Cholera mapping model description

Javier Perez Saez

2024-01-30

Contents

Aim: This report aims at describing the mapping model as it currently stands.The overall aim of the model is to infer cholera incidence and incidence rates at subnational level, combining data from a variety of spatial and temporal resolutions. We here expand on the approach followed in Lessler et al. (2018).

1 Data pre-processing

Briefly, data pre-processing consists in aggregating data in time per location period by collating successive observation of the same place within a given time slice and observation collection. Other pre-processing steps include the definition of censored observations (described below), and dropping of 0-censored observations which do not contribute information to the model.

2 Base model

2.1 Overview

Suppose we are interested in modeling yearly cholera incidence (modeling time resolution of 1 year), for a period of T years (T time slices of 1 year each). Suppose that all available observation data consists of number of cases for distinct spatial locations that are all one year long (observation time resolution equals modeling time resolution). Our task is to model the space-time incidence rates over a spatial domain that covers the area of interest, and across the T years. To do so we define a modeling space-time grid at a given spatial resolution, and with time resolution of 1 years. We define the modeled cases c_i corresponding to observation y_i as:

$$c_i = \sum_{\mathcal{S}_{s,t}} \lambda_{s,t} \times \phi_{i,s} \times pop_{s,t},$$
$$\log(\lambda_{s,t}) = \gamma + w_{s,t} + \eta_t,$$

where $\mathcal{S}_{s,t}$ is the set of space-time grid cells intersecting observation i, $\lambda_{s,t}$ is the incidence rate in space-time grid cell s, t, $\phi_{i,s}$ is the population-weighted spatial fraction of space-time grid cell s, t covered by the observation, and $pop_{s,t}$ is the total population of space-time grid cell s, t. Grid cell incidence rates are modeled using a log-link as the sum of an offset term γ (typically the expected incidence rate), a spatial random effect $w_{s,t}$ and a yearly random effect η_t. We expand on these random effects below.

Observations are then linked to modeled cases through an observation model $\mathcal{M}(c_i, y_i)$. As an example, we can assume that observations follow a Poisson distribution with mean c_i:

$$y_i \sim Poisson(c_i).$$

We expand on the observation model below.

2.2 Spatial random effects

Spatial random affects can in principle take distinct values in space and time. We here restrict the model to only account for a space grid cell random effect w_s. We model these using a DAGAR prior. This prior assumes spatial auto-correlation between grid cells controlled by an overall parameter ρ. The spatial random effect w_s follows a normal distribution with mean μ_s and standard-deviation w_s conditional on ρ:

$$w_s \sim \mathcal{N}(\mu_{w_s}, \sigma_{w_s}),$$
$$\mu_{w_s} = \frac{\rho}{(1 + (nn_s - 1)\rho^2)} \sum_{u \in \Omega_s} w_u,$$
$$\sigma_{w_s} = \xi_{\sigma_w} \times \sqrt{\frac{(1 - \rho^2)}{(1 + (nn_s - 1)\rho^2)}},$$

where n_s is the number of neighbors of cell s, and Ω_s is the set of neighboring cells of s. From this formulation we can see that a same marginal standard deviation can be obtained with different values of the auto-correlation parameter ρ and scaling constant ξ_{σ_w}. This will impact our choice of prior in as described below.

2.3 Temporal random effects

We here assume that temporal random effects are independent, from a common standard-normal prior. To improve model identifiability, we impose a zero-sum constraint on the temporal random effects, enforcing a marginal standard-normal prior following Goodman (2018). Briefly, the approach employs a QR decomposition on the covariance matrix of the yearly random effect to obtain a set of random variables with a zero-sum constraint and marginal standard deviations of 1. In practice this means defining a vector of raw parameters of size $T-1$ on which priors are set, and computing the T^{th} element following the approach in Goodman (2018).

3 Accounting for partial observations

In practice, observations to not necessarily span an integer number of modeling time slices. We therefore need to account for partial time coverage of observations. We here propose two alternatives which differ in the way they make assumptions about how incidence varies during the year.

3.1 Homogenous incidence within modeling time slices

If we are willing to assume that incidence hazard rates are constant in time within a given time slice, we can then scale modeled incidence by the fraction $\Phi_{i,t}$ of the time slice covered by a given observation (we keep subscript t as an observation my span multiple time slices). The modeled cases corresponding to observation y_i is then:

$$c_i = \sum_{\mathcal{S}_{s,t}} \lambda_{s,t} \times \Phi_{i,t} \times \phi_{i,s} \times pop_{s,t},$$

with the rest of the terms described as above.

3.2 Inhomogenous incidence and censored observations

If we are unwilling to assume homogeneous incidence hazard rates as above we cannot use the temporal fraction correction. A conservative approach that does not make any assumption on how incidence rates vary within the time slice consists in treated non-full observations (observations that do not cover a whole modeling time slice) as right-censored. In other words we say that the number of cases in the whole time slice corresponding to the observation would be at least as large as the observation, had the observation covered the whole time slice. We do so by modifying the observation model using the complementary-CDF of the corresponding distribution we choose:

$$\mathcal{L}(y_i) = Pr(Y \geq y_i | c_i).$$

As an example, assuming a Poisson observation model as above, we have that:

$$\mathcal{L}(y_i) = 1 - CDF_{Poisson}(y_i, c_i).$$

4 Observation over-disperion

In the examples above we assumed a Poisson observation model. We however expect that observations across information sources for a given time slice and location to be over-dispersed. We account for over-dispersion by using a negative-binomial likelihood:

$$y_i \sim NB(c_i | \tau_{l[i]}),$$

where τ is the over-dispersion parameters controlling the variance: $var = \mu + \mu^2/\tau$. To account for difference in reporting across spatial scales, we allow for a different over dispersion parameter for each observation's administrative level $l[i]$.

5 Final model formulation

Putting together the previous sections, our final model formulation is:

$$\begin{aligned} &\text{Process model:} \\ c_i &= \sum_{\mathcal{S}_{s,t}} \lambda_{s,t} \times \phi_{i,s} \times pop_{s,t}, \\ \log(\lambda_{s,t}) &= \gamma + w_s + \eta_t. \end{aligned}$$

$$\begin{aligned} &\text{Observation model:} \\ Pr(y_i|c_i) &= \begin{cases} NB(c_i|\tau_{l[i]}), & \text{if } \Phi_{i,t} \geq a \\ 1 - CDF_{NB}(y_|c_i, \tau_{l[i]}) & \text{if } \Phi_{i,t} < a, \end{cases} \end{aligned}$$

where $a \in (0, 1)$ is the threshold of time fraction above which we consider an observation as full.

6 Priors

We use the following priors:

$$\begin{aligned} &\text{Spatial random effects:} \\ w_s &\sim DAGAR(\boldsymbol{w}, \rho, \xi_{\sigma_w}), \\ \rho &\sim beta(5, 1.5), \\ Pr(\xi_{\sigma_w}) &= \theta \times f_{\mathcal{N}}(\xi_{\sigma_w}|10, \sigma'_{w,1}) + (1-\theta) \times f_{\mathcal{N}}(\xi_{\sigma_w}|0, \sigma'_{w,2}), \\ \theta \ &\sim beta(2, 1), \\ \sigma'_{w,1} &\sim \mathcal{N}^+(0, 2), \\ \sigma'_{w,2} &\sim \mathcal{N}^+(0, 0.5), \end{aligned}$$

$$\begin{aligned} &\text{Temporal random effects:} \\ \eta_{[1:T-1]} &\sim \mathcal{N}\left(0, \frac{1}{\sqrt{1 - \frac{1}{T}}}\right), \end{aligned}$$

$$\begin{aligned} &\text{Over-dispersion:} \\ \tau_0 &= 100, \\ \frac{1}{\tau_{>0}} &\sim \mathcal{N}^+(0, 1). \end{aligned}$$

To account for the possible bi-modality of ρ in high and low regions, we use a mixture prior on the marginal standard deviation scaling factor ξ_{σ_w}. The mixture consists of two normal distributions with $f_{\mathcal{N}}$ denoting the normal probability density, one centered on 10 and one on 0, with mixture parameter θ and inferred prior standard deviations $\sigma'_{w,1}, \sigma'_{w,1}$. Set priors on ρ and θ which favor slightly high auto-correlation. The temporal random effect prior is set on the $T-1$ raw values, with the last value being computed with the QR method to ensure the zero-sum constraint (Goodman (2018)). Finally we set the negative-binomial over-dispersion parameter to 100 for the administrative level 0 (national level), corresponding to a moderate amount of over-dispersion, and set the prior for other administrative level parameter on the inverse of the parameter following Gelman (2020).

7 Simplified models

The full model described in the previous section relies on the availability of sub-national data to infer parameters related to spatial random effects. In some instance countries in our database did not have sufficient sub-national observations to draw inference with the full model. We therefore optionally ran to additional sets of models, one without spatial random effects, and the other without the mixture prior on the DAGAR variance parameter. A summary of what model formulation was retained for each country and time period is given in Supplementary Table XX.

7.1 No spatial random effects model

In the absence of subnational data, or if all subnational data was composed of zeros, we ran a simplified model without spatial random effects. The main change is therefore that the cell-level grid rate only depend on an intercept and a yearly random effect:

$$\log(\lambda_{s,t}) = \gamma + \eta_t.$$

7.2 No mixture prior on DAGAR variance

If only a limited amount of spatial data was available, defined as XXX, we ran a model that did not have a mixture on the variance parameter, the priors of the model therefore reduced to:

$$\xi_{\sigma_w} \sim \mathcal{N}^+(5, 0.5),$$

where the prior mean and standard deviation where chosen as to have a marginal variance of XX assuming a spatial autocorrelation parameter ρ of 0.9.

References

Gelman, Andrew (2020). *Prior Choice Recommendations.* en. URL: https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations (visited on 05/22/2023).

Goodman, Aaron (Sept. 2018). *Test: Soft vs Hard sum-to-zero constrain + choosing the right prior for soft constrain.* en. Section: Modeling. URL: https://discourse.mc-stan.org/t/test-soft-vs-hard-sum-to-zero-constrain-choosing-the-right-prior-for-soft-constrain/3884/31 (visited on 05/22/2023).

Lessler, Justin et al. (2018). “Mapping the burden of cholera in sub-Saharan Africa and implications for control: an analysis of data across geographical scales”. In: *The Lancet* 391.10133, pp. 1908–1915.