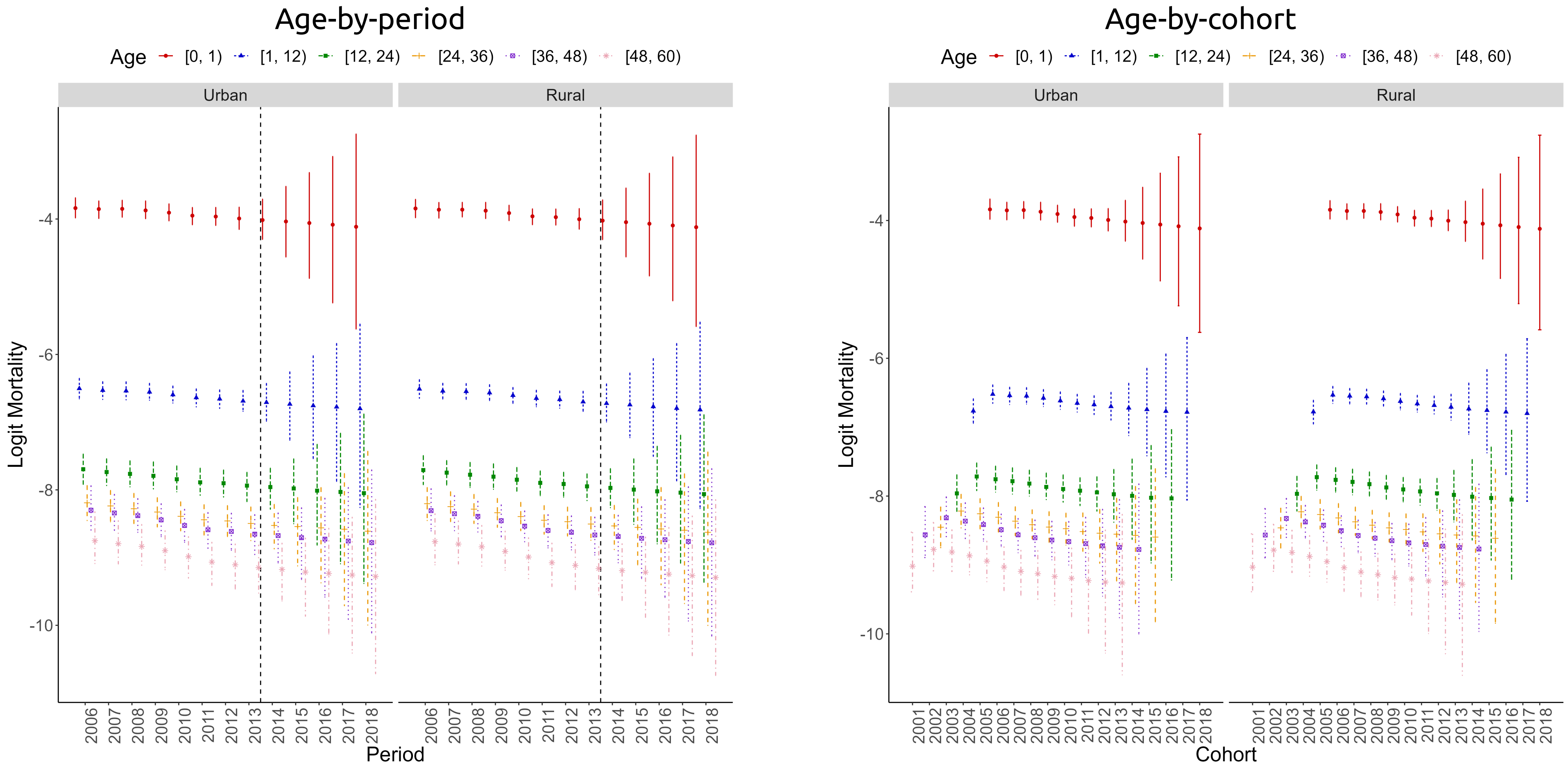




Age-Period-Cohort



Age-Period-Cohort



Model Validation

- Cross validation against direct estimates:
  - Direct estimates are "Gold Standard".
  - Compare submodels of APC model (AP and AC).
- Age-period model:
  - Similar to cluster-level models.
  - Only include age and period (time/year).
- Age-cohort:
  - Want to model cohort without identification.
  - Omit period in favour of cohort.

Model Validation

- Cross validation against direct estimates:
  - Direct estimates are "Gold Standard".
  - Compare submodels of APC model (AP and AC).
- Age-period model:
  - Similar to cluster-level models.
  - Only include age and period (time/year).
- Age-cohort:
  - Want to model cohort without identification.
  - Omit period in favour of cohort.

Model	MAE	MSE	IS	Cov.
AP	70.1	77.4	145.5	88.4
AC	71.3	80.2	155	88.4
APC	69	73.9	142.6	88.4

[1] MAE and MSE: x10^(-2)

[2] Coverage: %

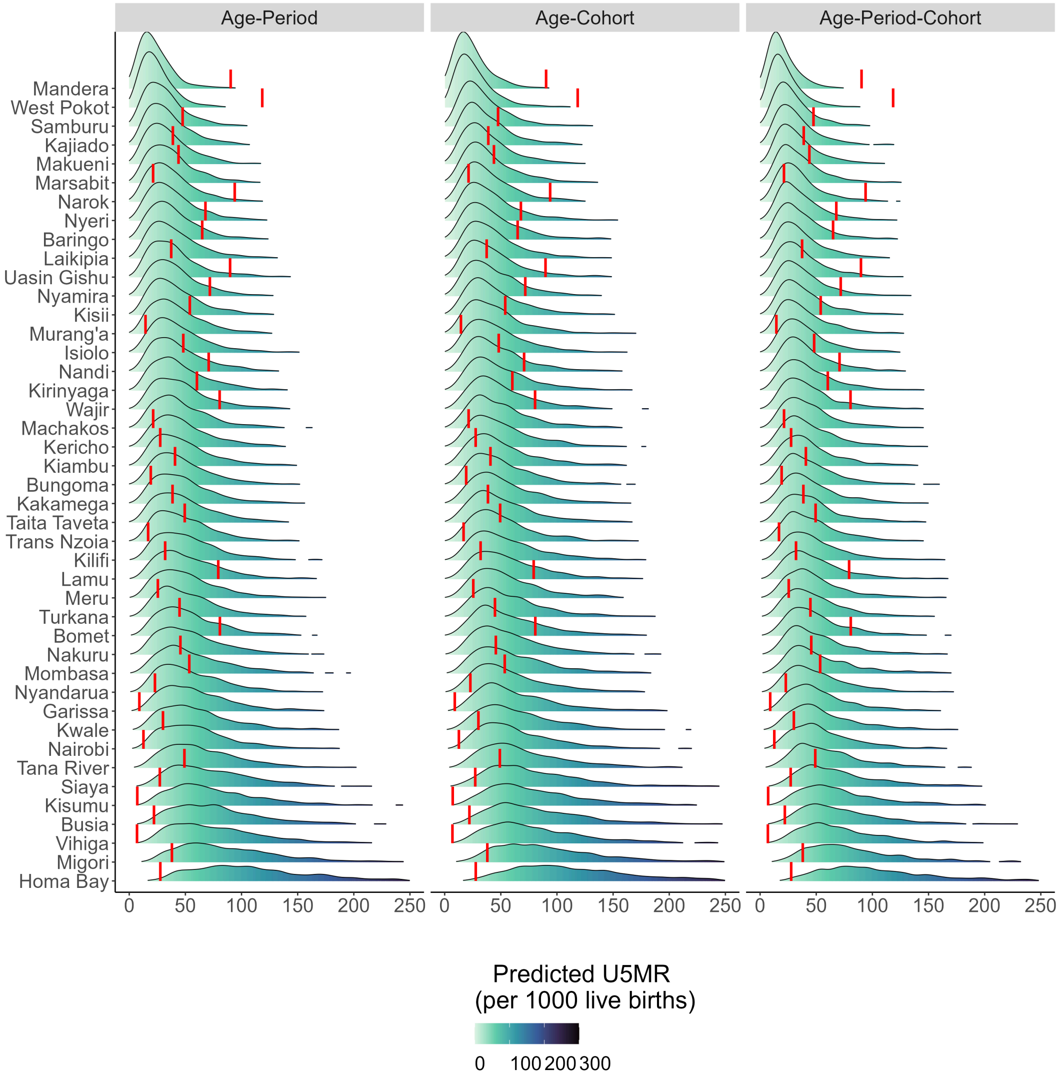
Model Validation

- Cross validation against direct estimates:
  - Direct estimates are "Gold Standard".
  - Compare submodels of APC model (AP and AC).
- Age-period model:
  - Similar to cluster-level models.
  - Only include age and period (time/year).
- Age-cohort:
  - Want to model cohort without identification.
  - Omit period in favour of cohort.

Model	MAE	MSE	IS	Cov.
AP	70.1	77.4	145.5	88.4
AC	71.3	80.2	155	88.4
APC	69	73.9	142.6	88.4

[1] MAE and MSE: x10^(-2)

[2] Coverage: %



Conclusion

Conclusion

1. Included cohort alongside age and period in U5M:
   - U5MR literature novelty - inclusion of cohort.
   - APC literature novelty - accounting for complex survey design and application to U5MR.
2. Is it reasonable to include cohort in context of U5M?
   - Yes, the model comparison to direct estimates show APC model is in-line with literature.
   - Up to the users discretion whether to include cohort.

Thanks For Listening, Any Questions?

: c.gascoigne@imperial.ac.uk

: connorgascoigne.github.io

: connorgascoigne

Preprint on Arxiv:

2014 Kenyan DHS

Stratification:

- 47 counties (admin-1).
- Nairobi and Mombasa are both entirely urban counties.
- 92 stratum ([45 $\times$ 2] + Nairobi + Mombasa).

First sampling stage:

- 1,612/96,251 primary sampling units (PSUs) selected.
- Defined by 2009 Kenyan Census.
- Selected using probability proportional to size.
- Of 1,612, 617 urban and 995 rural.

Second sampling stage:

- 40,3000 households sampled in total.
- 25 households per PSU.

Geospatial jittering:

- For confidentiality.
- 2km radius for urban jitter.
- 5km radius for rural jitter.
- For an additional 1% of rural, 10km radius.

Prior Distributions

Fixed effects:

- $\beta_1, \beta_2, \beta_3, \beta_4 \sim \mathrm{N}\,(0, 1000)$.

Temporal (non-linear) effects:

- $\nu_{\tilde{a}} | \tau_\nu \sim \mathrm{RW2}\,(\tau_\nu), \eta_p | \tau_\eta \sim \mathrm{RW2}\,(\tau_\eta)$ and $\xi_c | \tau_\xi \sim \mathrm{RW2}\,(\tau_\xi)$.

Spatial effect:

- $S_{r[s_k]} | \tau_s, \phi \sim \mathrm{BYM2}\,(\tau_s, \phi)$ where $\phi \in \{0, 1\}$.

Spatio-temporal effect:

- $\delta_{p,r[s_k]} | \tau_\delta \sim \mathrm{N}\left(0, \tau_\delta^{-2}\right)$.

Overdispersion:

- $d \sim \mathrm{N}\,(0, 2.5)$.

Hyperpriors:

- Precisions: $\mathrm{PC}\,(\tau_\star > 1) = 0.01$.
- Spatial mixing: $\mathrm{PC}\,(\phi > 1/2) = 2/3$.

Hazard for month a:

$$\mathrm{h}(a) = p(A = a | A \geq a) = h_a$$

Survival beyond month a:

$$\begin{aligned}\mathrm{Survival}(a) &= p(A > a) \\ &= p(A > a | A \geq a) \times \\ &\quad p(A > a-1 | A \geq a-1) \times \\ &\quad \cdots \times p(A > 1 | A \geq 1) \\ &= [1 - \mathrm{h}_a][1 - \mathrm{h}_{a-1}] \times \cdots \times [1 - \mathrm{h}_1].\end{aligned}$$

Discrete Hazards:

- Constant hazard within each age group.
- $\mathrm{h}_{48} = \cdots = \mathrm{h}_{59} = \mathrm{h}_{52.5}$.
- $\mathrm{h}_{36} = \cdots = \mathrm{h}_{47} = \mathrm{h}_{41.5}$.
- $\ldots$

U5MR:

$$\begin{aligned}\mathrm{U5MR} &= 1 - \mathrm{Survival}(a = 59) \\ &= 1 - [1 - \mathrm{h}_{59}][1 - \mathrm{h}_{58}] \times \cdots \times [1 - \mathrm{h}_0] \\ &= 1 - [1 - \mathrm{h}_{52.5}]^{12}[1 - \mathrm{h}_{41.5}]^{12} \times \\ &\quad [1 - \mathrm{h}_{29.5}]^{12}[1 - \mathrm{h}_{17.5}]^{12} \times \\ &\quad [1 - \mathrm{h}_6]^{11}[1 - \mathrm{h}_0] \\ &= 1 - \prod_{i=1}^{6}\left[1 - \mathrm{h}_{\tilde{a}[i]}\right]^{z[i]}\end{aligned}$$

Explaining The U5MR Formula

Discrete time survival analysis | **Under Five Mortality Rate**

$$\text{U5MR}_{p,r} = 1 - \prod_{i=1}^{6} \left[1 - \text{expit}\left(\text{logit}\left[\pi_{\tilde{a}(i),p,\bar{c},r}\right]\right)\right]^{z[i]}$$

Where:

- $\tilde{a} = 0.5, 6, 17.5, 29.5, 41.5, 52.5$ mid points of each discrete hazard (age group).
- $z = 1, 11, 12, 12, 12, 12$ lenght of number of months in each discrete hazard.

Predictive Distribution | Assessment Criteria

- Compare model estimates (APC, AP, AC) against direct estimate at subnational level.
- U5MR distribution we calculated:
  – Contains sampling variability.
  – Does not contain variability from complex survey design.
- How to make model and direct estimates more comparable?
  – Add variation from survey design to model estimates.

Let $\widehat{Y}_r = \text{logit}\left(\text{U5MR}_r\right)$:

$$\widetilde{Y}_r = \widehat{Y}_r + \text{N}\left(0, \hat{V}_{\text{r}}^{\text{Des}}\right)$$

where $\hat{V}_{\text{r}}^{\text{Des}}$ is the complex design variance for admin-1 region r.

Model Validation

Predictive Distribution | **Assessment Criteria**

For $n = 1, \ldots, N$ draws from U5MR posterior,

$$\text{MAE} = \frac{1}{R}\frac{1}{N}\sum_{r=1}^{R}\sum_{m=1}^{N}\left|\widetilde{Y}_r^{(n)} - Y_r\right|$$

$$\text{MSE} = \frac{1}{R}\frac{1}{N}\sum_{r=1}^{R}\sum_{m=1}^{N}\left[\widetilde{Y}_r^{(n)} - Y_r\right]^2$$

$$\text{IS}_\alpha\left(Y\right) = \frac{1}{R}\sum_{r=1}^{R}\left[\left(u_r - l_r\right) + \frac{2}{\alpha}\left(l_r - Y_r\right)\mathbb{I}\left(Y_r < l_r\right) + \frac{2}{\alpha}\left(Y_r - u_r\right)\mathbb{I}\left(Y_r > u_r\right)\right]$$

- Interval Score (IS):
  - Gneiting and Raftery, *Journal of the American Statistical Association*, 2007.
  - Proper Scoring rule.

National Results

- Comparison against "gold standard" direct estimates.
- Include, Fay-Herriot, Age-Period and Age-Cohort estimates.