PREDICTION FEVER

Predicting the Flu via Regional Effects

Shannon Gallagher
Statistics Advisors[†]: Ryan Tibshirani & Bill Eddy
External Advisor[‡]: Roni Rosenfeld

November 6, 2015

Carnegie Mellon University
[†]Department of Statistics
[‡] MIDAS & Department of Computer Science

How bad is the flu going to be this year?

OUTLINE

- Background
- Data
- ADA Goals
- Modeling
  - Empirical Bayes with Regional Effects
  - Target Model
  - Posterior Biasing
- Final Results
- Future Work

BACKGROUND

Influenza (flu) is destructive

The flu is...

- old
- deadly
- costly
- stochastic

WE WANT TO MITIGATE THE EFFECTS OF THE FLU THROUGH PREDICTION

With accurate predictions, the flu is

- old
- ~~deadly~~ $\rightarrow$ manageable
  - Resource allocation
  - Alert health officials
  - Issue warnings
- ~~costly~~ $\rightarrow$ feasible
  - Fewer sick days
  - More awareness
- ~~stochastic~~ $\rightarrow$ forecasted

DATA

THE CDC COLLECTS DATA VOLUNTARILY FROM PHYSICIANS

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: Cross Section of Available data.
wILI - Weighted Influenza Like Illness

FLU CURVES USUALLY HAVE ONE PROMINENT PEAK

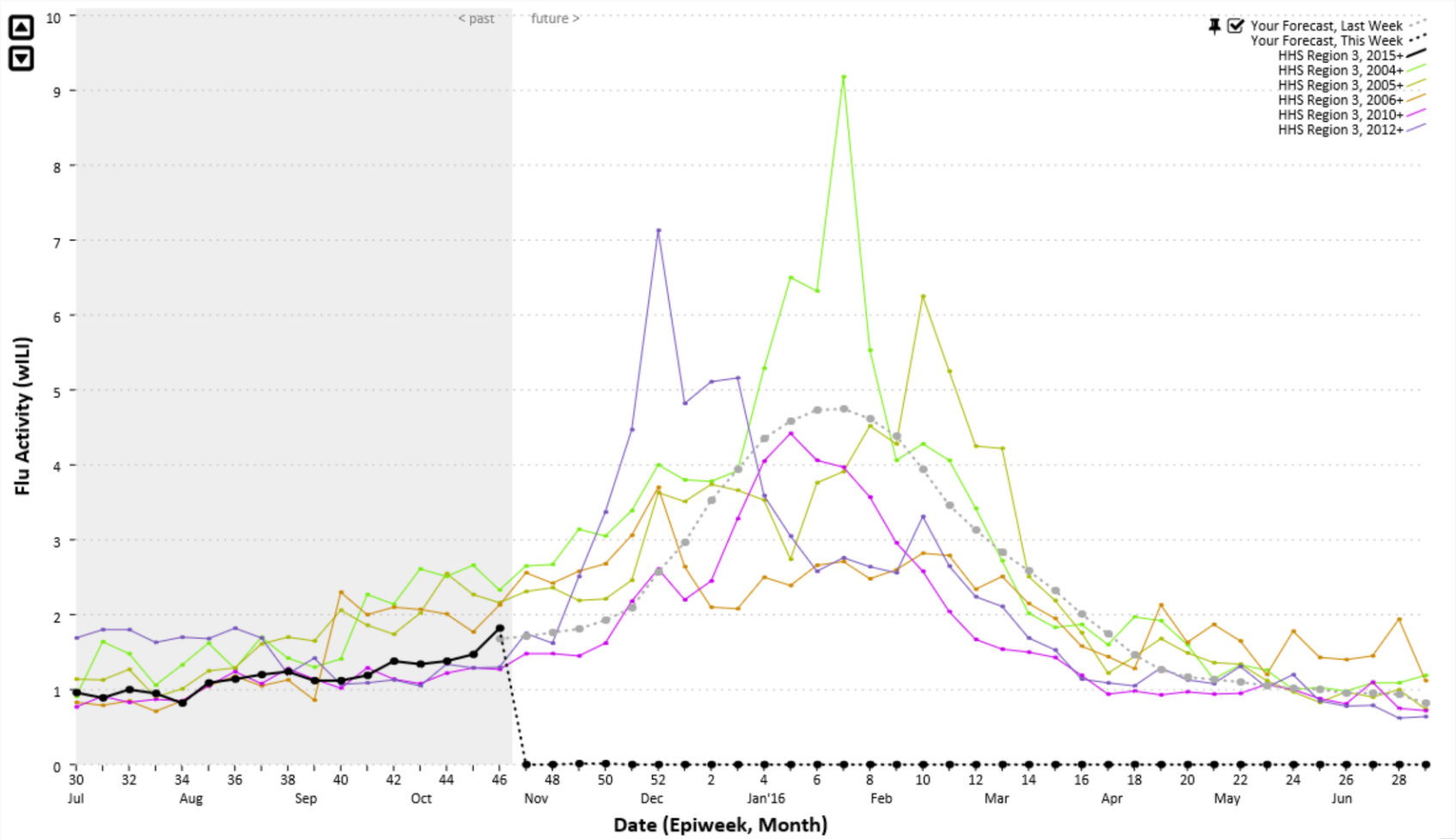


Figure: Examples of wILI curves. From David Farrow's FluV.

epicast.org

ADA GOALS

WE WANT TO PREDICT THE WILI FOR THE REMAINING WEEKS OF A SEASON

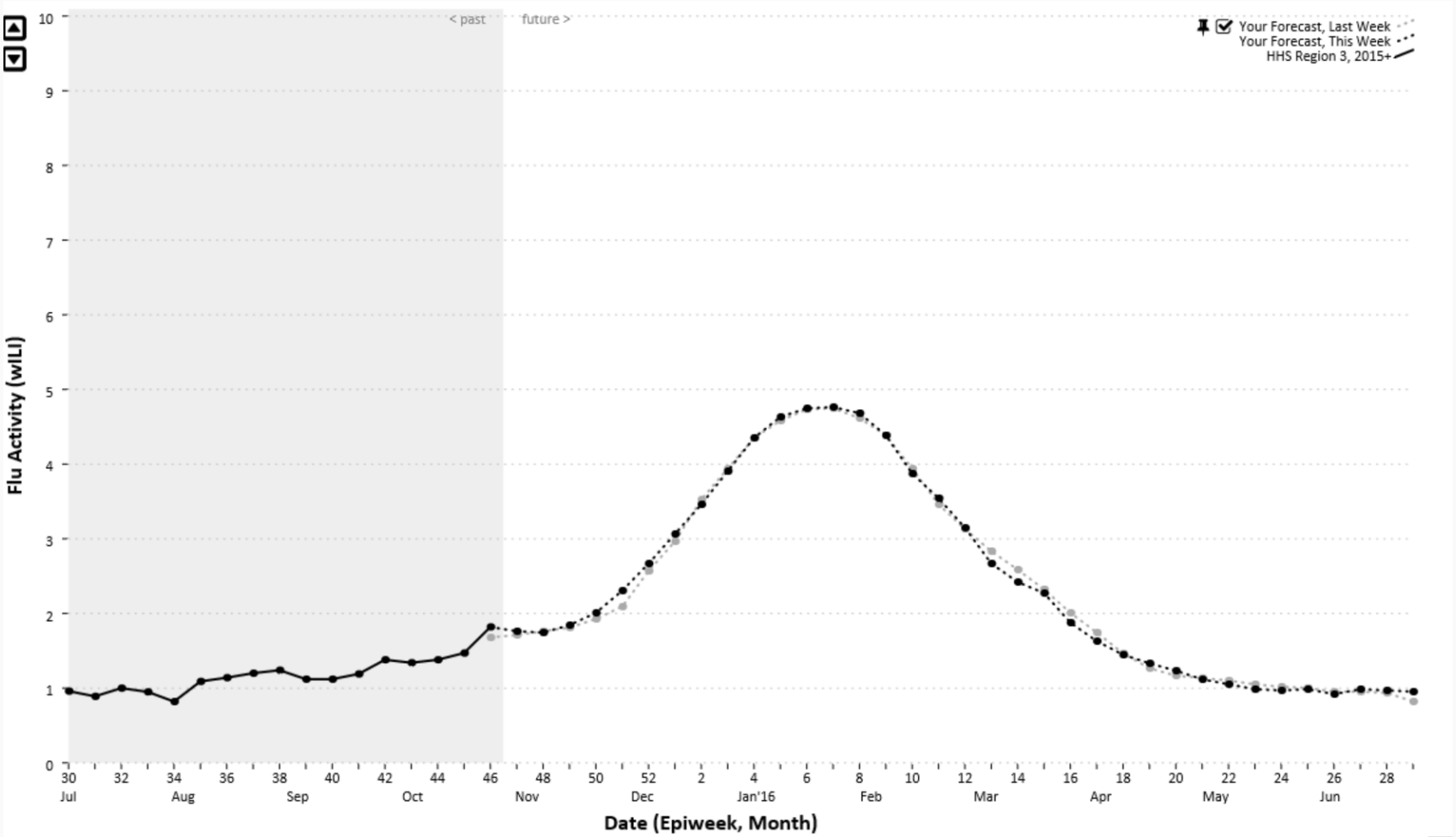


Figure: Examples of wILI curves. From David Farrow's FluV.

`epicast.org`

FOR EACH OF THE 10 CDC REGIONS

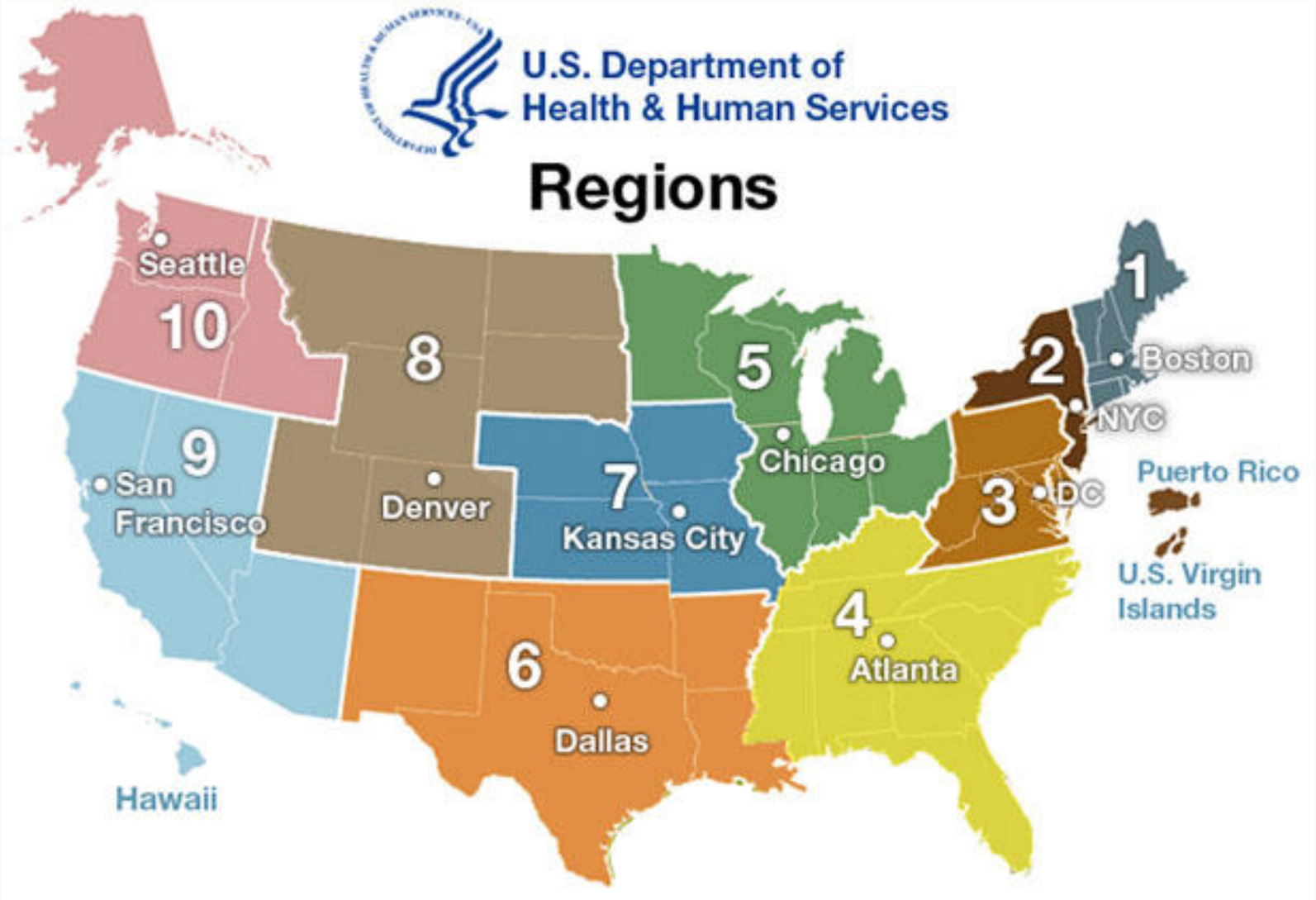


Figure: From cdc.gov

WE HAVE EVIDENCE OF REGIONAL DEPENDENCIES

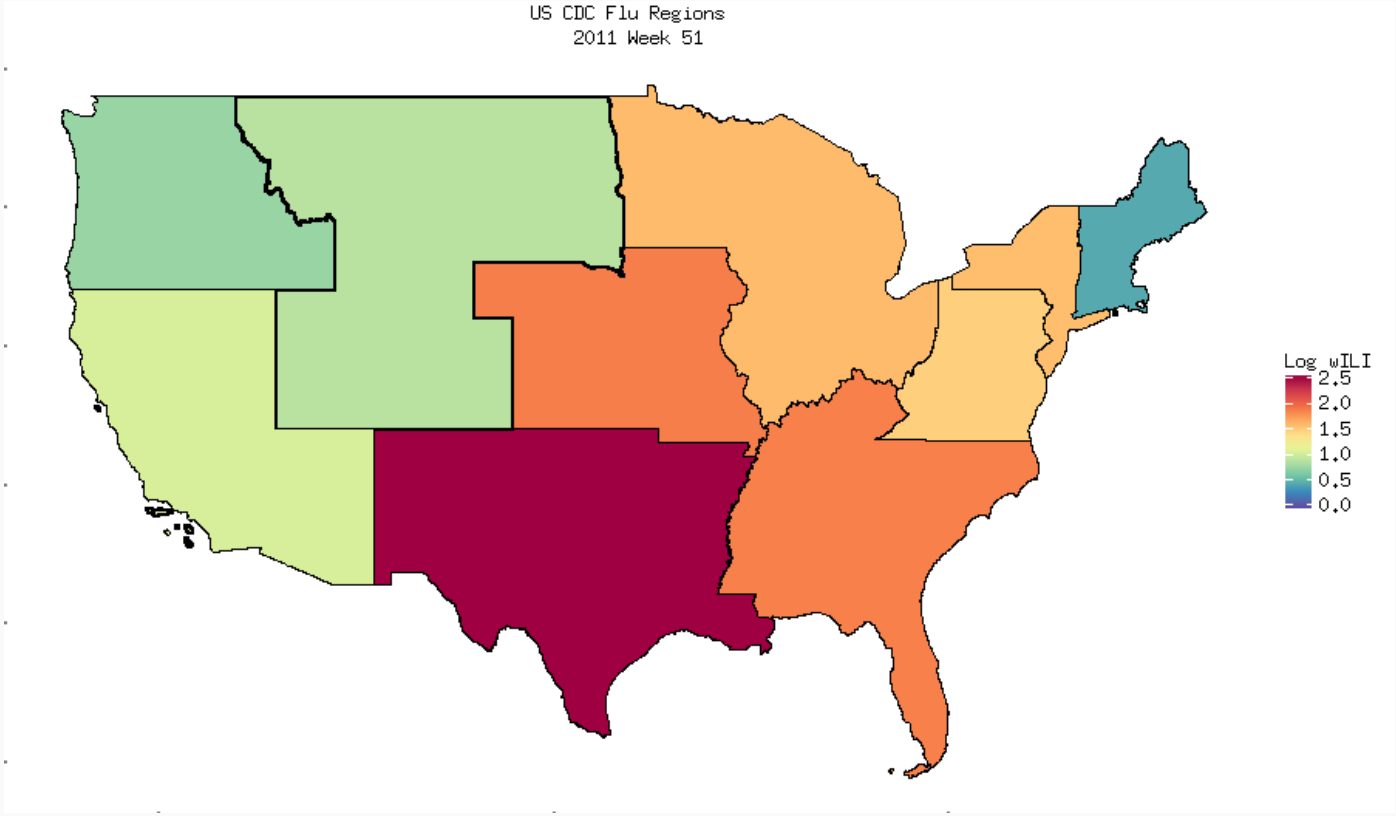


Figure: 2001 Week 51. This map is generally representative of the other seasons, when the peak flu season begins. Region 6 (Texas +) is hit hard and first. Regions 1 (NE) and 10 (PNW) are barely effected. PA is generally in the middle.

FOR EACH REGION AND ANY AMOUNT OF DATA, WE WANT TO

1. Allow for regional dependencies
2. Predict the wILI values for the remainder of the season

3. Predict specific targets
    - Peak Week
    - Peak Height/Incidence
4. Produce distributional forecasts

5. Explain our model to health care professionals

MODELING

WE WANT TO BUILD UPON PRIOR WORK ON EMPIRICAL BAYES

Figure: From `delphi.midas.cs.cmu.edu`.

Empirical Bayes (EB) Framework is based on the assumptions:

1. A flu curve will look like a past curve + modifications + noise
2. The regions are independent from one another

WE WANT TO BUILD UPON PRIOR WORK ON EMPIRICAL BAYES

Figure: From `delphi.midas.cs.cmu.edu`.

Empirical Bayes (EB) Framework is based on the assumptions:

1. A flu curve will look like a past curve + modifications + noise
2. ~~The regions are independent from one another~~

MODELING SUMMARY

1. Create EB model with regional effects for full curve
2. Validate model through cross validation
3. ~~Rejoice over great results~~
4. Create another model focusing on targets
5. Combine two models (Posterior Biasing)
6. Cross validate
7. Finish

THE FULL MODEL RELIES ON UNIFORM PRIORS.

$$Y_t^{(r,s)} \sim N(\mu_t^{(r,s)}, \sigma^2)$$

where

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

for week t, region r, season s and priors:

$a_s \sim \text{Unif}(2, 10)$ -seasonal scaling

$b_s \sim \text{Unif}\{-6, -5, \ldots, 6\}$ -seasonal shifting

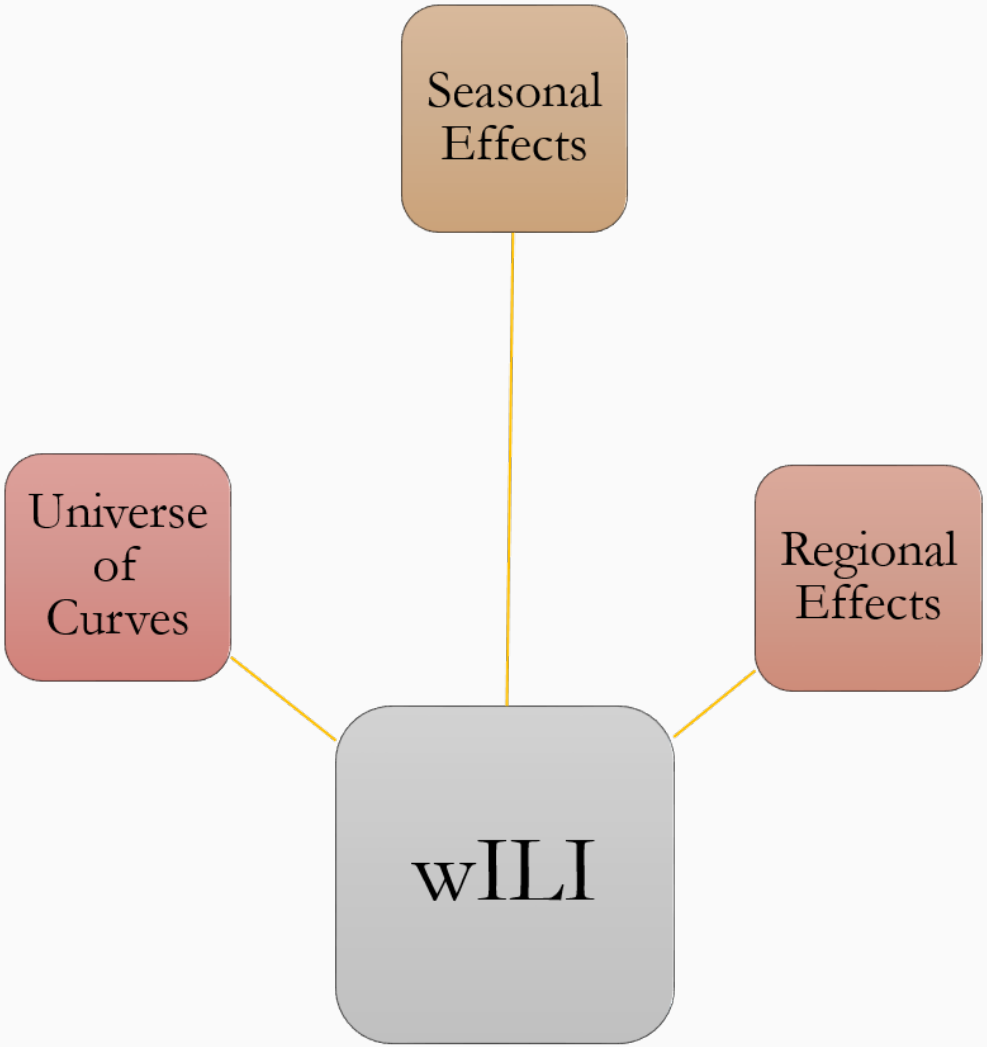
$\alpha_r \sim \text{Unif}(0.25, 1.25)$ -regional scaling

$\beta_r \sim \text{Unif}\{-3, -2, \ldots, 3\}$ -regional shifting

$f \sim \text{Unif}\{\hat{F}\}$ -smoothed observed curves

$\sigma^2 \sim \text{Unif}(0.5, 2.5)$ -variance

THE MODEL RELIES ON SEASONAL VARIABLES



FITTING A MODEL ACROSS REGIONS IS EXPENSIVE

For given season, given weeks known, the posterior is

$$\mathcal{P}(\text{fut. wILI}|\text{obs. wILI, params}) \propto \frac{1}{\sigma} \exp\left\{ \frac{\sum_{\text{regions r}} \sum_{\text{obs. weeks t}} (y - \hat{y})^2_{r,t}}{2\sigma^2} \right\}$$

FITTING A MODEL ACROSS REGIONS IS EXPENSIVE

For given season, given weeks known, the posterior is

$$\mathcal{P}(\text{fut. wILI}|\text{obs. wILI, params}) \propto \frac{1}{\sigma} \exp\left\{ \frac{\sum_{\text{regions r}} \sum_{\text{obs. weeks t}} (y - \hat{y})^2_{r,t}}{2\sigma^2} \right\}$$

Recall the priors for the regional variables are

$$\alpha_r \sim \text{Unif}(0.25, 1.25) \qquad \text{-regional scaling}$$
$$\beta_r \sim \text{Unif}\{-3, -2, \ldots, 3\} \qquad \text{-regional shifting}$$

$\implies$ If we choose just 5 values for each parameter, then summing over the regions is 25^{10} possible combinations!

FITTING A MODEL ACROSS REGIONS IS EXPENSIVE

For given season, given weeks known, the posterior is

$$\mathcal{P}(\text{fut. wILI}|\text{obs. wILI, params}) \propto \frac{1}{\sigma} \exp\left\{ \frac{\sum_{\text{regions r}} \sum_{\text{obs. weeks t}} (y - \hat{y})^2_{r,t}}{2\sigma^2} \right\}$$

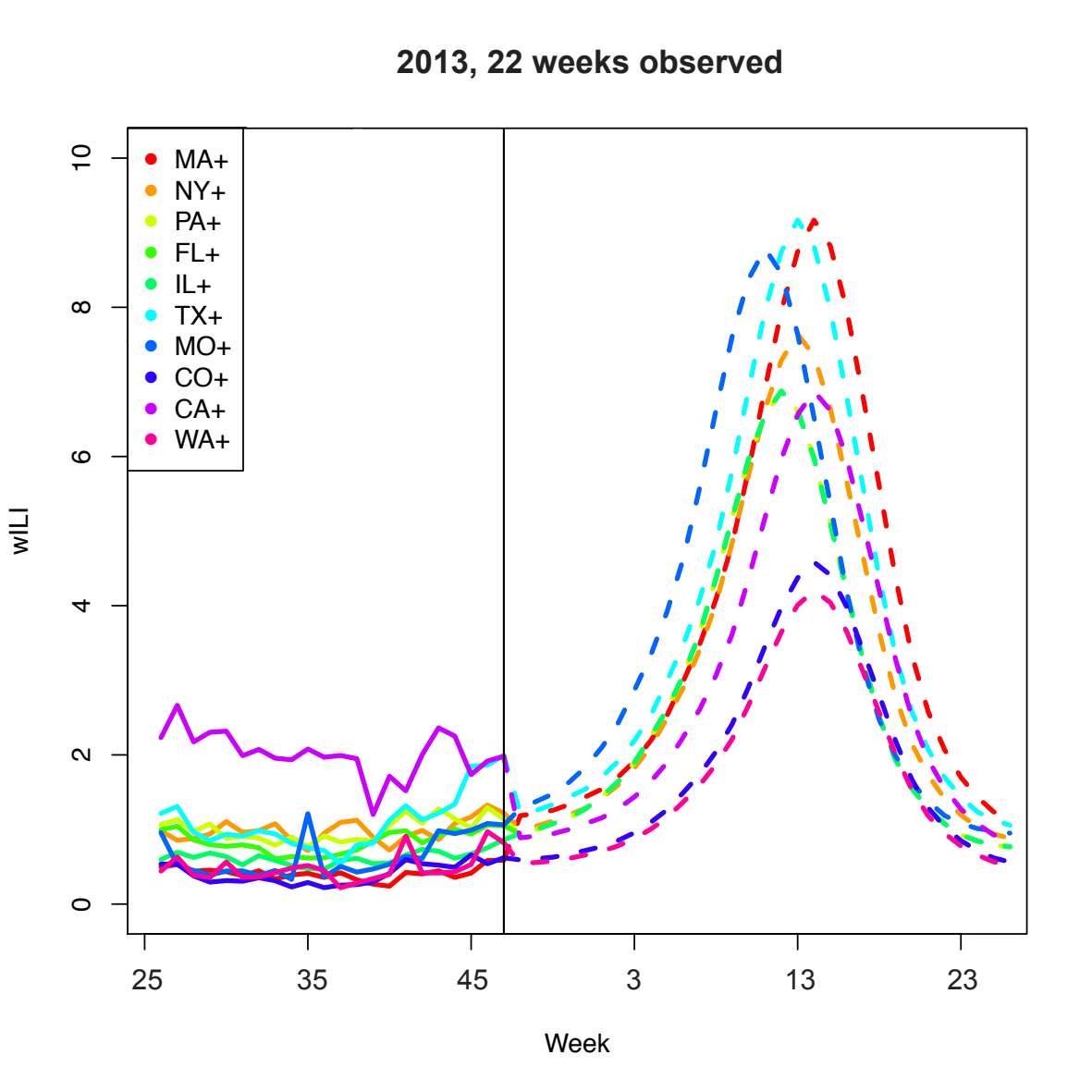
Recall the priors for the regional variables are

$$\alpha_r \sim \text{Unif}(0.25, 1.25) \qquad \text{-regional scaling}$$
$$\beta_r \sim \text{Unif}\{-3, -2, \ldots, 3\} \qquad \text{-regional shifting}$$

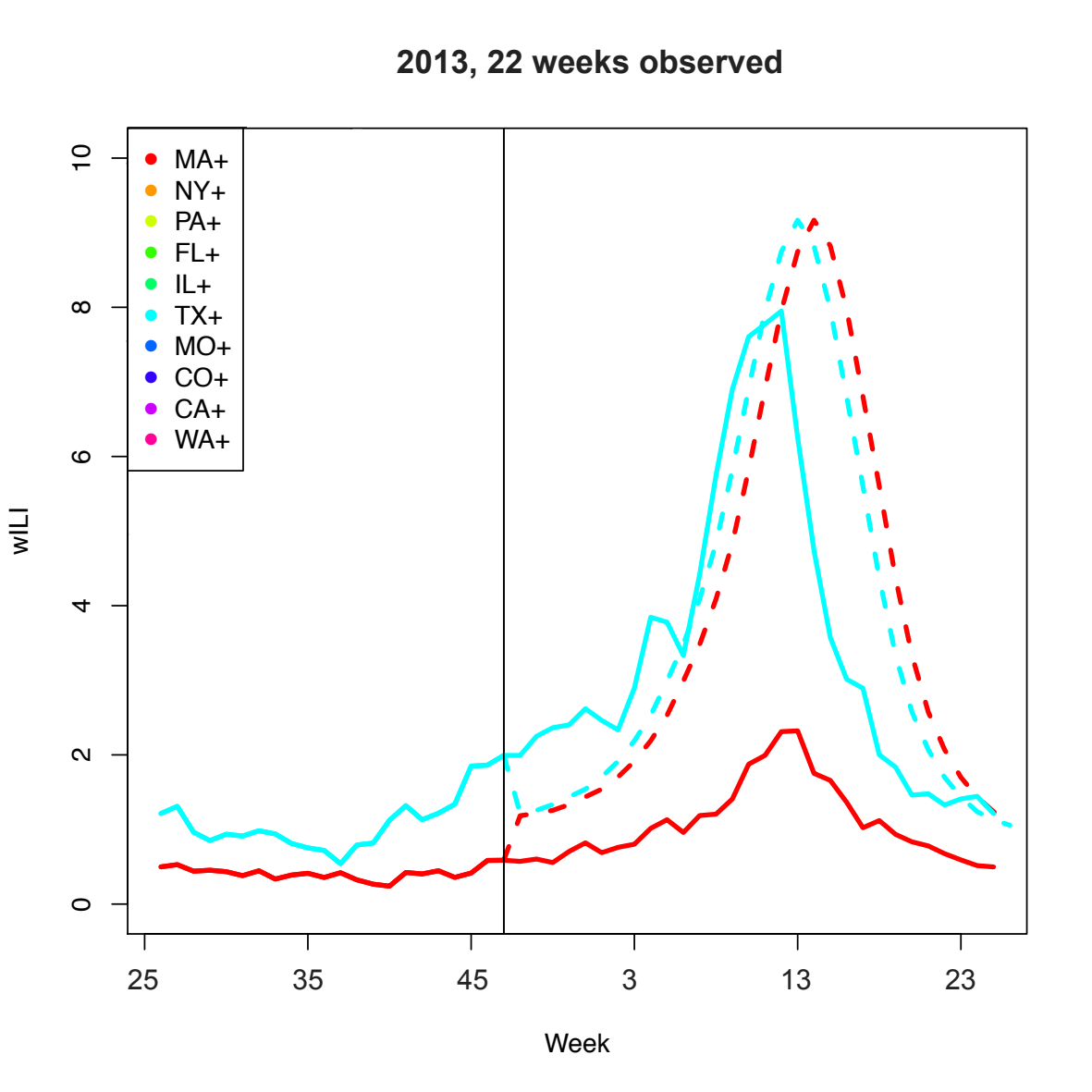
$\implies$ If we choose just 5 values for each parameter, then summing over the regions is 25^{10} possible combinations!

Approximation: Fix the regional variables through min. error.

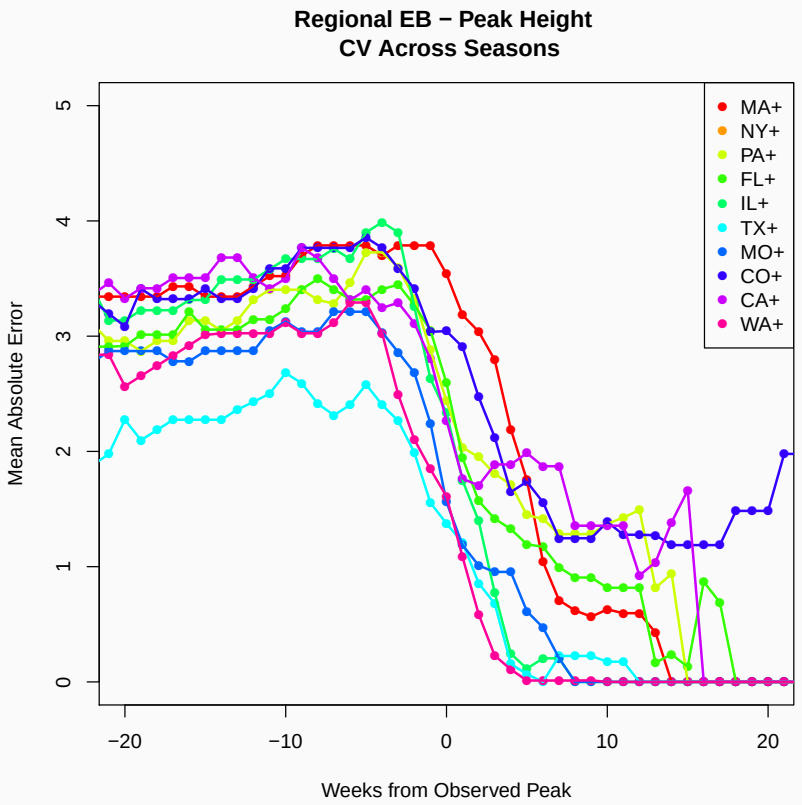
PREDICTIONS FROM EB WITH REGIONAL EFFECTS LOOK LIKE FLU CURVES



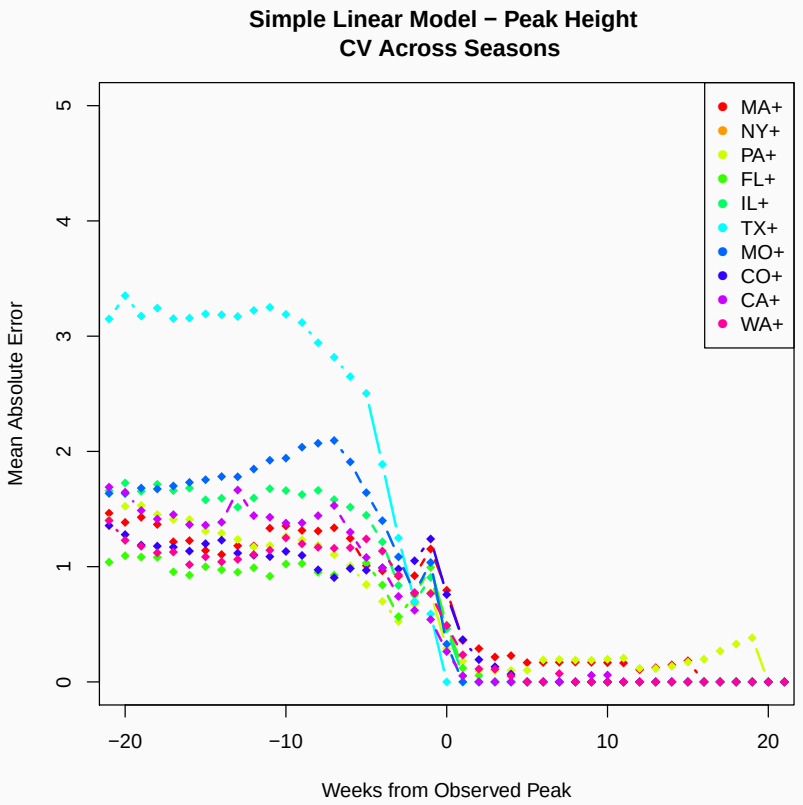
BUT THE APPROXIMATIONS IN IMPLEMENTATION HURT US.



ESTIMATING A POINT IS SIMPLER THAN ESTIMATING A CURVE



(a)



(b)

Figure: 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.

WE TRAIN OUR DATA ON THREE DIFFERENT SETS OF DATA

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

- Model I
  - Complete independence. Train separate model for each region. (All $\beta_{j,i} = 0$)
- Model II
  - United Model. Same model for all regions. (All $\beta_{j,i} = 0$, $\alpha_i^{(r)} = \alpha_i^{(r')}$ for all r, r′)
- Model III
  - Semi-Dependence. Train separate model for each region and impose sparsity penalty on the $\beta_{j,i}$s, $j \in \{0, 1, 2\}$.

VARIABLE SELECTION REVEALS REGIONAL DEPENDENCIES

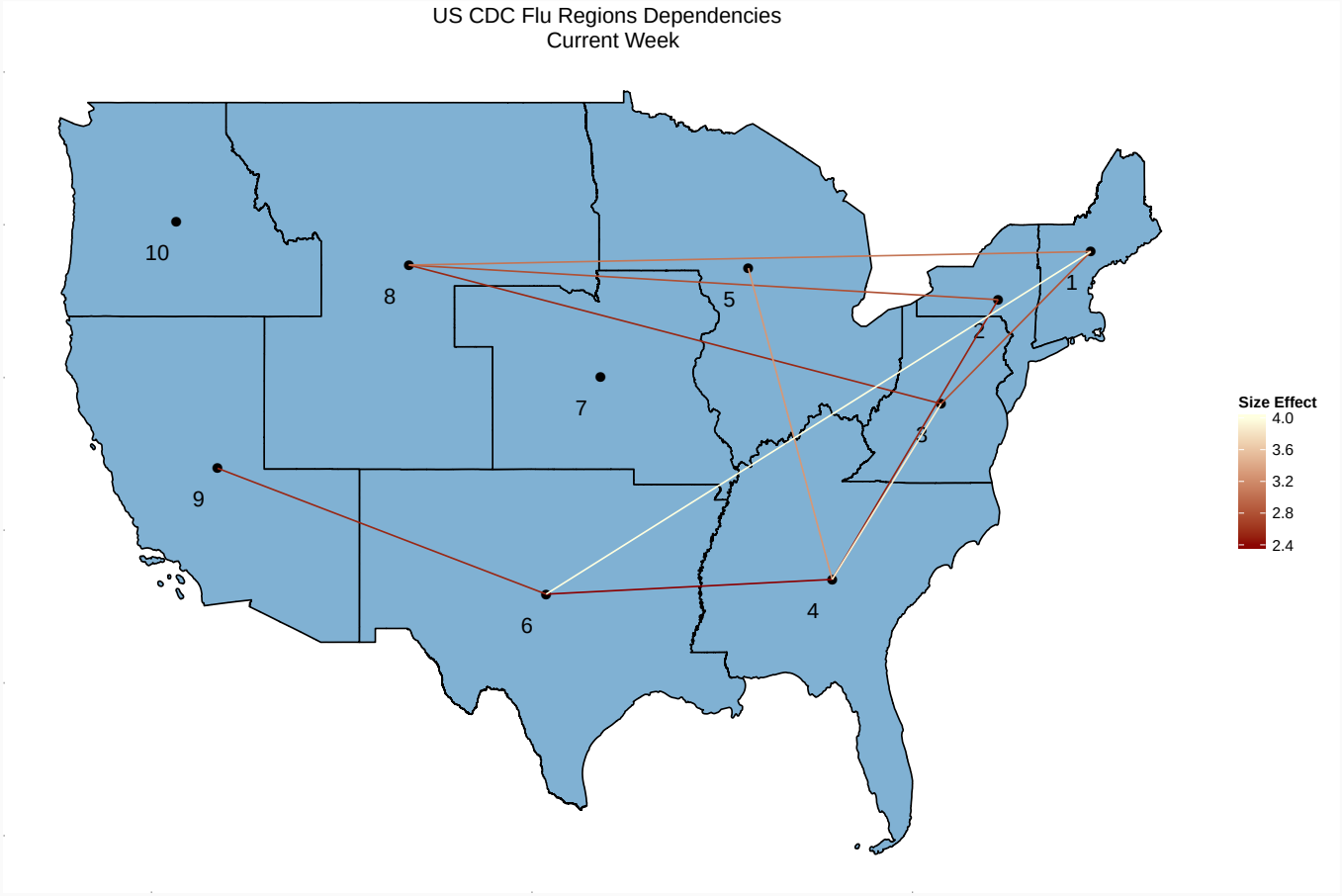


Figure: Arrows colored by size of effect. A line between regions indicates that one of the region's was used in the other region's regression model.

WE ENDED UP FITTING A PARTIALLY LINEAR ADDITIVE MODEL

- Spline fit on the lagged values of the region in question
- Incorporation of regional effects
    - Mixture of Model I and Model III - chosen by lowest CV error

ULTIMATELY, WE WANT TO BIAS ESTIMATES/SHRINK POSTERIOR

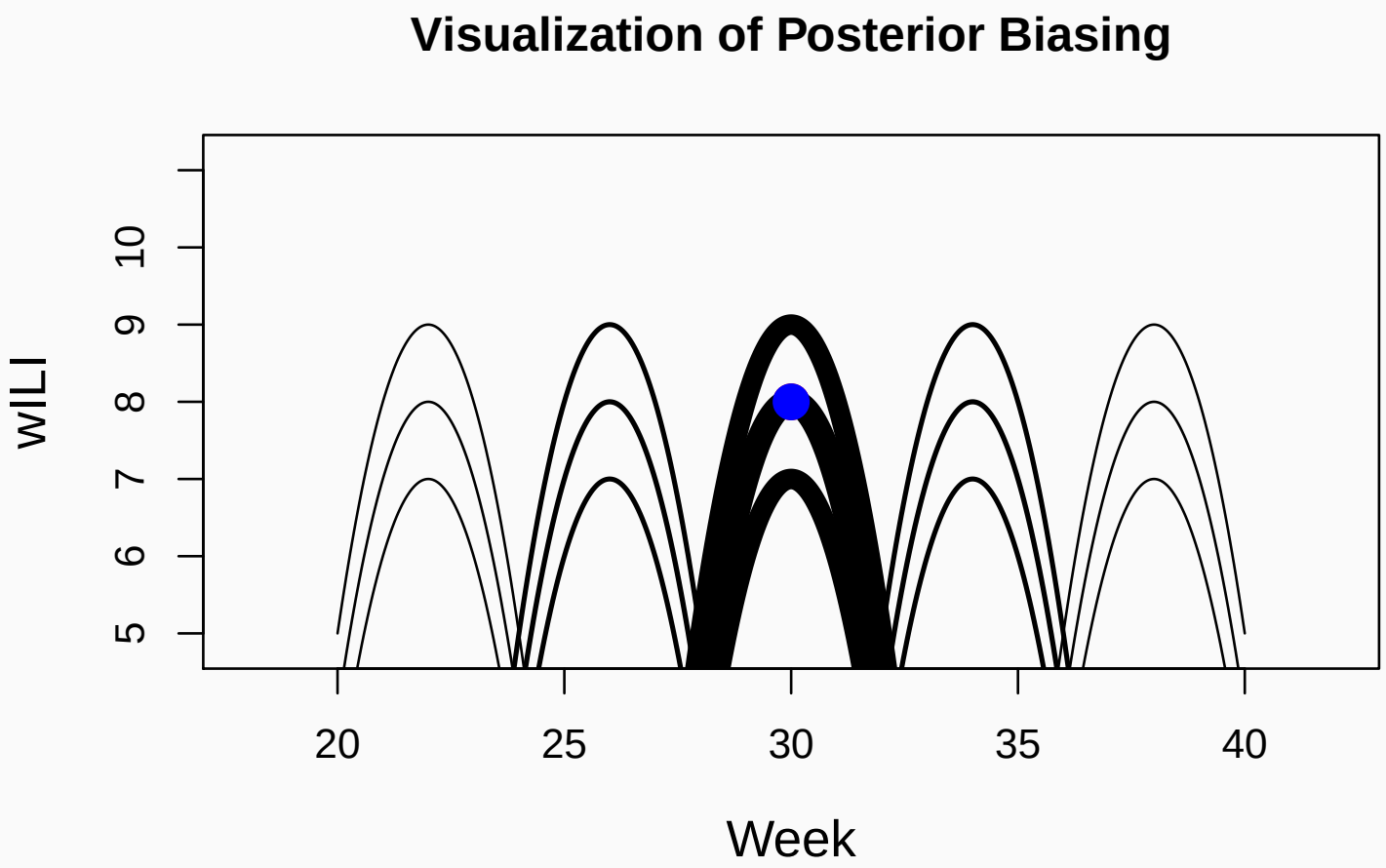


Figure: Image depicting of weighting curves whose peak values are closer to our estimated values. A thicker line represents a larger weight. The blue dot is our estimated value of the peak height and week.

POSTERIOR BIASING YIELDS IMPROVED RESULTS

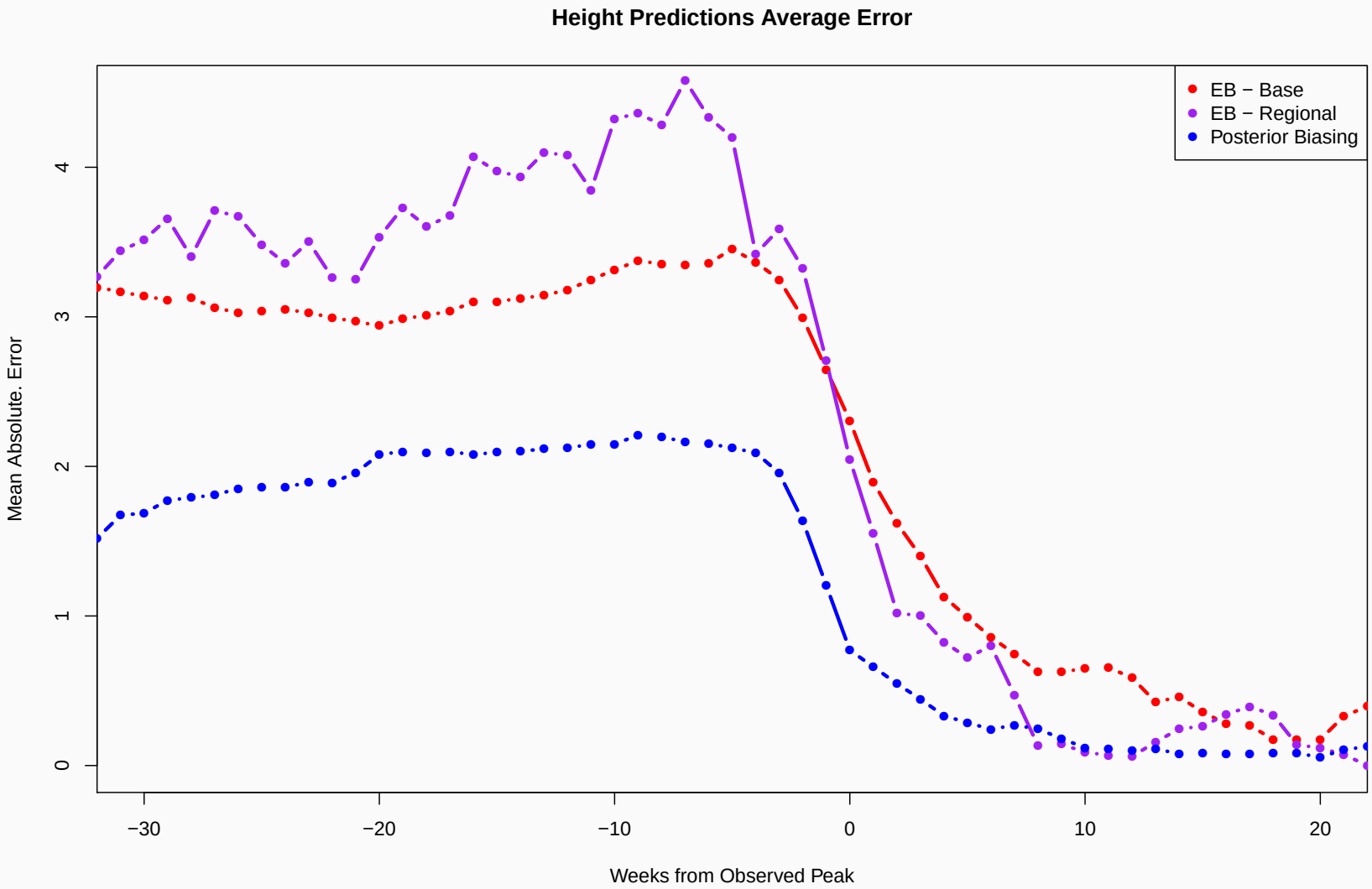


Figure: CV averaged over the different seasons for the different model types.

FINAL RESULTS

WRAPPING UP EVERYTHING

Final Results

Model	Season CV	CV St. Dev.	Pros	Cons
EB	179	40	Baseline	Independ.
EB w/ Reg.	187	48	Regional Effects	Curse of Dim.
Target	57	7	Simpler	No Curve
Post. Bias.	91	28	Flexible	Interpret.

WE ACHIEVED OUR GOALS WITH POSTERIOR BIASING

1. Allow for regional dependencies - Targets use regional data
2. Predict the wILI values for the remainder of the season - Competitive Model

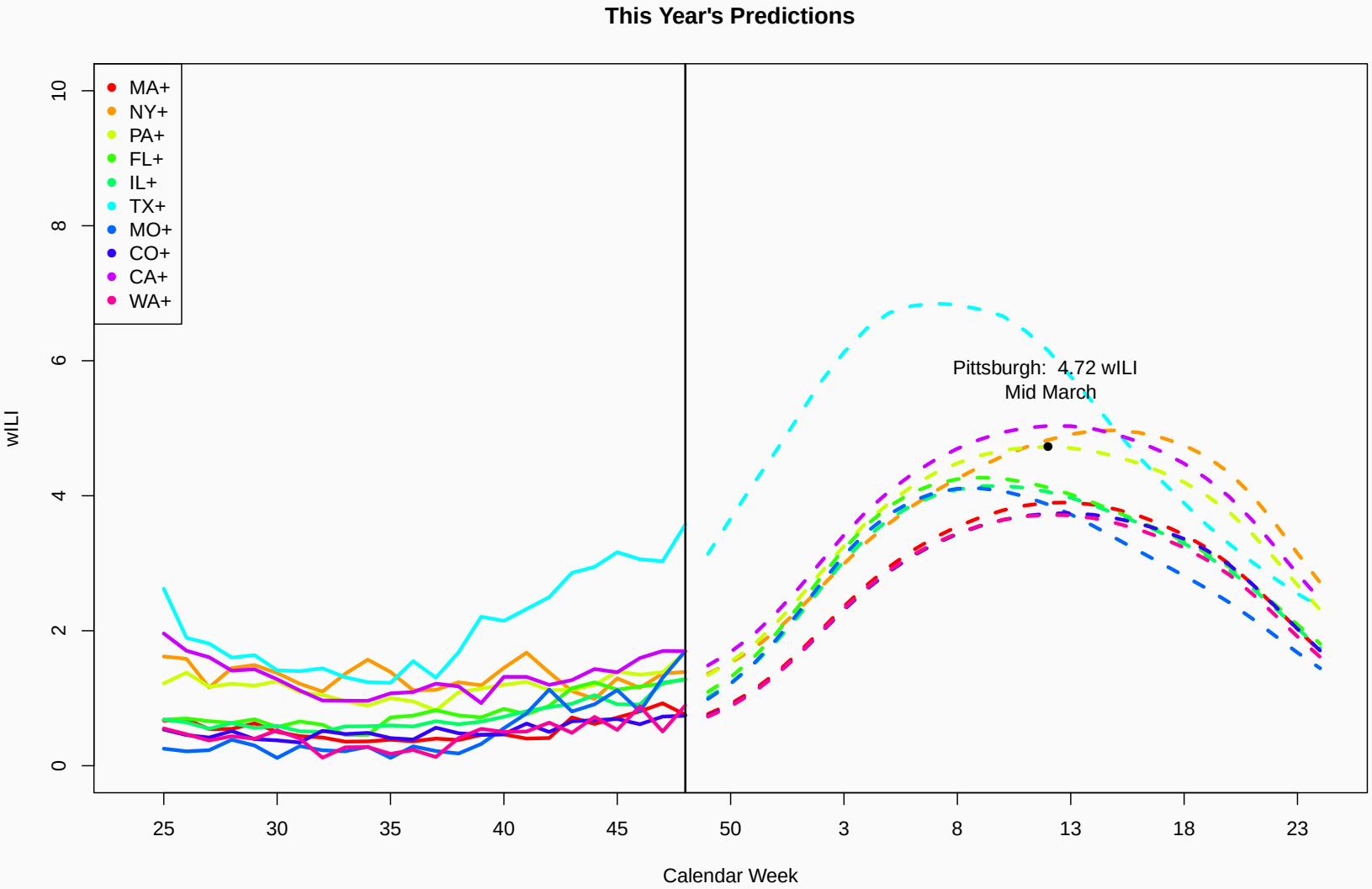
3. Predict specific targets - biased toward targets
    - Peak Week
    - Peak Height/Incidence
4. Produce distributional forecasts - We have entire posterior

5. Explain our model to health care professionals - will talk to other MIDAS researchers

LOOKING FORWARD

- Submit predictions to CDC Flu Prediction Contest
- Quantile Regression
- More regional data

THANK YOU


This Year's Predictions
MA+
NY+
PA+
FL+
IL+
TX+
MO+
CO+
CA+
WA+
wILI
0
2
4
6
8
10
25
30
35
40
45
50
3
8
13
18
23
Calendar Week
Pittsburgh: 4.72 wILI
Mid March