Modeling Flu Probabilities in the United States

Vadim von Brzeski
AMS 245

March 20, 2013

Abstract

In this paper we construct a model for the probability of the flu index being above a certain value for a given state in the United States at a given point in time. We find that this probability has a periodic temporal component as well as a spatial component in the form of local neighborhoods.

1 Introduction and Data Exploration

In this paper we construct a model for the probability of the flu index being above a certain value for a given state in the United States at a given point in time. The data source we work with is the Google Flu Trends data, available at www.google.com/flutrends. This dataset consists of values of the "flu index" (a positive integer value proportional to the number of cases of flu reported) for each state in the United States, including the District of Columbia, for a total of 494 weeks starting from September 28, 2003, to March 10, 2013. Not every combination of (week,state), i.e. $Y(t, s)$, has a recorded value; 1309 (5.4%) values are missing, and those are recorded as -1 in the data.

We transform the dataset raw Y as follows:

- Since we will ultimately be building a spatial model, we only consider the contiguous states, i.e. we remove Alaska and Hawaii from the dataset.
- We deal with the missing values in a very simple way : we average over the states's neighbors; i.e. if $Y(t, s)$ is missing, we set $Y(t, s) \leftarrow \sum_{s'} Y(t, s')/k$, where s' iterates over the "neighbors" of state s. If all neighbors happen to also have missing values, we average over all the states for the given time period. We (manually) construct the neighbors of a state s by looking for states s' that share a border with state s, even if that border is just a "corner" as in the case of Colorado and Arizona. Following this scheme, on average, each state has 4.4 neighbors. A sample of our neighborhood map is shown in Table 1.
- Finally, we transform the real values into binary value as follows : if $Y(t, s) > 7500$, then $Y(t, s) \leftarrow 1$; else $Y(t, s) \leftarrow 0$.

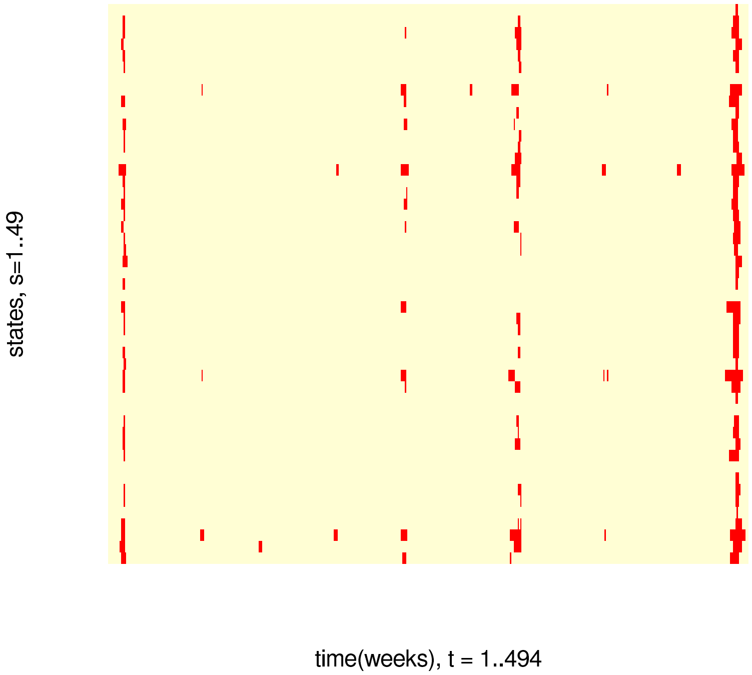
We are thus left with a data matrix consisting of $T = 494$ rows and $S = 49$ columns, with a total of $S * T = 24206$ entries, of which only 472 (1.9%) are 1, and the rest are 0.

An image of our data is shown in Figure 1 below. As we can see, there are weeks during which many, if not most, of the states have value of 1, i.e. a flu index above 7500, specifically the weeks of 2003-12-21 ($t = 13$), 2009-10-18 ($t = 317$), and more recently 2013-01-06 ($t = 485$), 2013-01-13 ($t = 486$), where we have used the date format : YYYY-MM-DD.

Table 1: Sample neighborhoods for four states.

State	Neighbors
Alabama	Mississippi,Georgia,Tennessee,Florida
Arkansas	Louisiana,Texas,Oklahoma,Missouri,Tennessee,Mississippi
California	Oregon,Nevada,Arizona
Colorado	New Mexico,Arizona,Utah,Wyoming,Nebraska,Kansas,Oklahoma

Figure 1: Binary data set organized into a matrix Y of T weeks by S states; Y^T is shown below. Red spots correspond to $Y(t,s) = 1$.



2 Model Definition

We construct a spatial model for the binary variable $Y(t,s)$ that includes known covariates (including time), spatially structured variability, unstructured variability and uses a link function based on the distribution of a Student random variable with fixed number of degrees of freedom; in our case we set the d.f. $\nu = 4$ to get a distribution with heavier tails. The model is specified below:

$$\begin{aligned} y_{t,s}|\pi_{t,s} &\overset{\text{iid}}{\sim} \mathit{Bernoulli}(\pi_{t,s}) \quad t = 1..T, \quad s = 1..S \\ F_\nu^{-1}(\pi_{t,s}) &= \beta^T \mathbf{x}_{t,s} + \gamma_s \quad = \omega_{t,s} \\ \gamma_s &= z_s + \epsilon_s \\ \epsilon &\sim N(\mathbf{0}, \frac{1}{\tau_\epsilon}\mathbf{I}) \\ \mathbf{z} &\sim N(\mathbf{0}, \frac{1}{\tau_z}\mathbf{W}^{-1}) \end{aligned}$$

In the above model, F_ν represents the cumulative distribution function of a t_ν random variable. $\mathbf{x}_{t,s}$ is a vector of the following covariates:

- $\mathbf{x}_{t,s}[1] = 1$ (intercept term)
- $\mathbf{x}_{t,s}[2] =$ (Year - 2002), i.e 1..11, to capture the long term trend
- $\mathbf{x}_{t,s}[3] =$ cosine (((Month - 1)*30.5 + Day) / 365), i.e. the cosine of the time into the year. This way we have a periodic covariate that reflects that week 52 of the year is right next to week 1 of the following year. January 1 would map to a value of ≈ 1; July 1 to a value of ≈ -1.
- $\mathbf{x}_{t,s}[4] =$ population density per sq. mile from http://simple.wikipedia.org/wiki/List_of_U.S._states_by_population_density

We can envision adding other covariates such as GDP of each state, number of doctors per capita, and so on, but for this paper we will stick to the above three.

In the above model, γ_s represents that spatial variability and is broken down into two components : structured variability in the form of $\mathbf{z}$ and unstructured variability (noise) in the form of ϵ.

We model $\mathbf{z}$ as a locally linear Gaussian Markov Random Field (GMRF), where the conditional distribution of each z_s given all other z_{-s} values depends only on the (local) neighborhood of (state) s, where the neighborhood δ_s is defined as above. Specifically, we have:

$$p(z_s|z_{-s}) \propto \exp\Big[-\frac{\tau_z}{2} \sum_{s' \in \delta_s} (z_s - z_{s'})^2 \Big]$$

This corresponds to the following precision matrix $\mathbf{W}$ in our model above :

$$W_{ij} = \begin{cases} n_i, & \text{if } i = j \\ -1, & \text{if } i \in \delta_j \\ 0, & \text{otherwise} \end{cases}$$

3 Model Fitting

Having defined the model, we now fit it using a combination of Gibbs and MCMC sampling. In order to do this, we first compute the full conditional distributions of all parameters and latent variables. The full joint posterior distribution of parameters and latent variable is as follows :

$$\begin{aligned} p(\beta, \tau_z, \tau_\epsilon, \gamma, \mathbf{z}|\mathbf{Y}) \propto & \prod_{s=1}^{S} \prod_{t=1}^{T} \Big[F_\nu(\beta^T \mathbf{x}_{t,s} + \gamma_s)\Big]^{y_{t,s}} \Big[1 - F_\nu(\beta^T \mathbf{x}_{t,s} + \gamma_s)\Big]^{1-y_{t,s}} \\ & \prod_{s=1}^{S} N(\gamma_s; z_s, \frac{1}{\tau_\epsilon}) \ \prod_{s=1}^{S} p(z_s|z_{-s}) \\ & N(\beta; \mathbf{0}, b_\beta \mathbf{I}) \ \ Gamma(\tau_z; a_z, b_z) \ \ Gamma(\tau_\epsilon; a_\epsilon, b_\epsilon) \end{aligned}$$

where the hyperparameters (of the prior distributions) b_β, $a_z, b_z, a_\epsilon, b_\epsilon$ were fixed.

The full conditionals (where available in closed form) are as follows :

$$\begin{aligned} p(\tau_\epsilon|\beta, \tau_z, \gamma, \mathbf{z}, \mathbf{Y}) &\propto Gamma(\tau_\epsilon; a_\epsilon, b_\epsilon) \ \prod_{s=1}^{S} N(\gamma_s; z_s, \frac{1}{\tau_\epsilon}) \\ &= Gamma(\tau_\epsilon; a_\epsilon + S/2, b_\epsilon + \sum_{s=1}^{S} (\gamma_s - z_s)^2/2) \end{aligned}$$

$$\begin{aligned} p(\tau_z|\beta, \tau_\epsilon, \gamma, \mathbf{z}, \mathbf{Y}) &\propto Gamma(\tau_z; a_z, b_z) \ \prod_{s=1}^{S} \tau_z^{1/2} \exp\Big[-\frac{\tau_z}{2} \sum_{s' \in \delta_s} (z_s - z_{s'})^2\Big] \\ &= Gamma(\tau_z; a_z + S/2, b_z + \sum_{s' \in \delta_s} (z_s - z_{s'})^2/2) \end{aligned}$$

$$\begin{aligned} p(z_s|z_{-s}, \beta, \tau_z, \tau_\epsilon, \gamma, , \mathbf{Y}) &\propto N(\gamma_s; z_s, \frac{1}{\tau_\epsilon}) \ \ exp\Big[-\frac{\tau_z}{2} \sum_{s' \in \delta_s} (z_s - z_{s'})^2\Big] \\ &= N\Big(\frac{\gamma_s \tau_\epsilon + k\tau_z \bar{z}_s}{\tau_\epsilon + k\tau_z}, \frac{1}{\tau_\epsilon + k\tau_z}\Big) \ , \quad k = |\delta_s| \end{aligned}$$

The remaining full conditionals are not available in closed form and must be sampled in a MCMC scheme :

$$p(\beta|\tau_z, \tau_\epsilon, \gamma, \mathbf{z}, \mathbf{Y}) \propto N(\beta; \mathbf{0}, b_\beta \mathbf{I}) \prod_{s=1}^{S} \prod_{t=1}^{T} \Big[F_\nu(\beta^T \mathbf{x}_{t,s} + \gamma_s)\Big]^{y_{t,s}} \Big[1 - F_\nu(\beta^T \mathbf{x}_{t,s} + \gamma_s)\Big]^{1-y_{t,s}}$$

$$p(\gamma_s|\beta, \tau_z, \tau_\epsilon, \mathbf{z}, \mathbf{Y}) \propto N(\gamma_s; z_s, \frac{1}{\tau_\epsilon}) \prod_{t=1}^{T} \Big[F_\nu(\beta^T \mathbf{x}_{t,s} + \gamma_s)\Big]^{y_{t,s}} \Big[1 - F_\nu(\beta^T \mathbf{x}_{t,s} + \gamma_s)\Big]^{1-y_{t,s}}$$

Finally, we would like to acquire samples of $\omega_{t,s}$ so we can get samples of the final probalities $\pi_{t,s}$:

$$\omega_{t,s} \sim \begin{cases} t_\nu(\beta^T \mathbf{x}_{t,s} + \gamma_s, 1)\mathbf{1}_{(-\infty,0]}, & \text{if } y_{t,s} = 0 \\ t_\nu(\beta^T \mathbf{x}_{t,s} + \gamma_s, 1)\mathbf{1}_{[0,\infty)}, & \text{if } y_{t,s} = 1 \end{cases}$$

i.e. truncated t_ν distributions centered at $\beta^T \mathbf{x}_{t,s} + \gamma_s$.

We found that the sampling performance was not too sensitive to the initial values of τ_z or τ_ϵ (or their hyperparameters for that matter), but was very sensitive to the initial values of β (due to the role of the proposal distribution in the MCMC step for β). Therefore we performed one extra step to set the initial values for β and to determine a decent variance for the (Normal) proposal distribution. We ran a quick logistic regression step where we modeled $logit(\pi_{t,s}) = \beta^T \mathbf{x}_{t,s}$, ignoring any spatial effect. We then used the estimates of the coefficients to set the initial value of β, and the variances in the proposal distribution for β were multiples of the standard errors of the coefficient estimates in the regression. This extra step proved crucial to achieving reasonable sampling runs.

As a last note, model fitting this way proved very computationally (time) intensive due to the number of iterations over the data that are required. After the initial code was working in R, we ported the R code into C++, relying on the GNU Scientific Library for matrix computation and random number generation. (The C++ code ran approximately 100 times faster.)

4 Results

The simulation results are shown in the figures below. Figure 2 shows the posterior distribution estimates for τ_z and τ_ϵ, and figure 3 shows the posterior estimate of β. Looking at the estimate of $\beta[Year]$, and the fact that its 95% interval lies above 0, we see that there is a long term increasing trend in the flu index over the years (higher values of $\beta[Year]$ lead to higher values of ω which lead to higher probalities π). Also, looking at $\beta[cos(t/T)]$, we see that larger values of t (winter weeks), drive higher probabilities for the flu index being above 7500.

The periodic nature of the flu index is shown in figure 4, where we plot the mean and interval estimates for ω for one state, California, as well as in figure 6. Figure 6 shows the probability of the flu index being above 7500 as a function of time (week) for four sample states. Some states like Alabama and Kentucky have higher probabilities of crossing the flu index threshold than some others like Colorado.

Figures 5 and 7 - 10 show the spatial side of the model. Figure 5 shows how the z components (from the GMRF, driven by neighboring states) differ for two sample states, Alabama and Colorado. The flu index probability is driven higher in Alamaba due its neighbors, whereas it is driven lower in Colorado due to its neighbors.

Finally, figures 7 - 10 show the evolution of the flu index probability over four weeks in the last seven weeks of 2012. The neighbor effect is clearly visible in the spread of flu from state to state. In figure 7, only two states in the south have a high probability values, approximately 74%, and are shown nearly red. Two weeks later, figure 8, things start heating up as two more neighboring states go red. Another two weeks, figure 9, the entire southeast is red, and other states across the U.S. start getting pinkish. Finally, in the last days of 2012, figure 9, the entire eastern and southern sections of the U.S. are in a flu epidemic.

Figure 2: Posterior density estimates for τ_z and τ_ϵ. The values of the mean and the 95% interval endpoints are shown for each plot.

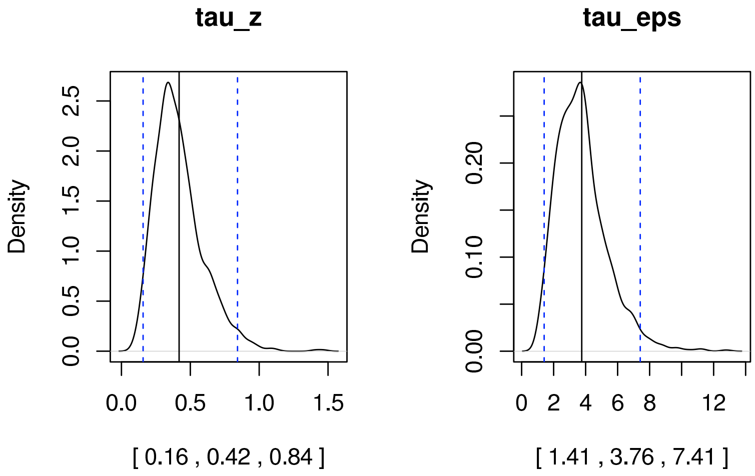
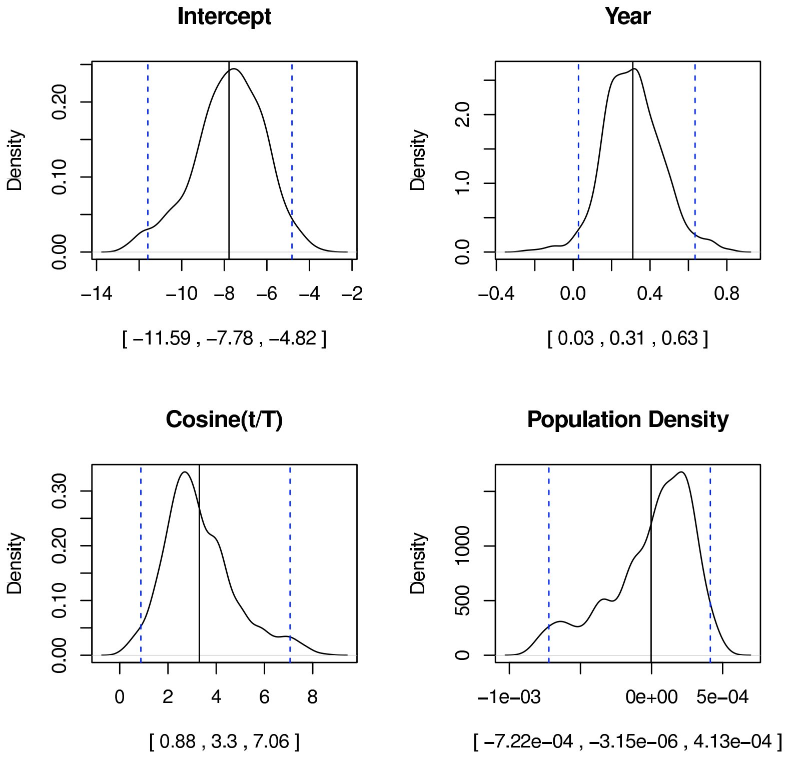


Figure 3: Posterior density estimates for β. The values of the mean and the 95% interval endpoints are shown for each plot.



5 Conclusion

In this paper we construct a model for the probability of the flu index being above a certain value for a given state in the United States at a given point in time. We find that this probability has a periodic temporal component as well as a spatial component in the form of neighborhoods of states. Our model predicts probabilities in the range of 0% to 82%, depending on the time and state, and shows how flu activity can spread from state to state over time. The remaining work to de done is to perform model validation via predictive distributions and compare some chosen statistics to the same statistics in the real data. However, we are unfortunately out of time.

Figure 4: Posterior density estimates for $\omega_{t,s}$ for s = California. The periodic nature of the model is clearly visible, as well as a long term upward trend. The spikes in the plots correspond to weeks of intensive flu activity driven by the spatial component of the model, γ_s.

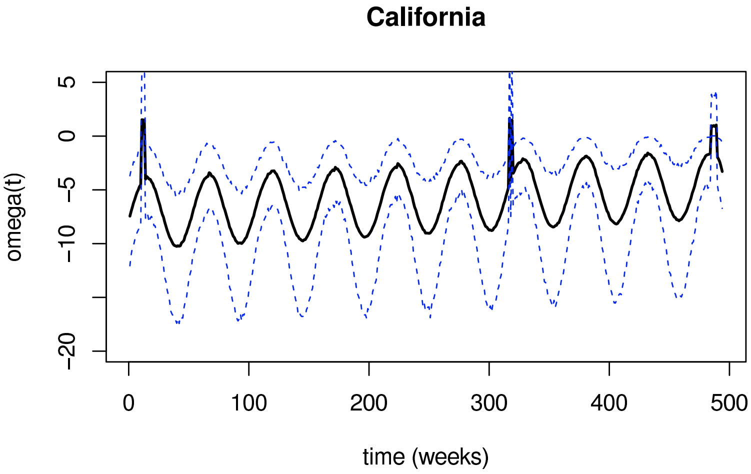


Figure 5: Posterior density estimates for a few z variables, corresponding the structured spatial effect. We can see that Alabama's flu index is more likely to be driven higher by its neighbors than Colorado's. The mean and the 95% intervals are shown below each plot.

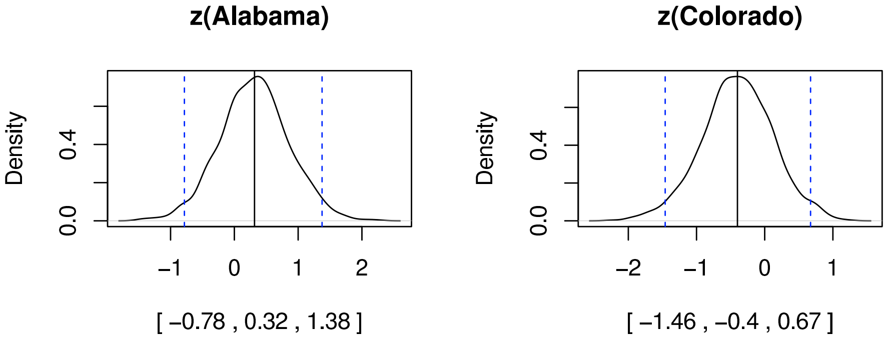


Figure 6: Probabily of flu index > 7500 ($\pi_{t,s} = F_\nu(\omega_{t,s})$) for four states s as a function of time $t = 1..494$ weeks. Black lines show the mean, blue lines show the 95% interval.

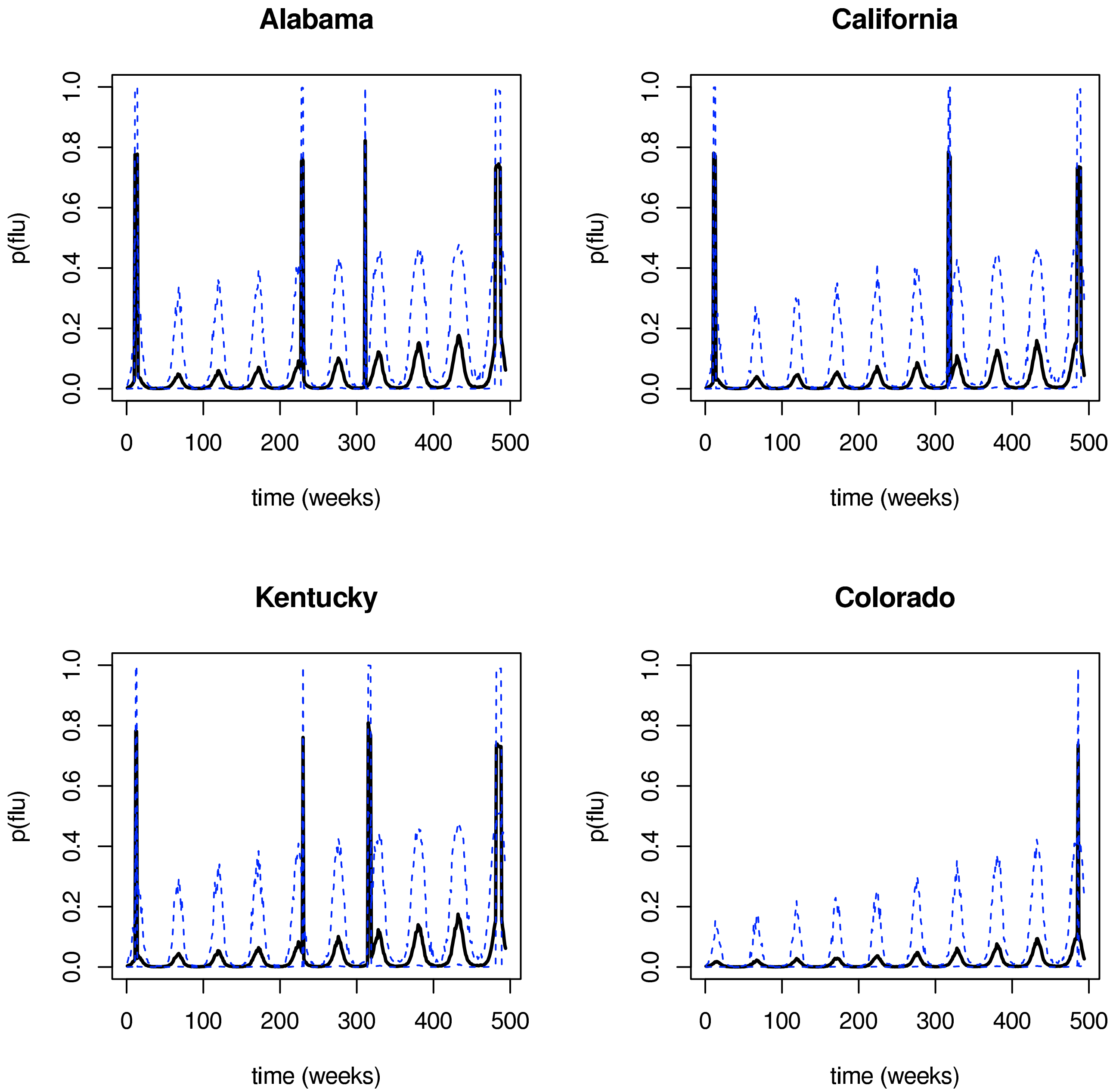


Figure 7: Probability of flu index > 7500 across the U.S. for the week of 2012-11-18.

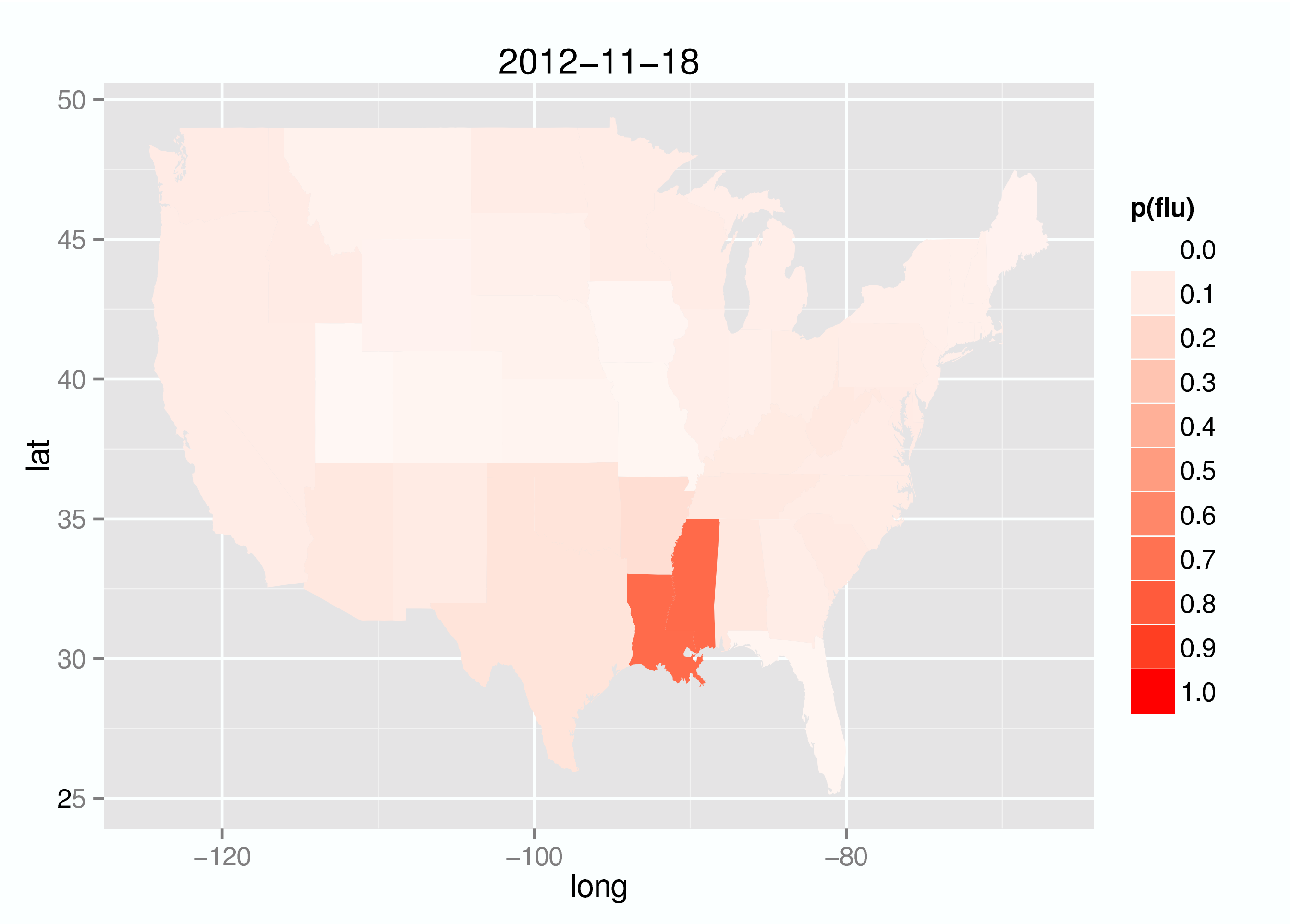


Figure 8: Probability of flu index > 7500 across the U.S. for the week of 2012-12-02.

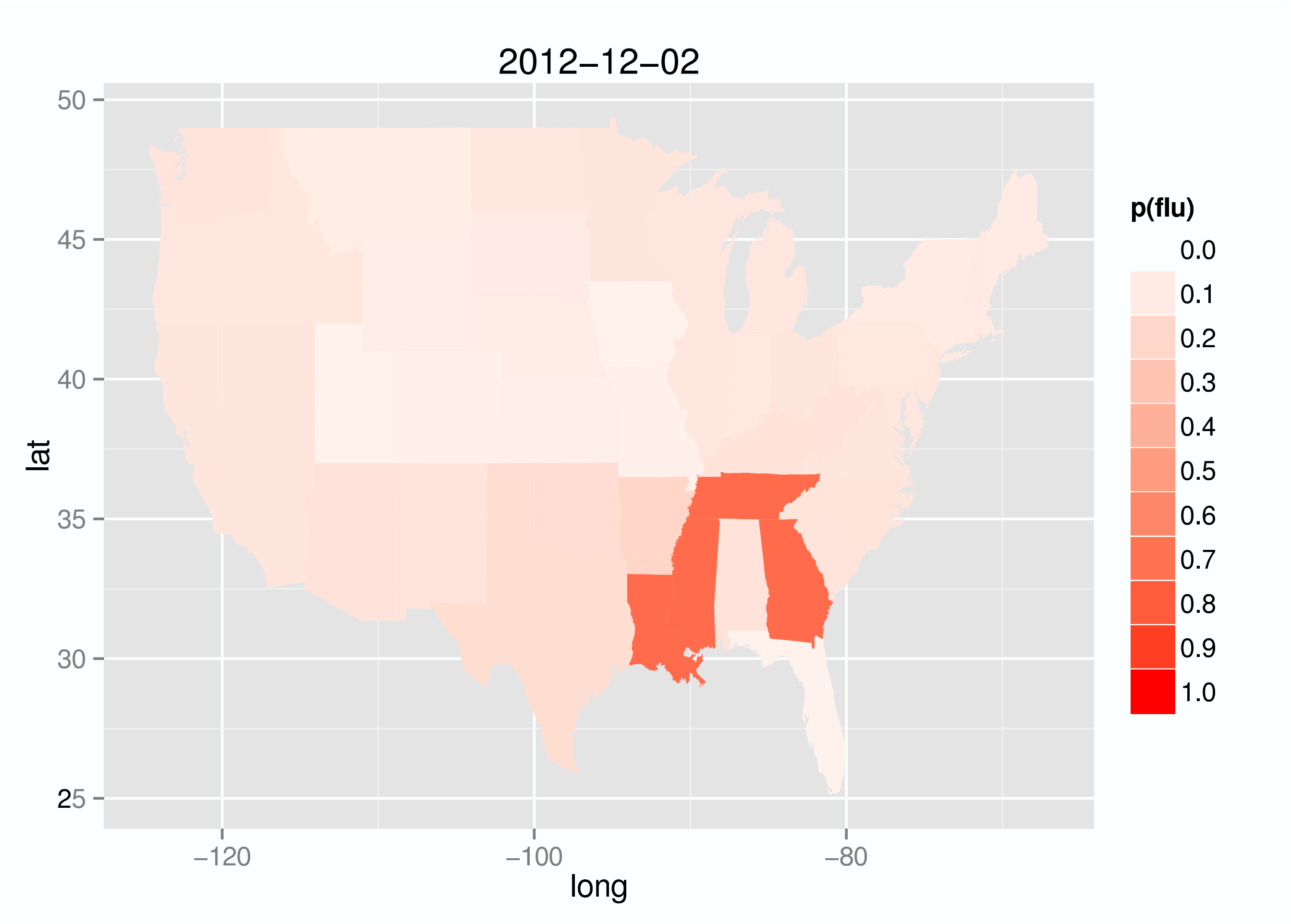


Figure 9: Probability of flu index > 7500 across the U.S. for the week of 2012-12-16.

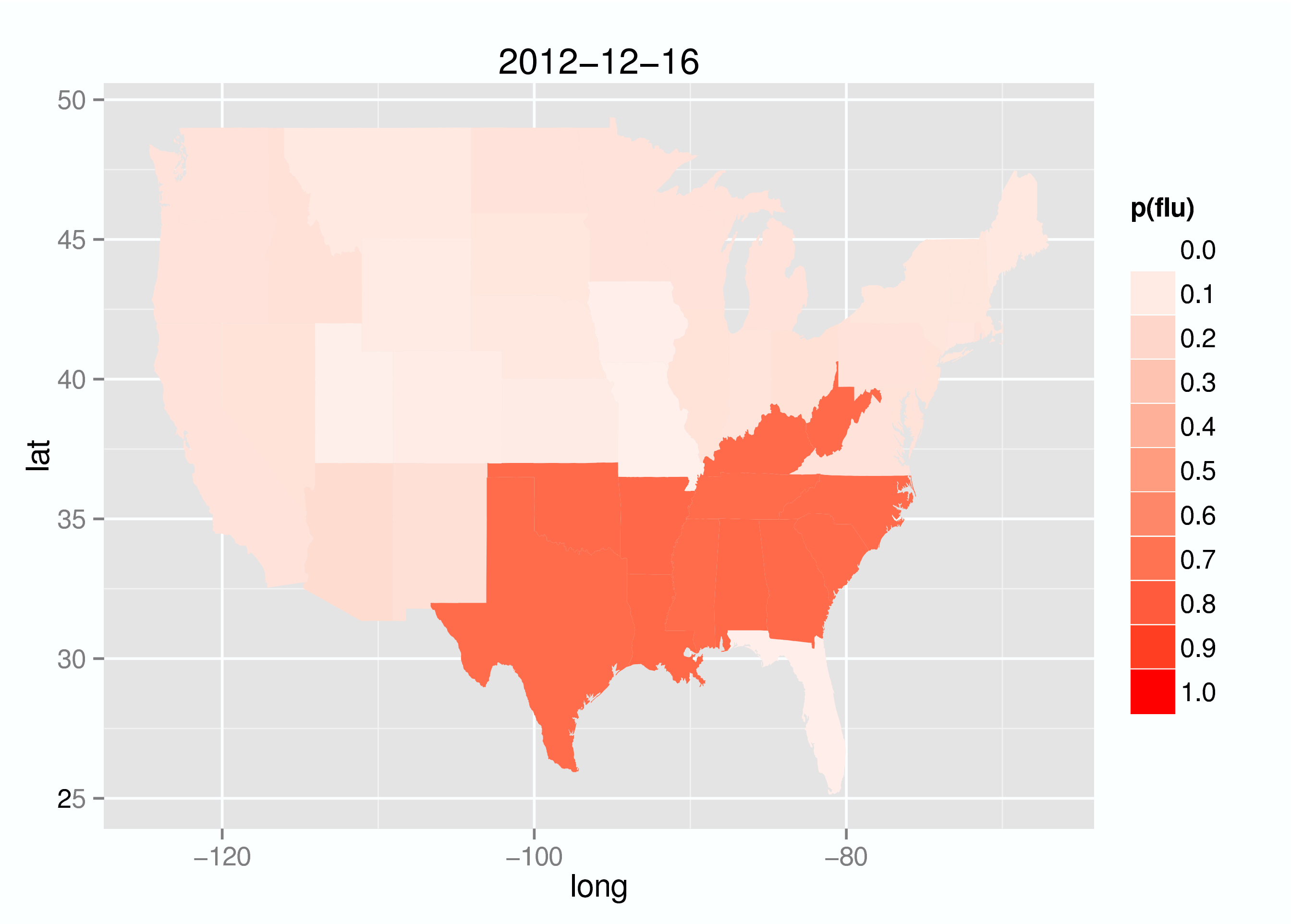
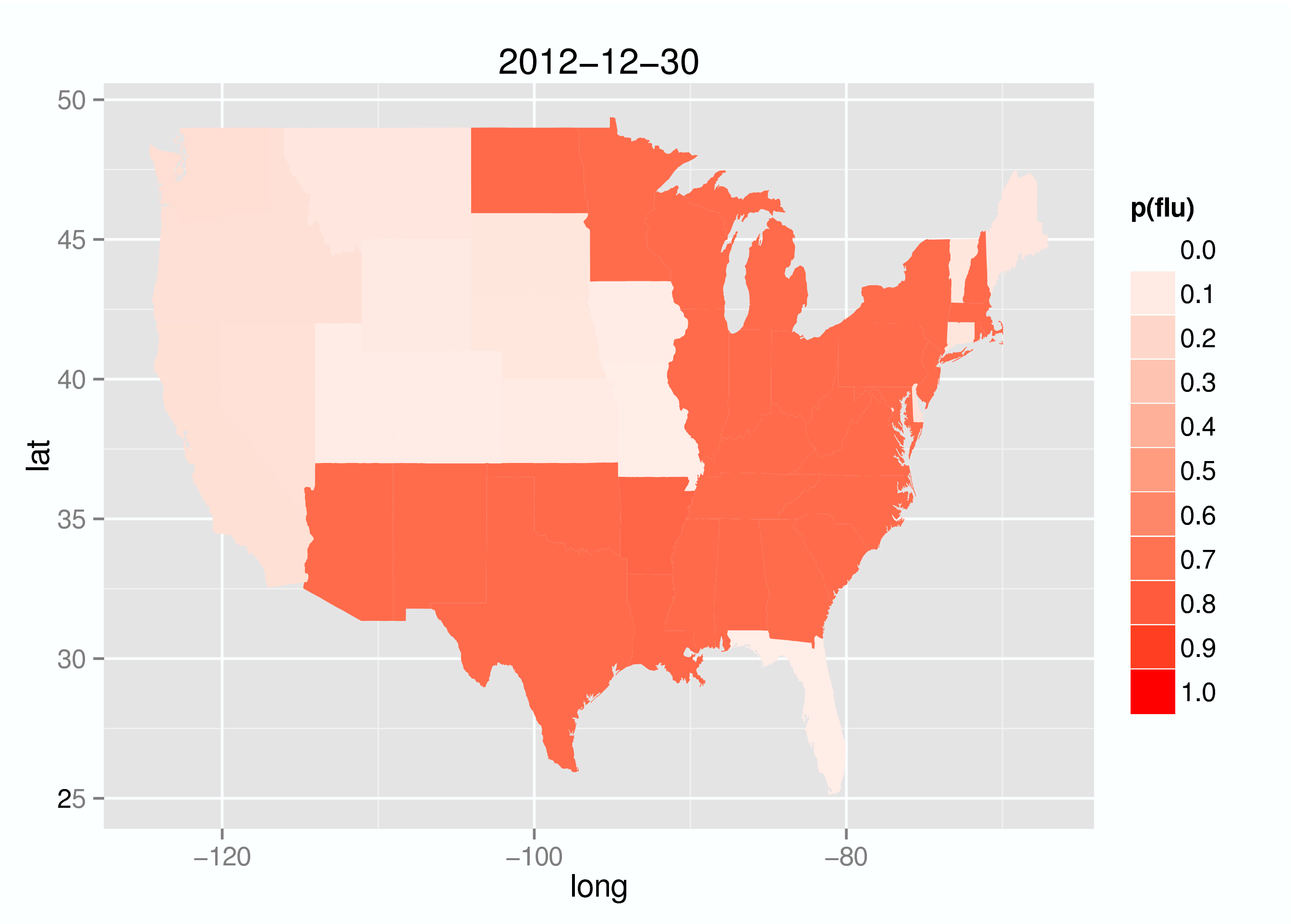


Figure 10: Probability of flu index > 7500 across the U.S. for the week of 2012-12-30.


2012−12−30
50
45
40
35
30
25
lat
−120
−100
−80
long
p(flu)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0