An R Package for Preferential Sampling Model for Spatial Prediction

March 6, 2024

1 Introduction

In practice, it is common that the selection of locations of sites where the pollutants are monitored are affected by the density of the pollutants. It is crucial to take the preferential sampling effect into account to accurately model the dispersion of the pollutant and to make predictions of pollutants either spatially or into the future.

Watson et al. (2019) proposed a framework that jointly modeling the distribution of an environmental process and a site-selection process, where the environmental process can be spatial, temporal, or spatio-temporal. By sharing the random effects between the two process, the joint model can detect the preferential sampling effects in site selection.

In this work, we develop an R package for this joint model framework for the purpose of making spatial predictions. We demonstrate this R package by applying it to the modeling and prediction of PM10 distributions in California.

2 Background

We consider a spatio-temporal environmental process Z_{st}, $s \in \mathcal{S}$, $t \in \mathcal{T}$. The space-time point is defined $(s,t) \in \mathcal{S} \times \mathcal{T}$, where $\mathcal{S}$ denoting the spatial domain of interest and $\mathcal{T}$ the temporal domain. Spatial network designer must specify a set of time points $T \subset \mathcal{T}$ at which to observe Z and at each time $t \in T$, a finite subset of sites $S_t \subset \mathcal{S}$ at which to do so.

The population of all site locations considered for selection at any time $t \in T$ is defined as $\mathcal{P} \subset \mathcal{S}$, and $\mathcal{P}$ is finite and should be specified a priori. A Bayesian model is introduced for the joint distribution of the response vector (Y_{st}, R_{st}). $R_{st} \in \{0,1\}$ is a binary response for the site selection process. By sharing random effects across the two processes, the stochastic dependence (if any) between $Y_{s,t}$ and $R_{s,t}$ and be quantified and subsequently the model can adjust the space-time predictions according to the preferential sampling effect detected. Furthermore, in the joint model, the factors affecting the initial site placement can be allowed to differ from those affecting the retention of existing sites in the network.

2.1 The joint model

We let $Y_i(t)$ denote the spatio-temporal observation process at site i, that is at locations $s_i \in \mathcal{P} \subset \mathcal{S}$, at time $t \in T$. We let $R_i(t)$ denote the random selection indicator for site $s_i \in \mathcal{P}$ at time t. We let $t_1, \ldots, t_N$ denote the N observation times, and let $r_{i,j} \in \{0,1\}$ denote the realization of $R_i(t_j)$, for $i \in \{1, \ldots, M\}$,

$j \in \{1, \ldots, N\}$, where $M = |\mathcal{P}|$. The general model framework is

$$
\begin{aligned}
(Y_{i,j} \,|\, R_{i,j} = 1) &\sim f_Y(\mu_{i,j}, \theta_Y), \quad f_Y \sim \text{density}, \\
g(\mu_{i,j}) = \eta_{i,j} &= \mathbf{x}_{i,j}^T \gamma + \sum_{k=1}^{q_1} u_{i,j,k} \beta_k(s_i, t_j), \\
R_{i,j} &\sim \mathsf{Bern}(p_{i,j}), \\
h(p_{i,j}) = \nu_{i,j} &= \mathbf{v}_{i,j}^T \alpha + \sum_{\ell=1}^{q_2} d_\ell \sum_{k=1}^{q_1} w_{i,j,\ell,k} \beta_k(s_i, \phi_{i,\ell,k}(t_j)) \\
&\quad + \sum_{m=1}^{q_3} w^\star_{i,j,m} \beta^\star_m(s_i, t_j), \\
\beta_k(s_i, t_j) &\sim \text{(possibly shared) latent effect with parameters } \theta_k, \\
&\quad k \in \{1, \ldots, q_1\}, \\
\beta^\star_m(s_i, t_j) &\sim \text{site selection only latent effect with parameters } \theta^\star_m, \\
&\quad m \in \{1, \ldots, q_3\}, \\
\Theta &= (\theta_Y, \alpha, \gamma, d, \theta_1, \ldots, \theta_{q_1}, \theta^\star_1, \ldots, \theta^\star_{q_3}) \sim \text{Priors}, \\
\mathbf{x}_{i,j} \in \mathbb{R}^{p_1}, \mathbf{u}_{i,j} &\in \mathbb{R}^{q_1}, \mathbf{v}_{i,j} \in \mathbb{R}^{p_2}, \mathbf{W}_{i,j} \in \mathbb{R}^{q_2 \times q_1}, \mathbf{w}^{\star T}_{i,j} \in \mathbb{R}^{q_3}
\end{aligned}
$$

This framework allows a range of different data types of Y to be modeled. In the linear predictor $\eta_{i,j}$, we include a linear combination of fixed covariates $\mathbf{x}_{i,j}$ with a linear combination of q_1 latent effects $\beta_k(s_i, t_j)$. These q_1 random effects can include any combinations of spatially-correlated processes (such as Gaussian [Markov] random fields), temporally correlated processes (such as autoregressive terms), spatial temporal processes and IID random effects. Note that we include the additional fixed covariates $\mathbf{u}_{i,j}$ to allow for spatially-varying coefficient models, as well as both random slopes and/or scaled random effects.

As for the site selection process $R_{i,j}$, the linear predictor $\nu_{i,j}$ may also include a linear combination of fixed covariates $\mathbf{v}_{i,j}$ with a linear combination of latent effects. In particular, the latent effects appearing in the observation process $Y_{i,j}$ are allowed to exist in the linear predictor of the selection process $R_{i,j}$. Note that the matrix $\mathbf{W}_{i,j}$ is fixed beforehand, and allow for q_2 linear combinations of the latent effects from the $Y_{i,j}$ process to be copied across. The parameter vector $\mathbf{d}$ determines the degree to which each shared latent effect affects the $\mathbf{R}$ process and therefore measure the magnitude and direction of stochastic dependence between the two models term-by-term. We allow q_3 latent effects, independent from the $Y_{i,j}$ process to exist in the linear predictor.

For added flexibility we allow temporal lags in the stochastic dependence. This allows the site-selection process to depend on the realized values of the latent effects at any time arbitrary time in the past, present or future. For example, if for a pollution monitoring network, site-selection were desired near immediate sources of pollution, then we may view as reasonable, a model that allows for a dependence between the latent field at the previous time step as a site-selection emulator. In this case, we would select as temporal lag function $\phi_{i,\ell,k}(t_j) = t_{j-1}$.

Also of interest is the possibility of setting $w_{i,j,\ell,m} = 0$ for some values of the subscripts to allow for the directions of preferentiality to change through time. For example, the initial placement of the sites might be made in a positively (or negatively) preferential manner but over time the network might be redesigned so that sites were later placed to reduce the bias. To capture this, it would make sense to have a separate PS parameter d estimated for time $t = 1$ and for times $t > 1$ to capture the changing directions of preferentiality through time. This can easily be implemented. Furthermore, we may wish to set $w_{i,j,\ell,m} = 0$ for certain values of the subscripts to see if the effects of covariates and/or the effects of preferential sampling differs between the initial site placement process and the site retention process.

2.2 A specific model

We build one model from the general framework introduced earlier. Let $t_j^\star$ denote the jth time-scaled observations that lie in the interval $[0, 1]$. The model for the observation process is

$$\begin{aligned}
(Y_{i,j} \,|\, R_{i,j} = 1) &\sim \mathcal{N}(\mu_{i,j}, \sigma_\epsilon^2) \\
\mu_{i,j} &= \gamma_0 + \gamma_1 t_j^\star + \gamma_2 (t_j^\star)^2 + b_{0,i} + b_{1,i} t_j^\star + \beta_0(s_i) + \beta_1(s_i) t_j^\star + \beta_2(s_i)(t_j^\star)^2 \qquad (1) \\
[\beta_k(s_1), \ldots, \beta_k(s_m)]^T &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma(\zeta_k)) \quad \text{for } k \in \{0, 1, 2\}, \quad \Sigma(\zeta_k) = \text{Matern}(\zeta_k) \\
[b_{0,i}, b_{1,i}] &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_b), \quad \Sigma_b = \begin{pmatrix} \sigma_{b,1}^2 & \rho_b \\ \rho_b & \sigma_{b,2}^2 \end{pmatrix}, \\
\theta &= (\sigma_\epsilon^2, \gamma, \zeta_k, \sigma_{b,1}^2, \rho_b) \sim \text{Priors}.
\end{aligned}$$

The model for site-selection process is

$$\begin{aligned}
R_{i,j} &\sim \mathsf{Bern}(p_{i,j}) \\
\text{logit}\, p_{i,1} &= \alpha_{0,0} + \alpha_1 t_1^\star + \alpha_2 (t_1^\star)^2 + \beta_1^\star(t_1) \\
&\quad + \alpha_{rep} I_{i,2} + \beta_0^\star(s_i) \\
&\quad + d_b[b_{0,i} + b_{1,i}(t_1^\star)] \\
&\quad + d_\beta[\beta_0(s_i) + \beta_1(s_i) t_1^\star + \beta_2(s_i)(t_1^\star)^2], \\
\text{for} j \neq 1 \quad \text{logit}\, p_{i,j} &= \alpha_{0,1} + \alpha_1 t_j^\star + \alpha_2 (t_j^\star)^2 + \beta_1^\star(t_j) \\
&\quad + \alpha_{ret} r_{i,(j-1)} + \alpha_{rep} I_{i,j} + \beta_0^\star(s_i) \\
&\quad + d_b[b_{0,i} + b_{1,i}(t_{j-1}^\star)] \\
&\quad + d_\beta[\beta_0(s_i) + \beta_1(s_i) t_{j-1}^\star + \beta_2(s_i)(t_{j-1}^\star)^2], \qquad (2) \\
I_{i,j} &= \mathbb{1}\left[\left(\sum_{\ell \neq i} r_{\ell, j-1} \mathbb{1}(\|s_i - s_\ell\| < c) \right) > 0 \right], \\
[\beta_0^\star(s_1), \ldots, \beta_0^\star(s_m)]^T &\sim \mathcal{N}(0, \Sigma(\zeta_R)), \Sigma(\zeta_R) = \text{Matern}(\zeta_R), \\
[\beta_1^\star(t_1), \ldots, \beta_1^\star(t_T)]^T &\sim \text{AR1}(\rho_a, \sigma_a^2), \\
\theta_R = [\alpha, d_b, d_\beta, \rho_a, \sigma_a^2, \zeta_R] &\sim \text{Priors}
\end{aligned}$$

The first component is the global effects of time on the log odds of selection. We also add first-order autoregressive deviation, $\beta_1^\star(t_j)$, from this global quadratic change. α_{ret} represents the "retention effect" reflecting how the probability a site is selected in a given year changes, conditioned on its inclusion in the previous year. Here, we share all parameters across the two processes and allow only a unique intercept to exist between the processes. α_{rep} captures the repulsion effect. $I_{i,j}$ denote an indicator variable that determines whether or not another site in the network placed within a distance c from site i was was operational at the previous time t_{j-1}. We choose the hyperparameter c to be 10 km.

This is a joint model with three processes: an observation process, an initial site-placement process and a site-retention process. We only allow for a unique intercept to exist across the two processes, sharing the remaining parameters. Only the pseudo-sites contribute a zero to the Bernoulli likelihood for the site-placement across all years. Only the sites that have been removed from the network in year j contribute a zero to the Bernoulli likelihood for the site-retention process at year j. This ensures that no site in the network was ever reinstalled after its removal.

The latent effects appearing in the observation process $Y_{i,j}$ are allowed to exist in the linear predictor of the selection process $R_{i,j}$. In particular, the two linear combinations of the latent effects, $b_{0,i} + b_{1,i}(t_1^\star)$ and $\beta_0(s_i) + \beta_1(s_i) t_{j-1}^\star + \beta_2(s_i)(t_{j-1}^\star)^2$, from the $Y_{i,j}$ process are copied across. The parameters d_b and d_β determine the degree to which each shared latent effect affects the R process and therefore measure the magnitude and direction of stochastic dependence between the two models term-by-term.

3 The implementation using inlabru

To implement the preferential sampling model defined by Eq. (1) and Eq. (2) in INLA, or **inlabru**, we are supposed to specify two models. One for the observation process in the Gaussian family and one for the site selection process in the Bernoulli family. Also, we want to share two linear combinations of latent factors between the observation model and the site selection model:

$$b_{0,i} + b_{1,i}(t_1^\star), \quad \text{and} \quad \beta_0(s_i) + \beta_1(s_i)t_{j-1}^\star + \beta_2(s_i)(t_{j-1}^\star)^2.$$

While both INLA and **inlabru** allow copying factors between models, each factor ('component' in **inlabru**) must be copied separately and therefore introduce one new scale parameter for each copied factor (by setting $fixed = FALSE$). In our model, however, we only want two scale parameters d_b and d_β for these two linear combinations of factors:

$$d_b[b_{0,i} + b_{1,i}(t_1^\star)] \quad \text{and} \quad d_\beta[\beta_0(s_i) + \beta_1(s_i)t_{j-1}^\star + \beta_2(s_i)(t_{j-1}^\star)^2],$$

where d_b and d_β are two scale parameters. This is not directly achievable using the *copy* feature in INLA or **inlabru**, and if we use the *copy* feature to copy each latent factor separately, there will be five (instead of two) new scale parameters introduced at each site and time point.

3.1 An alternative approach using auxiliary models

To copy the linear combinations of factors in implementing the model for black smoke data, Watson et al. (2019) introduced two auxiliary factors and two auxiliary Gaussian models in addition to the original joint model:

$$0 = -C_b + [b_{0,i} + b_{1,i}(t_1^\star)] \tag{3}$$

$$0 = -C_\beta + [\beta_0(s_i) + \beta_1(s_i)t_{j-1}^\star + \beta_2(s_i)(t_{j-1}^\star)^2] \tag{4}$$

where C_b and C_β are auxiliary latent factors. These individual factors, $b_{0,i}$, $b_{1,i}(t_1^\star)$, $\beta_0(s_i)$, $\beta_1(s_i)t_{j-1}^\star$, $\beta_2(s_i)(t_{j-1}^\star)^2$, are copied separately from the observation model Eq. (1) to the two auxiliary models, Eq. (3) and Eq. (4), with the argument $fixed = TRUE$.

By setting the precision parameter of the two factors C_b and C_β to be ≈ 0 and setting the precision parameter of the two Gaussian auxiliary models to be $\approx \infty$, the latent factors C_b and C_β duplicate of the two factor combinations:

$$C_b = b_{0,i} + b_{1,i}(t_1^\star) \quad \text{and} \quad C_\beta = \beta_0(s_i) + \beta_1(s_i)t_{j-1}^\star + \beta_2(s_i)(t_{j-1}^\star)^2.$$

Given the two auxiliary models, the new model for site-selection process copies C_b and C_β from Eq. (3) and Eq. (4) instead with the argument $fixed = FALSE$:

$$\begin{aligned}
& \operatorname{logit} p_{i,1} = \alpha_{0,0} + \alpha_1 t_1^\star + \alpha_2(t_1^\star) + \beta_1^\star(t_1) \\
& \qquad + \alpha_{rep} I_{i,2} + \beta_0^\star(s_i) \\
& \qquad + d_b C_b + d_\beta C_\beta, \\
\text{for} j \neq 1 \quad & \operatorname{logit} p_{i,j} = \alpha_{0,1} + \alpha_1 t_j^\star + \alpha_2(t_j^\star)^2 + \beta_1^\star t_j \\
& \qquad + \alpha_{ret} r_{i,(j-1)} + \alpha_{rep} I_{i,2} + \beta_0^\star(s_i) \\
& \qquad + d_b C_b + d_\beta C_\beta.
\end{aligned}$$

With the auxiliary models and factors, it is possible to copy the linear combination of factors without introducing too many scale parameters. However, this approach requires us to fit four, instead of two models in INLA(or **inlabru**), and in general, more auxiliary models and factors will be required if more linear combinations of factors need to be shared between the observation process and the site selection process.

4 The Preferential Sampling Model

The population of sites considered for selection should also be selected carefully. Different choices of the population leads to different conclusions about the PS effect. In one case the population is all sites that have been monitored at some times $t \in T$, and the estimate of the mean value of the PM10 can be interpreted as the network average. By using this population, the model help us detect the effect of PS on estimates of the density of PM10s across all sites ever observed. In the other case, we include all vertices of the mesh grid that are inside the border in the population and we treat those unobserved vertices as pseudo site locations. These pseudo sites are placed at a density of approximately 3 km throughout SOCAB region, and in this case, the estimate of the mean value of the PM10 in this case can be interpreted as the PM10 density across the SOCAB region. Since we are uniformly cover the SOCAB region, this population help us detect if the observed sites are preferentially selected and the effect of PS on estimating the mean of PM10 over the entire SOCAB region.

A Bayesian model is introduced for the joint distribution of the response vector (Y_{st}, R_{st}), where R_{st} is a binary response for the site selection process. By sharing random effects across the two processes the stochastic dependence between the observation and the site selection can be detected and adjust the predictions. In particular, the site selection process is allowed to use information from both spatially varying Gaussian processes and spatially-uncorrelated site-specific effects to determine the site selection probabilities each year. We fit the same preferential sampling model for the two populations.

5 PM10 in California

The annual concentration of PM10s from 1965 can be download from the website (`https://www.epa.gov/outdoor-air-quality-data`) of the U.S. Environmental Protection Agency (EPA). We download the annual records of PM10 in California between 1985 to 2022. The raw data set include locations, year, and some summery statistics of measurements of all sites in California. The complete information of the data set can be found in the EPA website (`https://aqs.epa.gov/aqsweb/airdata/FileFormats.html#_annual_summary_files`). The raw data set downloaded also include records of other air pollutants, but we keep only the PM10 records.

We keep the annual mean of PM10 measurements to represent the PM10 level at each site. Sometimes exceptional events happened and can affect the measurements of air pollutants, but the local agency has no control over. A wildfire is an example of an exceptional event. We use the summary statistics which remove the affects of extreme events.

The site locations of these sites can be seen from Fig. Note that each measurement site might has multiple monitors planted in close but different locations. We combine the measurement of different monitors of each site by taking the arithmetic average of bothe the locations and PM10 measurements.

The decline trend in concentrations of PM10s from 1965 to can be seen from Fig. The sites are added to the network and dropped. It can be seen from the plot that sites remained in the network until the end are those with higher measurements.

5.1 The PM10 Data

A few data cleaning steps were carried out before fitting the models. Due to the right skewness of the PM10 observation distribution, we applied the natural logarithmic transformation to the values to make the observation more Gaussian in shape. Before the log transformation, we firstly divide each value by mean of all recorded values to make the response dimensionless. We scale the Eastings and Northings coordinates and the unit is 100 km. We scaled the years to lie in the interval $[0, 1]$ to stabilize the temporal polynomials used in later analysis.

5.2 Data Preprocessing

In order to make sure the the assumptions on the distributions of data is reasonable, some data cleaning and preprocessing is required before we fit the PS model. Due to the right skewness of the PM10 observations, we applied the natural logarithmic transformation to the values to make the observations Gaussian distributed.

To make the fitted model interpretable, we then subtract the logarithmic transformation of the mean value so that the data is dimensionless.

5.3 Map projection and Mesh Grid

The site locations in the data set are recorded as latitude and longitude under different coordinate reference systems (CRS). In order to better represent the distance between sites, we project all site locations to the UTM (Easting/Northing) coordinates with the measurement unit being kilometer.

The border map of SOCAB region is also projected to the same CRS as the site locations, and we keep the sites only in the SOCAB region.

To increase the numerical stability in model fitting, we rescale the Eastings and Northings coordinates of sites and the SOCAB border by 10, and each unit distance represent 10 km.

We create the mesh grid using the function $mesh_2d_inla$.

The same mesh is used in both implementations.

6 Model Fitting

We fit the same model on two populations using R-inlabru package. Inlabru is built upon the R-INLA package with simplified syntax. The R-INLA package apply the SPDE approach to add the This enables the rapid computation of approximate Bayesian posterior distribution of the model parameters and random effects. The R-INLA packages approximates the Gaussian Markov random field by solving an SPDE on a triangulation grid.

References

Watson, J., Zidek, J. V., and Shaddick, G. (2019). A general theory for preferential sampling in environmental networks. *The Annals of Applied Statistics*, 13(4):2662 – 2700.