Preferential Sampling Effect of PM10 in SOCAB Region

April 4, 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 the south coast air basin(SOCAB) region 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. In practice, the network designers need to specify a set of time points $T \subset \mathcal{T}$ at which to measure the pollutant of interest, and a finite subset of sites $S_t \subset \mathcal{S}$ at which to do so.

To model the spatial and temporal distribution of the pollutant, we use a discrete approximation to the environmental process, where we model the statistical distribution of the pollutant on a fixed grid containing finite points. This discrete grid is treated as the population, which contains all site locations considered for selection at any time $t \in T$, and it should be specified a priori.

To study the preferential sampling effect in selecting the site locations, a Bayesian model is introduced for the joint distribution of the response vector (Y_{st}, R_{st}). Where Y_{st} is the observation process of the pollutant, and $R_{st} \in \{0,1\}$ is a binary response for the site selection process. The idea behind the joint modeling framework is that by sharing random effects across the two processes, the stochastic dependence between $Y_{s,t}$ and $R_{s,t}$ and be detected and quantified. As a result, the model can adjust the space-time predictions according to the preferential sampling effect detected.

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 $M = |\mathcal{P}|$. Let $t_j^\star$ denote the jth time-scaled observations that lie in the interval $[0,1]$.

The model for the observation process is

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

The model for site-selection process is

$$
\begin{aligned}
R_{i,j} &\sim \mathsf{Bern}(p_{i,j}) \\
\operatorname{logit} p_{i,j} &= \alpha_{0,1} + \alpha_1 t^\star_j + \alpha_2 (t^\star_j)^2 + \beta^\star_1(t_j) \\
&\quad + \alpha_{ret} r_{i,(j-1)} + \alpha_{rep} I_{i,j} + \beta^\star_0(s_i) \\
&\quad + d_b[b_{0,i} + b_{1,i}(t^\star_{j-1})] \\
&\quad + d_\beta[\beta_0(s_i) + \beta_1(s_i) t^\star_{j-1} + \beta_2(s_i)(t^\star_{j-1})^2], \qquad (2) \\
I_{i,j} &= \mathbb{1}\left[\left(\sum_{\ell \neq i} r_{\ell,j-1} \mathbb{1}(\|s_i - s_\ell\| < c) \right) > 0 \right], \\
[\beta^\star_0(s_1), \ldots, \beta^\star_0(s_m)]^T &\sim \mathcal{N}(0, \Sigma(\zeta_R)), \Sigma(\zeta_R) = \text{Matern}(\zeta_R), \\
[\beta^\star_1(t_1), \ldots, \beta^\star_1(t_T)]^T &\sim \text{AR1}(\rho_a, \sigma^2_a), \\
\theta_R = [\alpha, d_b, d_\beta, \rho_a, \sigma^2_a, \zeta_R] &\sim \text{Priors}
\end{aligned}
$$

In the linear predictor $\mu_{i,j}$ for the observation process, we include a linear combination of fixed covariates with a linear combination of latent effects. These random effects include combinations of spatially-correlated processes, and IID random effects. Note that this framework allows for random slopes.

As for the site selection process $R_{i,j}$, the linear predictor may also include a linear combination of fixed covariates with a linear combination of latent effects. The first component is the global effects of time on the log odds of selection. We also add first-order autoregressive deviation, $\beta^\star_1(t_j)$, 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. $I_{i,j}$ 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.

To detect the preferential sampling effect, 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, two linear combinations of the latent effects from the $Y_{i,j}$ process are copied across. The parameter vector $[d_b, d_\beta]$ determines the degree to which each shared latent effect affects the site selection process and therefore measure the magnitude and direction of stochastic dependence between the two models.

For added flexibility the model includes 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. For a pollution monitoring network in reality, it is reasonable to allows for a dependence between the latent field at the previous time step as a site-selection emulator.

3 The implementation using inlabru

We take the Bayesian approach and perform posterior inference of the latent effect and parameters. Instead of applying MCMC, we use the integrated nested Laplace approximation method Rue et al. (2009) to approximate the posterior marginals of the random effects and parameters. 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.

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. 1. 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 both the locations and PM10 measurements.

The decline trend in concentrations of PM10s from 1985 to 2022 can be seen from Fig. 2. 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. The trend of variance of log(PM10) can be seen from 3.

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 taking the log transformation, we firstly divide each value by mean of all recorded values to make the response dimensionless. We scale the East and North coordinates and the unit is 10 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.

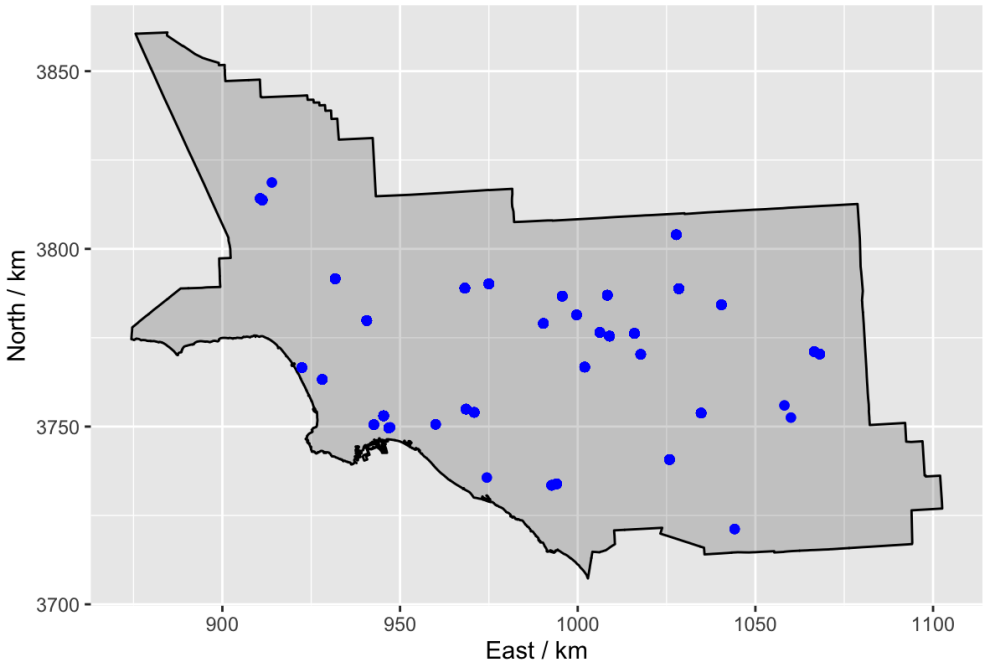

Figure 1: The sites in the SOCAB region. Each blue dot represent a site. If a site has multiple monitors, the site location is determined by taking the average of coordinates of all monitors.

We create the mesh grid using the function $mesh_2d_inla$.
The same mesh is used in both implementations.

6 Model Fitting

In spatial statistics it is common to formulate mixed-effects regression models in which the linear predictor is made of a trend plus a spatial variation The trend usually is composed of fixed effects or some smooth terms on covariates, while the spatial variation is usually modeled using correlated random effects. Spatial random effects often model (residual) small scale variation and this is the reason why these models can be regarded as models with correlated errors.

Lindgren et al. (2011) describe an approximation to continuous spatial models with a Matérn covariance that is based on the solution to a stochastic partial differential equation (SPDE). A Gaussian spatial process with Matérn covariance is a solution to SPDE This approximation is computed using a sparse representation that can be effectively implemented using the integrated nested Laplace approximation (INLA, Rue et al., 2009)

INLA focuses on models that can be expressed as latent Gaussian Markov random fields (GMRF)

INLA can handle models with more than one likelihood. By using a model with more than one likelihood it is possible to build a joint model with different types of outputs and the hyperparameters in the likelihoods will be fitted separately.

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

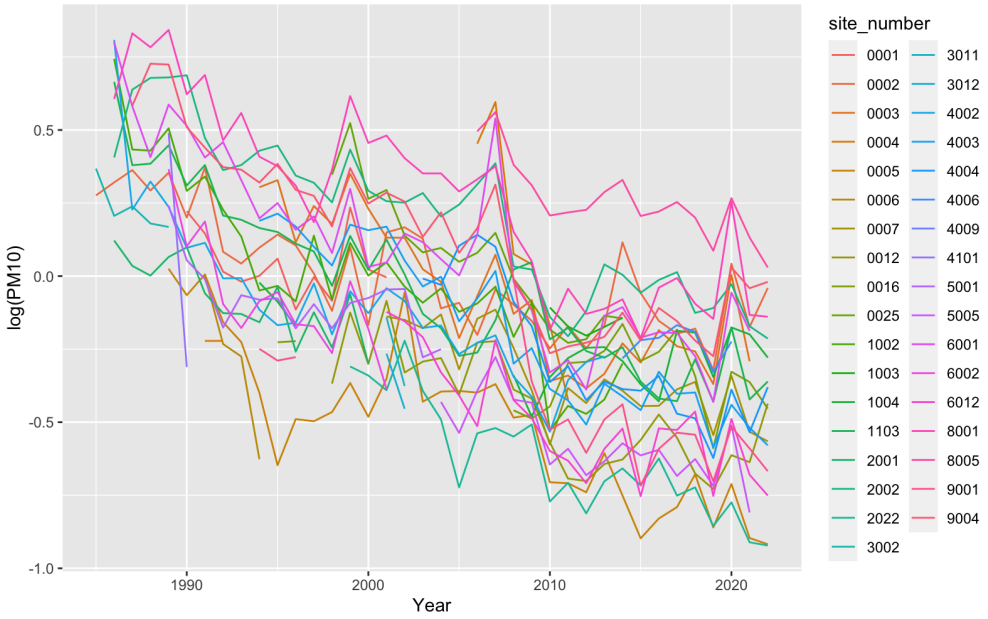

Figure 2: The sites in the SOCAB region. Each curve represents a measurement history of one site.

a triangulation grid. The goal of inlabru is to facilitate spatial modeling using integrated nested Laplace approximation via the R-INLA package.

The data set **PM10s_SOCAB** has 35×38 (number of site $\times$ number of years) rows, where each row represent the measurement of one site in a given year. If a site were not measured in some years, the measurement was noted as missing. The following variables:

- **annual_mean**: The logarithmic transformation of annual mean value of PM10s, which acts as the response variable in the regression model.
- **slc**: A dummy variable (0 or 1) indicating whether a site was selected in each year. This is the response variable in the site selection model.
- **slc_lag**: A dummy variable indicating whether a site was selected in last year.
- **site_number**: The indicator of each site, which is used as the group indicator for random effects
- **locs**: The North/East coordinates (unit 10 km) of sites.
- **year**: The year in which a observation was recorded.
- **time**: The standardized years.
- **repulsion_ind**: A dummy variable indicating whether there was other sites in the close neighbor of a site last year.
- **zero**: A vector of zeros that is used as the auxiliary variable in the auxiliary models.

We fit the joint model using the R-inlabru package. Our joint model includes four likelihoods, which includes a Gaussian likelihood for the PM10 concentration process, a binomial likelihood for the site selection process, and two auxiliary Gaussian likelihoods that are introduced to share linear combinations of random effects across the PM10 concentration process and the site selection process.

According to the syntax of the R-inlabru package, all components (including the shared ones) of all likelihoods need to be firstly claimed at once and then used in defining the likelihoods. Each effect is defined using a user-assigned name, the variable, and the random distribution. For example

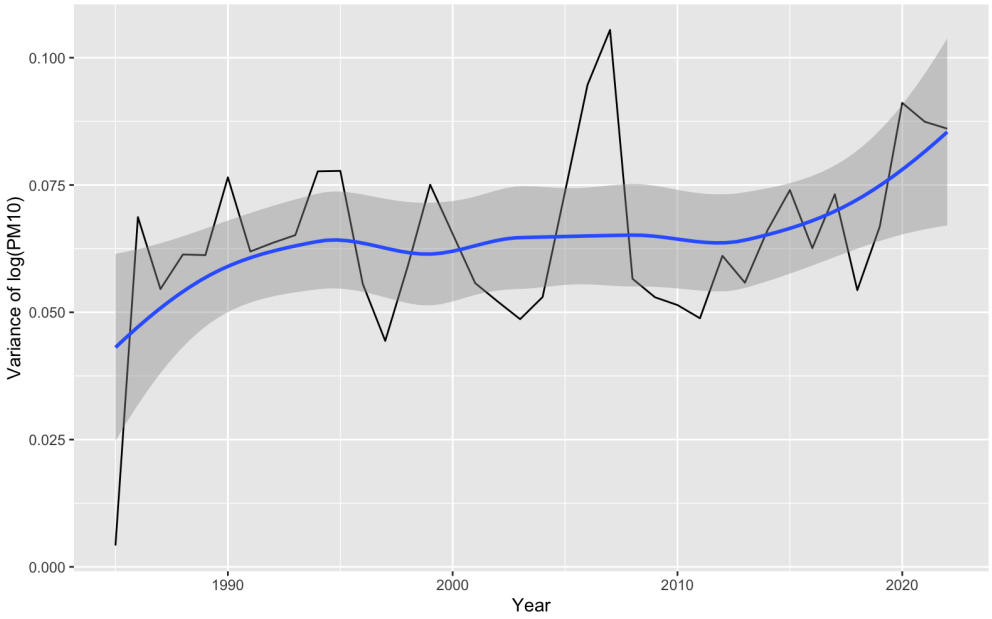

Figure 3: The sites in the SOCAB region.

```
components <- ~
  # Components for observation model
  intercept_obs(1) +
  time_1_obs(time) +
  time_2_obs(time^2) +
  random_0_obs(site_number, model = "iid2d", n = no_sites*2, constr=TRUE) +
  random_1_obs(site_number, weights = time, copy = "random_obs0") +
  spatial_0_obs(locs, model = spde_obj) +
  spatial_1_obs(locs, weights = time, model = spde_obj) +
  spatial_2_obs(locs, weights = time^2, model = spde_obj) +
  # Components for site selection model
  intercept_slc(1) +
  time_1_slc(time) +
  time_2_slc(time^2) +
  lag_slc(slc_lag) +
  repuls_slc(repulsion_ind) +
  ar_slc(year, model='ar1', hyper=list(theta1=list(prior="pcprec",param=c(2, 0.01)))) +
  spatial_slc(locs, model = spde_obj) +
  share_aux1(site_number, copy = "comp_aux1", fixed = FALSE) +
  share_aux2(site_number, copy = "comp_aux2", fixed = FALSE) +
  # Components for the first auxiliary model
  random_0_aux1(site_number, copy = "random_obs0", fixed = TRUE) +
  random_1_aux1(site_number, weights = time, copy = "random_obs11", fixed = TRUE) +
  comp_aux1(site_number, model = 'iid',
            hyper = list(prec = list(initial = -20, fixed=TRUE))) +
  # Components for the second auxiliary model
  spatial_0_aux2(locs, copy = "spatial_0_obs", fixed = TRUE) +
  spatial_1_aux2(locs, weights = time, copy = "spatial_1_obs", fixed = TRUE) +
  spatial_2_aux2(locs, weights = time^2, copy = "spatial_2_obs", fixed = TRUE) +
  comp_aux2(site_number, model = 'iid',
            hyper = list(prec = list(initial = -20, fixed=TRUE)))
```

The observation likelihood We assume that the concentration of PM10s in SOCAB follow a normal distribution. The mean value of the concentration is modeled as a combination of a fixed effect and random effects. The fixed effect is a quadratic function of time, and the random effects include spatially correlated process, which is a Gaussian Markov random field, and a site-specific random effect.

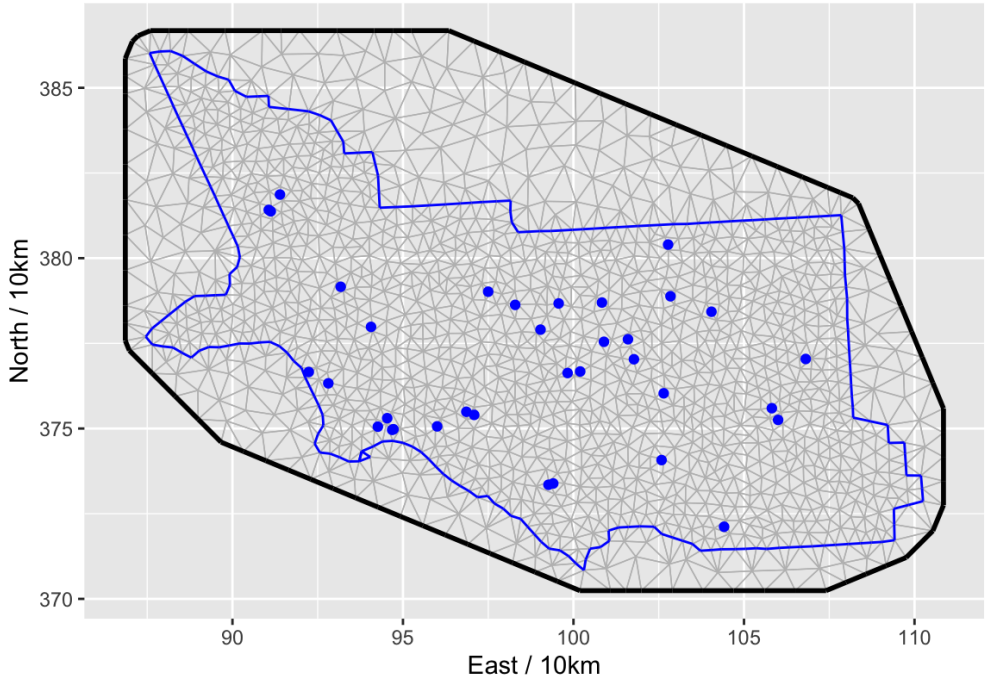

Figure 4: The meshgrid in the SOCAB region.

```
like_obs <- like(formula = annual_mean ~ intercept_obs + time_1_obs + time_2_obs +
                                         random_0_obs + random_1_obs +
                                         spatial_0_obs + spatial_1_obs + spatial_2_obs,
                 family = "gaussian",
                 data = PM10s_SOCAB)
```

The site selection is assumed to follow a Binomial distribution, which is connected to the linear predictor via a logistic transformation. The fixed effect include a quadratic term and terms indicating a site was selected a year before, and a variable indicating the presence of other sites within certain distance from a site. The random effects including a spatially correlated effect, and a temporally correlated effect. In order to detect the preferential sampling effect, the random effects in the observation model are added to the observation process. The shared effects across two models allow for the stochastic dependence between the two models.

```
like_slc <- like(formula = slc ~ intercept_slc + time_1_slc + time_2_slc +
                                 lag_slc + repuls_slc + ar_slc + spatial_slc +
                                 share_aux1 + share_aux2,
                 family = "binomial",
                 Ntrials = rep(1, times = length(PM10s_SOCAB$slc)),
                 data = PM10s_SOCAB)
```

While the R-INLA package and the R-inlabru package allows for components sharing across models, sharing of the linear combination of multiple components is not straightforward. To share a linear combination of components between the two likelihoods, we introduce an auxiliary model with zero mean and a vary large variance, and an auxiliary variable to copy the (negative) joint effect of the linear combination of multiple effects.

```
like_aux1 <- like(formula = zero ~ random_0_aux1 + random_1_aux1 + comp_aux1,
                  family = "gaussian",
                  data = PM10s_expand)
```

```
like_aux2 <- like(formula = zero ~ spatial_0_aux2 + spatial_1_aux2 + spatial_2_aux2 +
                                   comp_aux2,
                  family = "gaussian",
                  data = PM10s_SOCAB)
```

```
bru_options_set(bru_max_iter = 20,
                control.inla = list(strategy = "gaussian", int.strategy = 'eb'),
                control.family = list(
                    list(),
                    list(),
                    list(hyper = list(prec = list(initial = 20, fixed=TRUE))),
                    list(hyper = list(prec = list(initial = 20, fixed=TRUE)))),
                bru_verbose = T)
fit_bru <- bru(components,
               like_obs, like_slc, like_aux1, like_aux2)
```

7 Preferential Sampling Effects

References

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(2):319–392.

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.