A Bayesian model for estimating under-5 mortality age schedules in small areas

José H C Monteiro da Silva

Graduate Group in Demography, University of Pennsylvania

October 29, 2025

1 Introduction

The probability of dying under age 5, also known as the Under-5 Mortality Rate (U5MR) is a key indicator of a population's health and well-being (UNICEF, 2025). For that reason, reducing under-5 mortality by 2030 is part of goal 3.2 of the Sustainable Development Goals of the United Nations. Recently, scholars have moved forward in understanding under-5 mortality by trying to model it in more depth, by addressing the detailed age schedule of under-5 mortality (Guillot et al., 2022a; Okonek et al., 2024; Verhulst et al., 2022) instead of focusing on the single U5MR measure.

These previous works show that under-5 mortality age schedule has a regular shape (Guillot et al., 2022a), similarly to the overall mortality age schedule. Hence, this regularity favors the use of relational models to estimate mortality patterns in a data-sparse context (e.g., small areas). As more and more countries experience a decline in under-5 mortality, the stochasticity of deaths might incur in very irregular and sometimes unreasonable age-specific morality curves. Therefore, new methods are needed to address these issues and measure the true underlying mortality curve.

By taking advantage of the regularity of the mortality schedule below age 5, in this work I propose a methodology for estimating under-5 mortality curves in small areas or data-scarce contexts. I extend a previous Bayesian strategy used for modeling the overall mortality curve in small areas (Alexander et al., 2017) to 22 age groups below age 5.

2 Methods

For estimating the underlying mortality curve in small areas, we assume for our model likelihood that $D_{i,j,t}$, the number of deaths at age group i, area j, and time t follows a Poisson distribution.

$$D_{i,j,t} \sim Poisson\left(E_{i,j,t} \times m_{i,j,t}\right), \tag{1}$$

where $E_{i,j,t}$ is the person-years of exposure at age group i, area j, and time t and $m_{i,j,t}$ is the age-specific mortality rate at age group i, area j, and time t.

Further, we model the age-, area- and time-specific mortality rates $m_{i,j,t}$ as:

$$\log\left(m_{i,j,t}\right) = \beta_{1jt}P_{1_i} + \beta_{2jt}P_{2_i} + \beta_{3jt}P_{3_i} + \epsilon_{ijt}, \tag{2}$$

where P_{1_i}, P_{2_i} and P_{3_i} are the first three principal components of a set of standard under-5 mortality curves, and ϵ_{ijt} is an age-, area- and time-specific random effect.

These three principal components create a baseline age structure of mortality that represent the regularities of the under-5 mortality curve, and they are obtained via a singular value decomposition (SVD) on a $C \times N$ matrix of logged mortality rates M, where C is the number of standard life tables and N is the number of age groups (22 in this case, [0,7) days, [7,14) days, [14,21) days, [21,28) days, [28 days, 2 months], [2,3) months, [3,4) months, [4,5) months, [5,6) months, [6,7) months, [7,8) months, [8,9) months, [9,10) months, [10,11) months, [11,12) months, [12,15) months, [15,18) months, [18,21) months, [21 months, 2 years), [2,3) years, [3,4)

years, [4,5) years).

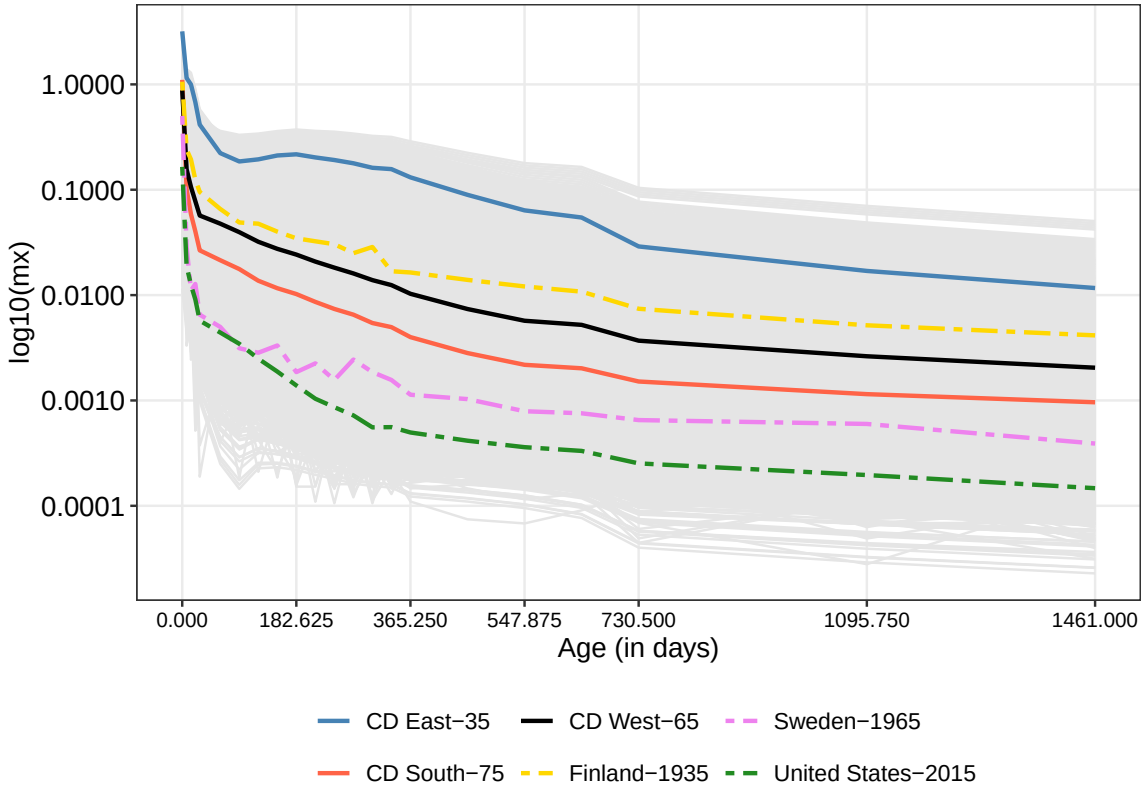
$$M = UDV^t \tag{3}$$

The three components P_1, P_2, and P_3 ($N \times 1$ vectors) are the first three columns of the V ($N \times N$) matrix. The singular value decomposition was applied to three different sets of under-5 mortality schedules (M matrix). The first two sets, which represent 489 mortality schedules) were derived from the Coale-Demeny (North, South, East, West) (Coale and Demeny, 1966; Coale and Guo, 1989) and the United Nations model life tables (Chilean, Far East Asian, Latin, South Asian, General) (United Nations, 1982). These single age model life tables do not have the detailed information below age 5 that we need for our modeling approach, therefore, to derive the detailed 22 age groups below age 5 for each level of these model life tables, I used the respective child, ${}_1q_4$, and infant, ${}_1q_0$, mortality rates of each of these model life tables as inputs to the log-quadratic under-5 mortality model (Guillot et al., 2022a)[1]. In addition to these two sets of under-5 mortality schedules derived from model life tables, I also included 1175 country-year under-5 mortality schedules from the Under-5 Mortality Database (U5MD) (Guillot et al., 2022b)[2]. I chose these three different sets to accommodate a wider range of under-5 mortality schedules. Figure 1 shows all under-5 mortality schedules and some highlighted curves that were used for deriving the three principal components.

[1]These two sets of Coale-Demeny and UN model life tables contains 729 mortality schedules. After applying the log-quad model to each of these, I excluded those that had the parameter k out of the recommended range (between -1.1270 and $+1.5047$) (Guillot et al., 2022a), and then we end up with 489 mortality schedules.

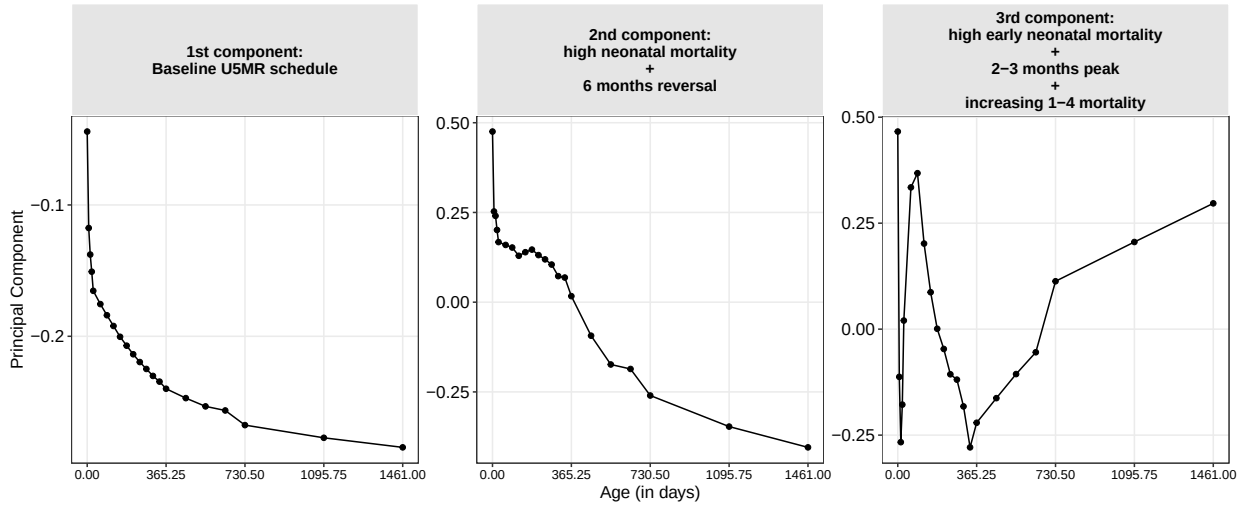
[2]I selected only schedules for both sexes and without flags for data quality issues, which represented about 1219 schedules. I further removed 44 country-years that had zeros for mortality rates, resulting then in 1175 schedules.

Figure 1: Under-5 mortality schedules used as inputs to the SVD decomposition for deriving the three principal components. Sources: Coale and Demeny (1966); Coale and Guo (1989); United Nations (1982); Guillot et al. (2022b)



The three selected principal components, which correspond to about 89% of the variability of the M matrix, are shown in Figure 2. The first component, which responds for about 82.8% of the variance of M, represents the baseline under-5 mortality curve with decreasing mortality from day 0 to the fifth year of life. The second component, 4.6% of the variance of M, reflects a high early neonatal mortality and a reversal in the declining trend around the age of 6 months. Finally, the third component, which represents about 1.6% of the data variance, reflect three distinguished patterns: high early neonatal mortality, a peak in mortality around 2-3 months, and an increasing mortality from ages 1 to 5 years.

Figure 2: Principal components of logged mortality schedules.



Then, for each area- and time-varying $\beta_{p_{jt}}$ parameter corresponding to each principal component p I assign the following time-specific normal priors:

$$\beta_{p_{jt}} \sim N\left(\mu_{\beta_p t}, \sigma_{\beta_p t}\right). \tag{4}$$

The idea of these priors under this hierarchical approach is that for each time (year), the β coefficients will shrink towards the mean represented by the larger geographic area (country or state, for example).

Finally, for the $\mu_{\beta_p t}$ I assign weakly informative normal priors differing by component and for $\sigma_{\beta_p t}$ I also assign the weakly informative truncated normal priors:

$$\mu_{\beta_1 t} \sim N\left(30, 1000\right) \tag{5}$$

$$\mu_{\beta_2 t} \sim N\left(0, 200\right) \tag{6}$$

$$\mu_{\beta_3 t} \sim N\left(0, 200\right) \tag{7}$$

$$\sigma_{\beta_p t} \sim N^{+}\left(0, 100\right) \tag{8}$$

Finally, for the age-, area-, and time-specific random effects coefficient ϵ_{ijt}, I assign a normal

prior centered on 0 but with age-varying standard deviation σ_i, for which I assign a weakly informative truncated normal prior.

$$\epsilon_{ijt} \sim N(0, \sigma_i) \tag{9}$$

$$\sigma_i \sim N(0, 100) \tag{10}$$

The model was coded in STAN using the integration with R for sampling.

3 Data

3.1 Test dataset

For testing the model, I use the assumption made in equation 1 to generate different death distributions from the same mortality curve. For that, I pool the under-5 death counts and exposures from the USA between 1980-1990 from the U5MD (Guillot et al., 2022b). Then, I use the same exposure distribution of this population and re-scale it using different population size scenarios for the under-5 population: 1) 1,000, 2) 5,000, 3) 20,000, 4) 50,000, 5) 100,000, 6) 1,000,000. I assume these 6 areas belong to the same larger area, and simulate mortality curves for each of them for 10 years.

3.2 Application - Brazilian intermediate regions

After testing and checking the fit of the model using the simulated dataset, I use data from Brazilian intermediate regions (areas composed by groups of municipalities) to apply the model. In this paper, I restrict the application to one single state, Pernambuco.

The death counts come from the publicly available vital statistics microdata from the mortality information system (SIM, from Portuguese *Sistema de Informações sobre*

Mortalidade) of the Brazilian Ministry of Health. I tabulate this information into the 19 possible age groups under age 5: [0,7] days, [8,14] days, [15,21] days, [22,28] days, [29 days, 2 months], [2,3) months, [3,4) months, [4,5) months, [5,6) months, [6,7) months, [7,8) months, [8,9) months, [9,10) months, [10,11) months, [11,12) months, [1,2) years, [2,3) years, [3,4) years, [4,5) years. I ungroup the death counts between ages 1 and 2, [1,2) years, into quarters ([12, 15) months, [15, 18) months, [18, 21) months and [21, 24) months) using a penalized composite link model (Rizzi et al., 2015) in order to make the data format the same number of age groups of the principal components (22 age groups).

The population exposures are calculated using the 2000, 2010, and 2022 national censuses. I first estimate the mid-year population by single age within the intercensal period assuming constant growth rates between censuses. Then, for the age groups with an age range below 1 year, I assume that the exposure is proportional to that age range (Guillot et al., 2022a). For example, for retrieving the population exposure for the age group 0-7 days, the mid-year number of children below 1 is then multiplied by 7/365.25 (average number of days in a year).

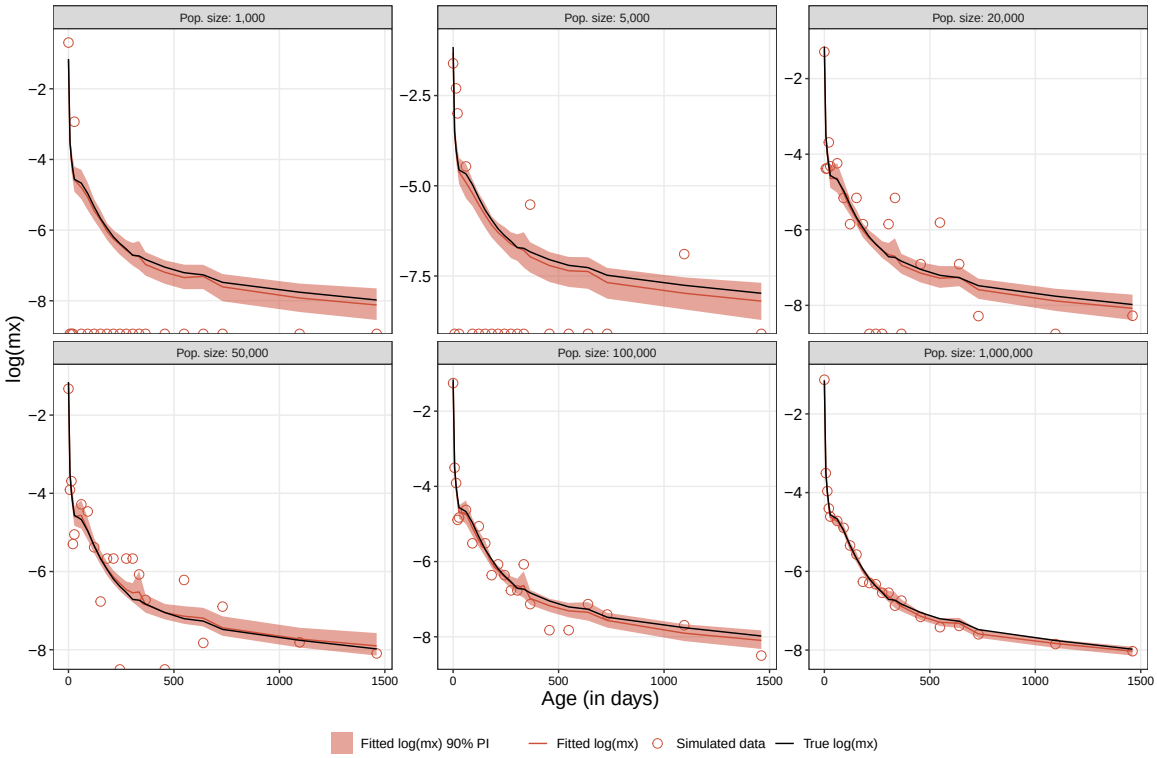
4 Preliminary Results

4.1 Simulated data

In Figure 3, we can see the comparison between the estimated underlying mortality curve for different simulated datasets with 6 different population sizes. We can see that the model performed very well in capturing the true mortality curve (black line) that was used to generate the simulated (observed) data (red dots). In all cases and population sizes, the true mortality curve lies within the 90% posterior interval of the under-5 mortality curves. As expected, the posterior intervals are larger for smaller population sizes, where the estimates shrink towards the mean value for the larger area. For larger populations (e.g., 1,000,000 children under-5),

we see that both observed, fitted, and the true mortality curves overlap.

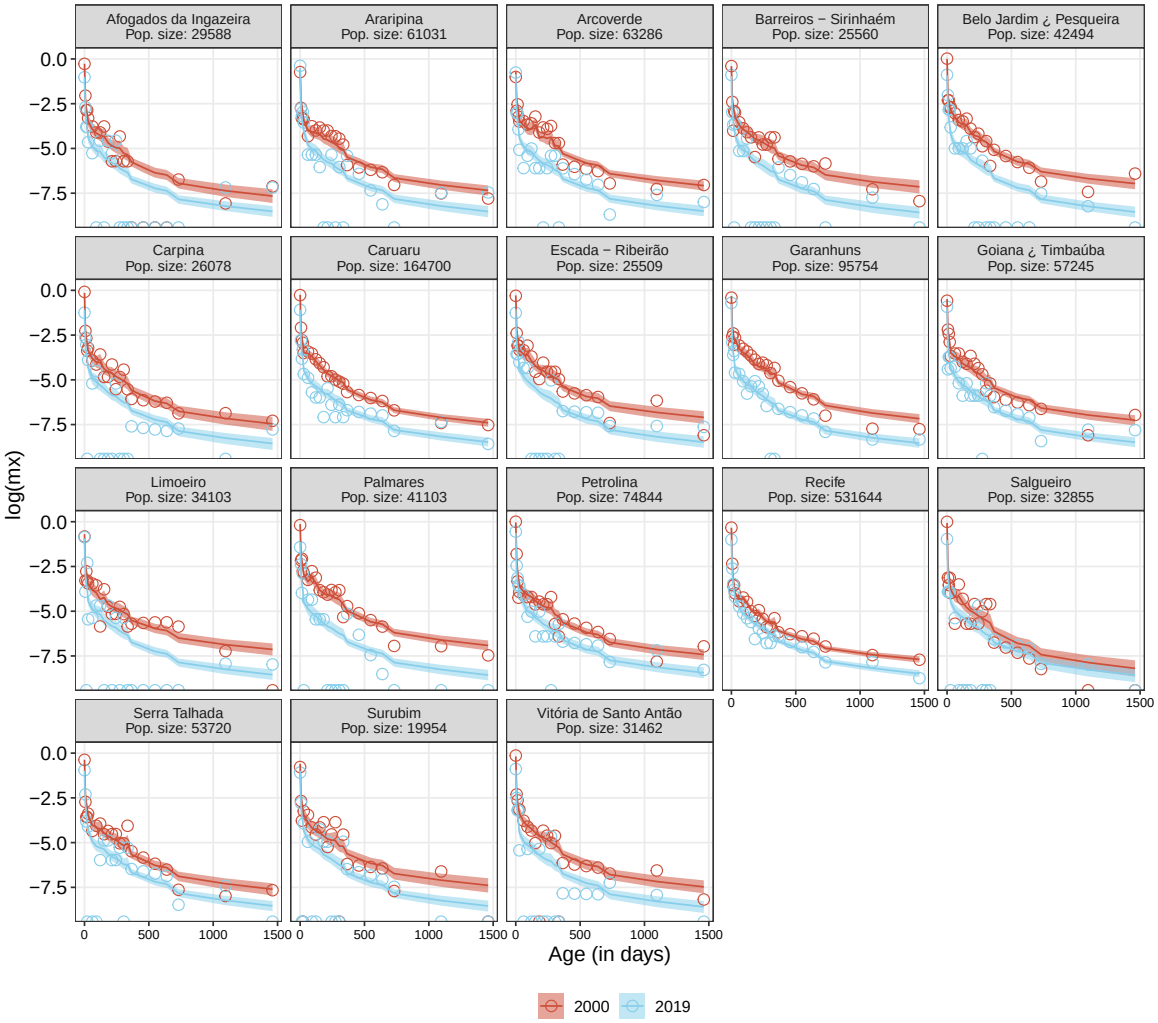
Figure 3: Estimated, observed, and true mortality curves from simulated data from areas with different population sizes of under-5 population.



4.2 Application - Pernambuco, Brazil

I then apply the model to the state of Pernambuco, Brazil. In Figure 4 we can see the fitted and observed values for years 2000 and 2019 for the intermediate regions of this Brazilian state. For all of these small regions, with the exception of Salgueiro, we can see a clear decline in infant mortality from 2000 to 2019. We can see that the estimation of the curve in 2019 using the raw data would be challenging due to the zeros in some of the age groups for several regions.

Figure 4: Estimated (lines and shaded 90% posterior intervals) and observed (circles) mortality curves for the intermediate regions of the state of Pernambuco, Brazil.



In addition, we can also evaluate the β parameters of the model. For example, the parameter β_1 reflects the overall level of under-5 mortality and we can assess it through time to evaluate if mortality has been declining. Increasing values of β_1 indicate a decrease in the overall level of under-5 mortality through time (due to the scale of the first principal component), as we can verify for all intermediate regions in Figure 5.

Figure 5: Estimated β_1 coefficients and 90% posterior intervals for the intermediate regions of the state of Pernambuco, Brazil.

5 Discussion

This work extended a previous Bayesian small areas mortality estimation approach (Alexander et al., 2017) to the under-5 mortality curve. Using three principal components, derived from model life tables and from the U5MD, that summarize the under-5 mortality information by age groups, I propose a model that was able to retrieve the true underlying mortality of a simulated dataset for different population sizes. I further applied the model to the Brazilian intermediate regions of the Pernambuco state.

Among the limitations, for now this model does not account for data quality issues.

Future developments on this model can incorporate the uncertainty of the under-5 mortality completeness by using prior information about death registration using previous estimates or values from other methods. Also, this model was developed by relying on detailed vital statistics information, and therefore it might not be well suited for countries without proper functioning civil registration and vital statistics systems. However, it could be extended to DHS surveys by retrieving the under-5 mortality information from the full birth histories questionnaires.

For the next steps, I will further explore different priors for the hyperparameters given that we had issues with convergence of the chains. One alternative is to use INLA instead of STAN for coding the model, using a different strategy to sample from the posterior distribution. Also, I might go down to the level of municipality in the application of the method to the case of Brazil.

References

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

Coale, A. J. and Demeny, P. (1966) *Regional model life tables and stable populations*. Princeton University Press.

Coale, A. J. and Guo, G. (1989) Revised regional model life tables at very low levels of mortality. *Population Index*, **55**, 613–643.

Guillot, M., Romero Prieto, J., Verhulst, A. and Gerland, P. (2022a) Modeling Age Patterns of Under-5 Mortality: Results From a Log-Quadratic Model Applied to High-Quality Vital Registration Data. *Demography*, **59**, 321–347.

Guillot, M., Romero Prieto, J., Verhulst, A. and Gerland, P. (2022b) Under-5 mortality database (U5MD). [Machine-readable database].

Okonek, T., Wilson, K. and Wakefield, J. (2024) A pseudo-likelihood approach to under-5

mortality estimation. URL: https://arxiv.org/abs/2310.11357.

Rizzi, S., Gampe, J. and Eilers, P. H. C. (2015) Efficient estimation of smooth distributions from coarsely grouped data. *American Journal of Epidemiology*, **182**, 138–147.

UNICEF (2025) Levels and trends in child mortality. *Tech. rep.*, United Nations Inter-agency Group for Child Mortality Estimation. URL: https://childmortality.org/wp-content/uploads/2025/03/UNIGME-2024-Child-Mortality-Report.pdf.

United Nations (1982) Model life tables for developing countries. *Population Studies*, **77**.

Verhulst, A., Prieto, J. R., Alam, N., Eilerts-Spinelli, H., Erchick, D. J., Gerland, P., Katz, J., Lankoande, B., Liu, L., Pison, G., Reniers, G., Subedi, S., Villavicencio, F. and Guillot, M. (2022) Divergent age patterns of under-5 mortality in south Asia and sub-Saharan Africa: a modelling study. *The Lancet Global Health*, **10**.