A Bayesian Approach to Model the $PM_{2.5}$ of Counties in Eastern Iowa

Jinu Susan Kabala

April 30, 2018

1 Introduction

$PM_{2.5}$ refers to "particulate matter" that has a diameter of 2.5mm. These particles are extremely small and can stay in the atmosphere longer than larger particulate matter[1]. Since the particles are extremely small, they can penetrate deep into our respiratory system and circulatory system and cause serious health issues. Our goal is to model the mean $PM_{2.5}$ concentrations of six counties in Eastern Iowa. Using a Bayesian model, we try to answer the following scientific questions of interest:
(i) What are the mean concentrations of $PM_{2.5}$ in a county in Iowa on a particular day?
(ii) Is there a relationship between the mean concentration of $PM_{2.5}$ and climate factors like wind and temperature?
(iii) What days are the mean concentrations of $PM_{2.5}$ the highest?
(iv) What counties are the mean concentrations of $PM_{2.5}$ the highest? Does the number of toxic sites influence the mean concentrations of $PM_{2.5}$?

2 Data

We collected data from three sources: EPA-Air data [3], NOAA Climate Data[5] and EPA-Toxic Release Inventory[4]. Table 1 shows a summary of the data we used to build our Bayesian model.

EPA-Air data [3] contains data that has been averaged from $PM_{2.5}$ pollutant samples measured on a 24-hours basis. The unit of measurement is micrograms/cubic meter. We noticed that the dataset does not contain information of all counties of Iowa and the most recent year available was 2017. We decided to use data for only one month and we randomly picked June 2017. For this paper, we use data from 6 counties in Eastern Iowa (Clinton, Delaware, Johnson, Linn, Muscatine, and Scott) which is shown in Figure 1. We chose these counties from the available data for 14 counties so as to capture the spatial effects. Since the data was not available for all days of June for all counties, we extracted the days for which there was no missing data. The eight days in June 2017 that we extracted data from are 06,09,12,15,18,21,24, and 27 which we annotate as time points 1, 2, 3, 4, 5, 6, 7 and 8 respectively. For the days chosen, we extracted maximum temperature, minimum temperature and average wind-speed from the NOAA Climate Data[5]. Since average temperature information was not available, we used only the maximum temperature.

3 Methods

3.1 Model Rationale

In Figure 3, we make a plot that shows the relationship between average wind-speed and maximum temperature with relation to average $PM_{2.5}$. These were measured in six counties of eastern Iowa (Clinton, Delaware, Johnson, Linn, Muscatine, and Scott) at time points {1,2,3,4,5,6,7,8} which corresponds to dates 06/{06,09,12,15,18,21,24,27}/2017 respectively. The data is grouped by counties as well as by time points to analyze relationships between average wind-speed and maximum temperature with relation to the average $PM_{2.5}$. We did not see any patterns and so, a simple linear model may not be sufficient.

In Figure 4, we made a plot to analyze the relationship between the six counties in Eastern Iowa selected, against: average $PM_{2.5}$, average wind-speed, and maximum temperature sampled at the eight time points. There does not appear to be much variability in average $PM_{2.5}$ values in the sampled counties when grouped by time. However, there does appear to be a variability in average wind-speed as well as maximum temperature values at these counties when grouped by time. We noticed that at time point 3, all counties record the highest average $PM_{2.5}$ value and maximum temperature value.

In Figure 5, we analyze the relationship between the time points against average $PM_{2.5}$, average wind-speed, and maximum temperature sampled at six counties in Eastern Iowa. The time plots show a clear pattern for average $PM_{2.5}$ when grouped by counties. They all exhibit a similar trend and the peak value for average $PM_{2.5}$ is observed at time point 3 in all six counties. A similar trend can be seen in the time plot for the maximum temperature where the temperature peaks at time point 3 for all six counties.

These plots motivated us to build a spatio-temporal model based on the average $PM_{2.5}$ data we had collected using maximum temperature and average wind-speed as explanatory variables.

3.2 Model

Data Model:

Let y_{ct} be the average $PM_{2.5}$ value for county c at time point t where $c \in$ {*Clinton, Delaware, Johnson ,Linn, Muscatine, Scott*} and $t = \{1, 2, 3, 4, 5, 6, 7, 8\}$ which corresponds to dates 06/{06,09,12,15,18,21,24,27}/2017. We transform the average $PM_{2.5}$ values on the log scale. Hence, we use $y_{ct}^{log} = log(y_{ct})$ when defining our data model, as $PM_{2.5}$ values are non-negative. $x_{temp[ct]}$ and $x_{wind[ct]}$ denotes the explanatory variable for maximum temperature and average wind-speed for county, c at time, t respectively. We fit four different models that account for complexity of space and time along with explanatory variables (maximum temperature, and average wind-speed) by changing the mean parameter of y_{ct}^{log}. The log transformed data, y_{ct}^{log} follows a normal distribution with mean parameter μ_{ct} and variance parameter σ^2. (Linear Model) specifies the mean parameter for a linear model that explains the data using the explanatory variables, maximum temperature and average wind-speed using β parameter for their coefficients. (Temporal Model) specifies the mean parameter for a temporal model that adds a temporal parameter, η and white noise term, ω for each time point, $t \in \{1, 2, 3, 4, 5, 6, 7, 8\}$. (Spatial Model) specifies the mean parameter for a spatial model that adds a spatial parameter, θ for each county, $c \in \{1, 2, 3, 4, 5, 6\}$ corresponding to {*Clinton, Delaware, Johnson ,Linn, Muscatine, Scott*}. (Spatio-Temporal Model) combines these models into a single model.

$$y_{ct}^{log} \overset{ind}{\sim} N(\mu_{ct}, \sigma^2)$$

$$\mu_{ct} = \beta_1 + \beta_2 x_{temp[ct]} + \beta_3 x_{wind[ct]} \tag{Linear Model}$$

$$\mu_{ct} = \beta_1 + \beta_2 x_{temp[ct]} + \beta_3 x_{wind[ct]} + \eta_t \tag{Temporal Model}$$

$$\eta_t = \phi \eta_{t-1} + \omega_t$$

$$\mu_{ct} = \beta_1 + \beta_2 x_{temp[ct]} + \beta_3 x_{wind[ct]} + \theta_c \tag{Spatial Model}$$

$$\mu_{ct} = \beta_1 + \beta_2 x_{temp[ct]} + \beta_3 x_{wind[ct]} + \eta_t + \theta_c \tag{Spatio-Temporal Model}$$

Hierarchical Model:

In order to account for spatial and temporal effects of our model, we introduce parameters θ and ω respectively. We use an Auto-Regressive model, AR(1) for the time-effect and a Conditional Auto-Regressive (CAR) model for the spatial effect. The noise term of the AR(1) model has a normal distribution with mean 0 and variance parameter σ_ω^2. The spatial parameter, θ follows a multivariate normal distribution of dimension, C with a mean vector of zeros and a variance-covariance matrix, $[\tau(D - \rho W)]^{-1}$. ρ is a parameter that controls spatial dependence, so if $\rho = 0$, it implies spatial independence. W is a C×C matrix such that an element in W, $w_{ii} = 0$ and $w_{ij} = 1$ if county, i is a neighbor of county, j. D is a C×C diagonal matrix where $d_{ii} = \sum_{j=1}^{C} w_{ij}$. C is the total number of counties. Figure 2 shows the six counties in this study as a connected graph, so our data contain no islands.

$$\omega_t \sim N(0, \sigma_\omega^2)$$

$$\theta \sim N_C(0, [\tau(D - \rho W)]^{-1})$$

Priors:

We need a prior for β_p where $p \in \{1, 2, 3\}$ which accounts for intercept and coefficients of the explanatory variables (maximum temperature and average wind-speed). The prior chosen for β_p is a normal distribution with mean and variance 0 and 0.5 respectively. The priors for standard deviation, σ and σ_ω are chosen to be half-Cauchy with location and scale values 0 and 5 respectively. $\phi \in [-0.999, 0.999]$ is chosen to be a normal distribution with mean 0 and variance 1. ρ follows a uniform distribution in the interval [0,1]. τ follows a gamma distribution with both shape and rate value of 2. All the priors used here are proper which ensures that the posteriors are also proper.

$$\beta_p \sim N(0, 0.5)$$

$$\sigma \sim Ca^+(0, 5)$$

$$\phi \sim N(0,1)$$
$$\sigma_\omega \sim Ca^+(0,5)$$
$$\rho \sim U(0,1)$$
$$\tau \sim Ga(2,2)$$

3.3 Model Fitting

We used Stan to generate MCMC samples from posterior distribution of parameters using the R package rstan. We first fit a linear model, then a temporal model followed by a spatial model and finally we ended up with a spatio-temporal model which is our final model. We use default values in Stan which is 4 chains. Each chain ran for 2000 iterations with a burn-in of 2000 with no thinning resulting in a total of 4000 post-warmup draws. The potential scale reduction factor is 1 for all the parameters β_p, ω_t, θ_c, σ, σ_ω, ϕ, τ, and ρ. The spatial model generated 600 samples for θ_c and the temporal model generated atleast 1000 samples for ω_t. The traceplots for all four models show well mixed chains and show no lack of convergence. The traceplots of the estimated parameters for the spatio-temporal model is shown in Figure 6. The potential scale reduction factor and the traceplots indicate that the Markov Chains converge.

We extracted samples from the posterior distributions of μ and σ of the fitted spatio-temporal model to generate $y^{(rep)}$ replications on the original scale of average $PM_{2.5}$ value as the original data sample. We did this by transforming it with the *exp* function. We used 20 $y^{(rep)}$ replications as shown in Figure 7 and randomly replaced one of these histograms with the one generated by the original $PM_{2.5}$ data. We did not detect any lack of fit of the model from this visual posterior predictive check.

3.4 Results

The histograms of posterior draws of the parameters with the curves of priors are shown in Figure 8. The priors for β, ϕ, τ and ρ overlap indicating that they are informative priors; whereas the histograms for σ and σ_ω do not overlap.

The 95 % posterior credible intervals for ω_t and θ_c are shown in Figure 9. The credible intervals ω_t for $t = \{2,3,4,5,6,7,8\}$ are narrow indicating good estimates for the temporal effects. The posterior median of ω_3 is the highest. Since the posterior credible intervals of $\omega_2, \omega_3, \omega_5, \omega_6$ does not contain zero, it indicates strong evidence of temporal effect at time points, t={2,3,5,6} corresponding to dates 06/{09,12,18,21}/2017. On the other hand, the 95% posterior credible intervals for θ_c are wider and cover zero values. The spatial effect is highest for θ_4 which is for Linn county and least for θ_2 which is for Delaware county.

Figure 10 shows box plots of posterior MCMC samples of temporal and spatial parameters, ω_t and θ_c respectively. The posterior median and standard deviation of spatial parameter θ is displayed in Figure 11. The posterior medians of θ_c are similar for Muscatine and Scott. The posterior median for all counties are very close to each other showing similar variability across their medians. This indicates little evidence of a spatial effect. The posterior medians of ω_t are not similar and exhibits different variability across time points. Hence, it shows strong evidence of temporal effects. Table 2 shows the posterior median and 95%

credible intervals around the median for the parameters β_p where p={1,2,3} which denotes the effect of the explanatory variable. The posterior median for β_2 and β_3 are not zero which suggests the explanatory variables (maximum temperature and average wind-speed) may have some effect on average $PM_{2.5}$ value.

We plotted the posterior median of μ parameter in Figure 12 which is our mean parameter for average $PM_{2.5}$ in the same scale as the original data. We had used log scale for building the model. This plot resembles the original data plot shown in Figure 4 and Figure 5 where the average $PM_{2.5}$ is plotted against time points across counties and also at counties across time. Table 3 shows the posterior median and 95% credible intervals around the median for the parameters the mean parameter μ_{ct}. The close resemblance to the original data suggests that this model is useful for estimating the mean concentration of $PM_{2.5}$ from the posterior μ_{ct}.

3.5 Discussion

In this section we try to answer the scientific questions that we posed in Section 1. From our results in the previous section, we can use the μ_{ct} parameter to give a good estimate of mean concentrations of $PM_{2.5}$ in the log scale. From the value of our posterior estimates of $\beta_p > 0$, we can say there is some evidence to suggest a relationship between the mean concentration of $PM_{2.5}$ and climate factors like wind-speed and temperature. We may need more data that spans several days to observe this relationship. In Figure 12, we can see at time point 3 (06/12/2017), the mean concentration of $PM_{2.5}$ is the highest by looking at the posterior estimates for μ_{c3}, $\forall c \in \{Clinton, Delaware, Johnson, Linn, Muscatine, Scott\}$. Also, we can see that for Linn county, the mean concentration of $PM_{2.5}$ is the highest by looking at the posterior estimates for $\mu_{[Linn,t]}$, $\forall t \in \{1, 2, 3, 4, 5, 6, 7, 8\}$ followed by Muscatine and Scott counties.

3.6 Improvements and Future Work

From the EPA-TRI data[4], among the counties we studied, Linn has the highest number of toxic sites of 18 followed by Scott with 15 toxic sites and Muscatine with 11 toxic sites which also has the highest posterior median for mean parameter, μ_{ct} as well. It would be interesting to investigate if there is a relationship between the mean concentration of $PM_{2.5}$ in these counties with the number of toxic sites. This would indicate a strong spatial effect. We feel that we do not have sufficient data to capture the spatial effect or the effect of explanatory variables like wind and temperature. Hence, we would like to check for the fit of the model with more data samples. Also, in our implementation of the CAR model, we assume that there are no islands in our spatial structure. We would like to fix this issue as well.

References

[1] Bliss Air,
https://blissair.com/what-is-pm-2-5.htm

[2] Iowa County Map,
https://www.digital-topo-maps.com/county-map/iowa.shtml

[3] US Environmental Protection Agency,Air Data,
https://aqs.epa.gov/aqsweb/airdata/

[4] US Environmental Protection Agency, Toxics Release Inventory (TRI) Program,
https://www.epa.gov/toxics-release-inventory-tri-program

[5] National Oceanic and Atmospheric Administration (NOAA) Climate Data,
https://www.climate.gov/

A Figures

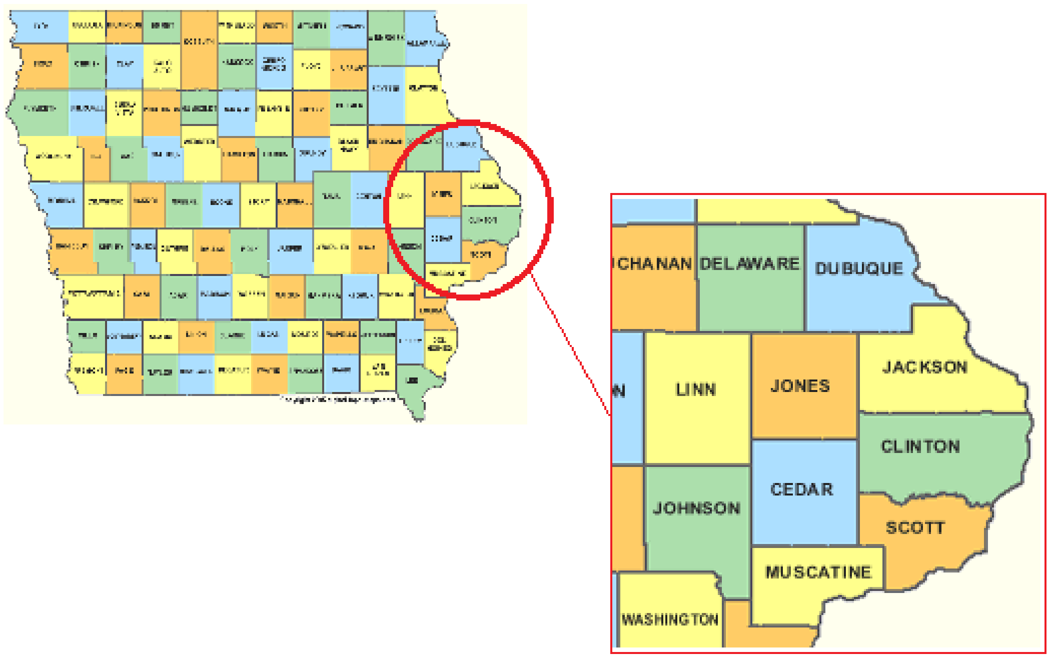

Figure 1: Map of counties in Eastern Iowa[2] that are part of the $PM_{2.5}$ study, specifically Clinton, Delaware, Johnson, Linn, Muscatine, and Scott.

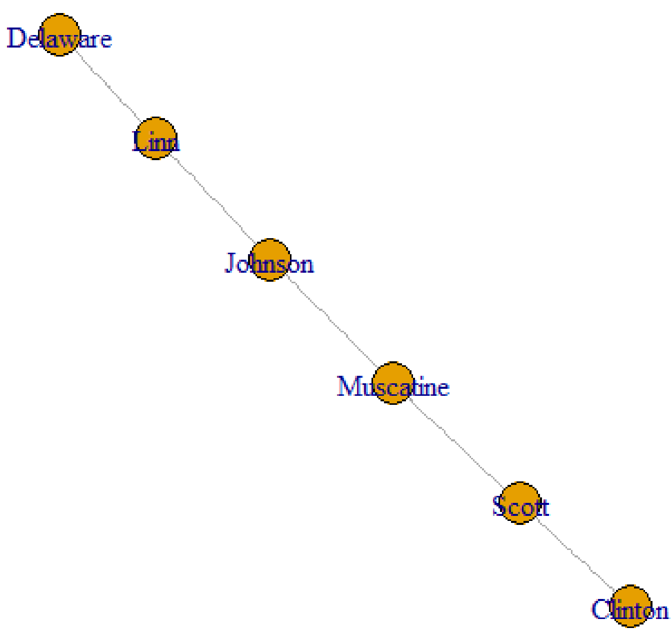

Figure 2: A bidirectional graph of the counties of Eastern Iowa that are part of the $PM_{2.5}$ study. An edge between two counties indicate a shared border between them.

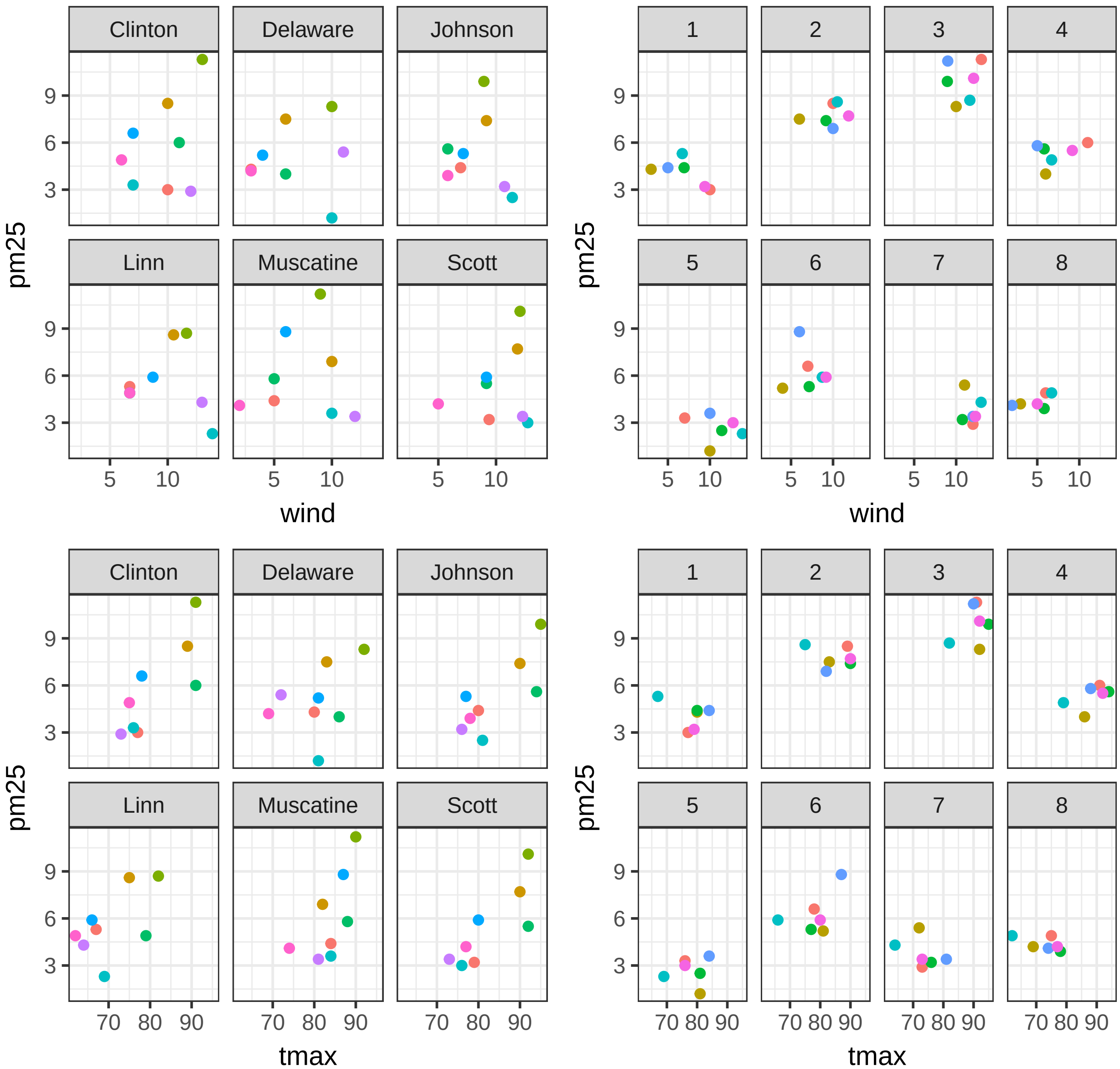

Figure 3: Plot that shows the relationship between the average wind-speed(wind) and maximum temperature(tmax) against the average $PM_{2.5}$(pm25) measured at the six counties in eastern Iowa(Clinton, Delaware, Johnson, Linn, Muscatine, and Scott) at time points {1,2,3,4,5,6,7,8} which corresponds to dates 06/{06,09,12,15,18,21,24,27}/2017 respectively. The data is grouped by counties as well as by time points to analyze relationships between average wind-speed(wind) and maximum temperature(tmax) with relation to average $PM_{2.5}$(pm25) samples.

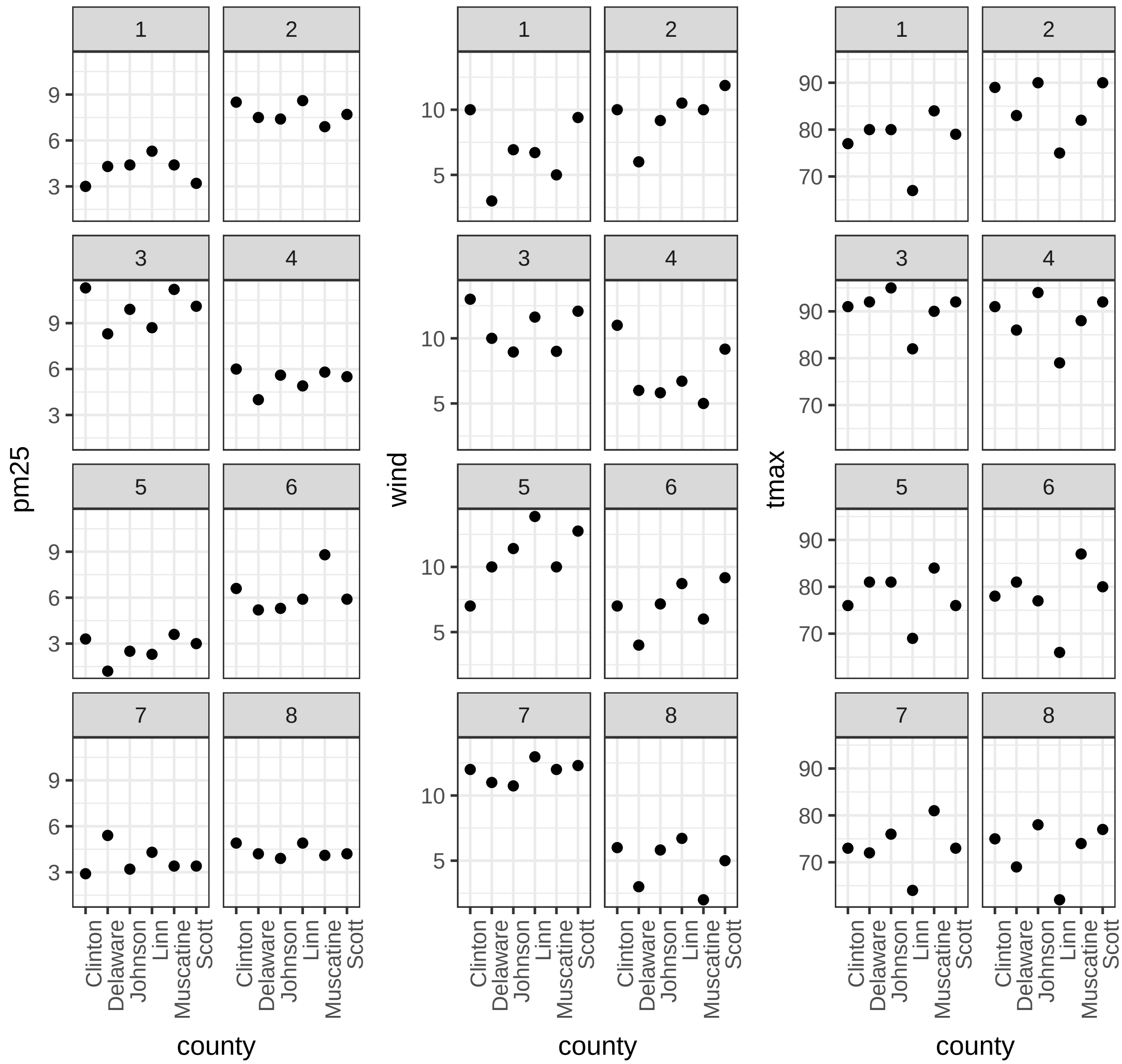

Figure 4: Plot that shows the relationship between the six counties in Eastern Iowa(Clinton, Delaware, Johnson, Linn, Muscatine, and Scott) against the average $PM_{2.5}$(pm25), the average wind-speed(wind) and the maximum temperature(tmax). These are sampled at time points {1,2,3,4,5,6,7,8} which corresponds to dates 06/{06,09,12,15,18,21,24,27}/2017.

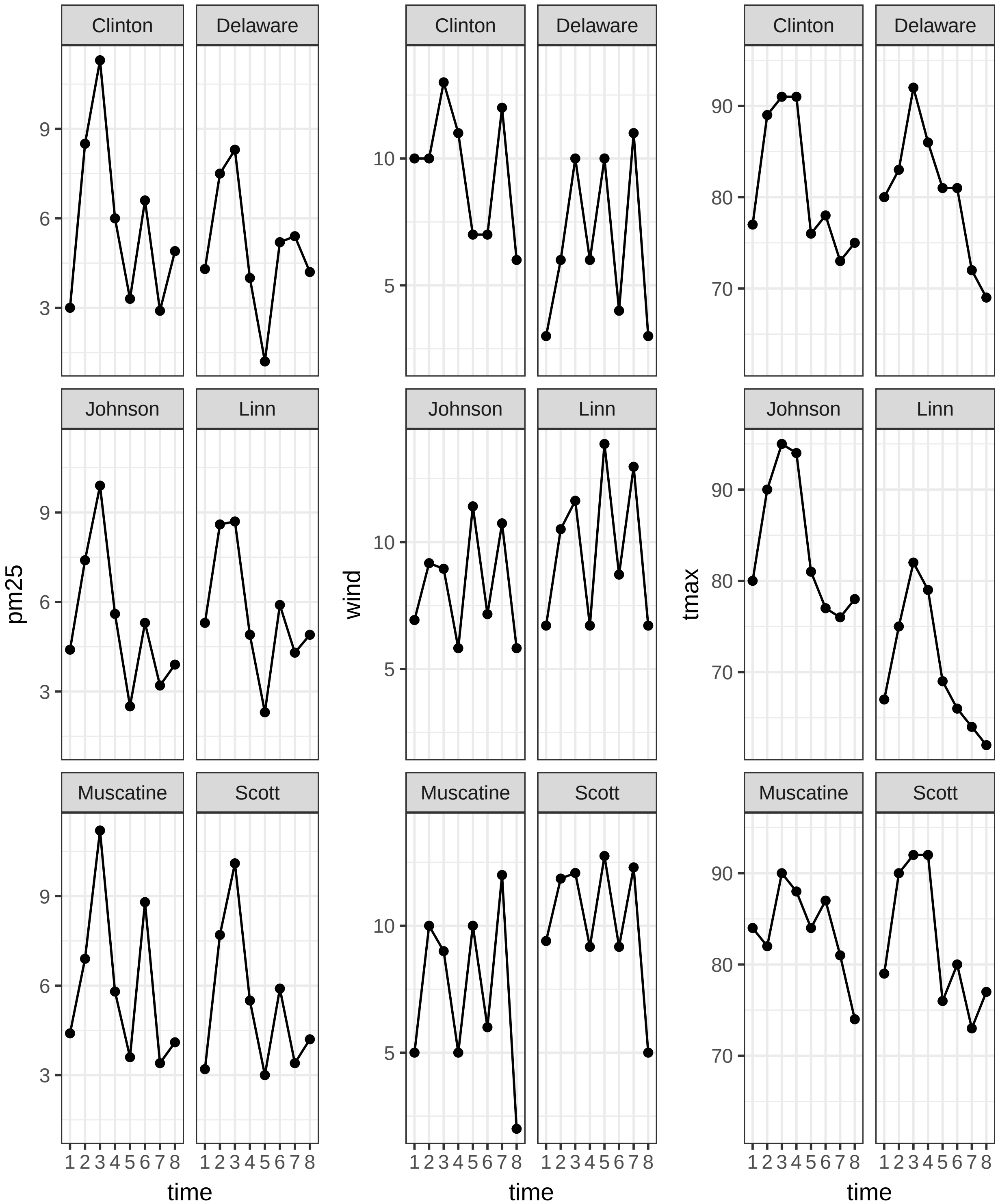

Figure 5: Plot that shows the relationship between the time points {1,2,3,4,5,6,7,8} which corresponds to dates 06/{06,09,12,15,18,21,24,27}/2017 against the average $PM_{2.5}$(pm25), average wind-speed(wind), and maximum temperature(tmax) sampled at six counties in Easter Iowa(Clinton, Delaware, Johnson, Linn, Muscatine, and Scott).

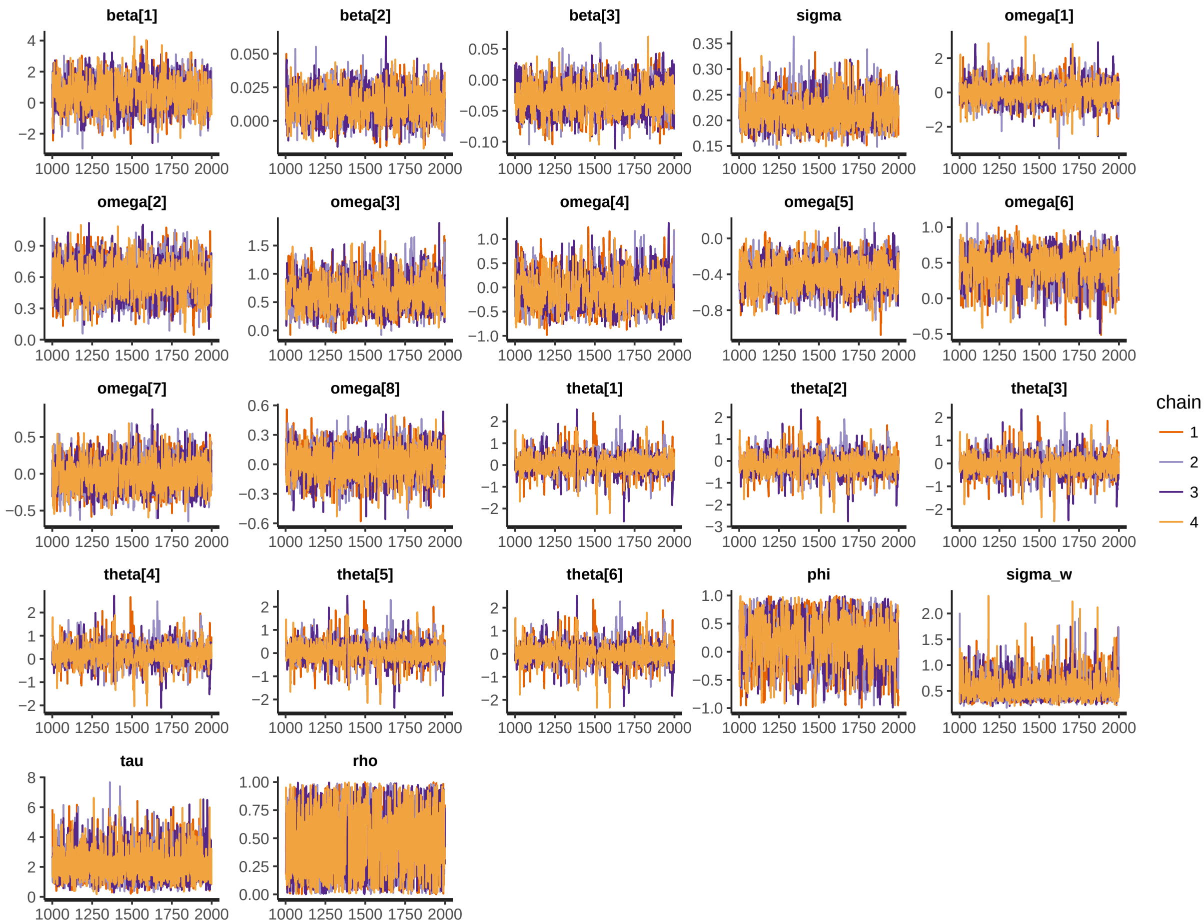

Figure 6: Traceplot of estimated parameters β_p, ω_t, θ_c, σ, σ_ω, ϕ, τ, and ρ

Figure 7: Posterior predictive plot for spatio-temporal model that shows the histograms of 19 replications(in red) of y indicating the estimated average $PM_{2.5}$ value. The histogram of the original $PM_{2.5}$ data is indicated in box 16(in blue).

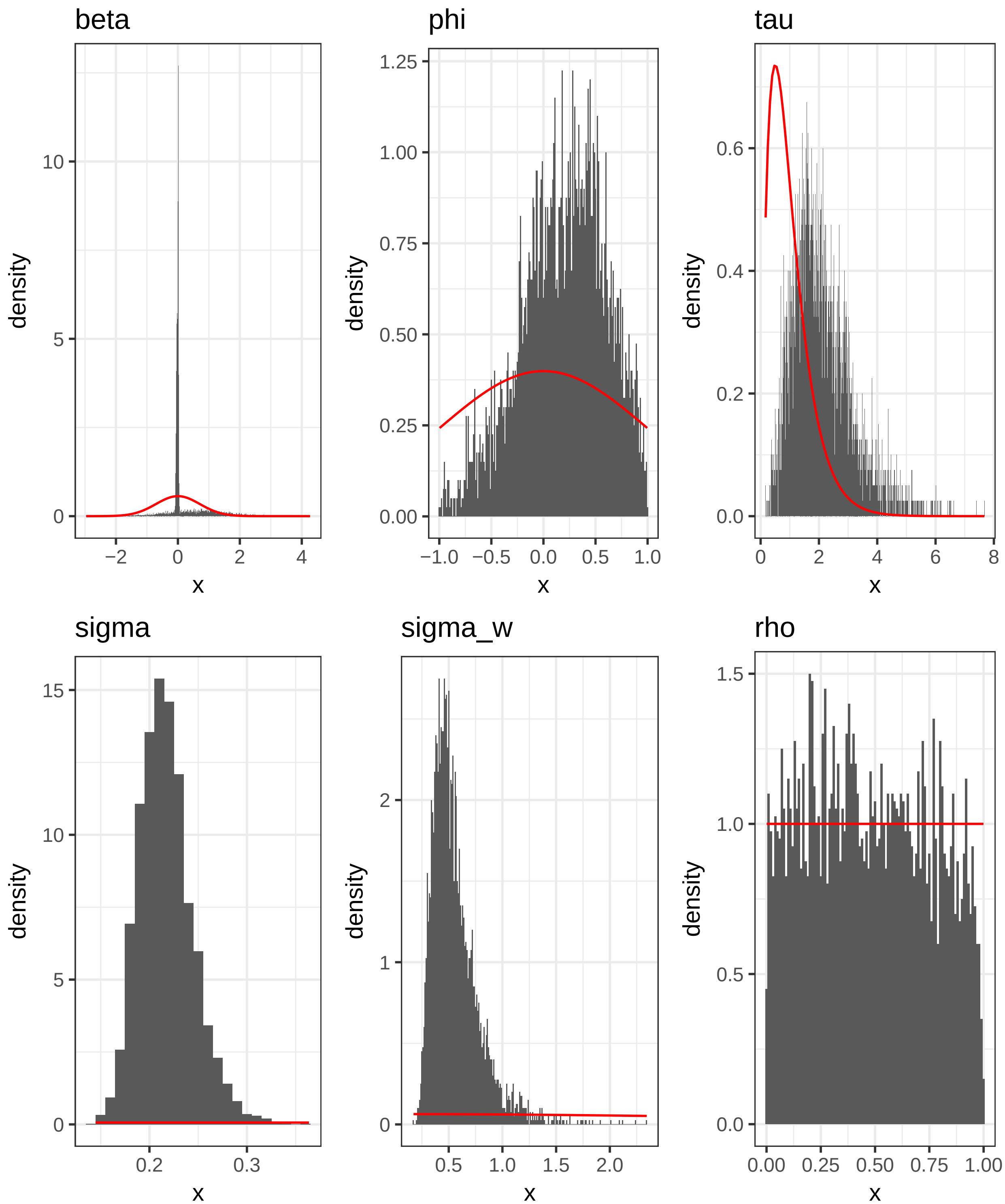

Figure 8: Comparison of histograms of posterior draws of parameters $\beta, \phi, \tau, \sigma, \sigma_\omega$, and ρ with respective curves of their prior distributions shown (in red) in each graph.

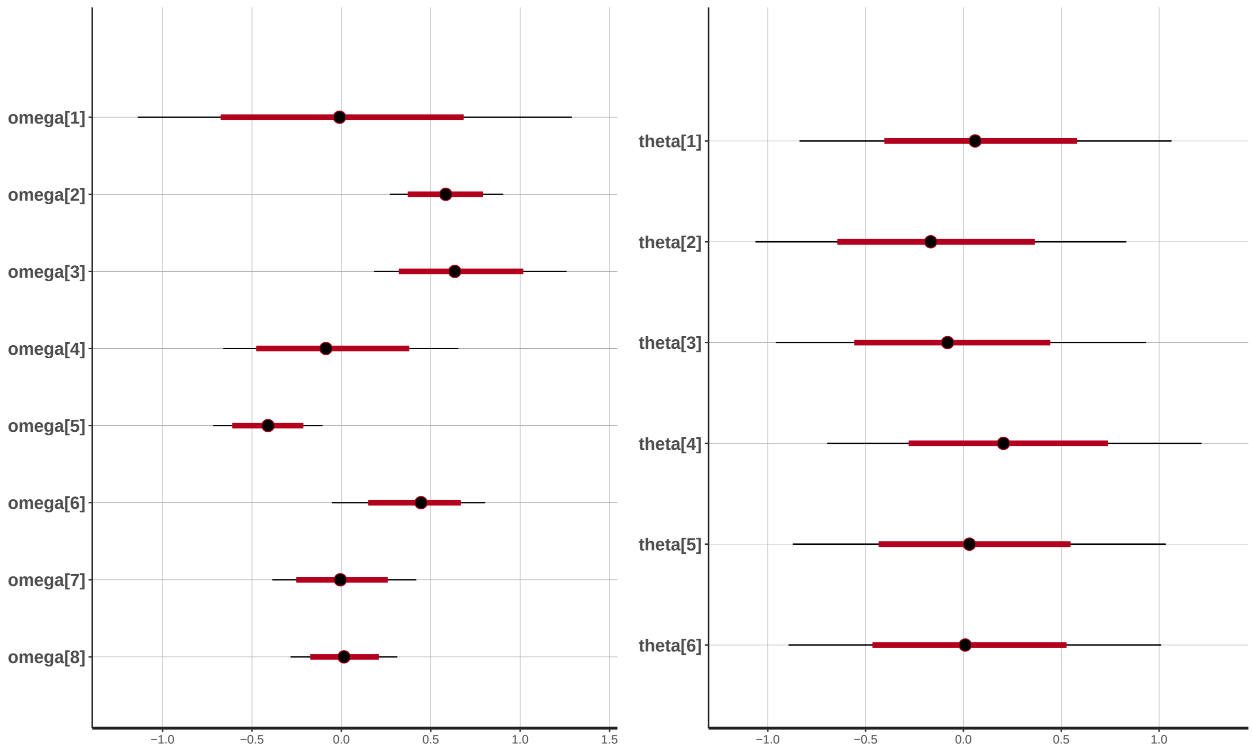

Figure 9: 95% Posterior Credible Intervals of temporal paramter, ω_t where t={1,2,3,4,5,6,7,8} corresponding to dates 06/{06,09,12,15,18,21,24,27}/2017; and 95% Posterior Credible Intervals of spatial parameter, θ_c where c={1,2,3,4,5,6} corresponding to six counties in eastern Iowa(Clinton, Delaware, Johnson, Linn, Muscatine, and Scott).

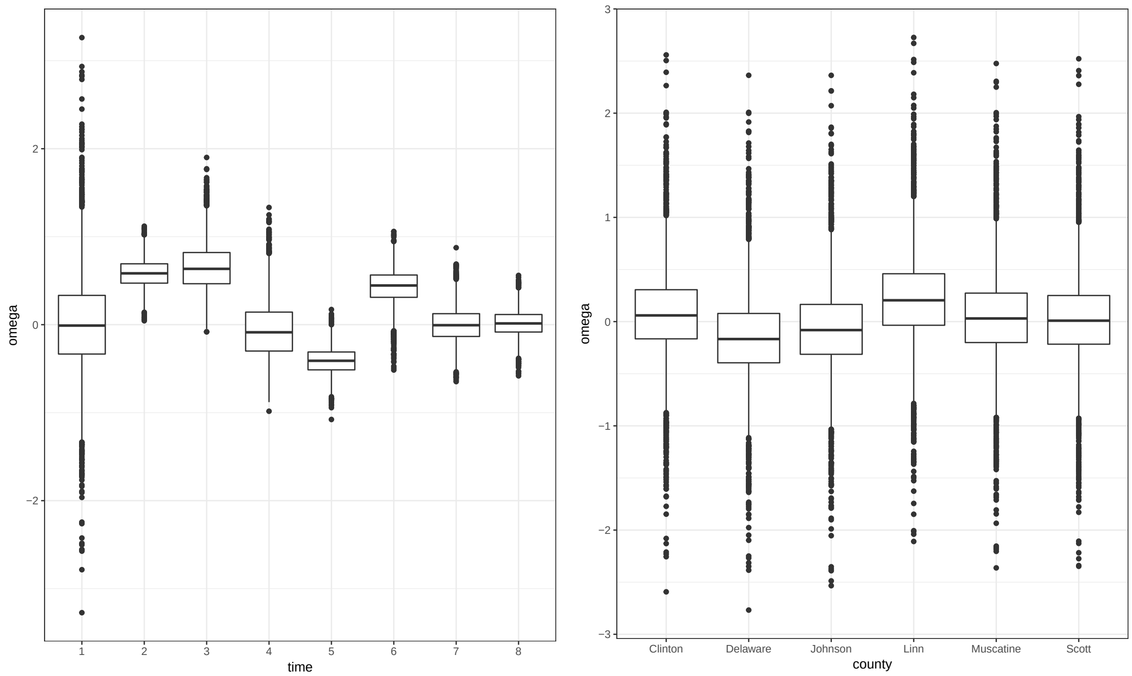

Figure 10: The box plot to the left is a comparison of posterior MCMC samples of temporal paramter, ω_t where t={1,2,3,4,5,6,7,8} corresponding to dates 06/{06,09,12,15,18,21,24,27}/2017. The box plot to the right is a comparison of posterior MCMC samples of spatial parameter, θ_c where c={1,2,3,4,5,6} corresponding to counties in Eastern Iowa(Clinton, Delaware, Johnson, Linn, Muscatine, and Scott).

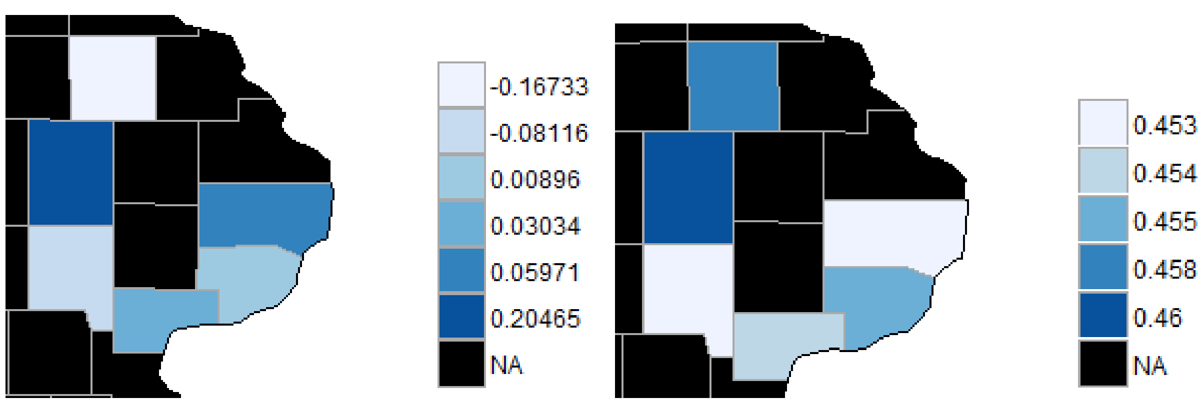

Figure 11: The county map of Eastern Iowa to the left shows (based off of the counties' color) the posterior median of θ_c, the spatial parameter for each corresponding county. The county map of Eastern Iowa to the right shows (based off of the counties' color) the posterior standard deviation of θ_c, the spatial parameter for each corresponding county.

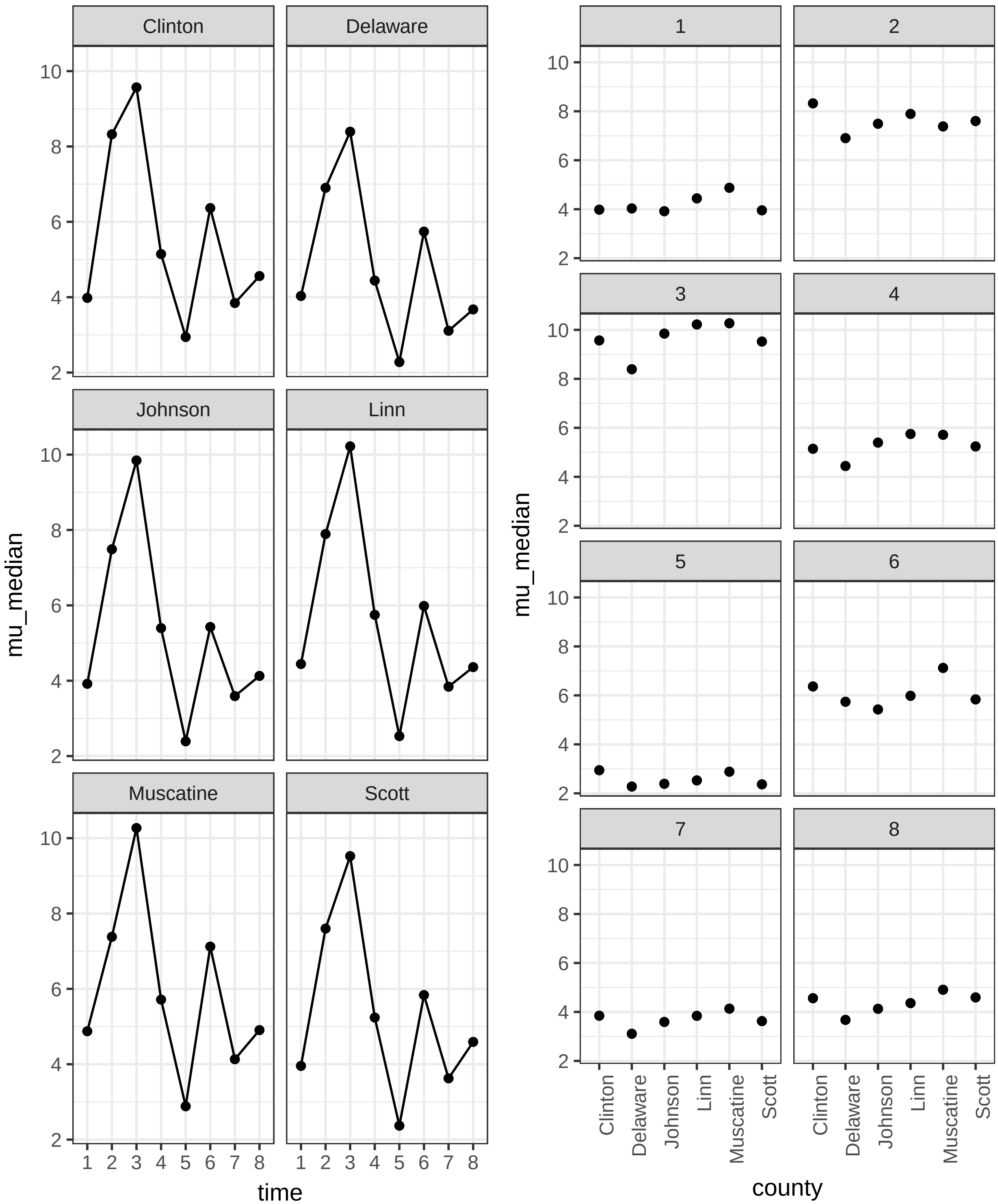

Figure 12: Plot of posterior median of mean paramter, μ_{ct}(mu_median) at time points t={1,2,3,4,5,6,7,8} corresponding to dates 06/{06,09,12,15,18,21,24,27}/2017 grouped by counties c={Clinton, Delaware, Johnson, Linn, Muscatine, Scott}; and plot of posterior median of mean paramter, μ_{ct} at county, c grouped by time, t.

B Tables

Table 1: Description of meta-data of samples used for $PM_{2.5}$ study. ToxicSites is not used by the model-only for evaluation.

Column Metadata	Description
County	Name of the Iowa County
Time	Day the pollutant was measured
pm25	Average $PM_{2.5}$ in micrograms/cubic meter
tmax	Maximum temperature for this day
wind	Average wind-speed during this day
ToxicSites	Number of toxic sites in the county

Table 2: Posterior median and credible interval of parameters $\beta_p, \sigma, \omega_t, \theta_c, \phi, \sigma_\omega, \tau$, and ρ where p={1,2,3},t={1,2,3,4,5,6,7,8} corresponding to dates 06/{06,09,12,15,18,21,24,27}/2017 and c={1,2,3,4,5,6} corresponding to counties in Eastern Iowa(Clinton, Delaware, Johnson, Linn, Muscatine, Scott)

	Posterior Median	Posterior 95% Credible Interval
β_1	0.63	[-1.3, 2.48]
β_2	0.01	[-0.01, 0.03]
β_3	-0.03	[-0.07, 0.02]
σ	0.21	[0.17, 0.28]
ω_1	-0.01	[-1.14, 1.29]
ω_2	0.58	[0.27, 0.91]
ω_3	0.63	[0.18, 1.26]
ω_4	-0.09	[-0.66, 0.65]
ω_5	-0.41	[-0.72, -0.11]
ω_6	0.45	[-0.05, 0.8]
ω_7	-0.01	[-0.39, 0.42]
ω_8	0.01	[-0.28, 0.31]
θ_1	0.06	[-0.84, 1.06]
θ_2	-0.17	[-1.06, 0.83]
θ_3	-0.08	[-0.96, 0.93]
θ_4	0.2	[-0.7, 1.22]
θ_5	0.03	[-0.87, 1.03]
θ_6	0.01	[-0.89, 1.01]
ϕ	0.23	[-0.69, 0.89]
σ_w	0.51	[0.28, 1.16]
τ	1.93	[0.63, 4.63]
ρ	0.46	[0.02, 0.95]

Table 3: Posterior median and credible interval of mean parameter μ_{ct} where county, c={Clinton, Delaware, Johnson, Linn, Muscatine, Scott} and time points t={1,2,3,4,5,6,7,8} corresponding to dates 06/{06,09,12,15,18,21,24,27}/2017

	Posterior Standard Deviation	**Posterior Median**	**95% Posterior Credible Interval**
$\mu_{[Clinton,1]}$	0.13	1.38	[1.13, 1.64]
$\mu_{[Clinton,2]}$	0.12	2.12	[1.88, 2.35]
$\mu_{[Clinton,3]}$	0.12	2.26	[2.02, 2.49]
$\mu_{[Clinton,4]}$	0.13	1.64	[1.38, 1.91]
$\mu_{[Clinton,5]}$	0.16	1.08	[0.76, 1.38]
$\mu_{[Clinton,6]}$	0.12	1.85	[1.62, 2.08]
$\mu_{[Clinton,7]}$	0.12	1.35	[1.12, 1.58]
$\mu_{[Clinton,8]}$	0.11	1.52	[1.3, 1.73]
$\mu_{[Delaware,1]}$	0.12	1.39	[1.15, 1.63]
$\mu_{[Delaware,2]}$	0.12	1.93	[1.68, 2.15]
$\mu_{[Delaware,3]}$	0.12	2.13	[1.9, 2.36]
$\mu_{[Delaware,4]}$	0.12	1.49	[1.25, 1.73]
$\mu_{[Delaware,5]}$	0.12	0.82	[0.59, 1.07]
$\mu_{[Delaware,6]}$	0.12	1.75	[1.51, 1.99]
$\mu_{[Delaware,7]}$	0.12	1.13	[0.89, 1.37]
$\mu_{[Delaware,8]}$	0.12	1.3	[1.05, 1.54]
$\mu_{[Johnson,1]}$	0.11	1.37	[1.14, 1.59]
$\mu_{[Johnson,2]}$	0.11	2.01	[1.78, 2.24]
$\mu_{[Johnson,3]}$	0.12	2.29	[2.04, 2.52]
$\mu_{[Johnson,4]}$	0.12	1.69	[1.44, 1.92]
$\mu_{[Johnson,5]}$	0.12	0.87	[0.64, 1.11]
$\mu_{[Johnson,6]}$	0.12	1.69	[1.45, 1.93]
$\mu_{[Johnson,7]}$	0.12	1.28	[1.05, 1.5]
$\mu_{[Johnson,8]}$	0.12	1.42	[1.19, 1.65]
$\mu_{[Linn,1]}$	0.11	1.49	[1.26, 1.72]
$\mu_{[Linn,2]}$	0.11	2.07	[1.85, 2.29]
$\mu_{[Linn,3]}$	0.11	2.32	[2.1, 2.55]
$\mu_{[Linn,4]}$	0.12	1.75	[1.52, 1.99]
$\mu_{[Linn,5]}$	0.12	0.93	[0.69, 1.17]
$\mu_{[Linn,6]}$	0.11	1.79	[1.57, 2.01]
$\mu_{[Linn,7]}$	0.11	1.35	[1.13, 1.57]
$\mu_{[Linn,8]}$	0.11	1.47	[1.24, 1.69]
$\mu_{[Muscatine,1]}$	0.12	1.58	[1.36, 1.82]
$\mu_{[Muscatine,2]}$	0.14	2	[1.72, 2.27]
$\mu_{[Muscatine,3]}$	0.12	2.33	[2.09, 2.57]
$\mu_{[Muscatine,4]}$	0.12	1.74	[1.5, 1.98]

Continued on next page

Table 3 – continued from previous page

	Posterior Standard Deviation	Posterior Median	95% Posterior Credible Interval
$\mu_{[Muscatine,5]}$	0.12	1.06	[0.83, 1.31]
$\mu_{[Muscatine,6]}$	0.13	1.96	[1.7, 2.23]
$\mu_{[Muscatine,7]}$	0.13	1.42	[1.16, 1.67]
$\mu_{[Muscatine,8]}$	0.12	1.59	[1.35, 1.83]
$\mu_{[Scott,1]}$	0.11	1.37	[1.15, 1.6]
$\mu_{[Scott,2]}$	0.12	2.03	[1.79, 2.26]
$\mu_{[Scott,3]}$	0.11	2.25	[2.02, 2.47]
$\mu_{[Scott,4]}$	0.12	1.66	[1.43, 1.88]
$\mu_{[Scott,5]}$	0.12	0.86	[0.62, 1.1]
$\mu_{[Scott,6]}$	0.12	1.76	[1.54, 1.99]
$\mu_{[Scott,7]}$	0.12	1.29	[1.05, 1.52]
$\mu_{[Scott,8]}$	0.12	1.52	[1.29, 1.76]