Space-Time Modeling of Health and Demographic Indicators

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

Overview

- The work I have been primarily involved with is subnational estimation of health and demographic indicators in low and middle income countries – currently working on producing Admin2 estimates of U5MR, as the official UN IGME estimates.
- In general, most of input data is from household surveys and censuses.
- I will briefly discuss: the modeling framework, aggregation, computing, visualization.
- In general, the data are available at different temporal and spatial scales, and (if relevant) for different age groupings.
- If we model at the lowest resolution, we can aggregate consistently to inform on the common linking parameters θ_{LINK}.
- Smoothing over space, time and age is feasible, but care required for identifiability, and to prevent computation blowing up.

Data Aggregation

- In (mostly) older surveys, GPS locations not available – to combine data we need to integrate over areas – Marquez and Wakefield (2020) develop a method to do this, based on a space-time Gaussian process (GP).
- Show the superiority over previous approaches used by IHME (Burstein *et al.*, 2019) or WorldPop (Utazi *et al.*, 2018).

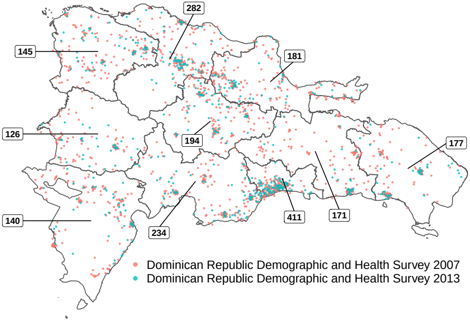


Figure 1: Application: Dominican Republic U5MR estimation (Marquez and Wakefield, 2020). The points represent the cluster locations for 2007 and 2013 DHS. Labeled counts are number of masked clusters in each region for the 2014 MICS.

Comparison of Temporal Models: Survey Weighted

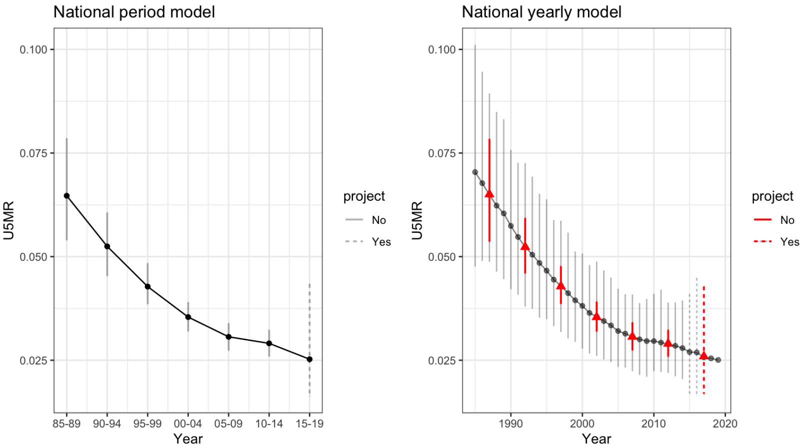


Figure 2: National 5-year periods (left) and yearly U5MR estimates for Ecuador – in both cases the data are available as 5-year averages, but for the estimates on the right we have an underlying 1-year time series model (a RW2 model). Methods described in Mercer *et al.* (2015) and Li *et al.* (2019).

Computing

There is computationally machinery to analyze complex stochastic systems:

- MCMC: Avoid if you can as harder to automate and for dependent data can be very computationally expensive.
- Integrated Nested Laplace Approximations (INLA) (Fong *et al.*, 2010).
- Template Model Builder (TMB) (Osgood-Zimmerman and Wakefield, 2020) – extensive simulations to examine accuracy of TMB and INLA.
- Provide code that allows results to be reproduced — we use the `SUMMER` package.

Modeling

- The majority of household surveys use a stratified, two-stage cluster design.
- The stratification is usually based on the cross of administrative areas and urban/rural.
- The clusters are households within enumeration areas.
- Paige *et al.* (2020) show that ignoring the stratification leads to bias, and ignoring the clustering, inappropriate uncertainty measures.
- Also show that a discrete spatial model was superior to a continuous spatial model, both described in Wakefield *et al.* (2019), when the target of inference was Administrative areas – complex is not always best!
- Generative models (full probability models that could simulate the raw data we see), are preferable as they offer a common framework for aggregate data and allowing modeling of biases.
- Are the biases associated with a particular data source constant over space or time?

Bias Modeling

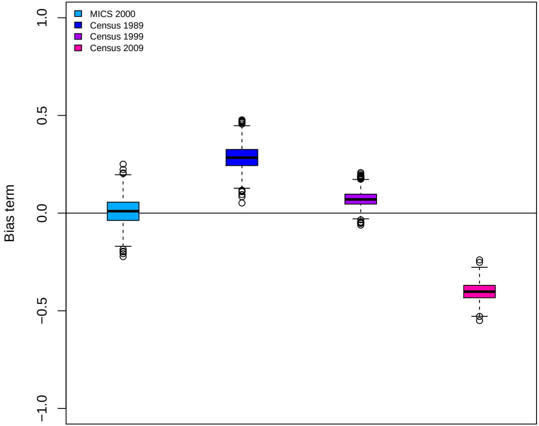


Figure 3: "Bias" in SBH data, as compared to FBH data, in Kenya, from Godwin and Wakefield (2020).

Modeling Strategy for U5MR: Incorporating Different Data Types

- To leverage information in countries without full VR systems – use data on full birth history (FBH), summary birth history (SBH), incomplete VR (IVR) systems.
- Have a coherent framework for all of the data sources, for example,

$$
\begin{aligned}
p(Y_{\text{FBH}}, Y_{\text{SBH}}, Y_{\text{IVR}} \mid \underbrace{\theta_{\text{FBH}}, \theta_{\text{SBH}}, \theta_{\text{IVR}}}_{\text{Data-Type Specific Parameters}}, \underbrace{\theta_{\text{LINK}}}_{\text{Common Parameters}}) \\
= p(Y_{\text{FBH}}|\theta_{\text{FBH}}, \theta_{\text{LINK}})\,p(Y_{\text{SBH}}|\theta_{\text{SBH}}, \theta_{\text{LINK}})\,p(Y_{\text{IVR}}|\theta_{\text{IVR}}, \theta_{\text{LINK}})
\end{aligned}
$$

- The common link parameters θ_{LINK} are the skeleton that holds everything together.
- We may approximate each of the constituent models, but make an effort to get the framework right – not all approaches do this...

Example: Modernizing the Brass Method

Auxiliary Data Approach (Wilson and Wakefield, 2020a):

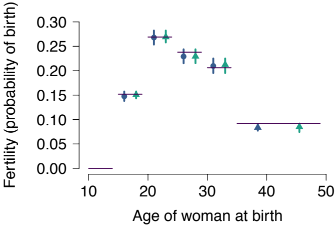
$$\Pr(\boldsymbol{D}_m | \boldsymbol{B}_m, \boldsymbol{c}_m, \boldsymbol{q}) = \sum_{\boldsymbol{B}_m} \sum_{\boldsymbol{D}_m} \underbrace{\Pr(\boldsymbol{D}_m | \boldsymbol{B}_m, \boldsymbol{q})}_{\text{Product of Binomials}} \times \underbrace{\Pr(\boldsymbol{B}_m | B_m, \boldsymbol{c}_m)}_{\text{Births Distribution}}.$$

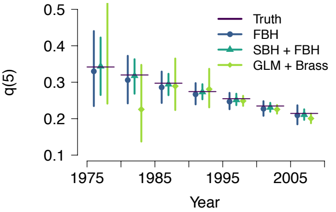
Requires MCMC and relatively slow.

Poisson Approximation (Wilson and Wakefield, 2020b):

$$D_m | B_m, \boldsymbol{c}_m, \boldsymbol{q} \sim \text{Poisson}\left(B_m \sum_{a=1}^{A_m} c_m(a) q(a) \right).$$

Computation (with TMB) went from > 7 days to 10 minutes.





Subnational U5MR Model

Sampling model:

$$\underbrace{Y_{a,c,k}(t)}_{\text{Deaths in age group } a} \mid h_{m,c,k}(t) \sim \underbrace{\text{BetaBinomial}}_{\text{Overdispersed binomial}} \left(\underbrace{n_{m,c,k}(t)}_{\text{At risk months}}, \underbrace{h_{m,c,k}(t)}_{\text{Monthly hazard}}, \underbrace{d}_{\text{Scale}} \right)$$

with

Hazard Model:

$$\underbrace{\text{logit}[\, h_{m,c,k}(t) \,]}_{\substack{\text{Year } t\text{, month } m\text{,} \\ \text{cluster } c\text{, survey } k}} \qquad \text{Hazard at start of month } m$$

$$= \log \text{BIAS}^{\text{HIV}}_{c,k}(t) \qquad \text{Adjustment for HIV}$$

$$+ \alpha_{a[m]} \qquad \text{Age specific intercepts: Fixed effects}$$

$$+ \beta_k \qquad \text{Survey fixed effects: With sum-to-zero constraint}$$

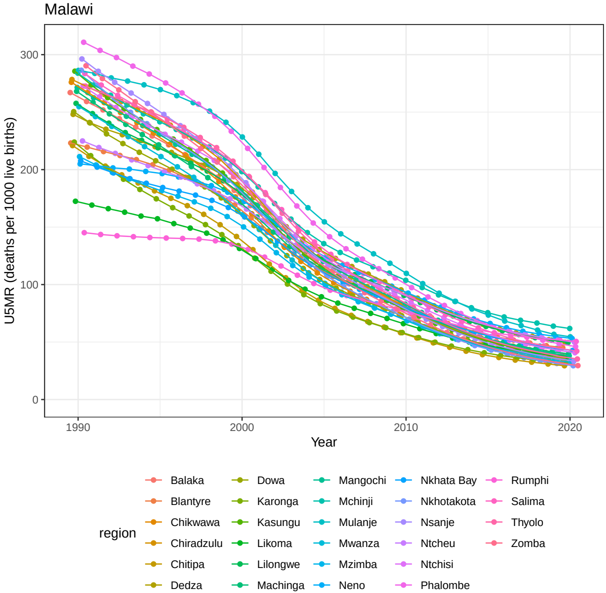
$$+ \epsilon(t) \qquad \text{IID shocks, random effects in year } t \;:\; \text{N}(0, \sigma^2_{\text{YEAR,IID}})$$

$$+ \gamma_{a^\star[m]}(t) \qquad \text{Yearly smooth, age band } a[m]\text{: } \text{RW2}(\sigma^2_{\text{YEAR,RW2}})$$

$$+ S_{i[c]} \qquad \text{Spatial term for area } i \text{ containing cluster } c\text{: } \text{BYM2}(\sigma^2_{\text{AREA}}, \phi_{\text{AREA}})$$

$$+ \delta_{i[c]}(t) \qquad \text{Interaction for year } t\text{, area } i\text{: } \text{ICAR-AR1}(\sigma^2_{\text{INTER}}, \rho_{\text{YEAR}})$$

Subnational U5MR estimates for Malawi, based on DHS data.



Subnational U5MR estimates

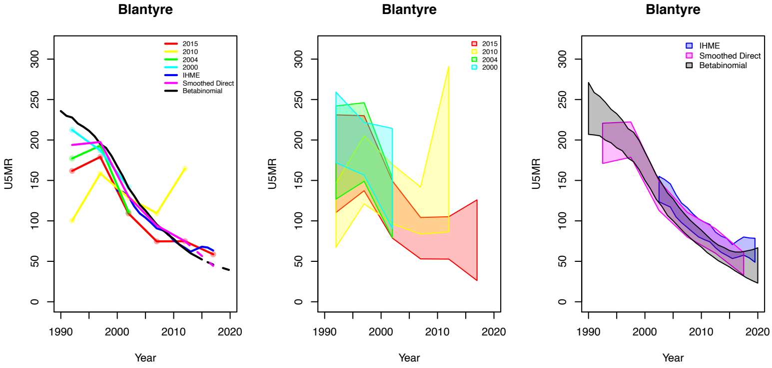


Figure 5: Subnational U5MR estimates for Blantyre, Malawi, based on DHS data, with comparison to IHME Local Burden of Disease estimates.

Comparison of Temporal Models, Fraction of under-5 Urban Population

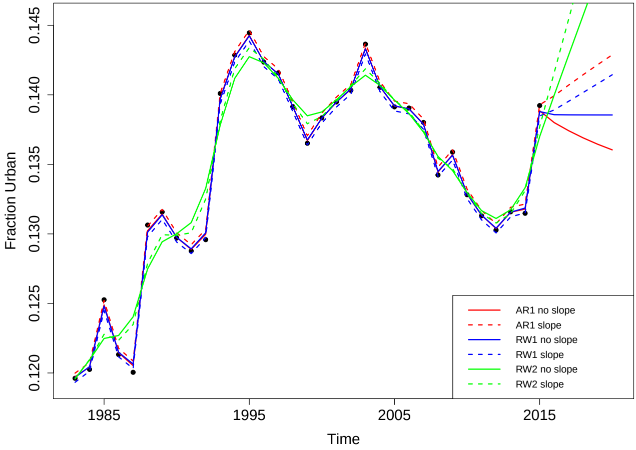


Figure 6: Posterior median fits from 6 models, on the original scale, with estimated common variance.

Comparison of Temporal Models: Survey Weighted

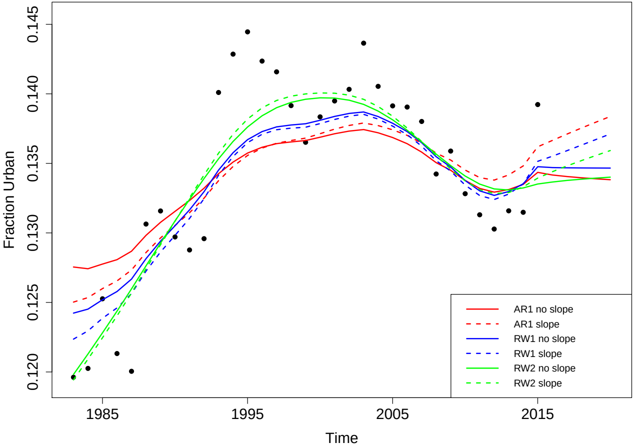


Figure 7: Posterior median fits from 6 models, on the original scale, with known variances.

Visualizing Uncertainty (Dong and Wakefield, 2020): Continuous Color Scale Maps

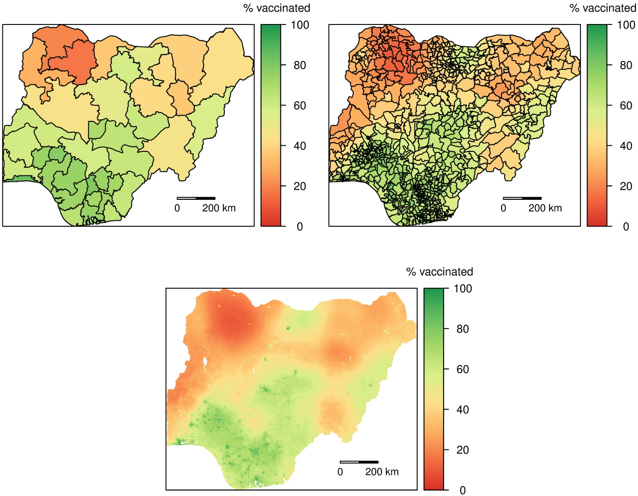


Figure 8: Posterior median MCV1 coverage at state, LGA and pixel level. .

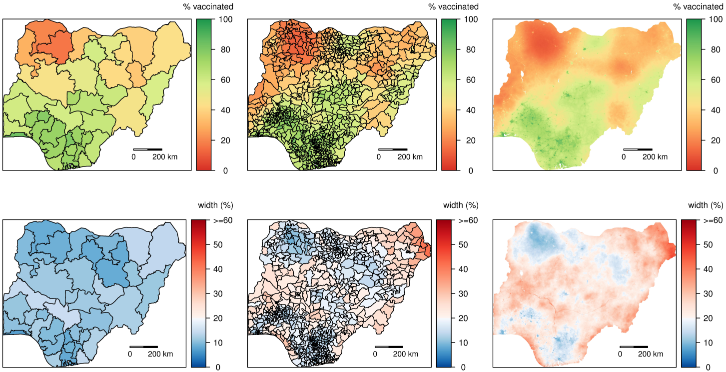


Figure 9: Top row: posterior median MCV1 coverage at state, LGA and pixel level. Bottom row: width of 90% credible interval at state, LGA and pixel level.

Hatching for Uncertainty

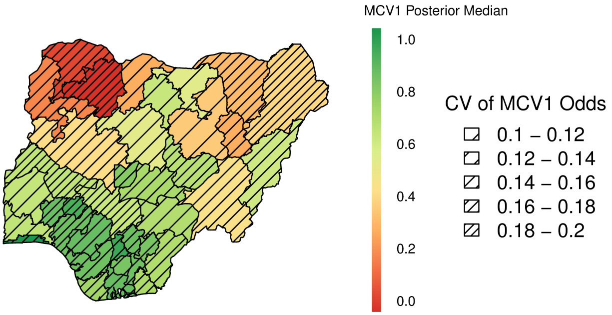
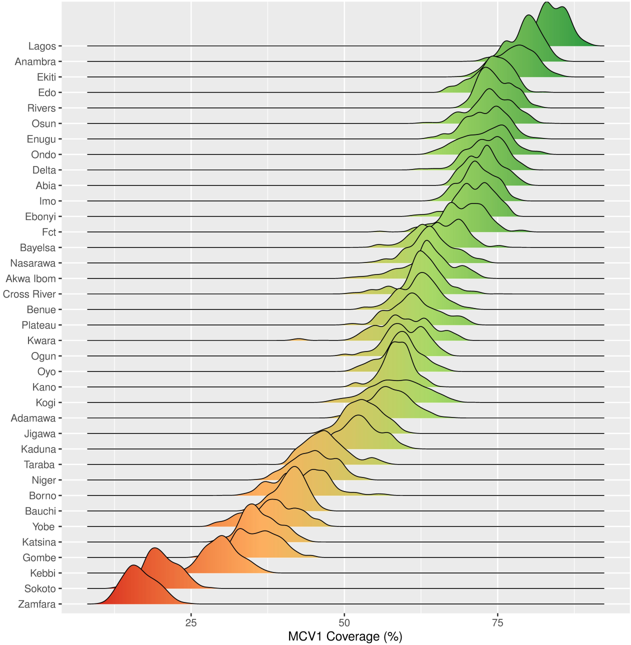
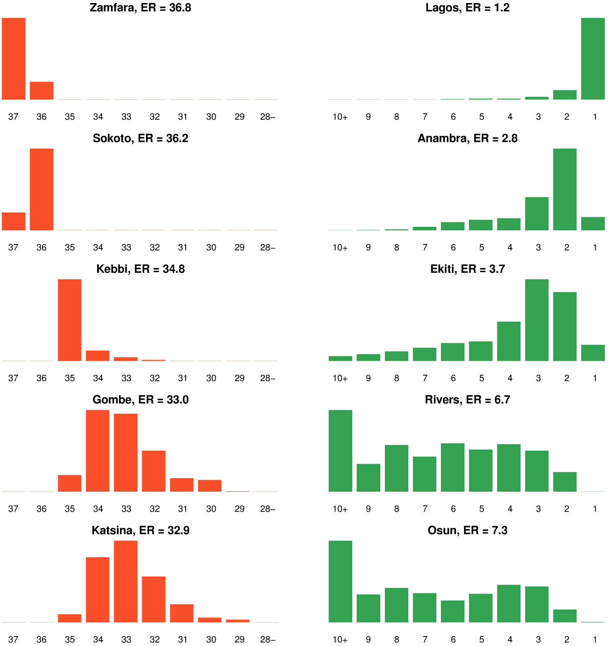


Figure 10: Denser hatching implies more uncertainty.

Ridgeplots for States



Rankings for States



Posterior Probability of Achieving a Threshold

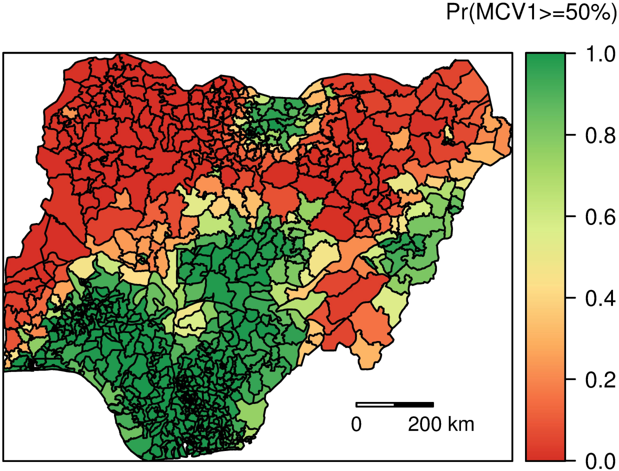


Figure 13: Posterior probability of MCV1 being at least 50% across Nigerian LGAs.

Discrete Color Scale Maps

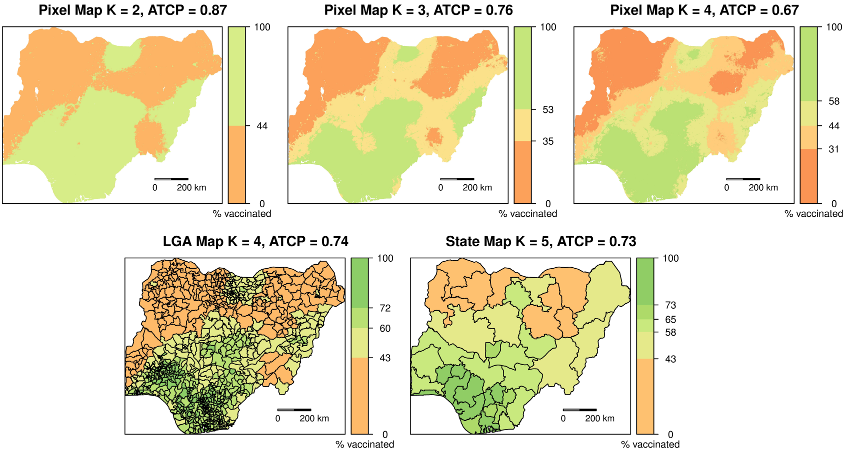


Figure 14: If we want a 70% chance that the probability of a randomly selected area bing of the right color, how many colors are we allowed? The ATCP is the Average True Classification Probability.

Discussion

- A statistical modeling framework is important – methods are published in statistical journals. Reproducible methods need more than a flowchart.
- Covariate modeling can be beneficial but covariates are often modeled themselves.
- Flexible machine learning techniques are appealing, but accounting for uncertainty is tricky.
- Model assessment is tricky, because one has to respect the dependencies in the data and the design (e.g., stratification).
- Short courses materials, papers, teaching videos, `SUMMER` package details, Shiny app here:

  `http://faculty.washington.edu/jonno/space-station.html`

Collaborators and Further Materials

UW: Jessica Godwin, Taylor Okonek, Johnny Paige, Austin Schumacher, Tracy Dong, Miranda Fix, Ziyu Jiang, Julianne Meisner, Orvalho Augusto, Avi Kenny, Serge Aleshin-Guendel, Yunhan Wu, Aaron Osgood-Zimmerman.

Yale: Zehang Richard Li.

Trondheim: Geir-Arne Fuglstad, Andrea Riebler.

OSU: Sam Clark, Erick Axxe, Eungang Choi, Yue Chu, Jon Muir, Jason Thomas.

UN: Dave Sharrow, Lucia Hug, Danzhen You.

Work supported by NIH and IUSSP.

Fixing up Zero Counts

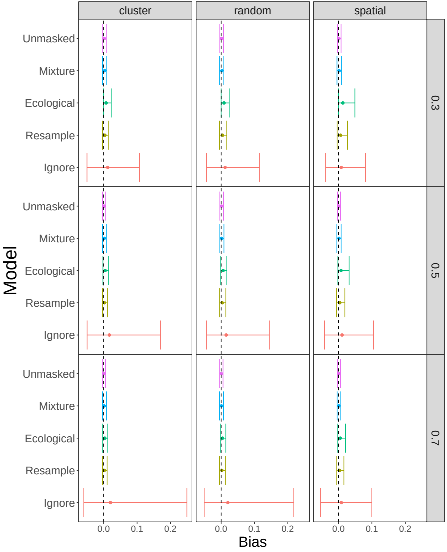


Figure 15: Comparison of mixture model based on a GP, with ecological and resampling methods, from Marquez and Wakefield (2020).

References

Burstein, R., Henry, N. J., Collison, M. L., Marczak, L. B., Sligar, A., Watson, S., Marquez, N., Abbasalizad-Farhangi, M., Abbasi, M., Abd-Allah, F., *et al.* (2019). Mapping 123 million neonatal, infant and child deaths between 2000 and 2017. *Nature*, **574**, 353–358.

Dong, T. and Wakefield, J. (2020). Modeling and presentation of health and demographic indicators in a low- and middle-income countries context. *Submitted*.

Fong, Y., Rue, H., and Wakefield, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics*, **11**, 397–412.

Godwin, J. and Wakefield, J. (2020). Space-time modeling of child mortality at the admin-2 level in a low and middle income countries context. *Statistics in Medicine*. Under Review.

Li, Z. R., Hsiao, Y., Godwin, J., Martin, B. D., Wakefield, J., and Clark, S. J. (2019). Changes in the spatial distribution of the under five mortality rate: small-area analysis of 122 DHS surveys in 262 subregions of 35 countries in Africa. *PLoS One*. Published January 22, 2019.

Marquez, N. and Wakefield, J. (2020). Harmonizing binomial outcome health data at disparate geographic levels. *Under revision*.

Mercer, L., Wakefield, J., Pantazis, A., Lutambi, A., Mosanja, H., and Clark, S. (2015). Small area estimation of childhood mortality in the absence of vital registration. *Annals of Applied Statistics*, **9**, 1889–1905.

Osgood-Zimmerman, A. and Wakefield, J. (2020). Template Model Builder (TMB), a flexible alternative to the Integrated Nested Laplace Approximation. *Submitted*.

Paige, J., Fuglstad, G.-A., Riebler, A., and Wakefield, J. (2020). Model-based approaches to analysing spatial data from complex surveys. *Journal of Survey Statistics and Methodology*. To appear.

Utazi, C. E., Thorley, J., Alegana, V. A., Ferrari, M. J., Nilsen, K., Takahashi, S., Metcalf, C. J. E., Lessler, J., and Tatem, A. J. (2018). A spatial regression model for the disaggregation of areal unit based data to high resolution grids with application to vaccination coverage mapping. *Statistical Methods in Medical Research*, **28**, 3226–3241.

Wakefield, J., Fuglstad, G.-A., Riebler, A., Godwin, J., Wilson, K., and Clark, S. (2019). Estimating under five mortality in space and time in a developing world context. *Statistical Methods in Medical Research*, **28**, 2614–2634.

Wilson, K. and Wakefield, J. (2020a). Child mortality estimation incorporating summary birth history data. *Biometrics*. To Appear.

Wilson, K. and Wakefield, J. (2020b). A model-based alternative to the Brass method. *Submitted*.