Prediction Fever: Modeling Influenza with Regional Effects

Shannon Gallagher
Primary Advisors: Bill Eddy and Ryan Tibshirani
External Advisor: Roni Rosenfeld

February 23, 2016

Abstract

We produce three models which incorporate regional data in order to forecast Influenza incidence in the United States that are updated as we learn more about the current season. In addition, we compare the models through Leave One Season Out Cross Validation. Although we are interested in producing predictions for each week of the remainder of the flu season, we focus on primary targets of peak incidence and peak week. The first model is an extension of an Empirical Bayes Framework described by Brooks et al. in [1], which allows for the incorporation of regional effects. The second model predicts the targets of peak week and peak height directly at the expense of producing a full curve of forecasts. The final model, which we call Posterior Biasing, combines the Empirical Bayes framework with emphasis on the targets, which creates peak incidence predictions that are approximately twice as good as those generated by Empirical Bayes alone. Moreover, we produce probabilistic forecasts through our posterior distribution and compare our Bayesian Credible Intervals with the Frequentist Coverage, which surprisingly have an empirical one-to-one correspondence.

Contents

1 Introduction

Influenza, commonly known as the flu, "is a contagious respiratory illness caused by influenza viruses [that] can cause mild to severe illness" [3]. The flu is responsible of the death of thousands of people in the US per year and in some extreme cases, can be responsible for the death of millions as was the case of Spanish Influenza in 1918-1919 [5]. Even today, the cost of the flu is high: the annual impact in the United States has been estimated at thousands of lives, 3.1 million hospitalized days, 31.4 million outpatient visits, and $87.1 billion in economic burden every year [4].

Due to the destructive and disruptive nature of the flu, we would like to better prepare for its incidence, and one such preparation is to predict certain aspects of the flu such as peak week and peak incidence or height. In addition, we emphasize the goals of the CDC contest, namely accurately predicting the peak incidence as well as the peak week.

Due to the destructive and disruptive nature of the flu, we would like to better prepare for its incidence. One such preparation is to predict certain aspects of the flu such as peak incidence and its corresponding peak week. In 2013, the U.S. Center for Diesease Control and Prevention (CDC) hosted a competition for research groups to foster awareness and research in the area of forecasting the flu. Many groups particpiated including the DELPHI group, a team of Carnegie Mellon researchers from the Computer Science and Statistics Departments, which utilized an Empirical Bayes (EB) Framework to produce point estimates and credible intervals for the above mentioned key values of the flu.

Our work extends and expands upon DELPHI's EB Framework and focuses on relaxing the assumption that CDC regions are indepdent from one another. As such we refer to our work as EB with Regional Effects and the original EB model as the base or original EB model. In addition, we emphasize the goals of the CDC contest, namely accurately predicting the peak incidence as well as the peak week. We incorporate this emphasis of predicting key targets in our model.

The rest of the paper is organized as follows. In Section 1.1, we explain in depth the data; in Section 1.2, we give evidence to the presence of regional effects of the flu; and we summarize our contributions in 1.3. We posit our methods for our three models in Section 2; we compare our results in Section 4.1; and finally we discuss the success of the model(s), future directions, and work in Section 5.2.

1.1 Data

The data is available from the CDC's website found in [2]. There are 10 regions in the United States shown in Figure 1. Since 1997, the data is released on a weekly basis for each region along with the aggregated national results. The primary quantity of interest is the percent weighted influenza like incidence or wILI. The wILI quantity is a measure of how many patients as a percentage have flu like symptoms as diagnosed by doctors and is adjusted for both changes in participation and state population [1]. We have wILI values ranging between 0 and as high as 14 for some regions.

Two other pieces of the data given by the CDC are the flu season threshold and the epidemic threshold. These values are determined by the CDC and are used to define many parts of the flu season such as the beginning and duration. For context, peak flu season is defined to be from October to May in extreme cases [2].

Other important variables include number of office visits and number of providers, which are incorporated into the wILI number. These values are determined by the CDC and are used to define many parts of the flu season such as the beginning and the duration. The data structure and variables of interest are summarized in Table 1.

Region	**Year**	**Week**	**wILI**
1	2015	25	0.55
2	2015	25	1.45
⋮	⋮	⋮	⋮
10	2015	25	0.44
⋮	⋮	⋮	⋮
1	2015	45	0.76
2	2015	45	1.57
⋮	⋮	⋮	⋮
10	2015	45	0.89

Table 1: Snapshot of the data and relevant variables. Cross Section of available data. wILI - Weighted Influenza Like Illness.

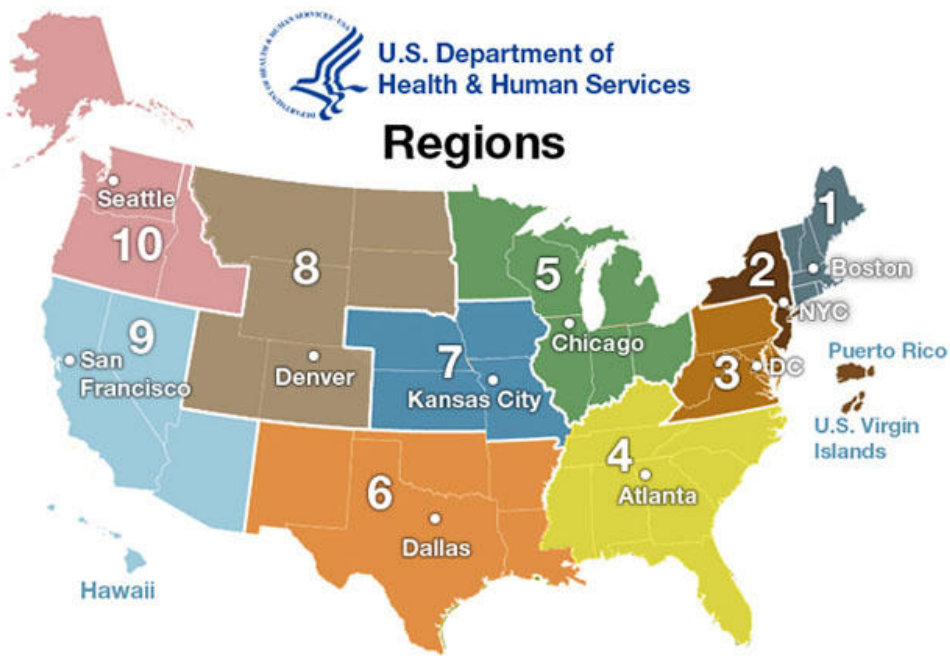
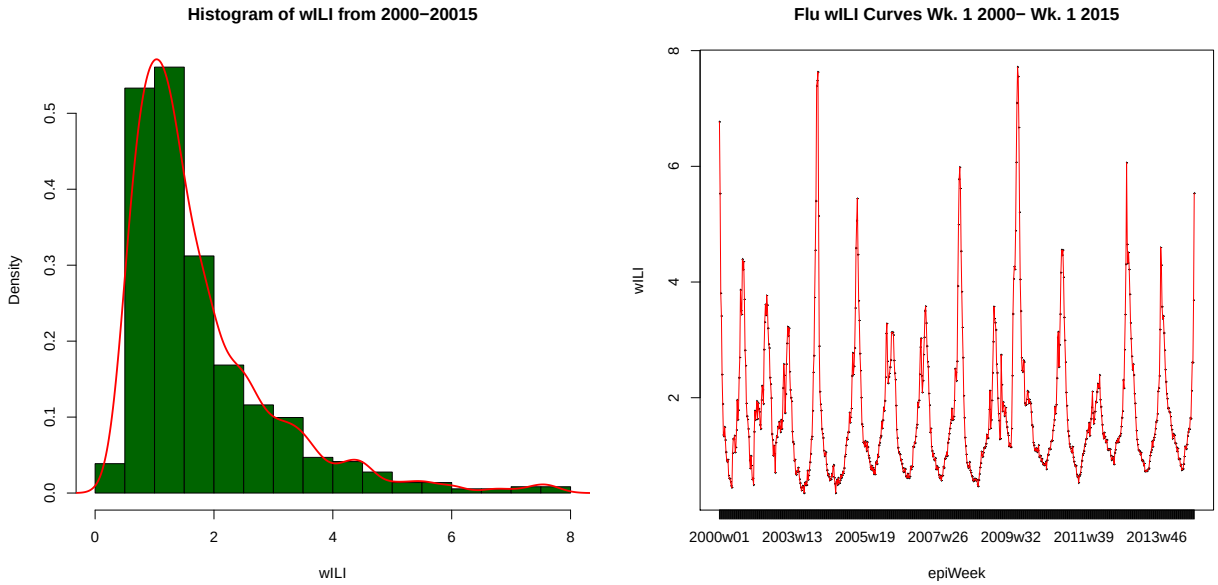


Figure 1: CDC Flu regions labeled 1-10. From `cdc.gov`.

In Figure 2a, we have a histogram of wILI values from a period ranging from Week 1, 2001 to Week 1, 2015 at the national level. We see that the most common wILI value on a week to week basis is around 1. We also note there appears to be a long, thin right tail. This tail indicates that the peak wILI value can range about anywhere



(a) (b)

Figure 2: Figure 2a: Histogram and overlayed density estimate line of wILI over a 15 year period. There seems to be a mode at 2 wILI, and the density seems to be right skewed with a long, thin tail. This skewness indicates that although high wILI values are common, they can also vary. Figure 2b: Time series of wILI from 2000-2015. In general, we see that wILI is period in nature although the peaks of the spikes can differ wildly from year to year.

from 4 to 8 which demonstrates that wILI curves can be highly variable from year to year.

Accordingly, in Figure 2b, we have the time series of wILI from 2000-2015. In the figure, we see there are periodic spikes, which correspond to the 'flu season.' However, we note that these spikes vary significantly in both height and width. For instance, we see that 2005 has a very high peak but is also very narrow. In contrast, the year 2012 has a low peak but has a long duration.

1.2 Regional Analysis

The CDC flu regions shown in Figure 1 are designed to have roughly the same amount of people per region. However, as we see in the figure, the boundaries are mostly arbitrary, and due to the nature of the flu, we do not expect the spread of the disease to be contained within the boundaries of each region. As such, we propose a model that aims to capture the interdependency among the regions.

To support the hypothesis that the regions are co-dependent, we generate multiple maps of the US for each season and week, coloring the different regions based on the wILI score. We have an example of one such image from the sequence of maps in Figure 3. In this map, the different regions are colored based on a log scale of wILI with colors closer to red representing high wILI values and colors closer to blue as low

wILI values. We truncated the log wILI values at 0 for visibility of the map. This image is representative of the regions and wILI as a whole. Although the seasons vary from year to year, there seems to be patterns among the regions within a season. For instance, Region 6 (TX+) which consists of Texas and other southern states, has the largest wILI in Figure 3. This is typical of this region to both have higher wILI than the other regions but also to begin the flu season before the others. The sequence of maps indicate that Region 6 may be a good indicator of the wILI for the other regions in the coming weeks. At the other extreme, we have regions 10 (WA+) and 1 (MA+), the Pacific Northwest and Northeast, respectively. In these regions, the flu seems to be milder than other parts of the United States.

We investigate these dependences further in Section 2.2.

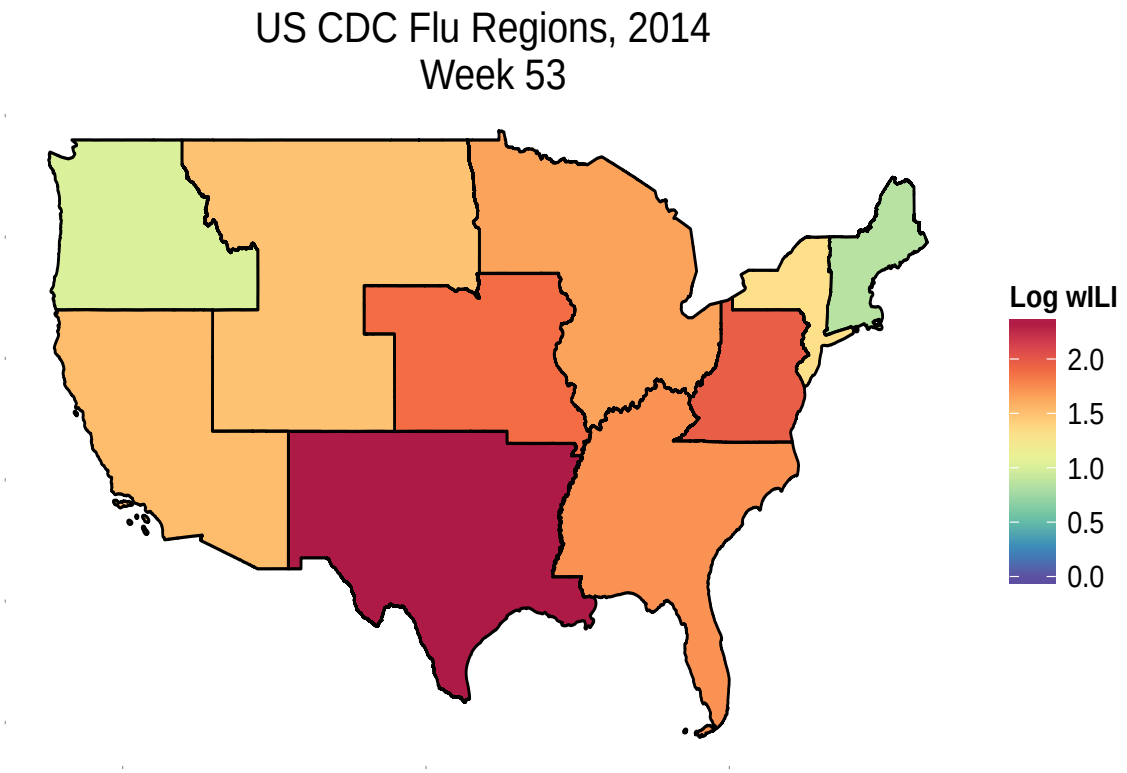


Figure 3: Colored wILI map of the United States for 2014 week 53. This week is about 2 weeks from the peak wILI for the different regions in 2014. What we see here is typical throughout the years. Region 6 in dark red is typical in this picture in both that region 6 tends to have higher wILI values than the other rergions but also seems to begin the the flu season before the other regions.

1.3 Summary of Contributions

The intuition behind the EB with regional effects model is that a flu or weekly incidence curve for the current season will look like a past season's curve with some transformations. The universe of observed incidence curves consists of smoothed splines of wILI for each season *and* each region. We extend this line of thought for the the extension of

regional effects with the added intuition that in some years the flu strikes with more intensity across all regions than in other years. Moreover, the peak weeks of the regions, although variable, tend to occur near one another. Hence, we postulate that there are seasonal parameters which tie the regions together. We describe our proposed EB with regional effects model more formally in Section 2.2.

From the EB with regional effects model we obtain weekly predictions for the current season's peak week, peak height, and flu season duration. Perhaps more importantly, we produce an entire posterior distribution for our predictions. Besides using the current season as a test of the accuracy of our predictions, we also perform cross validation on our methods to obtain mean absolute error of our model to compare to past results. Mean absolute error is used because of its use in the CDC flu forecasting contest.

Despite being interested in generating predictions via the form of curves for an entire season, the target goals consist only of peak week and peak height. In practice, it is easier to predict an estimate for a specific target rather than a curve, as is the case here. Thus, we combine a regression model for peak height along with base EB to bias our posterior towards accurate predictions for our targets. We call this method "Posterior Biasing" or PB.

Our method of posterior biasing arose from the fact that EB with regional effects did not work well in practice. EB with regional effects suffers from the 'curse of dimensionality' and we were forced to make approximations to generate forecasts. Instead, in order to generate a more sophisticated prediction season, we combined two simpler models together: base EB and target predictions for peak height and peak week. Together, the two form the method of PB, which works well in practice.

In addition to producing forecasts, we are interested in the interpretability of our predictions. Since epidemiologists are particularly interested in interpretable models and as Bayesian credible intervals can be particularly confusing towards that goal, we also analyze the frequentist properties of our generated models. We make no claim on what the properties should be, however.

2 Methods

2.1 Empirical Bayes

Empirical Bayes is an intuitive method. At a high level, we assume a future flu season will look similar to a past flu season, perhaps with some modifications. In practice, finding these modifications is difficult. Brooks et al. describe this implementation of Empirical Bayes in [1]. The steps are summarized as follows:

For each region:

1. Model past seasons' epidemic wILI values as smoothed curves added to noise
2. Create a prior of wILI curves by shifting and scaling past, smoothed flu curves
3. Form a posterior distribution of curves given we observe the first T weeks.
4. Use the posterior to predict key values and corresponding credible intervals

Their framework provides reasonably accurate predictions, in addition to assuming the following simplifying assumptions:

Empirical Bayes Assumptions:

1. The regions are independent from one another
2. The modifications appear in the form of shifting the curve in time and scaling the curve vertically.

In this paper, we focus on relaxing assumption 1, independence among the regions.

2.2 Regional Effects Extension

Extending EB to include regional effects arises from our suspicion that regions are, in fact, dependent on one another. As the CDC region boundaries are simply the union of some states' boundaries, and since the flu is transmitted through person to person contact, assuming the regions are independent from one another is suspect. We have evidence that the regions are not independent from one another as illustrated in Section 1.2.

Like the EB method for predicting the incidence of the flu, EB with regional effects (EBRE) allows for regional variations. EBRE allows past seasons' wILI curves to be both horizontally shifted and multiplicatively scaled. However, EBRE connects the regions through seasonal variables in shifting and scaling. These seasonal variables allow for flexibility in the intensity of a single season, and the regional parameters allow for regional variations. For instance, Region 6 (TX+) often has a larger wILI than the other regions. This is in accordance with prior beliefs that the flu spreads by way of the US-Mexico border [2]. By using seasonal in addition to regional parameters, we may have differing intensities and peak weeks in the flu over the years that are identified in every region while retaining region specific qualities.

The model we employ is as follows. First we scale the seasons' wILI values to lie between 0 and 10 based on the 'flu threshhold' for a given region as given by the CDC and shown in Appendix A.1. We 'pin' the values that are under the flu threshold, meaning we do not adjust these values. The pinned values are already small and if linearly scaled between 0 and 10, then the values would become distorted. Afterwards, we scale the remaining values between the flu threshold for the region and 10. In addition, we center the values around the peak week, relabeling it as week 0.

Let $r = 1, 2, \ldots, 10$ index the 10 CDC regions and $s = 1, 2, \ldots, 15$ index the 15 seasons, beginning in 2000. Let the wILI be denoted as

$$(y_1^{(r,s)}, \ldots, y_T^{(r,s)})^T \overset{iid}{\sim} (\mu_1^{(r,s)}, \ldots, \mu_T^{(r,s)})^T + N(\mathbf{0}, \sigma^2 \mathbf{I})$$

where T is the max number of weeks of the flu season, (either 52 or 53 depending on the season). Let

$$\mu_t^{(r,s)} = [a_s \cdot \alpha_r] \cdot f(t - b_s - \beta_r),$$

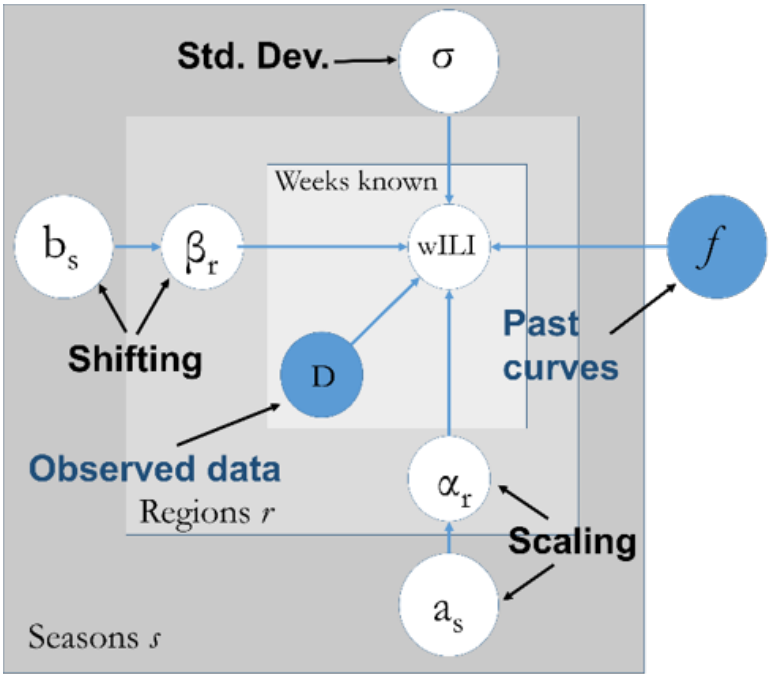


Figure 4: Plate model of EB with regional effects. We have regional and seasonal variables along with an observed universe of past curves and current wILI values in order to create a posterior of future wILI values for each region and each season.

with priors

$$
\begin{aligned}
a_s &\sim \text{Unif}(2, 10) \\
b_s &\sim \text{Unif}\{-6, -5, \ldots, 6\} \\
\alpha_r &\sim \text{Unif}(0.25, 1.25) \\
\beta_r &\sim \text{Unif}\{-3, -2, \ldots, 3\} \\
f &\sim \hat{F} \\
\sigma^2 &\sim \text{Unif}(0.5, 2.5),
\end{aligned}
$$

where $\hat{F}$ is the discrete uniform distribution over all smoothed curves over all regions and seasons, scaled between 0 and 10 as described previously. By $f(t)$, we indicate the wILI value of curve f at week t. Due to the shifting, it is possible to require a value that is less than 0 or greater than T. In this case, we set $f(t) = f(0)$ if $t < 0$ and $f(t) = f(T)$ for $t > T$. The parameter a_s is the latent peak for the fixed season. We scale this peak regionally through the regional scaling parameter α_r. We assume that the shape of the curve is uniformly distributed from the observed curves. Within the curve, we shift the time of the peak through the latent parameter b_s and again scale regionally with β_r. In words, the latent peak week depends on the season, and we adjust for time-invariant regional effects. Although σ^2 is not in the equation for $\mu_t^{(r,s)}$, it is used in the likelihood of the observed points and thus used to form the posterior of $\mu_t^{(r,s)}$ values. The model displayed as a plate model in Figure 4.

To fit a curve for a given season s_0 and week t_0, let $D_{s_0,t_0} = \{y_1^{(s_0,r)}, \ldots, y_{t_0}^{(s_0,r)}\}$ be the set of wILI values we have observed for season s_0 and up to and including week t_0.

Let $T_+ = \{t_0+1, t_0+2, \ldots, T_s\}$ be the weeks we would like to predict. The likelihood of the data D_{s_0,t_0} given the observed wILI values $y_t^{(r,s)}$ is

$$\mathcal{L}\left(D_{s_0,t_0}|y_{t_i}^{(r,s)},\ i \in T_+, r \in \{1,\ldots,10\}\right) \propto \frac{1}{\sigma}\exp\left(-\sum_{r=1}^{R}\sum_{t=1}^{t_0}\frac{\left(y_t^{(s_0,r)} - \mu_t^{(s_0,r)}\right)^2}{2\sigma^2}\right), \tag{1}$$

where T_{s_0} is the length of season s_0. Finally, as our prior consists of unform distributions, our posterior, λ, is also proportional to the likelihood,

$$\lambda\left(\mu_{t_i}^{(r,s)},\ i \in T_+, r \in \{1,\ldots,10\}|D_{s_0,t_0}\right) \propto \mathcal{L}. \tag{2}$$

To take advantage of the fact that we have previously observed data, we 'pin' the peak week and peak incidence to the observed week and incidence if the predicted peak week has already been observed. In this way, our error should tend to zero as we approach the end of the season.

We produce from our posterior both probalistic predictions for both $y_t^{(r,s)}$ and also form a posterior distribution for peak week and peak incidence. Finally, we predict those key values.

2.2.1 Implementation

In order to implement the EB model we utilize Importance Sampling as summarized by Tokdar and Kass [6]. The idea behind importance sampling as applied to our setting is as follows:

$$\begin{aligned} E_\lambda[\mu_{t,r,s}] &= \int \mu\lambda(\mu)d\mu \\ &= E_q[\omega(\mu_{t,r,s})\mu_{t,r,s}] \end{aligned}$$

where $\lambda(\mu)$ is our posterior density. However, as described in Equation 1, we do not know the normalizing constant of λ. Importance sampling says that instead of sampling from λ we can sample from $\omega(\mu) = \frac{\lambda(\mu)}{q(\mu)}$ where $q(\mu)$ is a proability density and $q(\mu) > 0$ when $y\lambda(\mu) \neq 0$. Thus we can estimate $Y_{t,r,s} = E_\lambda[\mu_{t,r,s}]$ through

$$\hat{Y}_{t,r,s} = \frac{1}{m}\sum_{j=1}^{m}\omega(\mu_{t,r,s}^{(j)})\mu_{t,r,s}^{(j)},$$

which by the strong law of large numbers converges to $Y_{t,r,s}$ and also converges to a normal density with mean $Y_{t,r,s}$ by the central limit theorem.

In practice, we estimate ω by iterating over a grid over each of the parameters in order to approximate the posterior values. Because we have to compare each set of regional variables, instead of making M^{10} comparisons where M is the combinations of regional variables α_r and β_r, we approximate by using α_r and β_r for $r = 1, \ldots, 10$ as the ones with the largest sum of weights across the seasonal variables.

2.3 EB Results

We work with 16 seasons of flu data for each of the 10 regions. We leave out the 2009 season when Swine Flu occurred because we believe it to be an outlier, a pandemic rather than an epidemic. In addition, in all of our results in this paper, we leave out Region 9 which seems not to follow the same model as the other regions.

We implement our methods using a super computer with 16 nodes and 64 cores per node which makes it possible to complete all of cross validation in the time it takes to run generate one posterior.

For EB with regional effects, we produce reasonable predictions as shown in Figure 5. Displayed in solid lines are the observed data up to and including week 22; the dashed lines are what we predict the rest of the season to be. In this figure, we see how the EB with regional effects method works. We see that the regions' predicted peak incidence are clustered around a wILI value of 8 and that the peaks are predicted to occur within a few weeks of one another, all which agree with current knowledge. Moreover, Region 6 (TX+) is predicted to have the largest wILI value, which is what generally occurs in practice. Similarly, we have Region 10 (WA+) with a much lower peak height than the others, which also seems to occur in nauture.

Although EB with regional effects does seem to agree with previously conceived notions about the different regions, when it comes to cross validation (CV), just does not compare to EB at predicting peak height and peak week, the targets posed by the CDC competition.

Like the CDC, we use mean absolute error to compare the quality of our predictions. In contrast to standard cross validation techniques, the output of our cross validation is a *curve*, with the x-axis being a time unit and the y axis being the mean absolute error. The reason for this particular output is that we expect to do better as we move further along in a season. As seasons' peak weeks do not occur in the same absolute time frame, we measure the error in weeks from the peak week. We note that although this time measurement is not possible in a new season–as we do not know the peak week– we find this CV useful to see how, on average, how our method performs.

Mathematically, the regional EB model requires fitting 2×10 more parameters, namely the regional shifting and scaling parameters, than than the base EB model. For the EB model, we have to fit a model for the fixed season and weeks known, a scaling and shifting parameter, a curve from the universe of past curves, and the standard error for the 10 regions, a total of 40 parameters. Computationally, EB with regional effects is more difficult to perform as we need to calculate all the different combinations of the regional parameters, which is required for the likelihood. These computations for the regional combinations alone are on the order of magnitude of $O(n^{10})$, where n is the max number of points tried in a parameter for importance sampling. This is infeasible, and instead we approximate the EB with regional effects method by fixing the regional variables, as described above. Once we make this approximation, then EB and EB with regional effects require the same order of magnitude of operations.

Strikingly, EB with regional effects is out performed by basic EB. This may be because of the approximations we made by selecting the minimal regional variables. We see this performance in Figure 6. In this figure, we have the average CV for the regions over the weeks. We see that EB does consistently better than EB with regional effects until the peak in which EB with regional effects slightly overtakes EB.

In fact, EB with regional effects seems to perform poorly as evidenced in 7. We

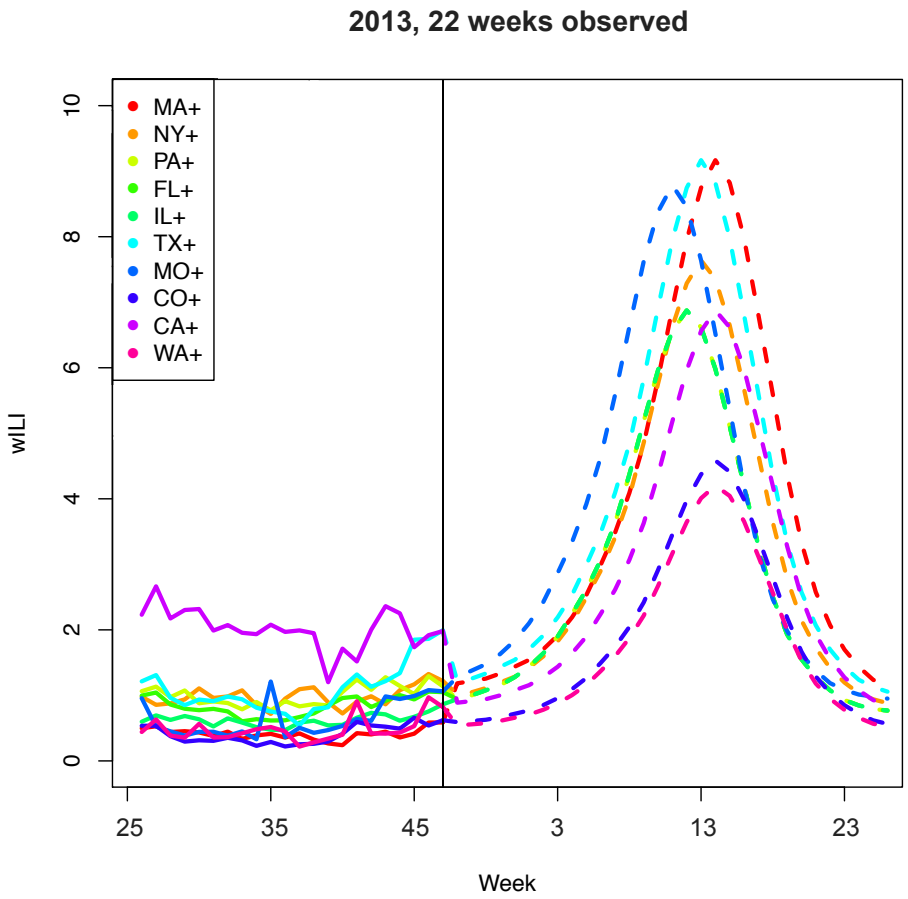


Figure 5: Visualization of predicted wILI curve for Season 14, with 22 weeks observed. Each region is a different color. Solid lines are the observed wILI values and the dashed lines are the predicted wILI values. We see that regional variables at work as the peak wILI and peak week seem to be clustered around one another.

see that our observed results and predicted results are close for Region 6 (TX+) but we truly 'miss the mark' for Region 1 (MA+) for 22 weeks observed in the 2013-2014 Flu season. Unfortunately, this seems to be the case for weeks known and season. The approximation to the impelementation of EB with regional effects becomes apparent like in the case of Region 1 we are not accurately representing the region.

Although some parts of our EB with regional effects model aligned with preconceived notions of regional dependence, it is clear we need to emphasize the goals of predicting the peak week and peak height. These targets are primary forecasting goals. As such, we first need to to create a model to estimate these targets.

3 Target Predictions

Generally, it is easier to estimate a few specific targets than to estimate a whole curve. As long as we have time invariant covariates, we can apply our favorite regression

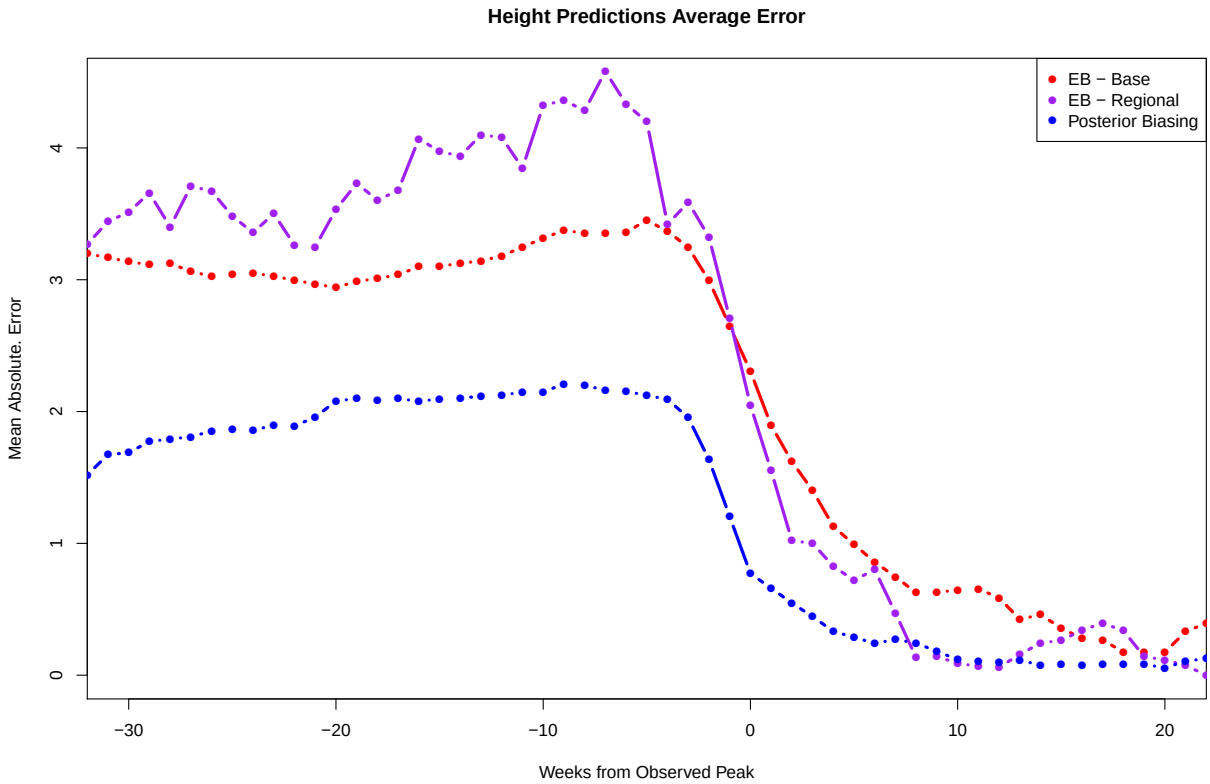


Figure 6: CV averaged over the different seasons for the different model types. Posterior Biasing performs about twice as well in predicting the height than the other two models.

method to estimate the peak height, generally surpassing our method of generating a whole curve. This is demonstrated by the fact that a simple linear model regressing on the three most current observed times t_0, t_{-1}, and t_{-2} and the mean of the rest of the points yields a much improved cross-validation curve than EB with regional effects alone. This is exemplified in Figure 14.

We attempted to find useful covariates by taking the average of different amounts of wILI values for values from weeks 1 to $t - nLag$ where t is the current weeks known and $nLag$ is the number of lags we use in the model. However, we found that none of these features was useful in predicting our target values. We were not completely unsuccessful as we found that using the current observed max week and current observed max wILI were useful covariates.

After some experimentation, we decided upon using 3 lags. The basic model we fit is the following:

$$\begin{aligned}\hat{\text{Target}}_{t,r,s} = \alpha_0^{(r)} Y_{t,r,s} + \alpha_1^{(r)} Y_{t-1,r,s} + \alpha_2^{(r)} Y_{t-2,r,s} + \alpha_3^{(r)} \text{Week} \\ + \alpha_4^{(r)} \cdot \text{Max Height Obs}_{t,r,s} + \alpha_5^{(r)} \cdot \text{Max Week Obs}_{t,r,s} \\ + \sum_{i \neq r} \left(\beta_{0,i}^{(r)} Y_{t,i,s} + \beta_{1,i}^{(r)} Y_{t-1,i,s} + \beta_{2,i}^{(r)} Y_{t-2,i,s} \right) + \epsilon,\end{aligned} \tag{3}$$

where $\text{Target}_{t,r,s}$ can refer to either the peak week or peak height as predicted at

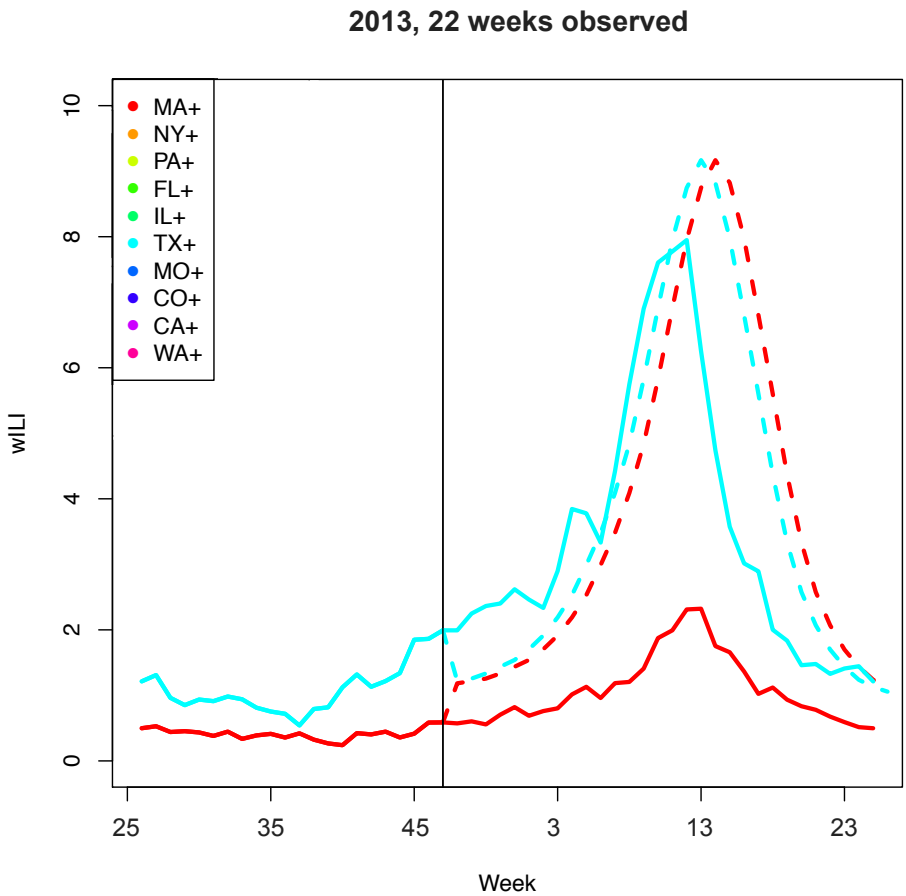


Figure 7: Predicted (dashed line) versus observed values (solid) for two regions during the 2013 season with 22 weeks observed. We see that our predictions for the Texas+ Region (Region 6) are fairly accurate, but we are off for the Massachusetts+ Region (Region 1).

time t in region r and season s. Here $Y_{t,r,s}$ is the wILI value for week t, region r, and season s.

The above model in Equation 3, is first used to fit a linear model, but we explain how we slightly modify the model to fit elastic nets and additive models. Again, for validation we use CV as described in Section 2.3.

When fitting linear and other models, we cannot pin our observed values as easily as when we were working with EB and had a clearly defined maximum. As such, in addition to fitting the predicted peak week or height, we also fit a logistic model $M(X)$ on the same covariates on whether the max week has occurred ($M(X) = 1$) or not ($M(X) = 0$). The final predicted target is then

$$Target_{r,t,s} = \hat{Target}_{r,s}\mathbf{1}(M(X) = 0) + \max_{\tau \in \{1,2,\ldots,\}} Y_{\tau,r,s} \cdot \mathbf{1}(M(X) = 1),$$

where $\mathbf{1}$ is the indicator function and $\hat{\text{Target}}_{r,t,s}$ is the estimation of $\text{Target}_{r,t,s}$ from the result of fitting Equation 3. We note that max becomes the `which max` when the target is peak week. In this manner, we pin down the max height and week. Given

that our logistic model tends to the 1 as the season ends, then our error will tend to 0. For the week prediction, we note that we also round to the nearest integer. We fit two types of models:

- Elastic nets. We only penalize the covariates with a 'β' coefficient. We also use this method for variable selection for the following method.
- Additive Models. We fit splines with 5 degrees of freedom on the lagged variables and keep the max observed height and max observed week as linear terms in the model.

In addition to fitting models other than linear ones, we can influence our model through our training set. Namely, we utilize three training sets:

- Design I: Complete Independence. We train a model for each region r. We set all of the βs equal to 0 and, $\mathbf{X}$, the design matrix contains only data from that region r.
- Design II: United Model. We train one model for all 10 regions and set the βs to 0. (All $\beta_{j,i} = 0$, $\alpha_i^{(r)} = \alpha_i^{(r')}$ for all r, r')
- Design III: Semi-Dependence. We train a model for each region r. We do *not* set the βs to 0.

These three design types represent different levels of dependence among the regions. The first type is that the models are completely independent from one another. The second type is that there is one model that can be fit across the whole US, a more stringent assumption. Finally, the third model says that we can aid our forecasting by using concurrent data from the other regions.

We had one more idea in that, perhaps, there are different models fit on different parts of the season. We decided to divy up the season into 3 parts - pre-season, in-season, and post season, and fit the model on each section of the season.

3.1 Targeting Results

We briefly mentioned that elastic nets can be used for model selection. With that in mind and Design III where we use concurrent information from the other regions, we can better analyze regional dependence.

We create an adjacency matrix by aggregating the standardized coefficients from fitting an elastic net with the one standard error rule. We say a region is dependent on another if the standardized coefficient of the latter region is large, on average. We display the top 10 largest effects (for visualization purposes only) on the map in Figure 8 looking only at the coefficients from the current weeks. For instance, we see that many regions' regressions are associated with Region 4 (FL+), two such regions from the North. Perhaps this an indication of northerners traveling south for the winter. Another interesting effect is that Region 1 (MA+) and Region 6 (TX+), which are physically distant from one another, strongly influence each other. Perhaps this is an indication of heavy air travel between these two regions.

In total we fit 8 different models, as after fitting elastic nets and finding the relevant variables, we find that fitting an additive model uniformly yields an improvement in error over time. We again compare the models through CV, this time only looking at

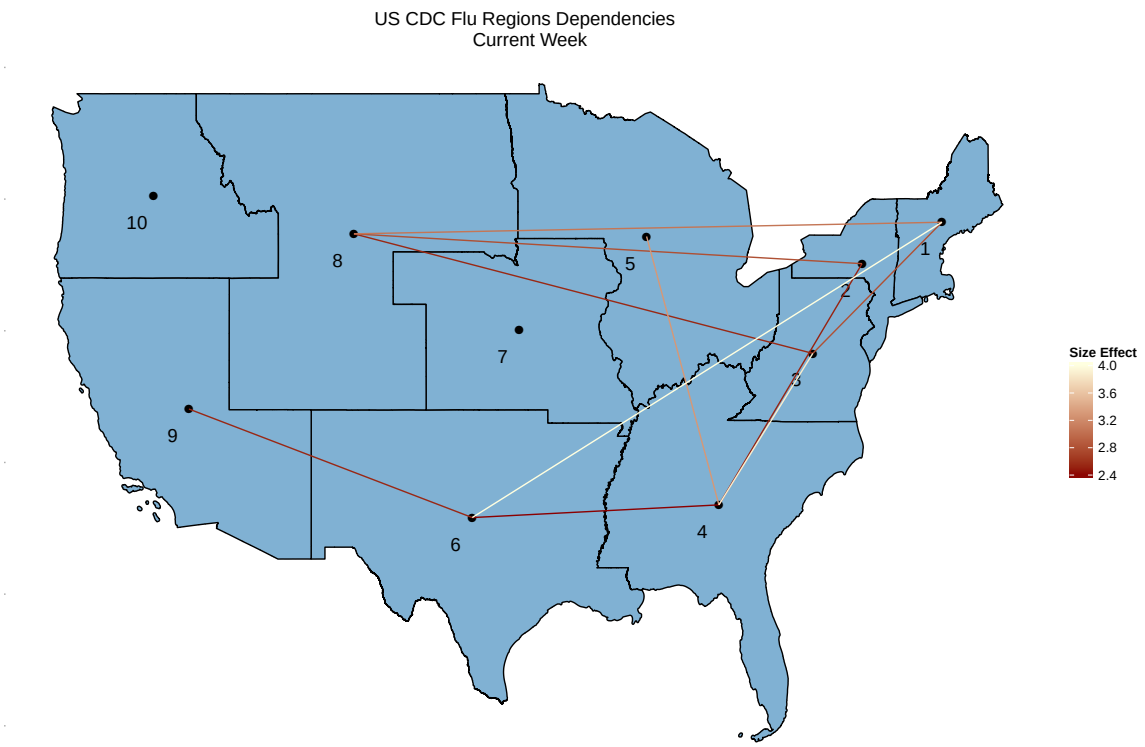


Figure 8: A map of the 10 CDC regions along with top 15 effect sizes from from other regions. The brighter the line, the more influence at least one of the regions has on the latter. This map indicates the effect that a region's current value has on another region's regression for Peak Height.

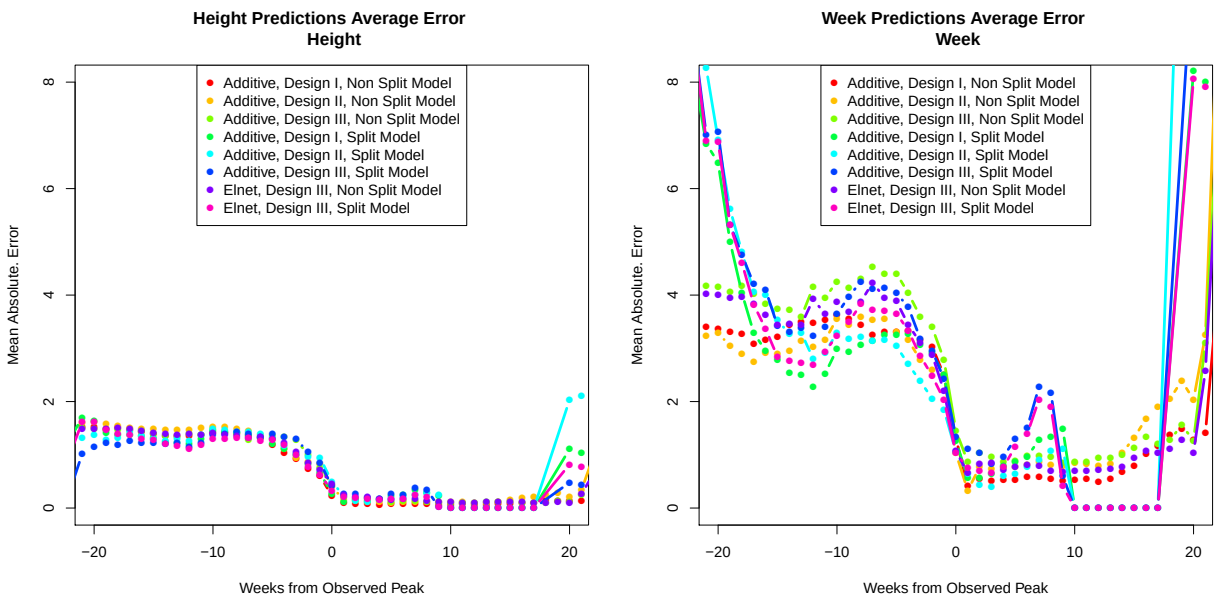
the average error over time amongst the regions. Surprisingly, none of the models were dramatically better than that of a simple linear model with 3 lags.

Displayed in Figure 9 are the CV curves for predicting peak height and peak week, respectively. For predicting the peak height, we see that the all the models produce approximately the same error curve, except at the extremes which generally reflect outliers more than any statistical meaning. The peak week CV curve is more interesting, as we see that fitting a different model in different parts of the season does not perform well, although two split models produce zero error. This can be attributed almost solely to fitting the logistic model on the post-season data. We see that at first the model does well but then begins to err extremely. The error does not seem to tend to zero for peak weeks, and this is due partially to a rounding error and also in part that for one of the regions, the logistic model for whether the peak has occurred is consistently wrong. Overall, we see that design types I and III both do well.

We use a linear combination of Design Type I (with weight .8) and Design Type III (with weight .2) because it had the the best CV curve and the smallest cumulative total absolute error.

We use the above model additive model in the rest of the analysis because it has the best CV curve and cumulative error total.

Regardless of the model, we see that the CV curves from targetting outperform that of EB, especially when many weeks away from the peak. However, these target



(a) (b)

Figure 9: CV curves from fitting the different model types on the data. For the height, we see there is close to no difference in the different model types except at the exteremes, which are generally not too relevant to the flu predicting process and generally only reflect outliers. For predicting the week, we find that splitting the season into different sections is not very helpful. Of these, Design Types I and III perform the best.

predictions as-is are not very useful. We lose interpretability, credible intervals, and the events leading up to the peak week. This is a high price to pay for a better CV curve. As such, we combine the strengths of our two models: distributions and interpretability from EB along with the accuracy of target models in what we call Posterior Biasing.

4 Posterior Biasing

Although generating point predictions is easier than predicting entire curves, we find value in generating a whole curve for a variety of reasons. By producing a whole curve, we can generate a whole host of other targets such as season beginning, duration of season, and average wILI during the flu season; create of a posterior can be used to produce credible intervals with ease; visualize the resulting curves; and incorporate our prior beliefs about the flu season into the model

As we have smooth wILI curves with average to good peak week predictions and regression models with adequate predictions for peak height, it makes sense to combine the two models. We mix the models using 'posterior biasing.' In essence, we are making the model more 'aware' of its target goals– namely predicting peak height. We bias the posterior by allocating more weight to wILI values that are closer to a region's predicted peak height. Specifically, we change Equation 2 to the below, with $R_+ = \{1, 2, \ldots, 10\}$,

$$\lambda\left(\mu_{t_i}^{(r,s)},\ i \in T_+, r \in R_+ | D_{s_0,t_0}\right) \propto \frac{1}{\sigma} \exp\left(-\sum_{t=1}^{t_0} \frac{\left(y_t^{(s_0,r)} - \mu_t^{(s_0,r)}\right)^2}{2\sigma^2}\right) \cdot \exp\left(\frac{|(\mu_{r,s} - \hat{\mu}_{r,s})|}{\omega}\right), \tag{4}$$

or as more simply described: weighting the posterior wILI values for each region by giving more weight to predicted values of the peak height $\hat{\mu}_{r,s}$ that are close to to the predicted value $\mu_{r,s}$ from the regression model. This weighting process is illustrated in Figure 10. In this way, we are departing from a traditional Bayesian's worldview by creating posteriors for each region that are biased towards the predicted peak height from our regression model. Also note, that although we are still using regional data, the incorporation of it is now in the $\hat{\mu}_{r,s}$ values, as we are no longer summing over the regional error.

We note that in Equation 4, we could also extend this posterior biasing to other targets or some combination thereof, but we focus on peak height for the remainder of the paper.

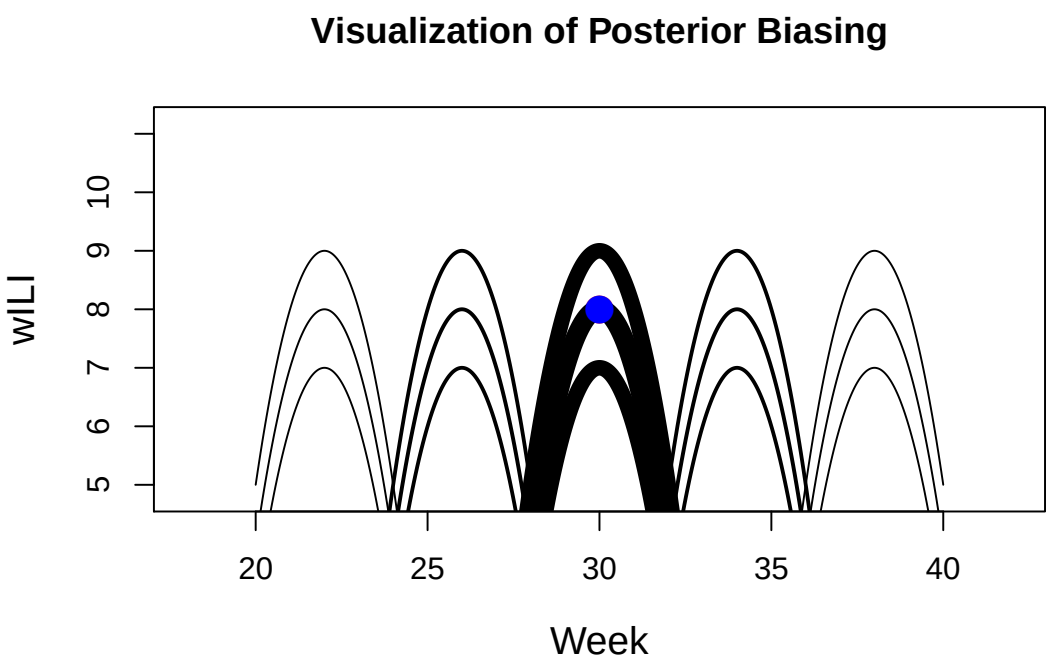


Figure 10: Image depicting of weighting curves whose peak values are closer to our estimated values. A thicker line represents a larger weight. The blue dot is our estimated value of the peak height and week.

4.1 Posterior Biasing Results

Even when we look at one value of ω, namely $\omega = 1$, we obtain better predictions for peak height than EB or EB with regional effects, which we display in the following Section 5.1. We have the CV curve for the 10 regions in Figure 11. Interestingly, we seem to have about constant absolute error for each region until about a month before the peak week is observed, at which the absolute error quickly drops to zero. Puzzlingly, Colorado's Region 8's absolute CV error does not drop to zero until after 20 weeks, which indicates that the logistic model is not accurately predicting whether we have achieved the peak week or not.

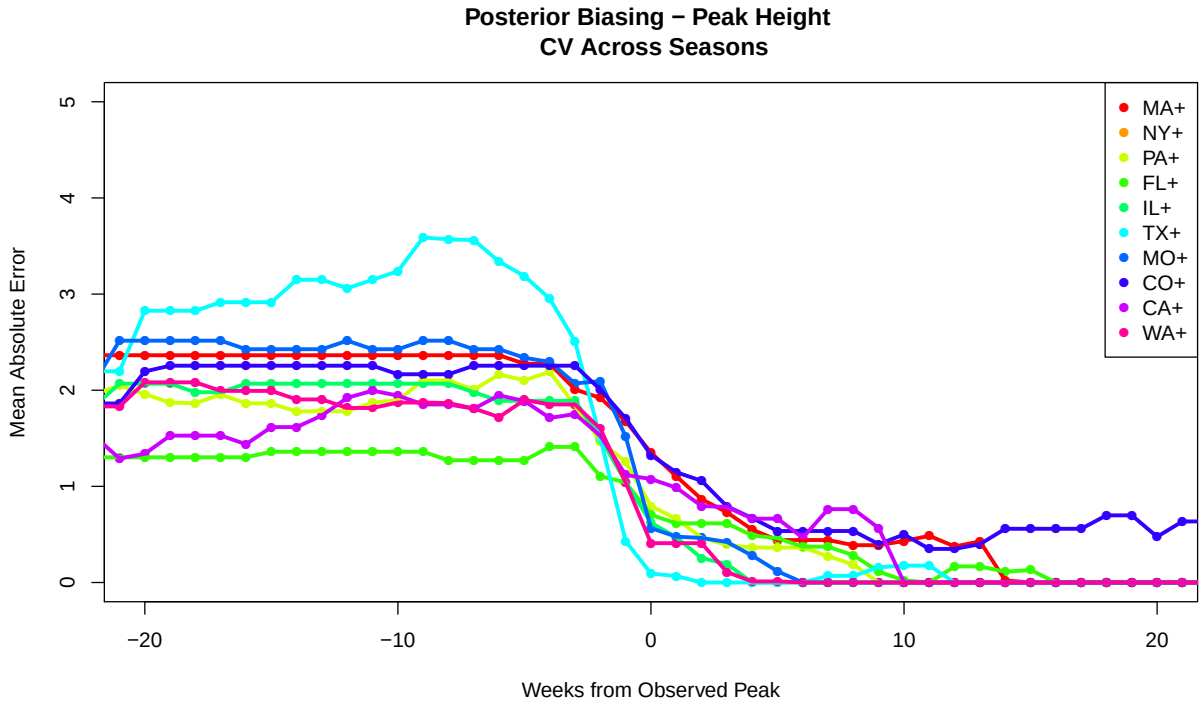


Figure 11: CV curve for Posterior Biasing for the different US regions. We see that from about 20 weeks prior to the peak to about 5 weeks prior we experience approximately the same error per region. However, around 5 weeks, we beging to more accurately predict the peak height and continue to do so until the error drops close to zero.

The predicted wILI curves for posterior biasing look like flu curves as shown in figure 12. We observed the regional effects in the model through Region 6 (TX+)'s peak is expected to be sooner and larger than the other regions. At the other extreme are Region 1 (MA+) and Region 10 (WA+) which have generally lower peaks that occur later than the flu season.

4.2 Analyzing the Frequentist Properties

Although we are working with a primarily Bayesian model, we are interested in the frequentist properties of the model for the sake of interpretation. The epidemiologists who would use these curves in their work are interested in how often our coverage

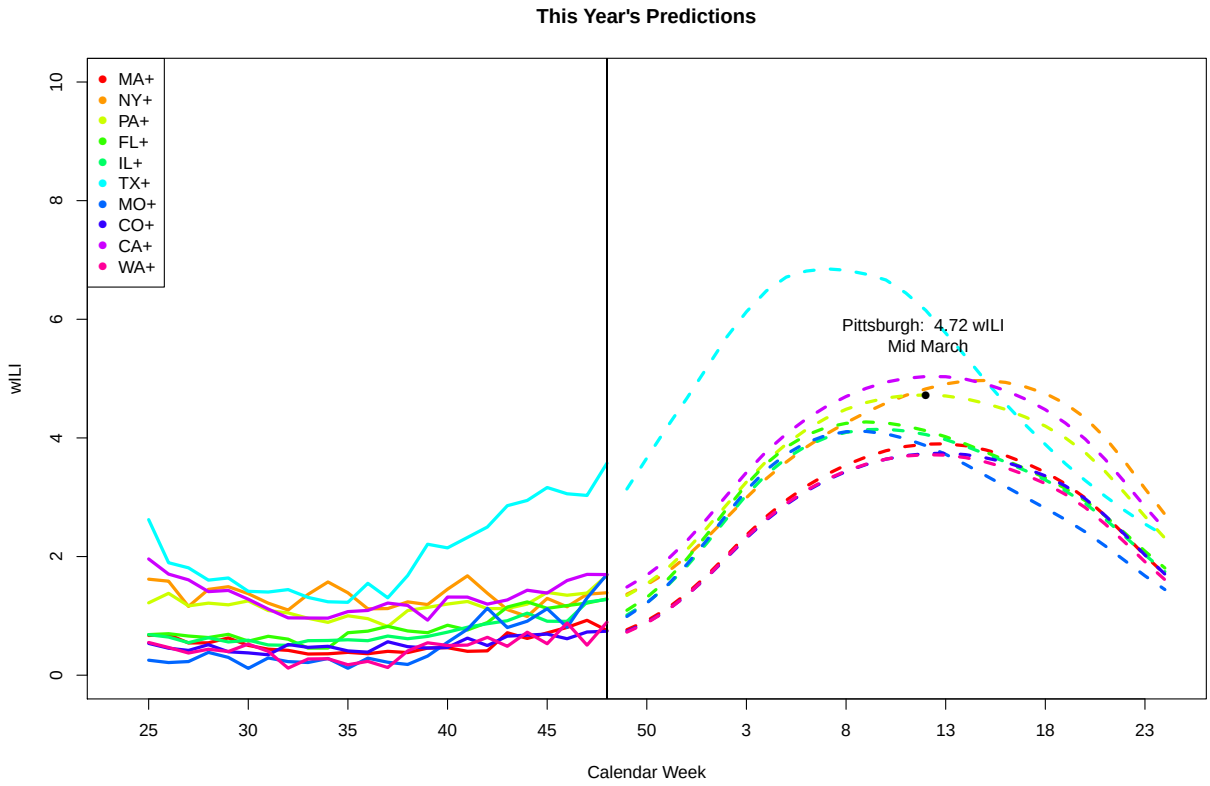


Figure 12: Current predictions for the (new) 2015-2016 Flu season with 25 weeks observed. We see regional effects in that the Texas region is predicted to be effected more intensely and sooner than the other regions. Pittsburgh, the author's current city, is predicted to have a peak of 4.72 wILI in mid March.

intervals enclose the true wILI value, and so it makes sense to look at these properties, even though we have no such expectations of the coverage as our model is Bayesian.

To analyze the frequentist coverage, we first calculate our Bayesian Credible Intervals (CI) for each run of cross validation for the different regions. For the intervals of 20 weeks to the peak and 20 weeks post the peak, we have over 100 observations at each week. We then calculate how many times the CIs captured the peak week or height at each week from the observed peak and plot the results in Figure 13.

Surprisngly, our 90% CI intervals have at least 90% frequentist coverage for both week and height from about 15 weeks from the observed peak onwards. This is surprising as usually we have no expectation that Bayesian CIs will correspond to frequentist coverage intervals. This result is useful for explaining our model to others as 90% of the time we expect our predictions to be contained in the given intervals. However, we discovered that the 90% CI intervals are very large so this may not be as useful as we originally thought.

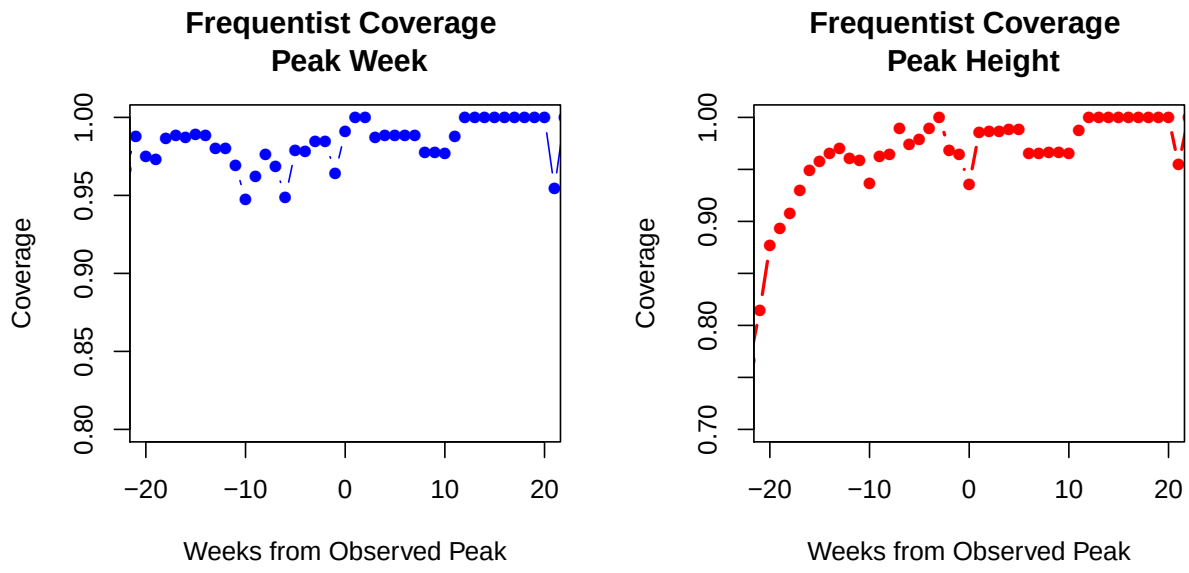


Figure 13: Frequentist coverage for our 90% Bayesian Credible Intervals for our Posterior Biasing model. Interestingly, we have at least 90% frequentist coverage across the CV runs after 15 weeks prior to the peak.

5 Discussion

5.1 Model Comparisons

We compare three full-curve models: EB as a benchmark, EB with regional effects, and Posterior Biasing. All three models are based on the same framework and so have similar interpretations. Both EB and EB with regional effects are true EB models but Posterior Biasing strays from the traditional EB model and so is accordingly harder to interpret. The three models can be differentiated as following

- EB - full independence
- EB with regional effects - regional dependencies
- Posterior Biasing - 'target awareness'.

Among the above three models, Posterior Biasing outperforms the other two when we are evaluating the performance of the predicted peak height. This result is displayed in Figure 6. We see that Posterior Biasing performs about twice as well as the other two models in predicting the height up until the peak week is observed, and is the clear best choice when it comes to predicting the peak height.

We see that the EB model with regional effects does worse than the benchmark EB model or about the same as it. Although the EB with regional effects model is flexible, and should at the very least perform as well as EB, the implementation of the model relied on simplifying approximations which hurt us here.

In addition to the full curve models, we also have the targetting model, which outperforms the other models in predicting the height. These results are summarized in Table 2. In the table, we see that although the Targeting model does outperform

Posterior Biasing, there is no distinguishable difference within one standard deviation. We prefer the Posterior Biasing model over the Targeting model due to its other useful properties: having an entire curve of predictions, probabilistic predictions, and a reliable frequentist interpretation of Bayesian CIs.

Final Results for Peak Height Predictions				
Model	Season CV	CV St. Dev.	Pros	Cons
EB	179	40	Baseline	Independence
EB with Reg.	187	48	Regional Effects	Curse of Dimensionality
Target	57	7	Simpler	No Curve
Post. Bias.	91	28	Flexible	Interpretability

Table 2: Summary of CV results for peak height. The Season CV is the cumulative absolute error throughout a season averaged over all the regions and cross validation runs. We see that the target model performs the best but does not have the full curve of predictions. Posterior Biasing, we see, performs about twice as well as EB and EB with regional effects. The unit for CV is wILI value.

Computationally for the *implemented models*, EB with regional effects and EB have about the same magnitude of operations. The Targeting model is by far the fastest as we are predicting only one point. Posterior Biasing combines both the EB and Targeting model and thus the time taken is about the same order of magnitude as EB. However, each of these models can be comfortably run on a PC, as long as we keep our prior grid size small. In fact, we observe that having just 5-10 values for each hyper parameter will produce about the same results as having more prior grid values.

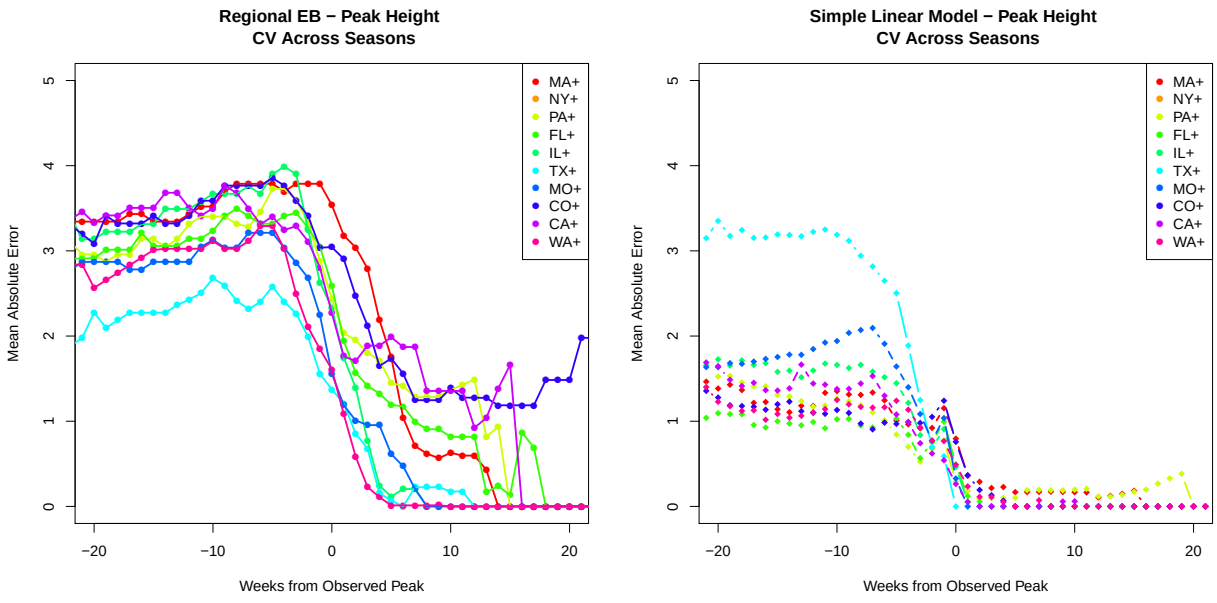
As we use regional effects in all but the base EB model, we conclude that the regions influence the others' peak heights and should be incorporated when modeling the flu in the US.

5.2 Final Comments and Looking Forward

Although our first idea of EB with regional effects failed to outperform base EB, we believe that is a more of a failure of implementing our model rather than evidence against the assumption that the regions are dependent upon one another. When incorporating regional effects into both our Targeting and Posterior Biasing models, we outperformed the height predictions of base EB.

Posterior Biasing is our preferred model because in addition to the fact that the total CV error is twice as small as the other full-curve models, posterior biasing also has the upsides of probabilistic predictions, empirically reliable frequentist coverage, and emphasis on the targets at hand.

That said, there is much room for this model to be improved in the future. We only used one value of ω, our tuning parameter, and could potentially find a better parameter value through cross validation, which will take some computer hours to evaluate. There is also the issue of biasing the posterior towards the peak week in addition to peak height. From this idea emerges the question: what is the optimal way to bias the posterior? We hope to explore this question in future work.



(a) (b)

Figure 14: Leave-one-season-out cross validation for EB with regional effects (a) and a targetted regression (b). The x-axis is the weeks from the observed peak and the y-axis is the mean absolute error. The max error for the targetted regression is just over 3, and is about 1 wILI off a few weeks before the observed peak. On the other hand, EB with regional effects is about 2-4 off a few weeks before the observed peak.

Perhaps most surprisingly is the fact that our 90% Bayesian CIs approximately correspond to 90% frequentist CIs when using Posterior Bayes, and is a point we would like to investigate looking foward.

We primarily assessed the usefulness of the predictions based upon the predictions for the peak week, but there are other ways to assess performance such as log scores of the probabilistic predictions of the targets or cumulative error for the remainder of the season, the former being especially useful if submitting our predictions to the CDC's competition.

Due to the fact that there are so many ways to evaluate the usefulness of our predictions exemplifies why it is key to rely upon the advice of epidemiologists and health care professionals who know which attributes of the predictions we care about most. We want to keep our models as scientifically interpretable as possible.

A Appendix

A.1 2015-2016 Epidemic Thresholds

We see the Epidemic Thresholds for the different regions for the 2015-2015 Flu Season in Table 3.

Region	1	2	3	4	5	6	7	8	9	10
Epidemic Threshold (%)	1.3	2.3	1.8	1.6	1.9	3.6	1.7	1.4	2.6	1.1

Table 3: Epidemic Threshold's for the 2015-2016 Season from the CDC. From `http://www.cdc.gov/flu/weekly/overview.htm`.

A.2 Computational Complexity

A difficulty in our method is the high dimensionality of our posterior. We have 2 seasonal parameters and 4 regional parameters. Since we treat the regions as dependent on one another, in order to fit our full model we must fit all combinations of regional variables, which would be of order $O(n^{10})$, which alone is nearly too expensive to run. Hence, we are forced to make an approximation of fixing the regional variables for each region by using the regional variables with the highest weights.

In importance sampling, we partition the prior into a discrete grid of values so depending on how fine we make our grid for each value of the parameters, we easily have to weight a grid of $O(n^8)$ values or more. This can become an issue, especially in `R`, with the available RAM a machine has and also from slow down in the model fitting process even though most of the operations are vector addition and matrix multiplication.

Not unexpectedly, relaxing the assumption of independence quickly leads us to the 'curse of dimensionality', and approximations and clever fitting methods must be used in order fit the originally intended model.

Fortunately, we are able to utilize parallelization in this model for the purposes of cross validation. Predictions for a given week and season are completely orthogonal to one another. We heavily use this fact to complete Leave One Season Out (LOSO) cross

validation where we generate $S \times T$ posteriors where S is the number of seasons and T is the number of weeks in a season. Parallelization reduces our total computation time from hundreds of real time computer hours to just a few.

References

[1] Logan C. Brooks, David C. Farrow, Sangwon Hyun, Ryan J. Tibshirani, and Roni Rosenfeld. Flexible modeling of epidemics with an empirical bayes framework. *PLoS Computational Biology*, 11(8), 2015.

[2] CDC. Flu View. Retrieved from `http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html`, 2014.

[3] CDC. Seasonal Influenza: Basics. Reterived from `http://www.cdc.gov/flu/about/disease/index.htm`, 2015.

[4] Hethcote HW. The Mathematics of Infectious Diseases. SIAM Review 42: 599-653, 2000.

[5] Knipe, D., Howley, P. Fields' Virology. Lippincott Williams & Wilkins, a Wolters Kluwer Business, 2007.

[6] Surya T Tokdar and Robert E Kass. Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60, 2010.