Understanding the spatio-temporal distribution of pollen in Toronto

Sara Zapata Marin

2022-11-15

Hello

For those who don't know me ...

- I come from Mexico and I have an undergrad in Physics Engineering
- During my Ph.D I worked with Alex M. Schmidt developing spatio-temporal models for environmental data
- She taught me everything I know about Bayesian statistics and spatial statistics
- Now I am working as a Postdoctoral researcher

What was my PhD about?

- Spatial statistical models for environmental processes.
  1. Pollen distribution in Toronto
  2. Distribution of Volatile Organic Compounds in Montreal

Art by Diane Obomsawin

Outline

1. Introduction
2. Pollen dispersion in Toronto
    - Accounting for the zeros
    - Temporal misalignment
3. Future work
4. Conclusions

Spatial data

Point-referenced

Measurements are taken at fixed locations (e.g., monitoring stations)

Areal

Measurements are summarized over a region (e.g., neighborhoods, provinces, counties, etc.)

Point pattern

The variables of interest are the locations of a random event (e.g. species distribution models)

Linear regression model

The basic linear regression approach is used to model the effect of independent variables on the value of a dependent variable,

$$\boldsymbol{Y} = \boldsymbol{X}'\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

- $\boldsymbol{Y} = (y_1, \ldots, y_n)'$ is a vector of observations,
- $\boldsymbol{X}$ is a $n \times k$ matrix of independent variables and an intercept,
- $\boldsymbol{\epsilon}$ the vector of errors for each observation
- and we assume that the errors, $\epsilon_i \sim N(0, \sigma^2)$ are independent.

How to accomodate for spatial structure?

- The data usually consist of measurements of $\boldsymbol{X}$ and $\boldsymbol{Y}$ taken at specific locations (s), so we have $Y(s)$ and $X(s)$.
- Sometimes, the locations are ignored and other estimation techniques are used to understand the effects of $\boldsymbol{X}$ on $\boldsymbol{Y}$.
- How to incorporate the spatial locations into our model?

Tobler's first law of geography

> *"everything is related to everything else, but near things are more related than distant things"*
>
> *--- Tobler W., (1970)*

Spatial autocorrelation

- Observations that are closer to each other tend to be more alike (**location similarity**).
- We want to account for spatial dependence in the error term and the dependent variable.
- **Spatial autocorrelation** describes the degree to which observations at specific locations are similar to each other.
- It is necessary to impose sufficient constraints such that a finite number of parameters that characterize the **spatial structure** can be efficiently estimated.

Spatial autocorrelation

- In **geostatistics** the covariance between locations can be expressed as a continuous function of the distance between them, such that,

$$Cov(Y(s_i), Y(s_{i'})) = \sigma^2 \rho(d_{ii'})$$

where $d_{ii'}$ is the distance between the two locations.

- So the model becomes,

$$Y(s) = \boldsymbol{X}'(s)\boldsymbol{\beta} + \epsilon(s) + \nu(s),$$

where $\epsilon(s) \sim N(0, \tau^2)$ and $\boldsymbol{\nu} = (\nu(s_1), \ldots, \nu(s_n))$ follows a multivariate normal distribution with mean zero and covariance $\sigma^2 \rho(\boldsymbol{d})$.

Spatio-temporal models

- In environmental processes we are not only interested in the spatial structure, but also how this surface changes over time.
- Sometimes, ongoing temporal measurements are collected at monitoring sites producing long-time series of data.
- One option is to assume that $Y_t(s)$ is described by a spatio-temporal model of the form,

$$Y_t(s) = \boldsymbol{X}_{\boldsymbol{t}}'(s)\boldsymbol{\beta_t}(s) + e_t(s)$$

where $\boldsymbol{\beta_t}(s)$ can be constant across time, space, or both.

The residual $\boldsymbol{e_t}(s)$ can be rewritten as the sum of two independent processes: **white noise** $\boldsymbol{\epsilon_t}(s)$ and a **spatio-temporal process** $\boldsymbol{\omega_t}(s)$.

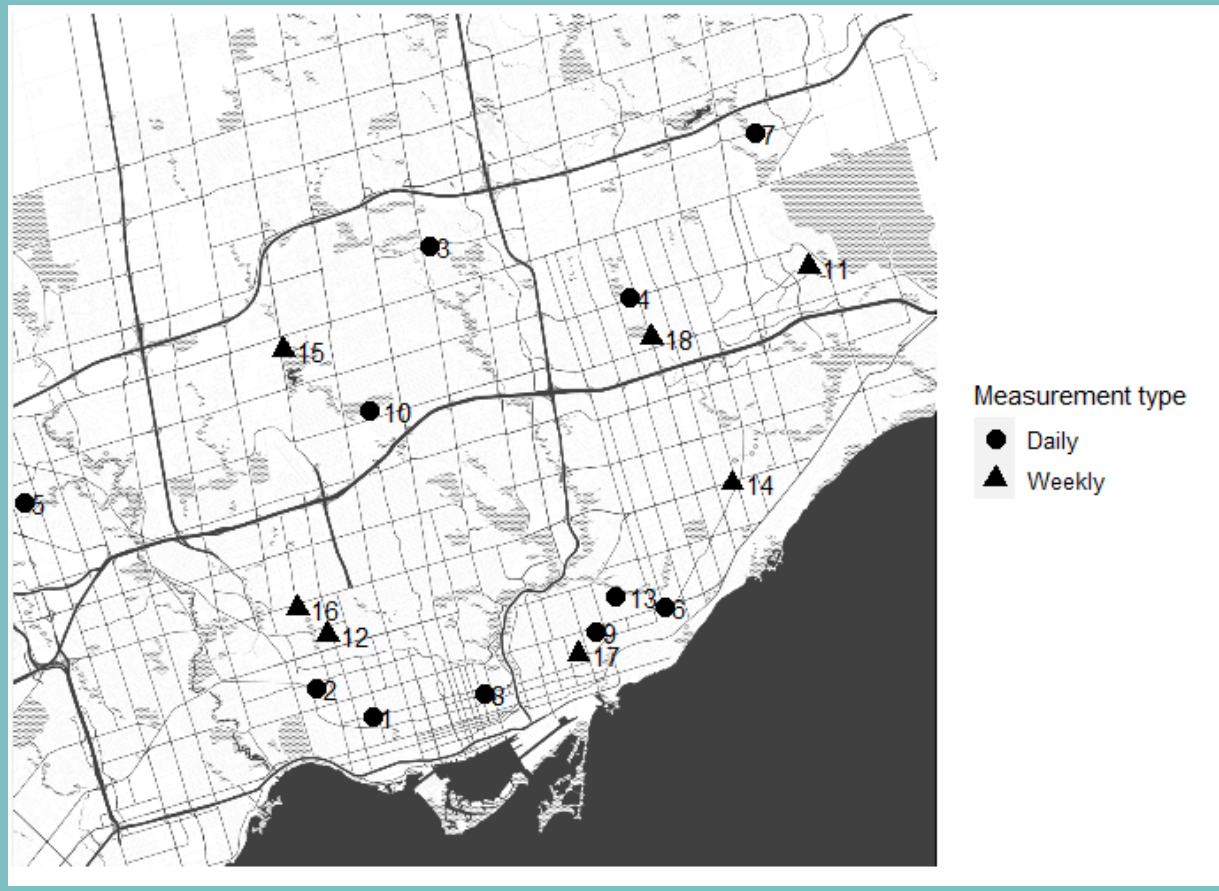
Spatio-temporal modelling of pollen concentration in Toronto, Canada

Motivation

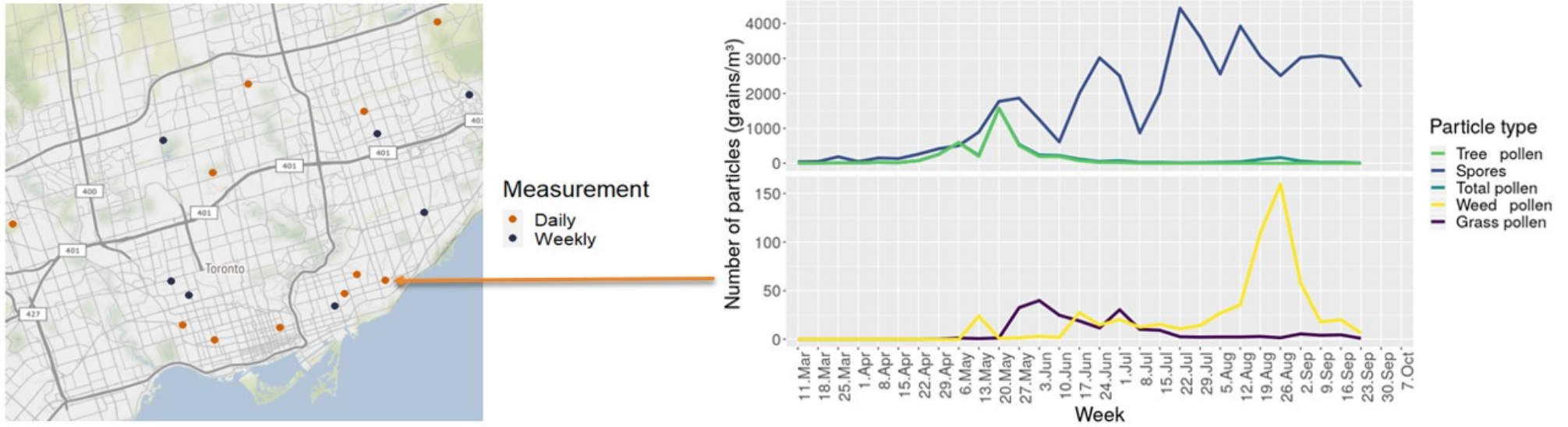
- Little is known about the spatial distribution of pollen concentrations within cities as pollen levels are typically recorded at a single monitoring site.
  - Weinberger et al., 2018: Intra-urban variation in tree pollen in NYC
  - Hjort et al., 2016: Land use regression model for grass pollen in the Helsinki area
- “Epidemiological studies using a single pollen station as a proxy for pollen concentrations are prone to significant measurement error if the study area is climatically variable.” (Katz et al. 2019, Katz and Batterman 2020)
- Further evidence is needed in order to better establish the spatio-temporal exposure to pollen within a study area

The data

- Daily and weekly measurements of pollen concentration in Toronto, for: grass pollen, weed pollen, tree pollen and total pollen.
- Local variables: distance to highways, road land use, buildings land use, etc.
- Weather-related variables: temperature, precipitation, humidity, wind speed and direction.



The data


Toronto
400
401
427
Measurement
Daily
Weekly
Number of particles (grains/m³)
4000
3000
2000
1000
0
150
100
50
0
11.Mar
18.Mar
25.Mar
1.Apr
8.Apr
15.Apr
22.Apr
29.Apr
6.May
13.May
20.May
27.May
3.Jun
10.Jun
17.Jun
24.Jun
1.Jul
8.Jul
15.Jul
22.Jul
29.Jul
5.Aug
12.Aug
19.Aug
26.Aug
2.Sep
9.Sep
16.Sep
23.Sep
30.Sep
7.Oct
Week
Particle type
Tree pollen
Spores
Total pollen
Weed pollen
Grass pollen

The data

Objective 1: Account for the high number of zeros

Methods

To account for the high number of zeros we propose a Hurdle-log-normal model which is a mixture between a Bernoulli distribution and a log-normal distribution defined as,

$$p(y_t(s) \mid \rho_t(s), \lambda_t(s)) = \begin{cases} \rho_t(s) & y_t(s) = 0, \\ (1 - \rho_t(s)) p(y_t(s) \mid \lambda_t(s)) & y_t(s) > 0, \end{cases}$$

where $y_t(s)$ is the pollen concentration at week t and location s.

- The positive values follow a lognormal distribution with parameters $\mu_t(s)$ and σ^2,

$$(y_t(s) \mid \mu_t(s), \sigma^2) \sim lognormal(\mu_t(s), \sigma^2),$$

Methods

Where $\mu_t(s)$ is defined as,

$$\mu_t(s) = \boldsymbol{\alpha x}(s) + \boldsymbol{\beta u_t}(s) + \boldsymbol{\gamma z}_t + \theta_t$$

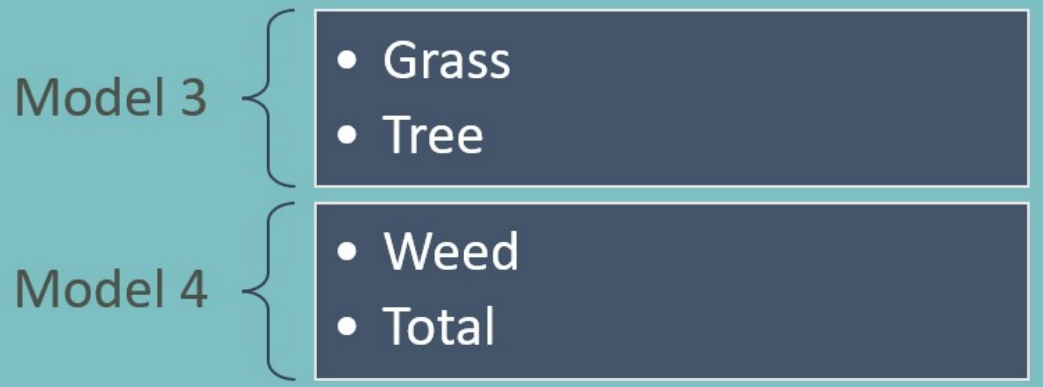
- $\boldsymbol{x}(s)$ is a vector of spatial covariates
- $\boldsymbol{u}_t(s)$ is a vector containing a vegetation measurement (TC greenness) and a ground composition measurement (TC brightness) for each site s at each time t.
- $\boldsymbol{z}_t$ is a vector of weather-related covariates
- θ_t is a time-varying mean which will capture the weekly overall temporal mean across sites and it follows

$$\theta_t = \theta_{t-1} + \omega_t, \qquad \omega_t(s) \sim N(0, W)$$

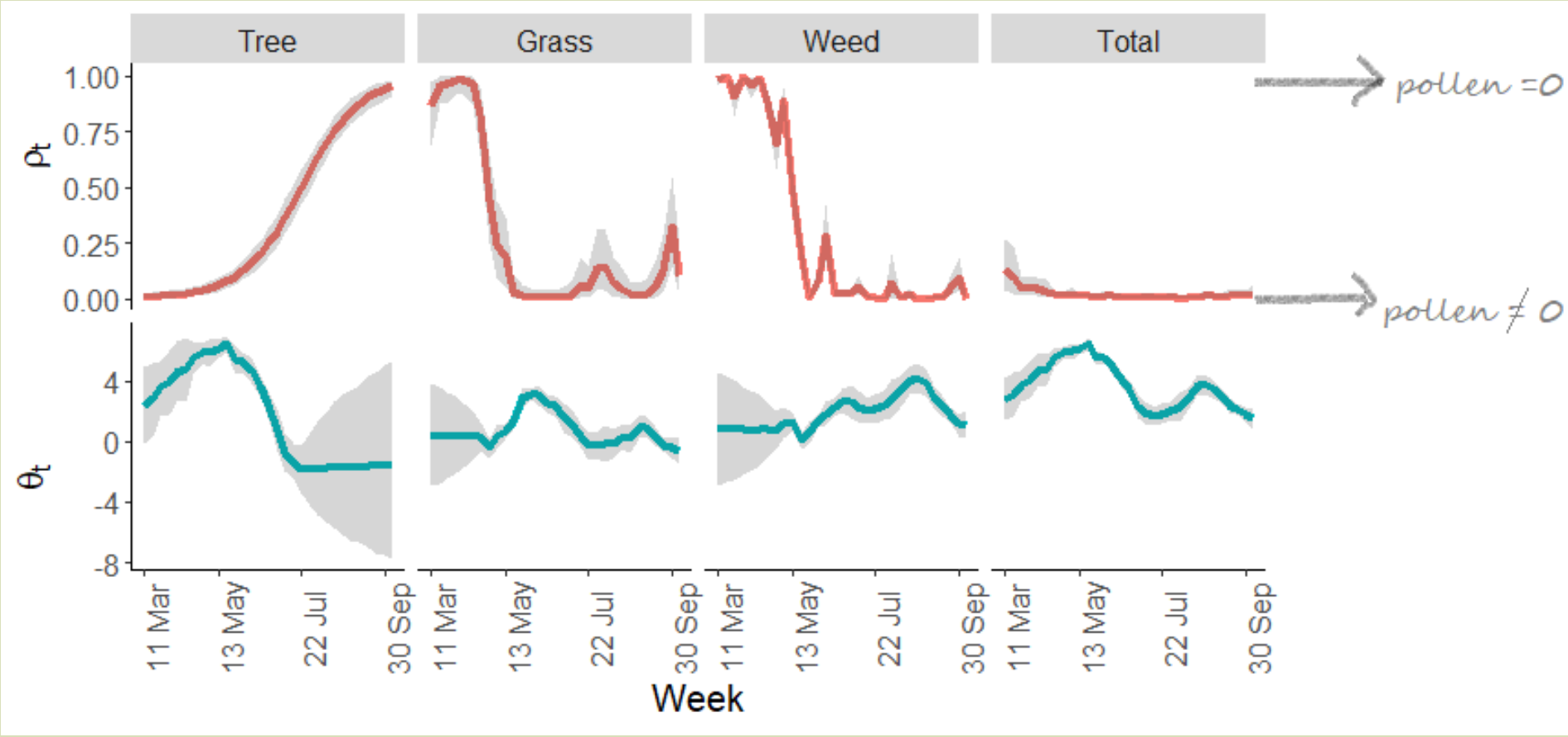
Methods

We tested 4 models with different definitions of $\rho_t(s)$

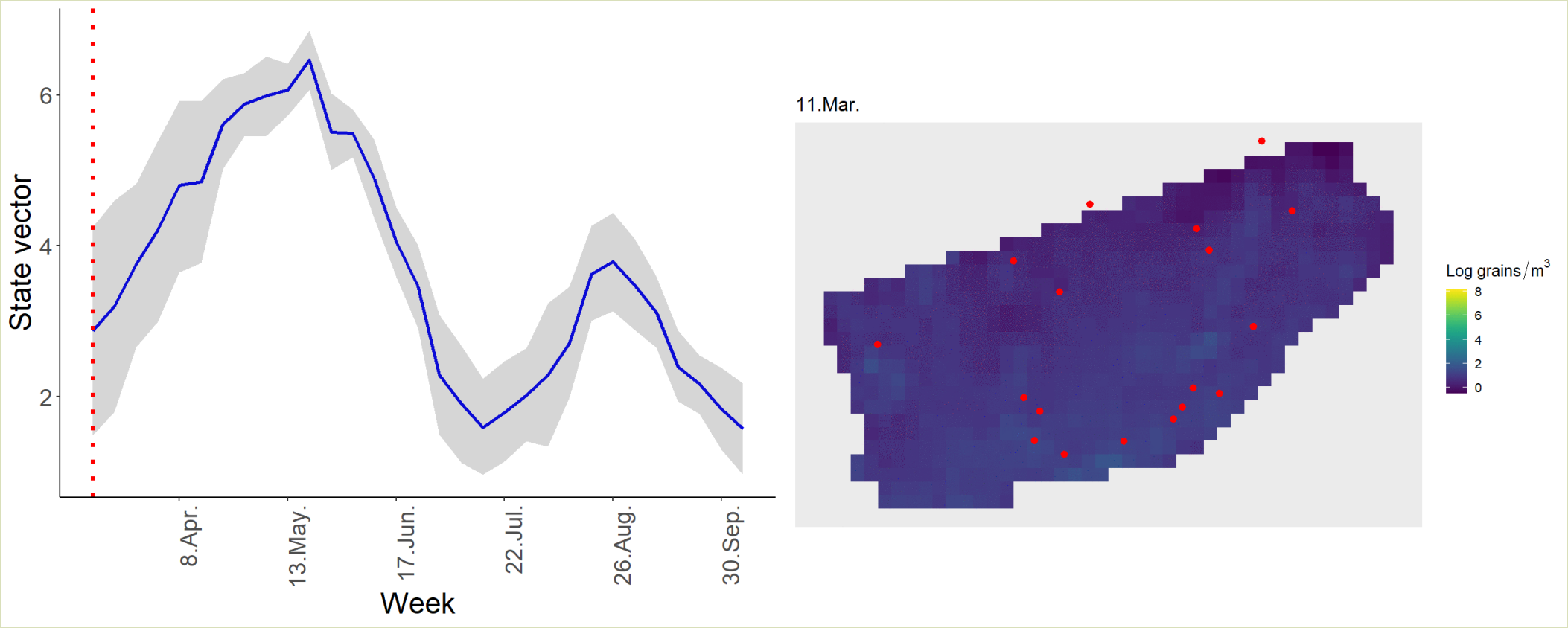
Model 1	$\rho_t = \rho$
Model 2	$\text{logit}(\rho_t) = \gamma_0 + \gamma_1 t + \gamma_2 t^2$
Model 3	$\text{logit}(\rho_t) = \gamma_t,\ \text{where}\ \gamma_t \sim N(\gamma_{t-1}, W_2^2)$
Model 4	$\text{logit}(\rho_t) = \gamma z_t$



Probability and state vector



Predicted surface



Objective 2: Temporal misalignment

Methods

Let $\{Y_t(s); s \in \mathbb{R}^d; t = 1, \cdots, T\}$ be a spatio-temporal process such that $\boldsymbol{Y}_t = (Y_t(s_1), \cdots, Y_t(s_n))'$, we assume that $\boldsymbol{Y}_t \sim N(\boldsymbol{\mu}_t, \boldsymbol{\Sigma})$ where,

$$\boldsymbol{\mu}_t = \mathbf{F}_\mathbf{t}'\boldsymbol{\theta}_\mathbf{t} + \boldsymbol{X}'\boldsymbol{\beta}$$

where the state equation is defined as,

$$\boldsymbol{\theta}_t = \mathbf{G_t}\boldsymbol{\theta}_{\mathbf{t-1}} + \boldsymbol{\omega}_\mathbf{t} \qquad \boldsymbol{\omega}_\mathbf{t} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{W})$$

- $\boldsymbol{F}_t$ is a known $q \times n$ matrix containing temporal covariates,
- $\boldsymbol{G}_t$ is a $q \times q$ evolution matrix,
- $\boldsymbol{X}$ is a matrix containing all land-use covariates and coordinates of the sites,
- $\boldsymbol{\Sigma} = \sigma^2 exp(-\boldsymbol{d}/\phi) + \tau^2\boldsymbol{I}$, where σ^2 is the partial sill, ϕ is the spatial range, and τ^2 is the nugget effect.

Methods

Let $\{Y_t(\boldsymbol{s}); \boldsymbol{s} \in \mathbb{R}^d; t = 1, \cdots, T\}$ be a spatio-temporal process such that $\boldsymbol{Y}_t = (Y_t(\boldsymbol{s}_1), \cdots, Y_t(\boldsymbol{s}_n))'$, we assume that $\boldsymbol{Y}_t \sim N(\boldsymbol{\mu}_t, \boldsymbol{\Sigma})$ where,

$$\boldsymbol{\mu}_t = \mathbf{F}_\mathbf{t}'\boldsymbol{\theta}_\mathbf{t} + \boldsymbol{X}\boldsymbol{\beta}$$

Based on conditional independence on $\boldsymbol{Y}_t$, the aggregated measurements are defined as,

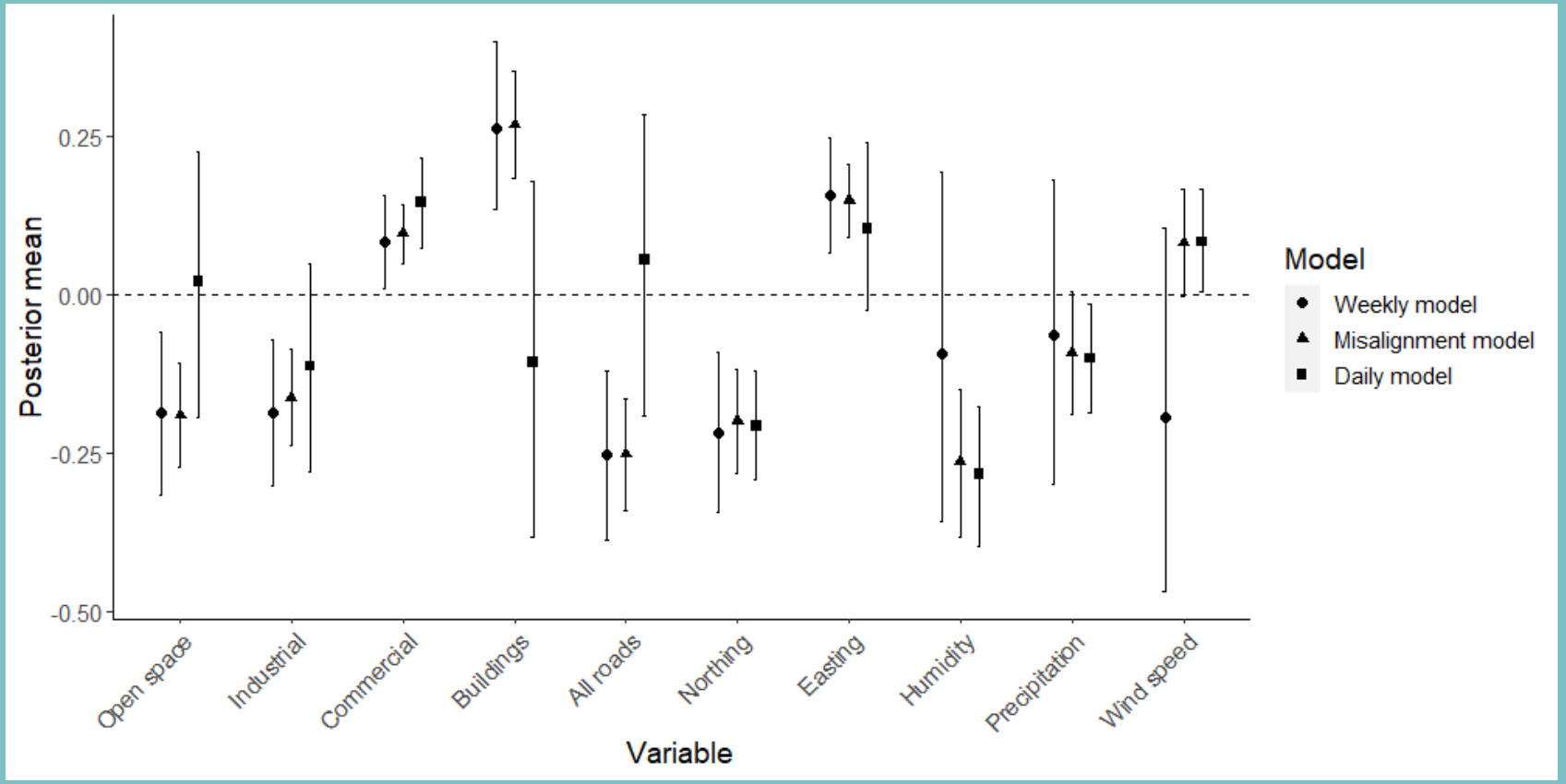
$$\boldsymbol{Z}_k = \frac{1}{r}\sum_{i=1}^{r} \boldsymbol{Y}_{rk+i} \sim N\left(\frac{1}{r}\sum_{i=1}^{r} \boldsymbol{\mu}_{rk+i}, \frac{1}{r}\boldsymbol{\Sigma}\right)$$

Methods

Using the properties of the multivariate normal distribution it is possible to obtain:

- fine-scale measurements at aggregated sites,
- fine-scale measurements at unobserved sites,
- and the aggregated values at unobserved sites.

Coefficient estimates



Aggregated estimated daily values at the weekly sites

Predicted surface

Conclusions

- We analyzed spatio-temporal variations in aeroallergen concentrations throughout the pollen season for different types of pollen within a city.
- We were able to propose a model to account for the temporal misalignment in the data.
- We estimated the unobserved daily measurements at the weekly sites.
- We found different estimates for the fixed effects depending on the temporal scale.
- The predicted surfaces will help future epidemiological studies to find possible associations between pollen levels and some health outcome like respiratory allergies.

Is that it?

Nope

Future work

- Aggregated Hurdle model:
  1. Is the probability constant or does it change across days?
  2. Is it another hurdle model with different probabilities?
- How to account for the temporal misalignment AND the high number of measurements equal to zero at the same time?

General conclusions

- Hierarchical Bayesian methods can easily accommodate more complex structures when analyzing the distribution of aeroallergens across space and time.
- This work shows how by considering a spatial or spatio-temporal structure, it is possible to learn about the spatial patterns and the temporal dynamics of aeroallergens dispersion.

Acknowledgements

- Dr Alexandra Schmidt (EBOH, McGill)
- Dr Scott Weichenthal (EBOH, McGill)
- Dr Daniel S.W. Katz (School of Public Health, University of Michigan-Ann Arbor)
- Dr. Tim Takaro (SFU)
- Dr. Jeff Brook (University of Toronto)
- Aerobiology Research Laboratory