POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

A spatio–temporal Bayesian model for the O_3 pollution problem in the Lombardia region

Bayesian Statistics
Mathematical Engineering - Statistical Learning

Matteo Coppellotti, Alessandro Ganelli, Francesco Maria Mancinelli, Leonardo Marchesin, Beatrice Sisti, Alessandro Venanzi

Tutors:
Alessandra Guglielmi
Michela Frigeri

Academic year:
2023-2024

Abstract: The degree of attention to the quality of our air is steadily increasing globally, therefore, addressing the complex and dynamic nature of pollution has become a critical research topic. This report illustrates a comprehensive analysis of ozone pollution levels in the Lombardia region, employing a Bayesian approach tailored for count data analysis to examine the frequency of days exceeding certain specified ozone concentration thresholds each month. By combining Bayesian statistical techniques with relevant atmospheric data, this project aims to uncover underlying patterns and discern the primary factors contributing to ozone pollution.

`https://github.com/venanzjones/Bayesian_Project`

Key-words: Count data, Poisson, Bayesian models

1. Introduction

In recent years, global warming has become one of the most debated topics and consequently more and more research efforts have been made to study the pollutants that may cause these changes. In this context, ozone is one of the most important among these and yet it is not being studied nearly enough.
O_3, the chemical formulation of the ozone, is usually associated with the ozone layer, which plays a crucial role in protecting life on Earth by absorbing the majority of the sun's harmful ultraviolet (UV) radiation. Nevertheless, this gas is also present in the air breathed every day, and its concentration is referred to as *ground-level ozone*; it is the result of many chemical reactions that involve different other pollutants and catalyzed by meteorological events such as high temperatures, making hot days in the proximity of big cities a critical hot spot for the proliferation of this type of ozone. This heightened awareness of the problems concerning air quality has pushed institutes dedicated to studying these metrics toward trying to approach the situation with a slew of different methods and techniques. In this particular case, the proposed framework is Bayesian modeling for count data, to capture and explain the different types of dependencies.
In particular, the objective is to describe two types of alert levels of ground ozone, which are:

- Information threshold: 1-hour average ozone concentration of 180 $\mu g/m^3$.
- Long-term objective: 8-hour average ozone concentration of 120 $\mu g/m^3$.

The resulting quantities to be analyzed are therefore made of count data representing the number of days in each month where these thresholds have been overcome at least once. This entails the need of defining two different models to handle these two indexes.

Many candidate variables have been considered to explain the variability in ozone levels and to further specialize the proposed models a spatial model will be implemented along side specific features integral to the accuracy end interpretability of the inference.

2. Dataset construction

The original dataset consists of the hourly concentrations of ozone (O_3) recorded by the ARPA Lombardia monitoring network. These data were gathered from 2010 to 2022 by a total of 51 monitoring stations across all the Lombardia region. The main focus is on a specific period, going from April to October, where the ozone pollution levels are higher. In fact, through a first inspection of the given data, the other months show very low concentration levels, hence they will not be considered in the study. In order to quantify the frequency of ozone pollution events exceeding the aforementioned specified thresholds, two distinct datasets are built containing the resulting computed count scores.

2.1. Handling missing values

As the original dataset has numerous missing values, informed decisions are taken when constructing the two datasets. This initial analysis aims to identify and address missing data. Visual inspection of plots reveals a pseudo-random distribution of missing values, making interpretation challenging. Consequently, a threshold of 10% is set for the total of NAs within the final datasets. As it is reported in the table, It's worth noting that the highest percentages of NAs are during the summer months. Therefore, it may be inferred that this is due to stations not being repaired promptly or operations due to the summer holidays.

Month	%NAs
April	6.85%
May	10.41%
June	11.53%
July	13.25%
August	9.77%
September	8.72%
October	8.11%

Table 1: % of NAs per month.

For the construction of the dataset *Count 180*, the following thresholds are set:

- A month with six or more days of missing data is considered missing in its entirety;
- A day is considered not valid if it lacks 6 or more hours of data and the reported values don't overcome 180.

For the months with 5 or less missing days, the count data is simulated using a linear interpolation between the observed values on the nearest preceding and succeeding non NA valued days. Finally, the number of days per month, per year and per station where the ozone concentration exceeds 180 $\mu g/m^3$ is counted.
As for the *Count 120* dataset, dealing with NAs is a much more challenging problem, as hourly averages are introduced. The average value is considered valid if there are at least 6 actual measured values. Once this new dataset of rolling averages is generated, the process for selecting days that exceed the threshold count is similar to that used for the dataset *Count 180*. The thresholds are tailored to the problem, as the aim is to push them down as much as possible to maximize the information gathered. At last, the dataset is built by counting the number of days per month, per year, and per station where the moving average for the ozone concentration exceeds the *long-term* threshold of 120 $\mu g/m^3$.

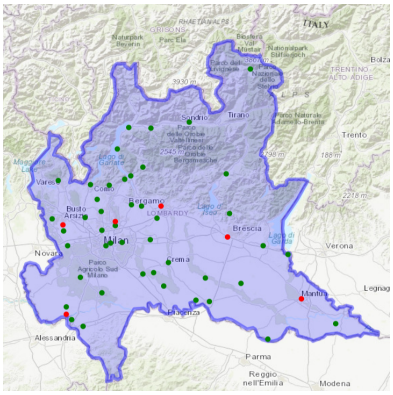

Figure 1: The 51 stations.

In both datasets, some stations exhibit significantly higher levels of missing data compared to others. Consequently, these stations are removed from the analysis, as inference becomes untenable when faced with a high percentage of missing data (around 30%) for a particular station in mixed-effects models. In particular, the removed stations are those with ID: `17288`, `17295`, `17297`, `20041`, `20154`, `30165`, and they are plotted in red in Figure 1.

2.2. Covariates identification

Firstly, all the available relevant data was taken into account as candidate covariates with the intention of providing as much information as possible to the models; then, according to an appropriate variable selection method, only a subset of the complete set of candidates is actually deemed significantly influential to the models.

We narrow our focus on meteorological factors, which are supposedly related to both the the presence of ozone and pollutants into the atmosphere. Through the open-source website `OpenMeteo` [8] we are able to collect the daily measurements of some relevant quantities by inputting the coordinates of every station and this data is later aggregated into monthly means:
The main focus was on meteorogical factors due to their supposed relevancy for both ozone production as well as presence of other pollutants in the atmosphere. Through the open-source website `OpenMeteo` [8] daily measurements of significant data are collected by inputting the coordinates of every station, to then aggregate these into monthly means:

- Mean temperature (°C);
- Total sum of precipitations (mm);
- Hours observing precipitations (h);
- Maximum wind speed (km/h);
- Total sum of shortwave radiations (MJ/m^2).

Information related to wind intensity was refactored using the Beaufort scale [6], defining also for each month the following values:

- The maximum number of consecutive days with wind greater or equal than 20 km/h (:= 4 Beaufort);
- Average wind speed in Beaufort scale.

Due to the fact that the ozone is considered as a secondary pollutant, meaning that it is not directly emitted into the atmosphere by factors like industries or traffic but rather formed through photochemical processes in the presence of primary pollutants (such as nitrogen oxides and volatile organic compounds), variables related to the anthropization of the area surrounding each station was also defined. The introduced features are:

- Population density of the station's town;
- Altitude, longitude, and latitude;
- Type of area.

In particular, the last one was determined using an image segmentation tool from Regione Lombardia [4] in order to identify and understand the area in the proximity of each station, classifying it in one of the following area types: {Urban, Rural, Industrial}.

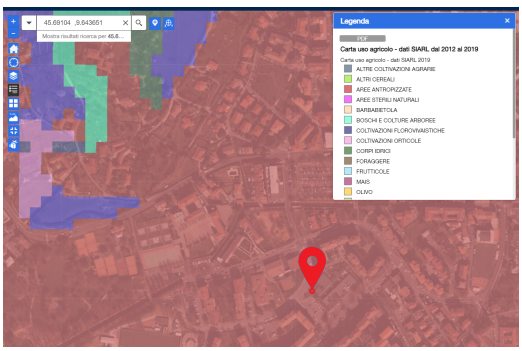

Figure 2: Area near station `17297`, Bergamo

3. Modeling

In this section the models that have been used for the purpose of the analysis are shown in a sequential manner: all the relevant features of the final models are added one at a time to the starting model to quantify and better understand the impact of each additional factor. Keeping in mind that there are two different quantities of interest originating from the two extracted datasets, a basic model is used as a common development starting point to then tailor, when needed, the subsequent changes to better fit the two different sets of count data. In fact, the final models for the two quantities of interest have different needs due to both the data and also interpretative reasons.

All models have been implemented in STAN, and were then compared using different metrics in order to choose the best performing one based on their ability to better predict new observations.

3.1. Base model

As a starting point the most common choice for counting data is to use the *Poisson* model, which represents a discrete quantity as a function of a parameter λ. This parameter is explained through a set of covariates, the ones explained in Section 2.2. The general model for n_{ijk} count is the following:

$$n_{ijk} \mid \lambda_{ijk} \overset{ind}{\sim} Poi(\lambda_{ijk})$$
$$log(\lambda_{ijk}) = \underline{x}_{ijk}^T \underline{\beta}$$
$$\underline{\beta} \mid \sigma_\beta^2 \sim \mathcal{N}_p(\underline{0}, \sigma_\beta^2 \mathbf{I})$$
$$\sigma_\beta^2 \sim inv - gamma(4, 2)$$
$$i = \text{April}, \dots, \text{October} \quad j = 2010, \dots, 2022 \quad k = 1, \dots, 45$$

It is the general conjugate model for a linear regression where the result of this regression is linked through the *log* function to the response variable, as in classical generalized linear model. The prior for σ_β is chosen such that the expected value is 2 and the variance is $0,\overline{2}$ from a pilot run of the model with a higher variance.

3.2. Variable selection

The Stochastic Search Variable Selection (SSVS) approach is used to filter the predictors, assuming a Spike-and-Slab prior, where the spike is a Gaussian distribution with very small variance. At its core, SSVS aims to identify a parsimonious subset of variables that exhibit a significant association with the response variable, while effectively filtering out irrelevant or redundant predictors. The Spike-and-Slab prior acts as a binary selector, with each predictor assigned a binary indicator indicating whether it should be included (non-zero coefficient) or excluded (zero coefficient) from the model. Consequently, the covariates choice is made via the posterior of the model index parameter. In particular, the Median Probability Model (**MPM**) strategy was used, picking all covariates with estimated marginal posterior inclusion probabilities larger than 0.5. Through this procedure, the following covariates have been dropped:

- For *Count 180* model: `max_consecutive_highwind_days`, `density`, `count_highwind`;
- For *Count 120* model: `max_consecutive_highwind_days`, `density`, `type_rural`, `type_urban`.

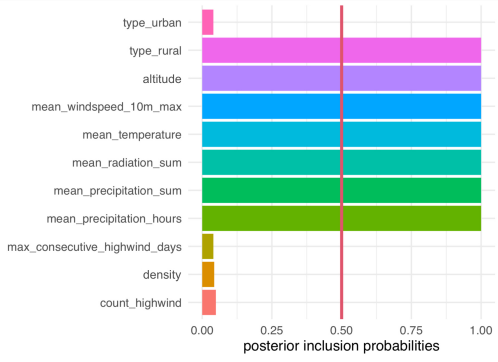

Figure 3: MPM for *Count 180*.

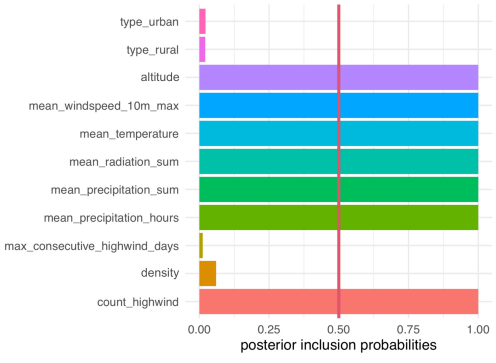

Figure 4: MPM for *Count 120*.

From now on the vector of coefficient $\underline{\beta}$ and the design matrix $\mathbb{X}$ have to be assumed cut with the removed variable out of the model.

3.3. Years' random effect

A random effect for the year factor is considered, and a variable is added with the name ξ representing the specific factor of the year.

$$n_{ijk} \mid \lambda_{ijk} \overset{ind}{\sim} Poi(\lambda_{ijk})$$
$$log(\lambda_{ijk}) = \underline{x}_{ijk}^T \underline{\beta} + \xi_j$$

$$\underline{\beta} \mid \sigma^2_\beta \sim \mathcal{N}_p(\underline{0}, \sigma^2_\beta \mathbf{I})$$
$$\underline{\xi} = (\xi_{2010}, \dots, \xi_{2022}) \mid \sigma^2_\xi \sim \mathcal{N}_p(\underline{0}, \sigma^2_\xi \mathbf{I})$$
$$\sigma^2_\beta \sim inv - gamma(4, 2)$$
$$\sigma^2_\xi \sim inv - gamma(4, 2)$$
$$i = \text{April}, \dots, \text{October} \quad j = 2010, \dots, 2022 \quad k = 1, \dots, 45$$

The prior for this random effect is assumed normal to maintain the conjugacy of the model, similar reasoning for the prior of its σ. Note that from now on the model is a typical *Generalized bayesian linear mixed effect model*.

3.4. Stations' random effect

The effects deriving from common station are here added as a random effect, analogously as the years' factor.

$$n_{ijk} \mid \lambda_{ijk} \overset{ind}{\sim} Poi(\lambda_{ijk})$$
$$log(\lambda_{ijk}) = \underline{x}^T_{ijk}\underline{\beta} + \xi_j + \eta_k$$
$$\underline{\beta} \mid \sigma^2_\beta \sim \mathcal{N}_p(\underline{0}, \sigma^2_\beta \mathbf{I})$$
$$\underline{\xi} = (\xi_{2010}, \dots, \xi_{2022}) \mid \sigma^2_\xi \sim \mathcal{N}_p(\underline{0}, \sigma^2_\xi \mathbf{I})$$
$$\underline{\eta} = (\eta_1, \dots, \eta_{45}) \mid \sigma^2_\eta \sim \mathcal{N}_p(\underline{0}, \sigma^2_\eta \mathbf{I})$$
$$\sigma^2_\beta, \sigma^2_\xi, \sigma^2_\eta \overset{\text{iid}}{\sim} inv - gamma(4, 2)$$
$$i = \text{April}, \dots, \text{October}, \quad j = 2010, \dots, 2022, \quad k = 1, \dots, 45$$

After introducing this effect, three of the covariates for the 180 model are no longer necessary, in fact the variable that controls if the station is in a rural area instead of being in an industrial area is easily incorporated in this random effect together with the altitude of the station, and the coefficient of these dummy variable and the last variable result in a distribution with very high variance, becoming non informative. They are from now on removed.

3.5. Spatial model

A spatial component is added to explicitly consider a spatial correlation between the observations. The general idea for introducing a spatial residual is that *everything is related to everything else, but near things are more related than distant things*. The relation of the observations is exploited by a covariance matrix which depends on the euclidean distance of the sites in which ozone is recorded. The most general form of covariance function is the Matérn covariance function:

$$C_\nu(d) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{d}{\rho} \right)^\nu K_\nu \left(\sqrt{2\nu} \frac{d}{\rho} \right)$$

where Γ is the gamma function, K_ν is the modified Bessel function of the second kind, ν, $\rho \geq 0$.
For the two different models the same class of covariance function has been used, and the exponential covariance, which corresponds to $C_{1/2}(d)$, is the optimal one for both quantities:

$$C_{1/2}(d) = \sigma^2 \exp\big(-\frac{d}{\rho}\big)$$

For what concerns the choice of ρ, according to the rule of thumb described in [5], we opt for $\rho = \frac{range}{3}$, which results in $21km$ as practical range.
The general form of the space model is the following:

$$n_{ijk} \mid \lambda_{ijk} \overset{ind}{\sim} Poi(\lambda_{ijk})$$
$$log(\lambda_{ijk}) = \underline{x}^T_{ijk}\underline{\beta} + \xi_j + \eta_k + w_k$$
$$\underline{\beta} \mid \sigma^2_\beta \sim \mathcal{N}_p(\underline{0}, \sigma^2_\beta \mathbf{I})$$
$$\underline{\xi} = (\xi_{2010}, \dots, \xi_{2022}) \mid \sigma^2_\xi \sim \mathcal{N}_p(\underline{0}, \sigma^2_\xi \mathbf{I})$$

$$\underline{\eta} = (\eta_1, \dots, \eta_{45}) \mid \sigma_\eta^2 \sim \mathcal{N}_p(\underline{0}, \sigma_\eta^2 \mathbf{I})$$

$$\underline{\mathrm{w}} = (\mathrm{w}_1, \dots, \mathrm{w}_{45}) \mid \sigma^2 \sim \mathcal{N}_p(\underline{0}, \sigma^2 \mathcal{H}) \qquad \mathcal{H} : (\mathcal{H})_{ij} = exp(\frac{1}{\rho} \cdot \mathrm{dist}_e(\mathrm{s}_i, \mathrm{s}_j))$$

$$\sigma_\beta^2, \sigma_\xi^2, \sigma_\eta^2, \sigma^2 \overset{\text{iid}}{\sim} inv - gamma(4, 2)$$

$$i = \text{April}, \dots, \text{October} \quad j = 2010, \dots, 2022 \quad k = 1, \dots, 45$$

The effect induced by the spatial relation will be then integrated out and induced in the coefficient for the stations. This upgrade is due to the fact that the coefficients η_k have all a posterior distribution centred on the value 0, indicating their significantly limited relevancy. Integrating the random effect these coefficient becomes again significant and can be compared with the others.

3.6. Model 180

For this model in particular a problem to be overcome is the high concentration of zeros in the observations, due to this kind of threshold being very difficult to be surpassed. This limitation has been tackled with the so called *Zero inflated Poisson*, which adds a variable θ indicating the probability for an observation to be null.
Being that ozone concentration highly depends on temperature, the concentration of zeros depends on the month in which an observation is gathered. The value of θ is assumed to be variable in each month, resulting in a vector of seven θ_i. The following is the complete model:

$$n_{ijk} \,|\, \lambda_{ijk} \overset{ind}{\sim} Poi(\lambda_{ijk})$$

$$\lambda_{ijk} = \theta_i \cdot 0 + (1 - \theta_i) \cdot \exp(\underline{x}_{ijk}^T \underline{\beta}) + \xi_j + \eta_k\}$$

$$\underline{\beta} \mid \sigma_\beta^2 \sim \mathcal{N}_p(\underline{0}, \sigma_\beta^2 \mathbf{I})$$

$$\underline{\xi} = (\xi_{2010}, \dots, \xi_{2022}) \mid \sigma_\xi^2 \sim \mathcal{N}_p(\underline{0}, \sigma_\xi^2 \mathbf{I})$$

$$\underline{\eta} = (\eta_1, \dots, \eta_{45}) \mid \sigma^2 \sim \mathcal{N}_p(\underline{0}, \sigma^2 \cdot \mathcal{H})$$

$$\mathcal{H} : (\mathcal{H})_{ij} = exp(\frac{1}{\rho} \cdot \mathrm{dist}_e(\mathrm{s}_i, \mathrm{s}_j))$$

$$\sigma_\beta^2, \sigma_\xi^2, \sigma^2 \overset{\text{iid}}{\sim} inv - gamma(4, 2)$$

$$\theta_1, \dots, \theta_7 \overset{\text{iid}}{\sim} \mathcal{B}\text{ern}(\frac{1}{2})$$

$$i = \text{April}, \dots, \text{October} \quad j = 2010, \dots, 2022 \quad k = 1, \dots, 45$$

Another important feature of this model is the complexity in building the credible intervals for the prediction of the data. In fact, allowing for a relevant probability of these samples being null, a significant quantity of zeros appears in the posterior distribution, leading the lower quantile to be always zero. This is not very informative, and so the proposed $(1 - \alpha)\%$ credible interval is of the following form.
Assuming that a fraction $\theta_i\%$ of the posterior distribution is put on zero, the points of the credible interval taking values different from zero are associated with a (1-θ) mass probability with respect to the whole credible interval. The probability associated with this interval should be $(1 - \alpha)(1 - \theta)$, so the practical approach is to select the $(1 - \alpha)(1 - \theta)\%$ quantiles of the posterior distribution outside zero:

$$\mathcal{I} = \{y \in \mathbb{R} : y \neq 0 \wedge \pi(y \mid \underline{y}) \geq k\}.$$

Summing all up:

$$\mathcal{CI} = \{0\} \cup \{y \in \mathbb{R} : \pi(y \mid \underline{y}) \geq k\}$$

$$\mathbb{P}(y \in \mathcal{CI}) = \theta + (1 - \theta)(1 - \alpha)\mathrm{q}_\pi((1 - \alpha)(1 - \theta)) \geq 1 - \alpha$$

.

3.7. Model 120

Regarding the model for the 120 threshold, the main problem to be tackled is the very high general level of the observations. In fact being this threshold lower than the one before, it happens in many situation that the maximum number of days in a month is reached, leading to coefficients that, in order to fit those kind of values, are such that some observations are estimated to be above the number of days in that month, which is obviously impossible by the nature of the problem. The first task is to try and lower this observations, that may reach even number such as 50 or 60, and for doing that a dummy variable that indicates, in a different way for different stations, whether it is the month of *July* or not. The rationale behind this is that basically the month for which that kind of problem happens is July, and assuming a dummy variable for that allows the model to adapt better in those observation and consequently the model has a lower estimate of the βs leading to the substantial elimination of prediction above the natural level. Despite this elaboration, very few data are still predicted above the number of days in that month, and to eliminate this problem a rejection sampling algorithm is implemented, where in the posterior estimate a sample is required until the estimate of the datum is not below the required number.
Hence the final model is:

$$n_{ijk} \mid \lambda_{ijk} \overset{ind}{\sim} Poi(\lambda_{ijk})$$

$$log(\lambda_{ijk}) = \underline{x}_{ijk}^T \underline{\beta} + \xi_j + \eta_k + \mathrm{w}_k + \gamma_i \cdot \delta_{\{i==\mathrm{July}\}}$$

$$\underline{\beta} \mid \sigma_\beta^2 \sim \mathcal{N}_p(\underline{0}, \sigma_\beta^2 \mathbf{I})$$

$$\underline{\xi} = (\xi_{2010}, \dots, \xi_{2022}) \mid \sigma_\xi^2 \sim \mathcal{N}_p(\underline{0}, \sigma_\xi^2 \mathbf{I})$$

$$\underline{\eta} = (\eta_1, \dots, \eta_{45}) \mid \sigma^2 \sim \mathcal{N}_p(\underline{0}, \sigma^2 exp(\mathcal{H}))$$

$$\mathcal{H} : (\mathcal{H})_{ij} = exp(\frac{1}{\rho} \cdot \mathrm{dist}_e(\mathrm{s}_i, \mathrm{s}_j))$$

$$\gamma_1, \cdots, \gamma_{45} \mid \sigma_\gamma^2 \overset{\mathrm{iid}}{\sim} \mathcal{N}_p(0, \sigma_\gamma^2)$$

$$\sigma_\beta^2, \sigma_\xi^2, \sigma_\gamma^2, \sigma^2 \overset{\mathrm{iid}}{\sim} inv - gamma(4, 2)$$

$$i = \mathrm{April}, \dots, \mathrm{October} \quad j = 2010, \dots, 2022 \quad k = 1, \dots, 45,$$

where in the posterior inference the pseudo-code is the following:

```
while y > max_month_i:
sample y ~ Poi(λ_ijk)
```

Other approaches exploited but resulted in poor performance are, for instance, the double Poisson suggested in [3], where two different parameters are set for controlling both the expected value and the covariance of the Poisson, but this makes the variance of the values to predict too large and result in very poor prediction. Another approach is to truncate the distribution of the predicted values to the maximum number of days in that month, but even this approach results in a worse performance than our model.

4. Posterior inference

The models described in Section 3 have been compared using different prediction-based scores, in order to evaluate their performance in predicting new observations. The metrics used are:

- Leave-one-out cross-validation (LOO) and the widely applicable information criterion (WAIC), which are methods for estimating pointwise out-of-sample prediction accuracy from a fitted Bayesian model using the log-likelihood evaluated at the posterior simulations of the parameter values;
- Mean square error (MSE) and mean absolute error (MAE) computed considering the distance between the median of the predictions and the corresponding observation;
- Percentage of the observations belonging to the 95% CI of the corresponding predictions for all data.

The results reported below show a gradual and constant improvement, along with the increase in complexity. This aspect seems to be more evident for the models for Count 180, where the advancements appears relatively larger.
It's important to point out that the values related to %CI have been computed considering just the training set of the models, which leads to overoptimistic results. The exactly percentages will be shown in Section 4.6.

Model	WAIC	LOO	MSE	MAE	%CI
Base	-5431.95	-5432.00	5.273	1.179	91.81%
Var sel	-5432.21	-5432.26	5.290	1.178	91.65%
Years	-4813.53	-4813.78	4.025	1.000	93.25%
Stations	-3838.39	-3838.98	2.260	0.723	96.98%
Spatial	-3837.95	-3838.47	2.244	0.721	96.90%
Spatial 2	-3838.12	-3838.63	2.265	0.722	96.90%
ZIP	**-3712.28**	**-3712.88**	**2.177**	**0.717**	**98.11%**

Table 2: Evaluation of models for Count 180

Model	WAIC	LOO	MSE	MAE	%CI
Base	-13593.08	-13593.17	29.608	3.760	80.16%
Var sel	-24873.87	-24873.78	55.945	5.126	55.49%
Years	-11159.35	-11159.49	19.577	3.261	86.20%
Stations	-10100.17	-10100.52	14.026	2.802	91.15%
Spatial	10100.81	-10101.11	14.030	2.797	91.17%
Spatial 2	-10101.17	-10101.48	14.068	2.799	91.20%
Rej. sampling	**-10056.66**	**-10057.39**	**13.483**	**2.739**	**91.51%**

Table 3: Evaluation of models for Count 120.

4.1. β

From this plot it is clear how overall the βs are almost all significant with the same sign both for the two models, the only one that is not significant for the 180 threshold and it is so for the 120 is the sum of precipitation (nevertheless, the point estimate for both of them is above zero).

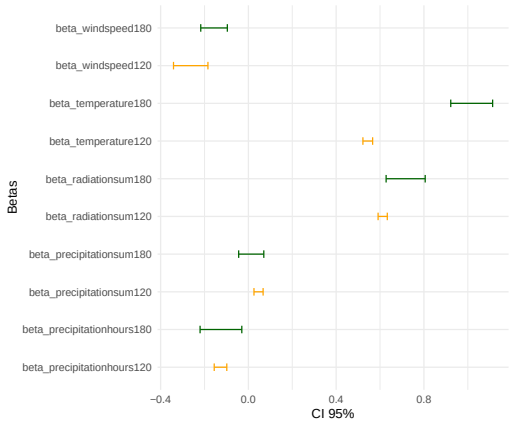

Figure 5: coefficients for the different models.

Regarding the others, as it might be expected the coefficients related to the atmospheric events like wind speed, precipitation sum and precipitation hours are significantly negative, while the coefficients for the temperature and the radiation are bigger than zero.
An important thing to point out is that precipitation sum is bigger than zero, in contrast with what expected, and this may be explained by the significantly negative coefficient for the precipitation hours.
Subsequently, it is important to remark that coefficients for the 120 threshold are lower despite having bigger values in the observations, this means that the other types of coefficients in the model will necessarily be higher than the others, and note that this effect may be incremented by the dummy coefficient that is added and could increase the observations in case of high values. Finally, the variance for the βs of the 120 threshold is lower than the others, and this may be explained by the higher bounds of the sigma for the β that allow for smaller variance, that we will find in Section 4.4.

4.2. ξ

The ξs coefficients represent the year-effect of the model and it is clear how in almost all the years the years the factor for the 120 model is higher than the other, in particular only these coefficient are repeatedly above zero significantly. As anticipated in Section 4.1 this is an expected behaviour, indeed these are the coefficients for which the high value of the 120 threshold make the estimate grow.

Regarding the 180 threshold, it seems that going on with the ages the years-dependent factor decreases, as

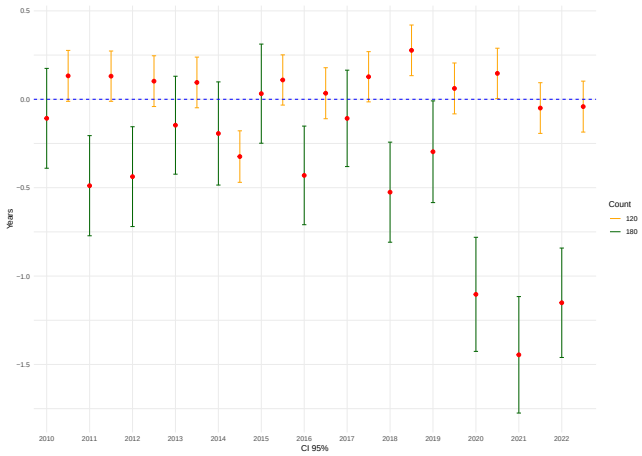

Figure 6: ξ coefficients for the different models.

it might not be expected. Nevertheless, note that these coefficient reflect only year-specific situations, not the overall trend, which may be increasing with years. In fact, the overall increment in the temperature may increase the values of very polluted days without the need of the year's coefficient, so it is not a so strange result.

4.3. η

The parameter η in both the models explains both the variability due to the spatial residual and the variability due to the specific station, and as remarked in Section 3.4 they account even for the specific type of zone in which the station is and for the altitude in which the station is. A scale is used in order to identify possible similarities: the trend appears to be similar, in fact in the neighborhood of big cities the factors increase, and the spatial trend, where near stations appear to have similar values, this is magnified even by the defined matrix for the spatial effect where stations that are near are set to be correlated.
In the Figures 8 and 7 the plot with the median estimates for the various stations.

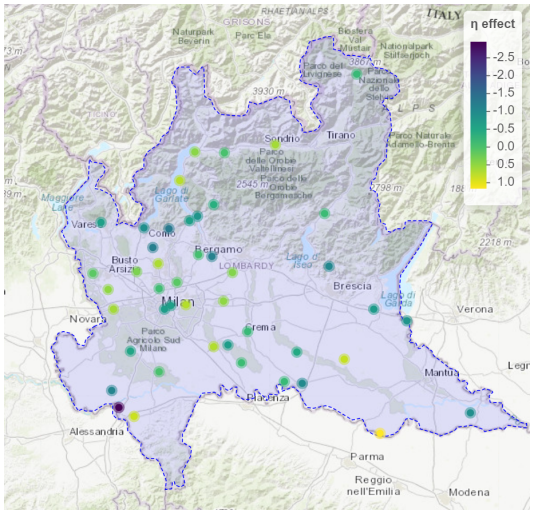

Figure 7: Effect on the different stations of the η for Model 180.

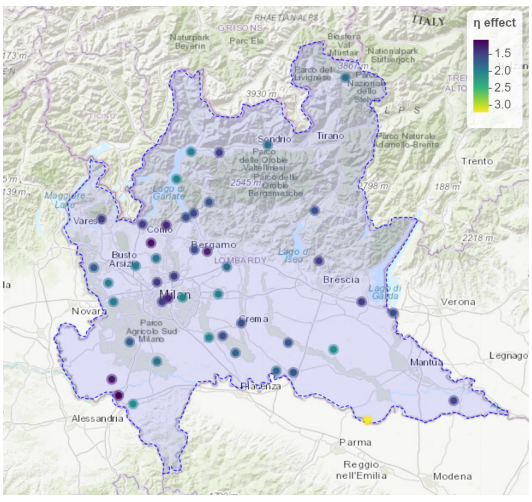

Figure 8: Effect on the different stations of the η for Model 120.

Moreover, the coefficients for the 120 model have very higher values with respect to the others, as anticipated

in Section 3.4, in fact these together with the coefficients for the year in Section 3.3 make the prediction for the 120 model higher than the other, as the data require.

4.4. Variance

The variance of each of the parameters is a very important quantity in the overall analysis, in fact it represent the importance of each of the component in the model. Given that in both the quantities of interest the variance parameters are indexed in the same way, it is possible to compare them and to make comparisons on them. Regarding the sigma, which is referred to the spatial residual together with the specific information of the station, it is significantly different from zero, meaning that the spatial analysis made is influential in the good prediction of the model, and for the two datasets the values are similar. Talking about the variance associated with the coefficients of the β, the model for the threshold 120 has a very higher variance with respect to the 120's one, and it is explainable with the very higher variance that the observations have in the two different datasets, being that in the 120 threshold there are a lot of high observations, that result in this higher variance. The ξs instead present two variances that differ in the two different models, in fact the variance regarding the 180 threshold is much higher with respect to the other variance, meaning that the values of the year-specific factor are more important for the information threshold with respect to the long term objective. There is only

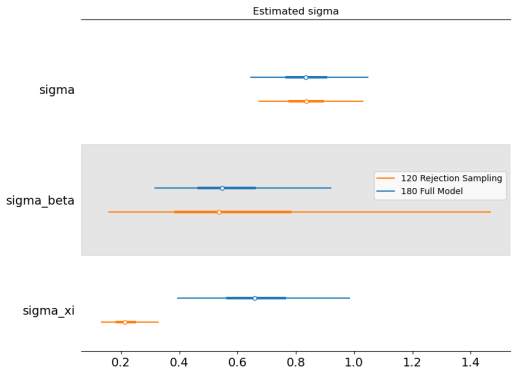

Figure 9: 95% CI for the σ.

a single variance out of this plot, which is the variance for the dummy variable associated with the month of July in the 120 threshold's model: this variance is significantly different from zero and its associated factors are fundamental for the goodness of fit of the model.

4.5. θ

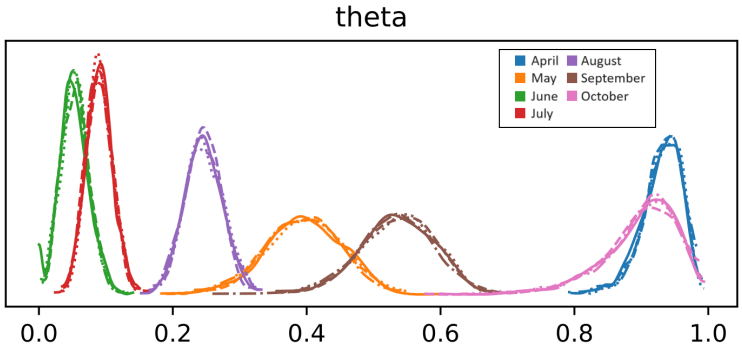

Figure 10: θ values for the ZIP model.

The posterior distribution of θ stands for the posterior probability of being zero in the corresponding month, according to the zero inflated Poisson model. As expected, the summer's months are all to the left part of the domain $[0, 1]$, in particular June is the month for which the probability of have a zero is the lower, while colder months are at the right of the figure and October and April are the two months with the higher probability of have a zero in the observation.

4.6. Final model predictions

In the following analysis a 70/30 split of the dataset into a train and test set was executed. The way this was done was by treating 30% of the non-NA observations of the dataset as missing values.

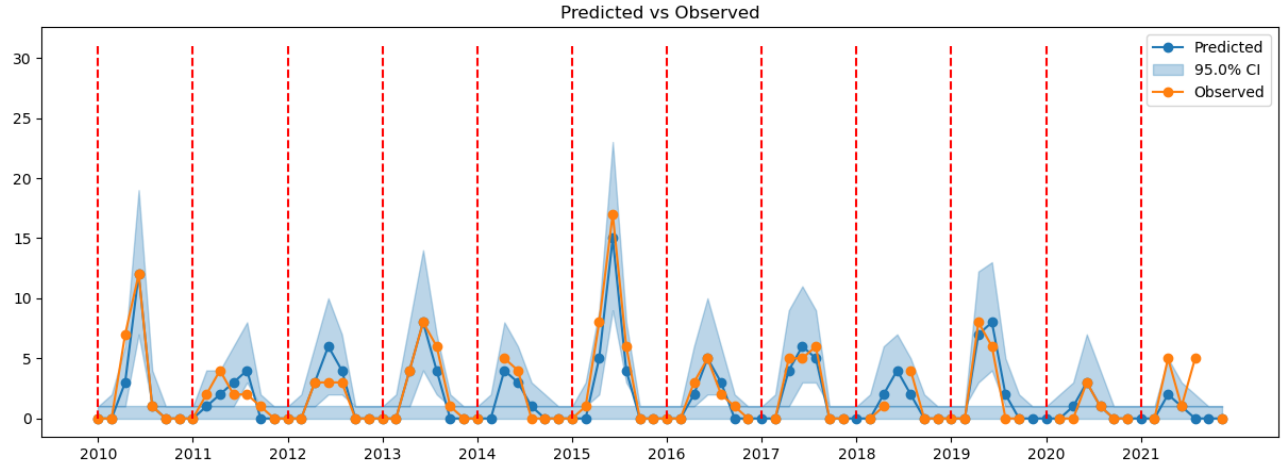

Figure 11: Predicted count values for the 180 model with 95% CI vs Observed values.

	MSE	MAE	%CI
ZIP Model	2.757	0.774	94.05%

Table 4: Perfomance of the ZIP model on the test set.

This is the by month prediction of the count data values obtained through the 180 zip model for the first station. The trend of the data is noticeably well captured by the model and this is reflected in the overall performance of the predictions. The credible intervals around the predictions are computed as mentioned above in section 3.6 and reflect the bi-modality of the posterior distribution of the predictions.

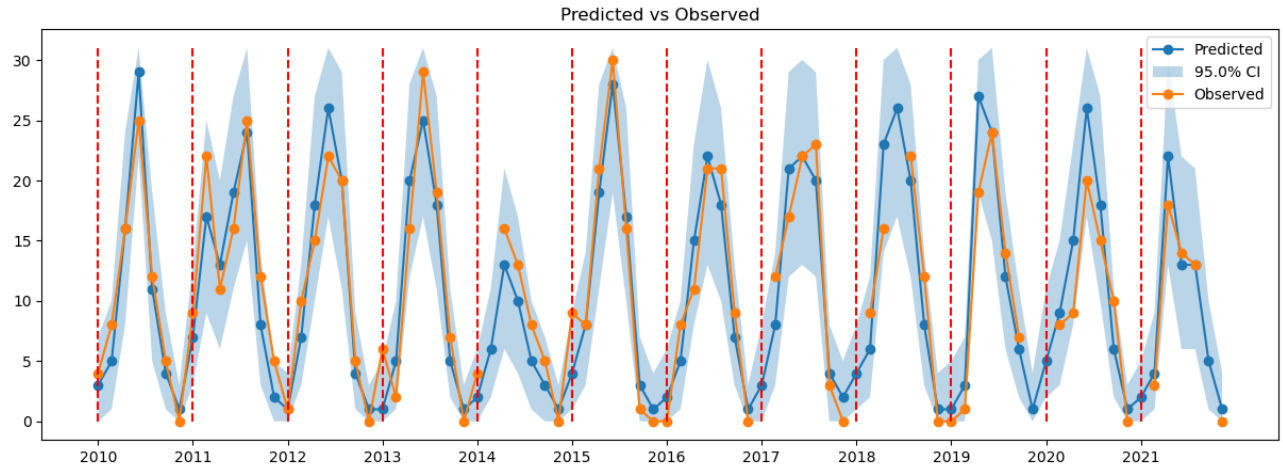

Figure 12: Predicted count values for the 120 model with 95% CI vs Observed values.

	MSE	MAE	%CI
120 Model	14.750	2.906	90.45%

Table 5: Perfomance of the 120 model on the test set.

This is the monthly prediction of the count data values for the first station obtained through the 120 model with the added dummy variables and using rejection sampling to sample the predictions. The observed values

reach as expected noticeably higher values but the trend of the count data is still very well captured. It can be noticed that due to rejection sampling the credible intervals are comparatively smaller when the predictions are near critically high valued points.

5. Conclusions

In these years attention to global warming has grown a lot, and many studies about pollution are everyday cited both on academic and in ordinary context. Understanding the precise causes of ozone pollution was the goal of this project, and overall the reached precision for the prediction of the two thresholds is quite satisfactory. The results about how ozone appears in the air and has very high levels are the ones expected, with high influence of temperature and radiation as they are catalysts for the development of this pollutant. Rain and wind act, as expected, as reducing factors for the ozone. This fact makes the situation even worse, considering the actual global warming situation that reduce the rain and increase the temperature.
The risk associated with this kind of pollutant should not be underestimated and in this direction attention to global effect of increasing of the temperatures might be taken seriously.

References

[1] Istituto Nazionale di Statistica. www.istat.it, 2023.

[2] Konstantinos Fokianos. Truncated poisson regression for time series of counts. *Scandinavian Journal of Statistics*, 2001.

[3] James E. Pustejovsky. www.jepusto.com/double-poisson-in-stan/, 2023.

[4] Geoportale Regione Lombardia. www.geoportale.regione.lombardia.it, 2023.

[5] Sujit K. Sahu. *Bayesian Modeling of Spatio-Temporal Data with R*. Chapman and Hall/CRC, 2021.

[6] National Weather Service. www.weather.gov/mfl/beaufort, 2023.

[7] Ming-Hung Tsai and Tsair-Wei Lin. Modeling data with a truncated and inflated poisson distribution. *Statistical Methods and Applications*, 2017.

[8] Patrick Zippenfenig. Open-meteo.com weather api, 2023.

A. Appendix

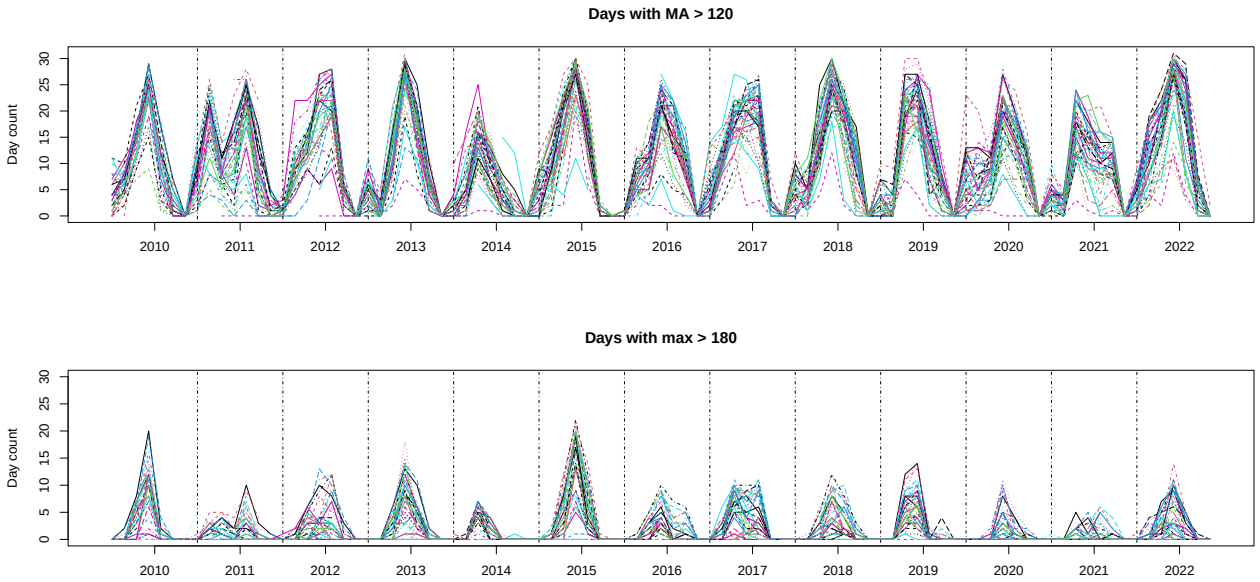

Figure 13: Plot of the two quantities of interests along the years.

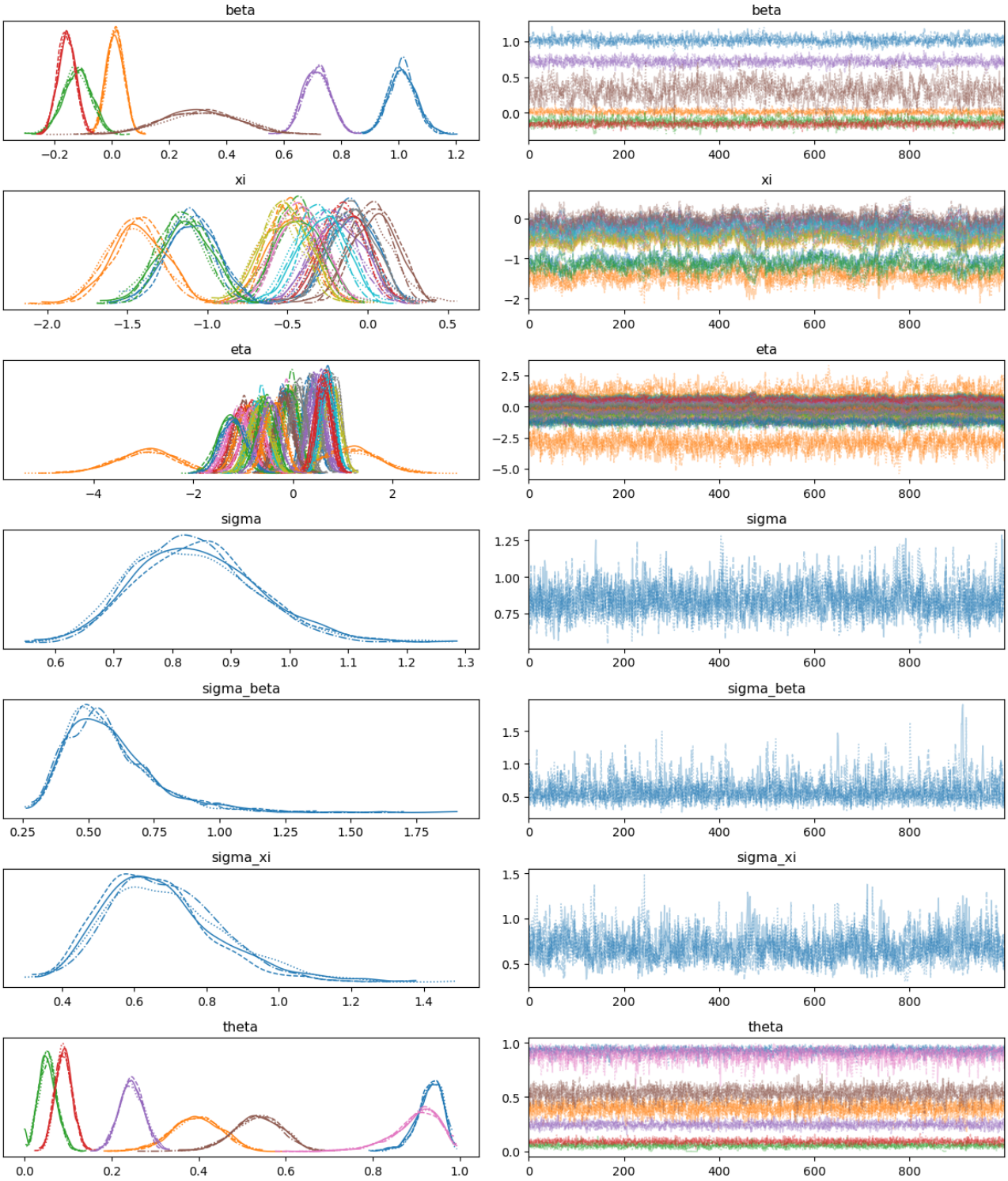

Figure 14: Traceplots for the Model 180.

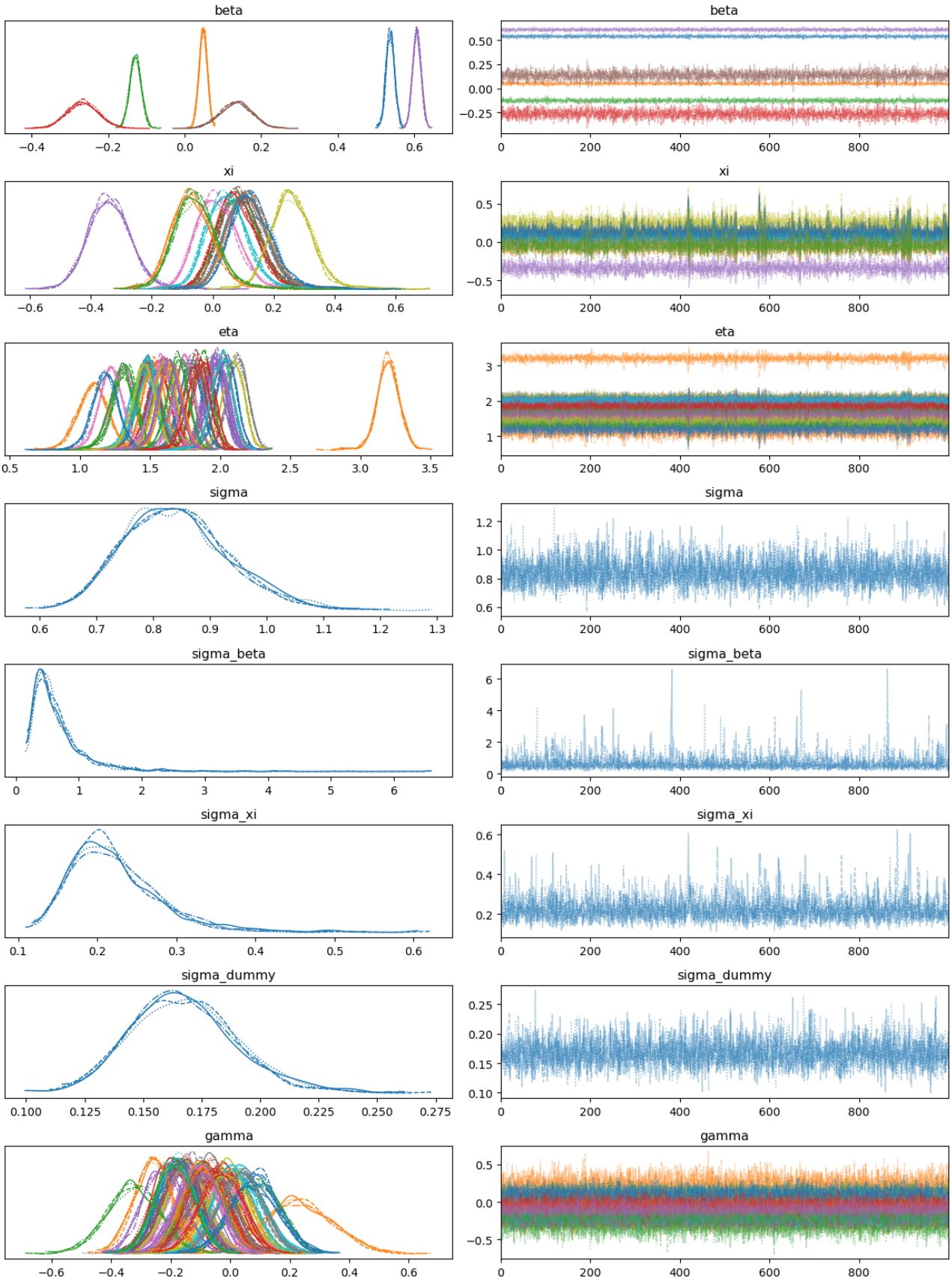

Figure 15: Traceplots for the Model 120.