Pollen Prediction

Bayesian Data Analysis **Aalto University**

Fenglei Li

Ognjen Stefanović

1. Dec 2024

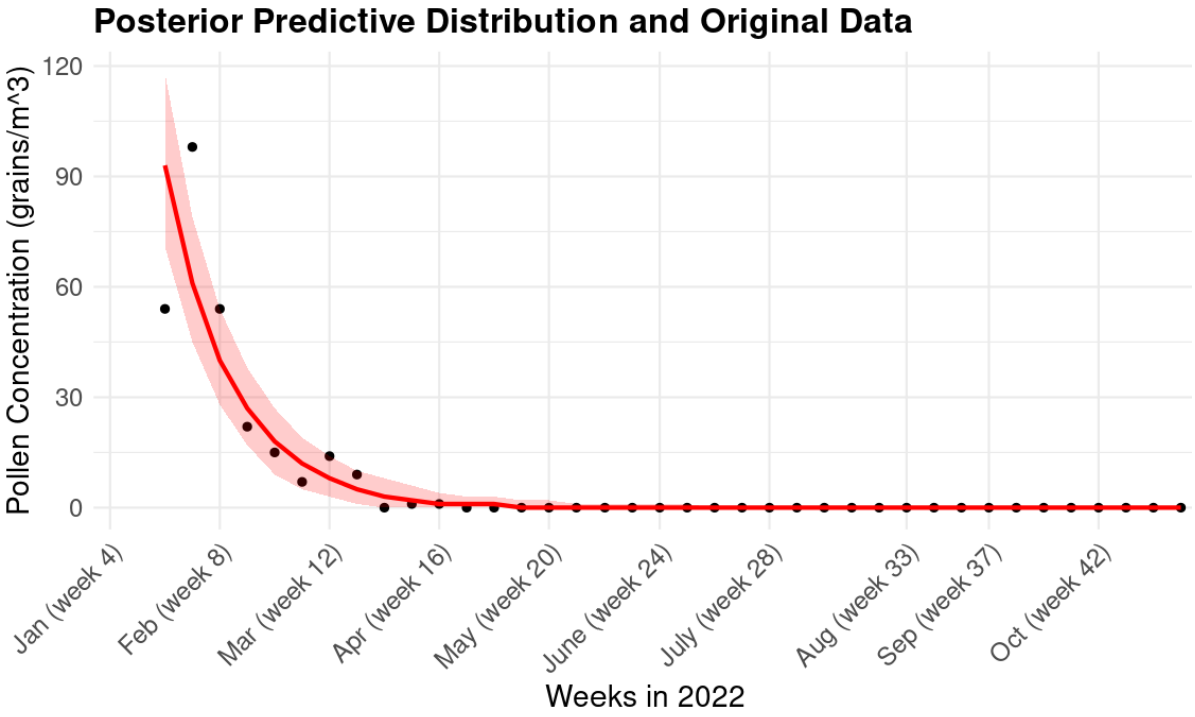


Figure 1: Poisson distribution fitted to measured pollen concentrations in Becej, Serbia in year 2022. The red area is the 95% confidence interval with the posterior predictive mean. This figure is an enlarged portion of Figure 6.

1 Introduction

Allergic reactions to pollen are a significant concern for individuals with respiratory sensitivities, particularly during peak pollen seasons. These allergic reactions can cause a range of symptoms, from mild irritation to severe health issues. To assist people in managing these conditions, it is important to predict pollen concentrations for specific times and locations. By accurately forecasting pollen levels, we can alert affected individuals in advance, helping them take appropriate precautions.

Our aim is to model the distribution of allergenic pollen concentrations using historical data. Since pollen levels vary seasonally, we use a hierarchical model where data is grouped according to the year. We compare two such hierarchical models. The results from the better model on the year 2022 are shown in Figure 1.

2 Data Description and Problem Analysis

The dataset Agency [2024] contains daily pollen concentration measurements for 26 different plant species spanning from 2016 to 2024. They are measured in 25 different cities across Serbia. The units of expression are integer number of pollen grains per m^3 air ($\frac{grains}{m^3}$). No measurements are available in December and January, likely due to minimal pollen concentrations during this period and the measuring devices being turned off.

To simplify, we performed the analysis for concentration of Alnus pollen from 2016 to 2024

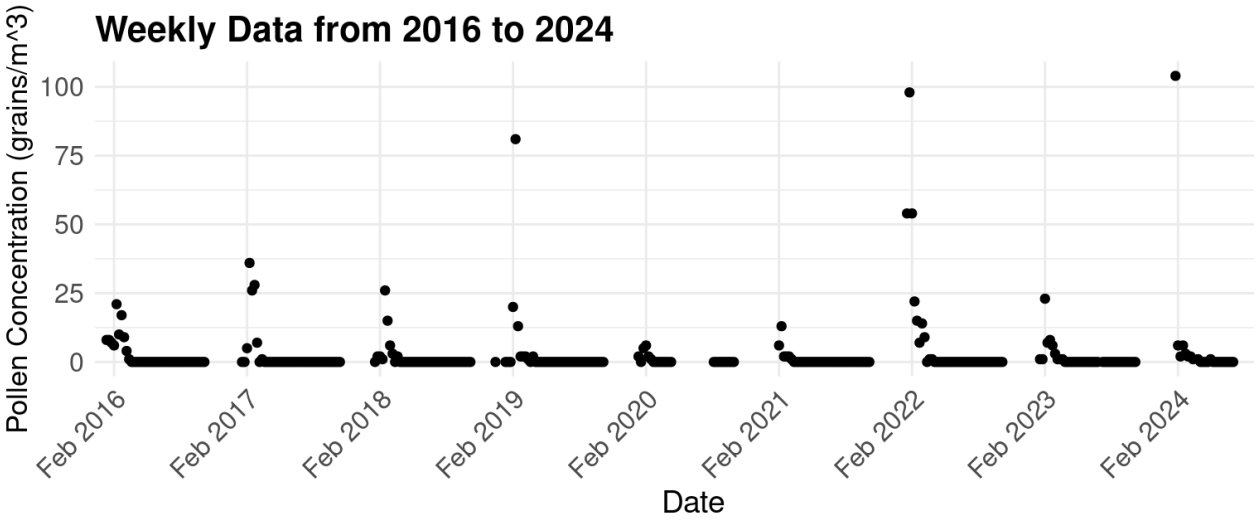


Figure 2: Weekly pollen concentration after data pre-processing

for the city of Becej, located in northern Serbia. Alnus pollen comes from the trees of genus Alnus, also known as alders. They belong to the birch family Betulaceae. It is commonly present in the Northern Hemisphere and typically causes hay fever Piotrowska-Weryszko [2013].

Inspired by Minić et al. [2020], we condense the data by averaging and then rounding pollen concentrations on a weekly basis. This reduced and smoothed the dataset. After pre-processing, the data has weekly pollen concentration measurements for nearly all 53 weeks in each year. The data is shown in Figure 2. All analysis is performed on weekly data, unless otherwise stated.

2.1 Related work

Minić et al. [2020] analyze the implementation of allergen immunotherapy with regard to air pollen concentration. They look at average weekly pollen concentration from 2015 to 2018. In the data analysis, they seek to determine whether the concentration of a particular pollen in any given year significantly deviates from the average. Our analysis goes one step further as we model the underlying distribution of pollen concentration for each year.

The yearly report on air quality in Serbia (Agency [2024]) gives a basic overview of pollen concentration. The pollen categories used are ragweed, birch and grass. The analysis summarizes the yearly total and maximum pollen concentration for each category in 26 cities. They provide an upper limit for acceptable allergenic pollen concentration for birch as $60\frac{grains}{m^3}$. Based on it, they count the number of days where the pollen exceeded this threshold during 2023. For example, in Becej on 12 days acceptable levels were exceeded during 2023. Our work focuses only on Becej and the Alnus birch pollen concentration levels.

Airborne pollen predictions in Málaga (Spain) were done by Hurtado et al. [2024]. The predictions were made using the Seasonal autoregressive integrated moving average (SARIMA) model and CNN-LSTM deep learning model. Our models are simpler than theirs as we use less parameters.

3 Description of the models

Pollen concentrations are modeled using a Poisson distribution. Several factors influenced our choice.

- By visually inspecting the data, the yearly trend appears to have an exponential decay over time.
- The data is discrete as measurements are taken each day (each week once pre-processed).
- Pollen concentration measurements are integers.
- The counts are number of events in a fixed interval of one week. A released grain is an event.
- Poisson distribution is usually used to model rare events, as the data contains many zeros.

We compare two hierarchical Poisson models. The observed pollen concentrations are denoted with y. The week index is $i \in \{1, 2, \ldots, 53\}$ and year index is $j \in \{2016, 2017, \ldots, 2024\}$.

Model 1 with parameters α_0 and α_j is given by

$$\alpha_0 \sim Prior_{userset}\left(\alpha_0\right) \tag{1}$$
$$\tau_\alpha \sim Prior_{userset}\left(\tau_\alpha\right) \tag{2}$$
$$\alpha_j \sim normal\left(0, \tau_\alpha\right) \tag{3}$$
$$y_{ij} \sim Poisson\left(\alpha_0 + \alpha_j\right) \tag{4}$$

In brms this is given by

$$y \sim 1 + (1|year) \tag{5}$$

Model 2 is with parameters α_0, α_j and β_j is given by:

$$\alpha_0 \sim Prior_{userset}\left(\alpha_0\right) \tag{6}$$
$$\tau_\alpha \sim Prior_{userset}\left(\tau_\alpha\right) \tag{7}$$
$$\alpha_j \sim normal\left(0, \tau_\alpha\right) \tag{8}$$
$$\tau_\beta \sim Prior_{userset}\left(\tau_\beta\right) \tag{9}$$
$$\beta_j \sim normal\left(0, \tau_\beta\right) \tag{10}$$
$$y_{ij} \sim Poisson(\alpha_0 + \alpha_j + \beta_j \cdot week_j) \tag{11}$$

or in brms as

$$y \sim 1 + (1 + week|year) \tag{12}$$

where we select the brms model family to be Poisson. The Poisson pmf with mean λ is given by:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{13}$$

4 Priors

The default prior used for the coefficients in equations (1), (2), (6), (7) and (9) is the student's t-distribution. We also tried out the normal, gamma and beta distributions as priors. The prior $normal\,(0, 10)$ was chosen because it is a common weak prior choice. Also we had no prior knowledge of the optimal coefficients' order of magnitude. When the second model converged and the coefficients took values between 0 and 5, we attempted with more informative priors $normal\,(0, 4)$ and $normal\,(0, 1)$.

The gamma prior was chosen because it is the conjugate prior for the Poisson distribution. We chose the beta distribution to experiment with a prior that has a bounded domain.

- Default prior: student's t: $\alpha_0, \tau_\alpha, \tau_\beta \sim Student_t(3, 0, 2.5)$
- beta $\alpha_0, \tau_\alpha, \tau_\beta \sim beta(1, 10)$
- gamma $\alpha_0, \tau_\alpha, \tau_\beta \sim gamma(1, 1)$, $\alpha_0, \tau_\alpha, \tau_\beta \sim gamma(5, 1)$, $\alpha_0, \tau_\alpha, \tau_\beta \sim gamma(10, 1)$
- normal $\alpha_0, \tau_\alpha, \tau_\beta \sim normal(0, 10)$ and $\alpha_0, \tau_\alpha, \tau_\beta \sim normal(0, 4)$

5 Code and MCMC options

The main brms code for optimization to run the model is shown below. The code for data pre-processing and plotting has been omitted but is available upon request.

Listing 1: Main code

```
formula <- bf(y ~1 + (1 + week | year), family = "poisson")

fit <- brm(formula,
           data = data,
           prior = priors,
           cores = 4, # default is 1
           chains = 4,
           iter = 4000, # default is 2000
           warmup = floor(iter/2),
           control = list(adapt_delta = 0.92) # default is 0.8
)
```

Each optimization has four chains, each chain running for 4000 post-warmup iterations.

We increased the number of iterations, although it slows down optimization. The justification is that we used 4 CPU cores instead of 1, which speeds up the the optimization.

We doubled the number of iterations compared with the default, which slows down the optimization. This is justified because we increased the number of CPU cores used from 1 to 4, which sped up the optimization.

We experimented with different values of adapt_delta (δ) in model 2, ranging from 0.8 to 0.99 with the weakly informative prior normal$(0, 10)$. For most experiments, the number

Table 1: Prior Sensitivity Analysis on predictive performance and convergence diagnostics for model $y_{ij} \sim Poisson(\alpha_0 + \alpha_j)$.

Prior on α_0, τ_α	δ	$\hat{R}$	ESS_{bulk}	ESS_{tail}	Divergence	Bayesian R^2
gamma(5,1)	0.8	1	617	821	No	0.03
gamma(10,1)	0.92	1	737	1029	Yes(1)	0.03
normal(0,10)	0.8	1	1257	2164	No	0.03

Table 2: Prior Sensitivity Analysis on predictive performance and convergence diagnostics for model $y_{ij} \sim Poisson(\alpha_0 + \alpha_j + \beta_j \cdot week_j)$. Adapt step is $\delta = 0.92$

Prior on $\alpha_0, \tau_\alpha, \tau_\beta$	$\hat{R}$	ESS_{bulk}	ESS_{tail}	Divergence	Bayesian R^2
student's t(3,0,2.5)	1	1200	1025	Yes (4)	0.597
beta(1,10)	1.59	7	11	Yes (1230)	/
gamma(10,1)	1.03	236	206	Yes(646)	0.601
gamma(5,1)	1	784	821	Yes (43)	0.599
gamma(1,1)	1	801	837	Yes (581)	0.596
normal(0,4)	1	1123	1190	Yes (3)	0.597
normal(0,10)	1	889	701	Yes (3)	0.598

of divergent transitions was less than 10. We opted to set $\delta = 0.92$ for the rest of the experiments as we found that it works for other priors as well.

With this setup we provide the optimally achieved results for both models.

6 Convergence diagnostic

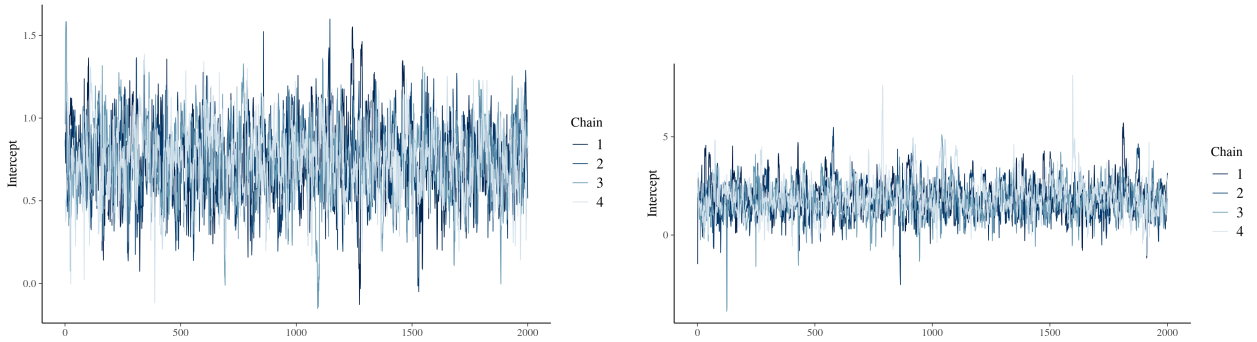
In tables 1 and 2 are shown convergence diagnostics for models 1 and 2, respectively. For model 1 we first tried out the gamma$(5, 1)$ prior (row 1), but the ESS values were low, indicating high correlation between MCMC samples. Hence we tried increasing adapt δ to 0.92 (row 2) and setting the prior to gamma(10, 1), but the ESS still remained low. Thus we attempted with the prior normal$(0, 10)$ (row 3) and this improved the ESS.

Table 2 shows the order of performed experiments. The default student's t prior (row 1) gave decent convergence diagnostic. As expected, for the beta prior the model did not converge and has low ESS because the beta is bounded between $[0, 1]$. The gamma$(10, 1)$ prior performed quite poorly as it puts minimal priors weights in the range $[0, 5]$, where the optimal parameters lie. Hence it has low ESS and high divergent transitions and $\hat{R} > 1.01$.

The gamma$(5, 1)$ performs better as it shifts prior weights closer to the optimal range of $[0, 5]$, though it still has 43 divergent transitions. The gamma$(1, 1)$ prior puts weight in the range $[0, 2]$ but it fails to capture the range $[2, 5]$, hence performs worse with more divergences.

The weakly informative priors normal$(0, 4)$ and normal$(0, 10)$ in table 2 perform better than other priors as they have only 3 divergences each.

For the models 1 and 2 with priors normal$(0, 10)$ we show that the chains are mixing in Figure 3.

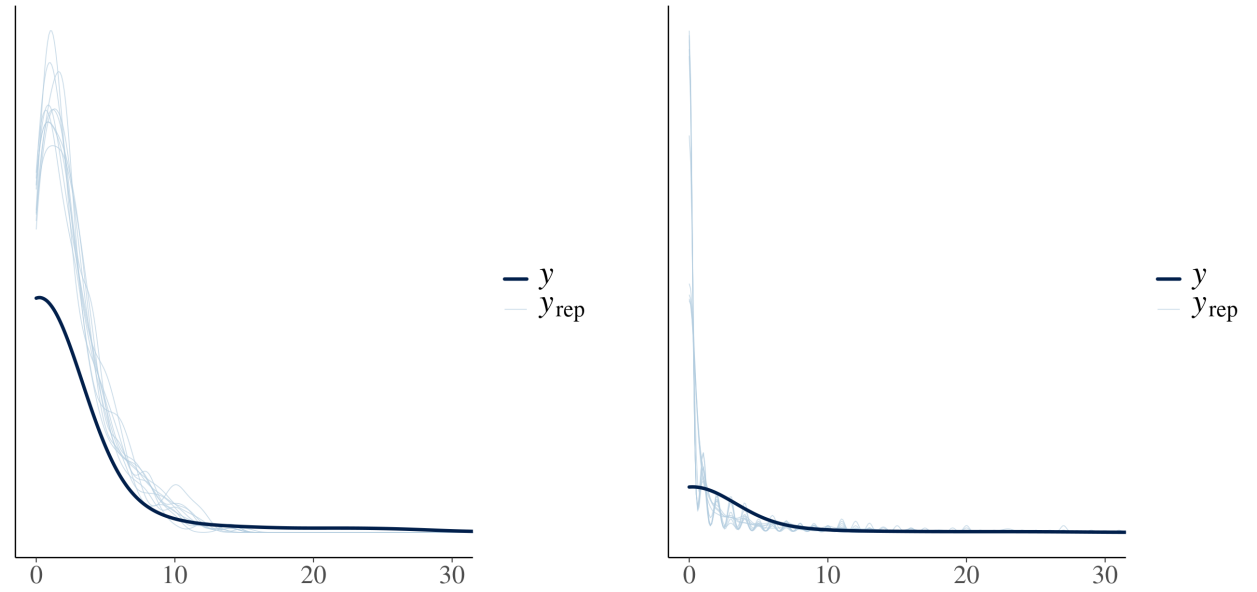


(a) Model 1

(b) Model 2

Figure 3: MCMC Trace for the intercept term α_0. The chains are mixing. Both models have priors of normal$(0, 10)$.

7 Posterior predictive checks



(a) pp_check of model 1, $y \sim 1+(1|year)$. Bayes $R^2 = 0.03$

(b) pp_check of model 2, $y \sim 1 + (1 + week|year)$. Bayes $R^2 = 0.6$

Figure 4: Posterior predictive Checks

For model 1 in Figure 4a the predictive distribution does not match the true kernel density estimate of the data.

For model 2 in Figure 4b the posterior predictive distribution somewhat overlaps with the shape of the true distribution, but the fit is not ideal. Given that the Bayes R^2 is only 0.6, the model's performance is relatively weak. As a result, it is not surprising that the posterior predictive checks show suboptimal results. The posterior samples with high peaks at 0 suggest too much days with no pollen than what the data actually has. The posterior

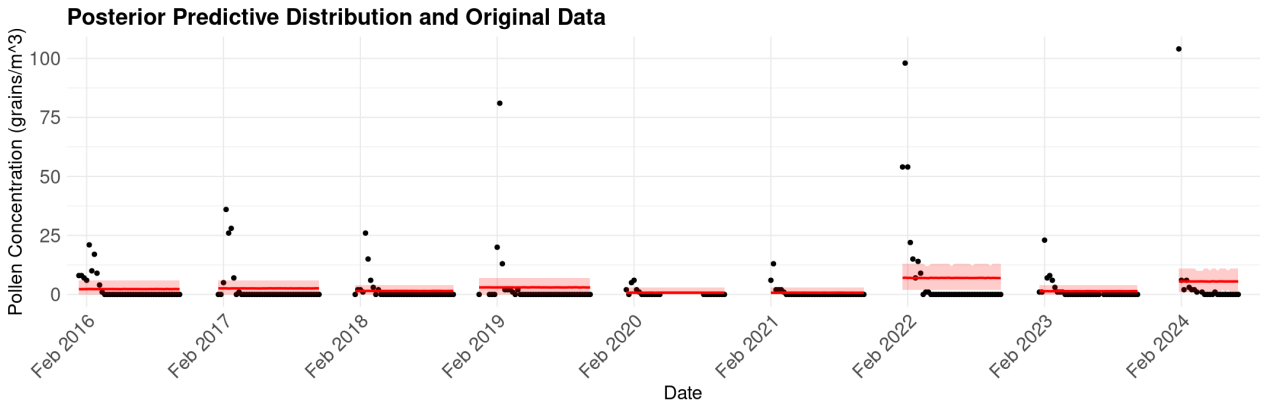


Figure 5: Best fit for the model $y_{ij} \sim Poisson(\alpha_0 + \alpha_j)$, where j is the index of the year. Bayesian $R^2 = 0.03$

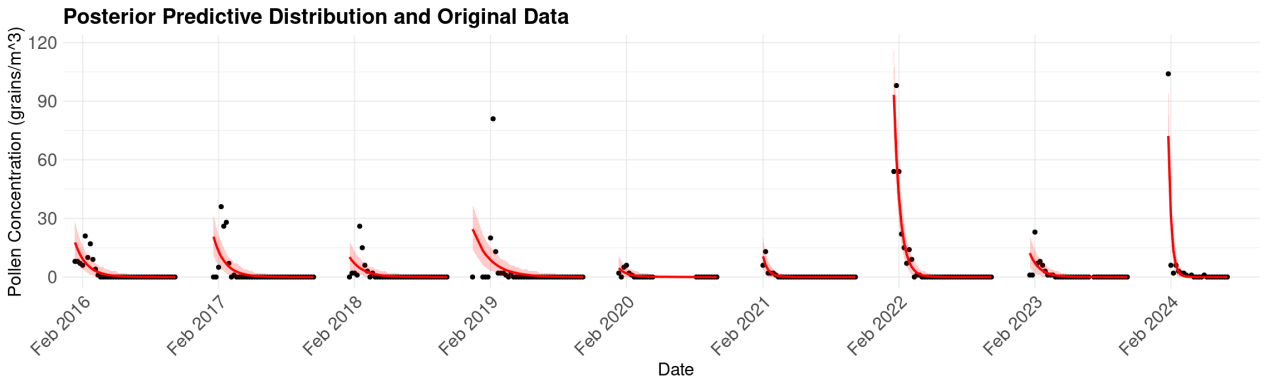


Figure 6: Best fit for the model $y_{ij} \sim Poisson(\alpha_0 + \alpha_j + \beta_j \cdot week_j)$, where j is the index of the year. Bayesian $R^2 = 0.6$

samples with high peaks at 0 may be due to the large number of zeros in the true data, causing the model to frequently output zero.

8 Sensitivity analysis

The sensitivity analysis for models 1 and 2 are shown in tables 1 and 2, respectively.

For both model, for any choice of prior the models underfit the data as indicated by the Bayesian R^2. The models are not expressive enough and hence are insensitive to prior choice. The only exception is the beta prior in table 2 as the parameters are bounded in the range $[0, 1]$, thus the model does not converge.

9 Model comparison

The best achieved fits for both models are shown in figures 5 and 6. Both chosen model have priors normal$(0, 10)$.

The Expected Log Predictive Density (ELPD) assesses the predictive performance of the

model, with higher (less negative) values indicating better predictive accuracy. The model 2 from figure 6 has $ELPD_2 = -779.4$, with a standard error of 162.7. For Model 1, the ELPD is lower with $ELPD_1 = -2061.9$, with a standard error of 395.6. This indicates that Model 1 has poorer predictive performance compared to Model 2. The larger standard error suggests greater uncertainty around this estimate, which further implies that the model's predictive power is less reliable.

However, both models exhibit problematic data with around ~ 6 points with $Pareto-\hat{k} > 0.7$ and ~ 6 points with $Pareto - \hat{k} > 1$. This affects the reliability of our models, although both models have ~ 300 well-behaved data points ($Pareto - \hat{k} \leq 0.7$).

10 Discussion of issues and potential improvements

The results indicate that the model's performance is suboptimal, as shown by a Bayesian R^2 of only 0.6. This suggests the model is underfitting the data, failing to capture essential relationships in the data. The remaining unexplained variance could be attributed to noise or unaccounted hidden variables.

Additionally, the original daily and processed weekly datasets contain many zeros. This presents a challenge. One potential method for improvement is to explicitly account for the generation of zeros in the model.

For the ELPD, some observations show that $Pareto - \hat{k} > 0.7$. This indicates potential issues with the approximation. We suspect that applying moment matching could help address this. Regarding the final results (see last rows in tables 1 and 2), although there are a few divergent transitions, they are minimal, and the R-hat value is 1.00, suggesting convergence. Resolving the error could potentially improve the overall model performance.

11 Conclusion

In this project, we aimed to model the distribution of allergenic pollen concentrations, specifically focusing on Alnus pollen in Becej, Serbia. We employed hierarchical Poisson models to capture the variability in pollen concentrations over time, leveraging data from 2016 to 2024. Two predictive models were proposed, with Model 2 ($y \sim 1+(1+week|year)$) outperforming Model 1 ($y \sim 1 + (1|year)$). Model 2 achieved Bayesian $R^2 = 0.6$ indicating there is room for improvement.

A range of priors were tested. The convergence diagnostics ($\hat{R} = 1.00$) indicate that the models are converging properly. However both models faced challenges with some data points exhibiting problematic $Pareto - \hat{k} > 0.7$ and $Pareto - \hat{k} > 1$ points and the presence of divergent transitions, although minimal. This raises concerns about the reliability of our leave-one-out cross-validation estimates. To address these issues, future work could explore alternative modeling approaches that more explicitly account for the many zeros present in the data.

12 Self-reflection

The project on modeling allergenic pollen concentrations using hierarchical Poisson models has been both challenging and rewarding. We did the three key steps in Bayesian Data Analysis.

1. Model the joint of observables and unobservables.
2. Compute the posterior (using MCMC).
3. Evaluate how good the model fits the data. If poorly go to 1.

We quickly realized that working with real-world data is quite different from theoretical problems. Since we had not initially decided on the best model to use, we spent a lot of time experimenting with different models and priors that failed. In hindsight, it may have been more efficient to first establish a clearer direction or hypothesis.

One of the key lessons from this experience was the importance of selecting appropriate models and priors. We learned that the selection process follows a progression from non-informative to slightly informative to more informative priors. However, we also faced the challenge of not knowing which specific model or prior would work. This uncertainty highlighted a key limitation in our approach and underlined the complexity of Bayesian analysis when prior knowledge is not readily available.

13 Use of AI in the project

We used AI to edit roughly one sentence in every second paragraph, when we struggled to re-phrase our poorly written sentence. All other text was written solely by us. We also used AI to help us with editing and generating our plotting and data pre-processing code. We did our best to utilize the code knowledge we learned during previous assignments and lectures.

References

The Serbian Environmental Protection Agency. Concentrations of pollen in air, 2024. URL `http://data.europa.eu/88u/dataset/kontsentratsije-polena-u-vazdukhu`. Data set.

Sandro Hurtado, María Luisa Antequera-Gómez, Cristóbal Barba-González, Antonio Picornell, and Ismael Navas-Delgado. e-science workflow: A semantic approach for airborne pollen prediction. *Knowledge-Based Systems*, 284:111230, 2024. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2023.111230. URL `https://www.sciencedirect.com/science/article/pii/S0950705123009802`.

R. Minić, M. Josipović, V. Tomić Spirić, M. Gavrović-Jankulović, A. Perić Popadić, I. Prokopijević, A. Ljubičić, D. Stamenković, and L. Burazer. Impact of tree pollen distribution on allergic diseases in serbia: Evidence of implementation of allergen immunotherapy to *Betula verrucosa*. *Medicina (Kaunas)*, 56(2):59, Feb 2020. doi: 10.3390/medicina56020059.

Krystyna Piotrowska-Weryszko. The effect of the meteorological factors on the alnus pollen season in lublin (poland). *Grana*, 52(3):221–228, 2013. doi: 10.1080/00173134.2013.772653. URL `https://doi.org/10.1080/00173134.2013.772653`.

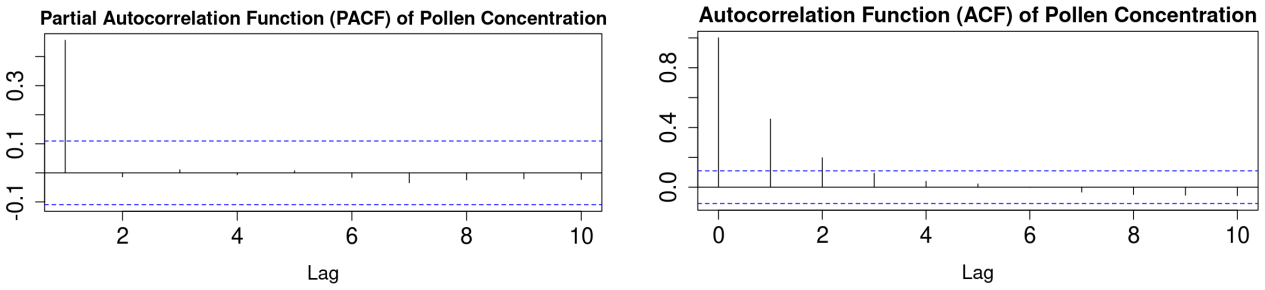
A Models

Aside from the Poisson model, other models were tried out as a baseline, such as the Autoregressive (AR) model and Moving Average (MA) model. The AR(1) model gave a starting point with Bayesian $R^2 = 0.2$ and the MA(2) did not converge.

A.1 Autoregressive Model

As a baseline we fit an AR(1) model. It only has three parameters to be fit. The optimal lag $p = 1$ value was determined from 7a. The AR model with Gaussian noise $\varepsilon \sim N(0, \sigma^2)$ is given by:

$$y_t = \mu + \phi y_{t-1} + \varepsilon_t \tag{14}$$



(a) The optimal value is lag $p = 1$ for the AR(p) model

(b) The optimal value is lag $q = 2$ for the MA(q) model

Figure 7: Partial ACF and ACF plots

The fitted model parameters are $\phi = (0.46 \pm 0.05)$, $\mu = (2.39 \pm 0.98)$ and $\sigma = (9.73 \pm 0.39)$. The Bayes $R^2 = (0.20 \pm 0.04)$. The posterior fit is shown in figure 8.

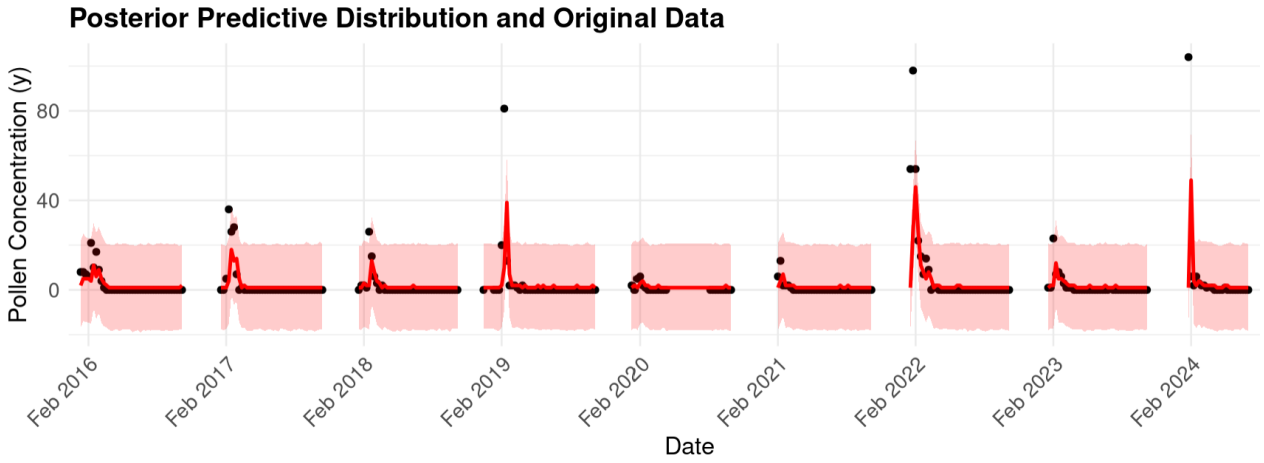


Figure 8: AR(1) Model Fit

A.2 Moving-Average Model

The MA(2) model did not converge with default priors. We chose $q = 2$ as the optimal coefficient as indicated from figure 7b. The fitted MA(2) model is given by:

$$y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} \tag{15}$$

but the convergence diagnostic $\hat{R}^2 = 1.60 > 1$ for all fitted parameters. We did not try to iterate and make this model converge as we were already satisfied with the AR(1) model as a simple baseline.

B MCMC Trace Plots for Optimal Models

Figure 9: MCMC Trace of Model 1

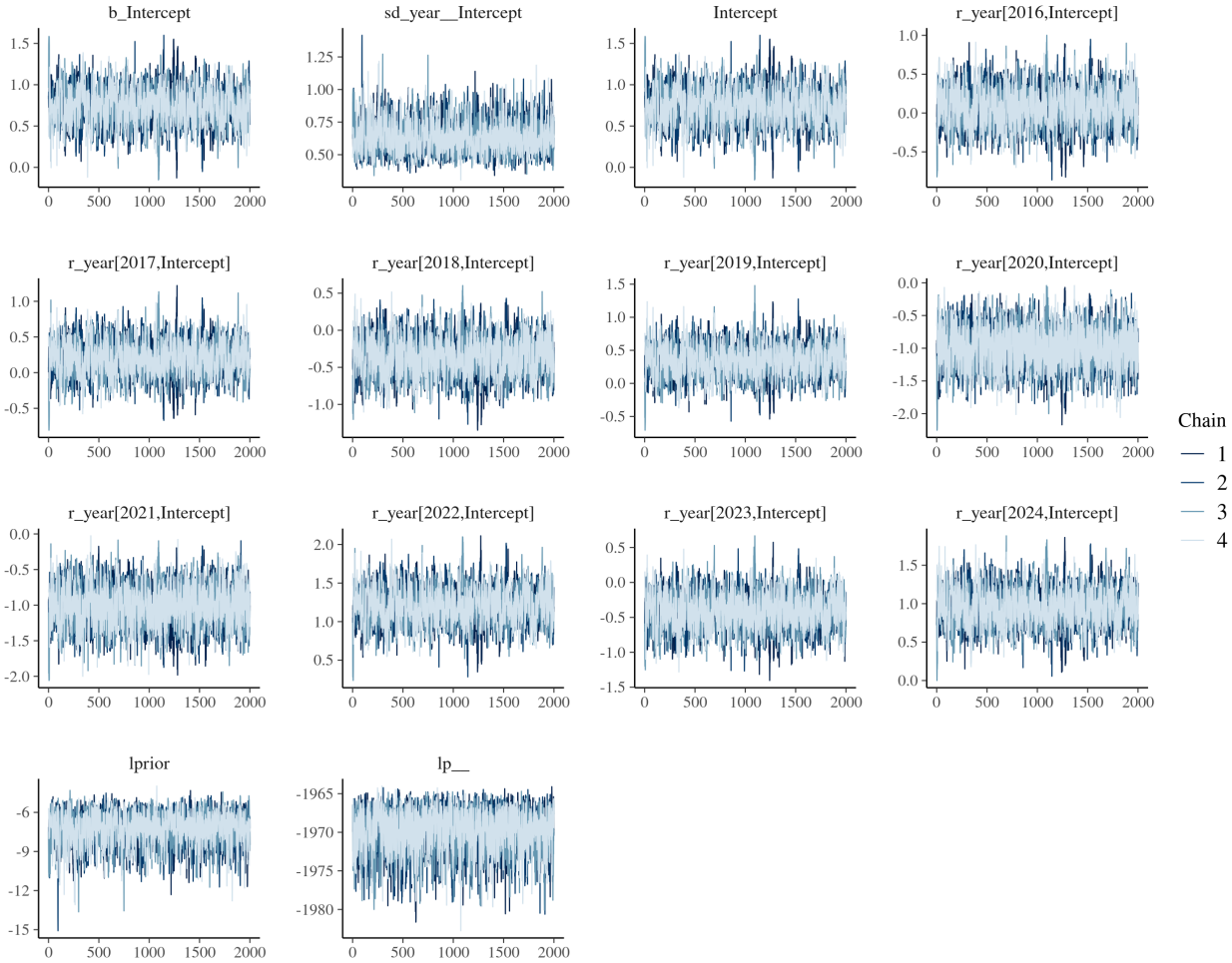


Figure 10: MCMC Trace of Model 2