Manuscript

Graham Casey Gibson

4/1/2019

Introduction

Flu Forecasting

Seasonal influenza hospitalizes over half a million people in the world every year~[**?**]. The United States alone, reported an approximate 80,000 influenza related moralities from this past 2017/2018 flu season, the virus targeting the more susceptible elderly and children~[]. In order to combat the flu, the CDC forecasts influenza percentages into the future, moving vaccines and resources to areas where the flu is expected to rise, hoping to attenuate an increase in hospitalization and moralities~[]. Real-time forecasting of influenza suffers from the reporting revision problem, where the estimated influenza like illness (wILI) data is revised as the season progresses. This presents a problem when forecasting because existing models assume that the data used to forecast from is up to date. Improving influenza forecasts by accounting for revisions will directly impact forecast accuracy, and therefore the reliability and usability of forecasts to public health officials.

Existing literature fouses on external data for nowcasting

Much of the current efforts are focused on using external data to improve the estimates of the unrevised data [**?**]. Although this has shown some benefit, the increase in accuracy is both limited and requires access to external data, which is not always available. It is also concievable that adjusting the current observed data by historical revisions may capture the true ILI better than a noisy internet based signal.

Historical Revisions

Thanks to the efforts of CMU [**?**] we have historical revisions for all weeks from the 2010/2011 to 2017/2018 seasons. We use this data to model revisions to ILI up to the 2014/2015 season, and use the remainder of the seasons as a test set.

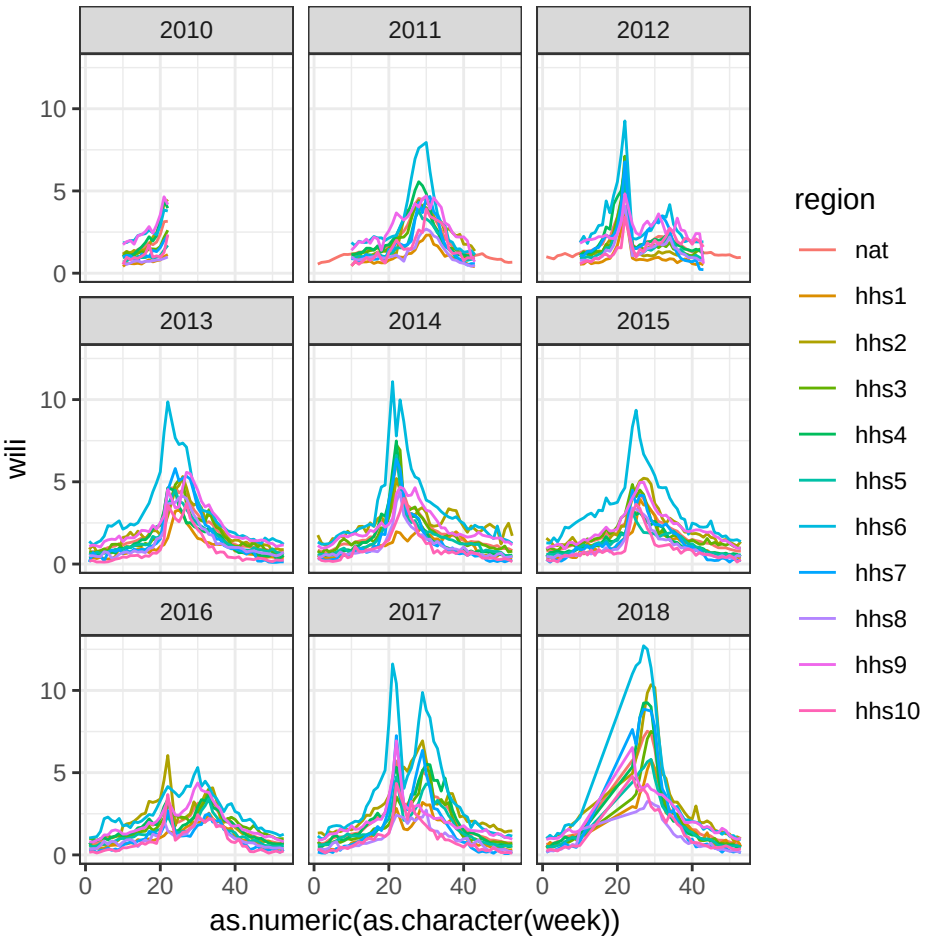
Problem statement

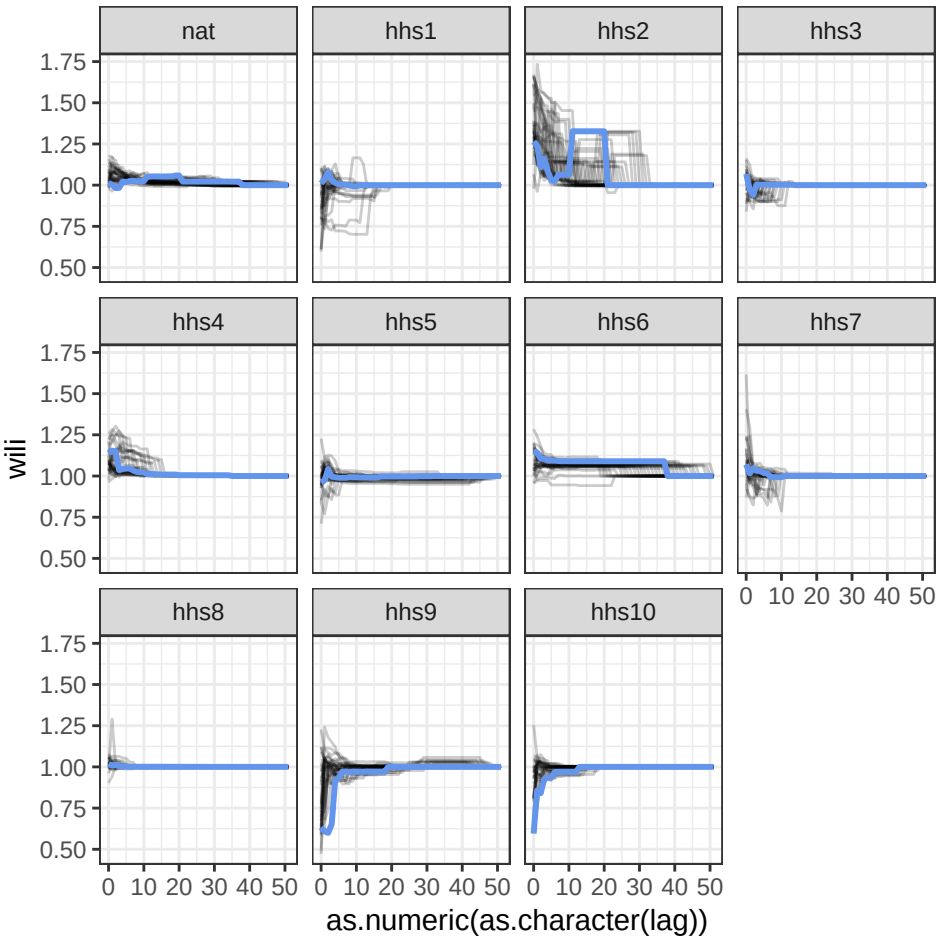
The above questions lead to the following three problem statements

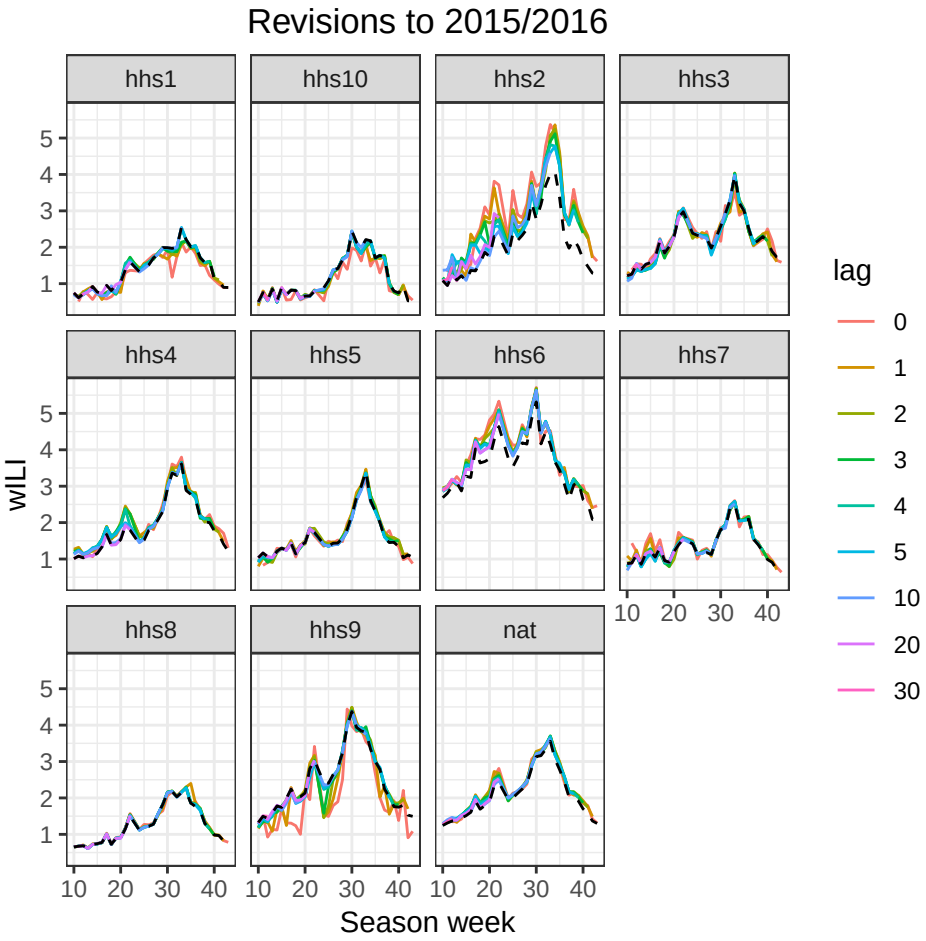
1) **Does using historical revisions to estiamte the true ILI increase forecast accuracy accross the CDC defined targets**.
2) **Do the results vary by choice of revision model?**
3) **Do the results vary by region and season?** In order for a method to be practical, we need to enfore some sort of lower-bound on accuracy, that is can we show that accuracy does not decrease.

Methods

Surveillance Data




nat
hhs1
hhs2
hhs3
hhs4
hhs5
hhs6
hhs7
hhs8
hhs9
hhs10
wili
1.75
1.50
1.25
1.00
0.75
0.50
0
10
20
30
40
50
as.numeric(as.character(lag))



Since 2013, the CDC has released wILI percentages per epidemic week and invited others to submit their own predictive models of the flu. The CDC combines these submitted predictive models to better inform 7 targets: 1,2,3,4 week ahead wILI percentages, the onset of the flu season (defined as the week where 3 consecutive weeks are above the baseline, the epidemic week where wILI peaks and that peak percentage~[]. During the season the CDC updates past epidemic week wILI percentages, collecting previously unreported data from clinical sites and updating the number of positive influenza tests and patients screened. A predictive model that accounts, not only for the dynamics of the flu, but for revisions throughout the season, will likely be less susceptible to past wILI changes and also more accurately predict future wILI percentages.

As we can see from Figure 1 the revised wILI data is highly seasonal and varies by region. Succesful models in the Flusight Network have historically takent his into account.

Data revisions

We denote the observed wILI value for a given region r season s and week w at week $w + l$ as.

$$Y_{r,s,w,l}$$

Borrowing from the notation of [**?**] we denote the finally revised data as

$$Y_{r,s,w,\infty}$$

where $l = \infty$ denotes the final revised value.

Reporting revision ratios

We use the ratio of currently reported wILI to finally observed wILI at a given region r, season s, week w, lag l as a central concept in our model.

$$a_{r,s,w,l} = \frac{Y_{r,s,w,l}}{Y_{r,s,w,\infty}}$$

As we can see from Figure 2, wILI values for a given r, s, w start off anywhere from 160% to 60% reported, relative to their final value. Although by lag 30 most wILI values are fully reported (ratio of 1), there is significant variability across epiweeks and regions.

Mean scale up

In order to estiamte the revised data $\hat{Y}_{r,s,w,\infty}$ we apply the simple estimator

$$\hat{Y}_{r,s,w,\infty} = \frac{Y_{r,s,w,l}}{\hat{a}_{r',s',w',l'}}$$

where we model $\hat{a}_{r,s,w,l}$ in a variety of methods descrbied below$

- $\hat{a}_{r,s,w,l} = a_{\cdot,\cdot,\cdot,l} = \frac{1}{N}\sum_{r,s,w} a_{r,s,w,l}$
- $\hat{a}_{r,s,w,l} = a_{\cdot,\cdot,w,l}$
- $\hat{a}_{r,s,w,l} = a_{r,\cdot,\cdot,l}$
- $a_{r,s,w,l} \sim N(\alpha + \beta_w + b_r, \sigma^2),\ \tilde{b} \sim N(0, \Sigma)$
- $\hat{a}_{r,s,w,l} = g(\alpha + \beta_w + \gamma_r)$

Sampling

Sample blue point from Fig. 4 and set

$$\hat{Y}^i_{r,s,w,\infty} = \frac{Y_{r,s,w,l}}{\hat{a}^i_{r,s,w,l}}$$

for $i \in 1:1000$. Now we have an empirical distribution of the form

$$P(Y_{r,s,w,\infty} = y) = \frac{1}{1000}\sum \mathbf{I}(\hat{Y}^i_{r,s,w,\infty} = y)$$

Applying scale ups to currently observed data

In order to apply both the scale up model to all of the currently observed data within a season up to week w, we develop the vector notation,

$$\vec{Y}_{r,s,w} = \{Y_{r,s,1,w-1}, Y_{r,s,2,w-2}, ..., Y_{r,s,w,0}\}$$

We then apply both the scale up method and the sampling method to the whole trajectory of currently observed data.

Forecasting Model

In order to forecast wILI across all HHS regions we use the SARIMATD method from the sarimaTD package []. We use the default options that include seasonal differencing and box-cox transformation of the underlying ili data to normality. See SARIMATD for further details.

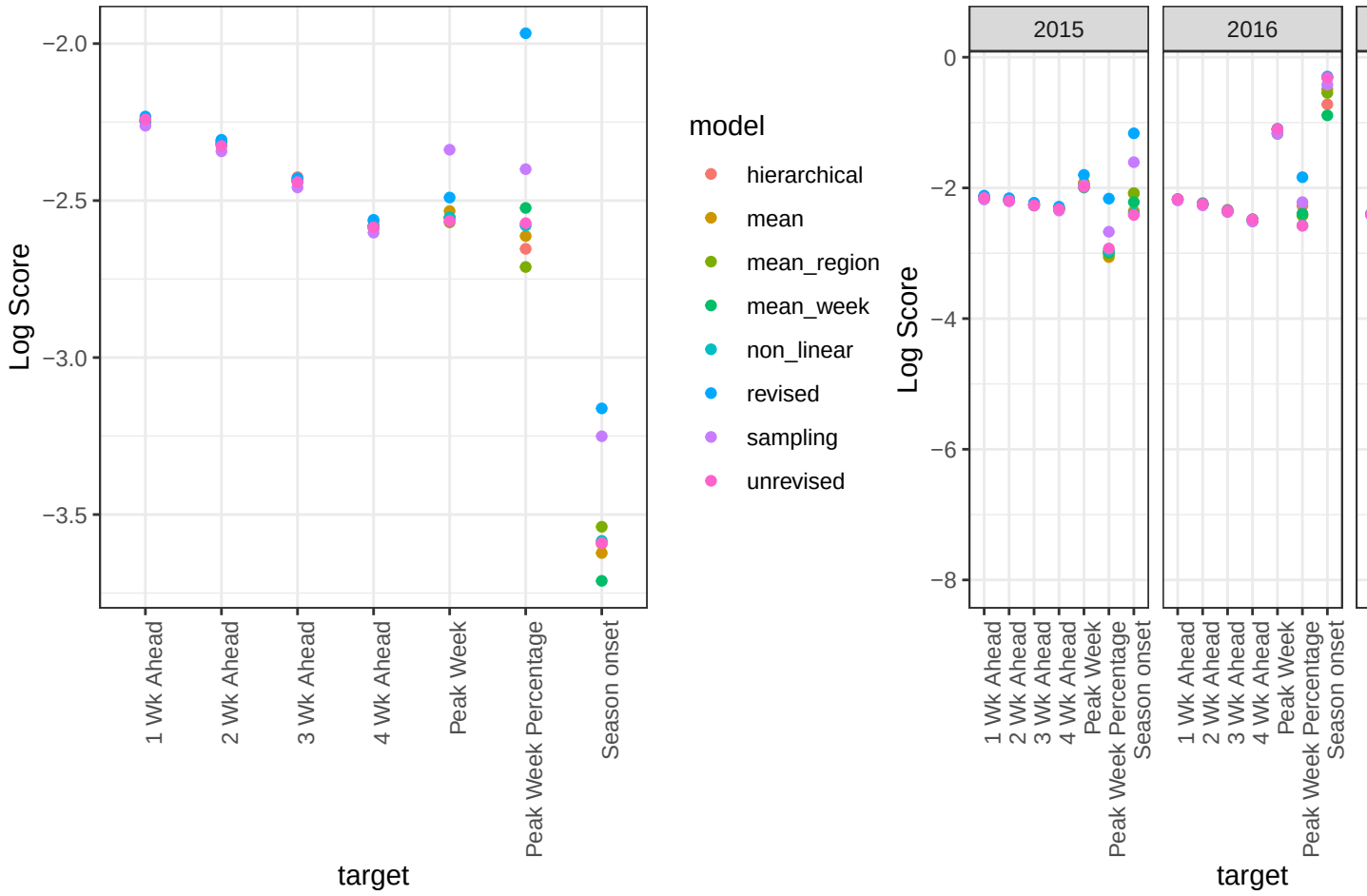
Evaluation

In order to evaluate our models for revision we choose to evaluate our estimates of the revised data ($\hat{Y}_{r,s,w,\infty}$) by forecasting the targets from the augmented data.

Region	Season	Week	Data Used	K-step Ahead Target	Season onset	Season peak week	Season peak
r	s	1	$Y_{r,s,1,0}$	$Y_{r,s,1+k,\infty}$	$w\ s.t. Y_{r,s,w,\infty} \geq$ onset	$argmax_w Y_{r,s,w,\infty}$	$max_w Y$
r	s	2	$Y_{r,s,1:2,1:0}$	$Y_{r,s,2+k,\infty}$	$w\ s.t. Y_{r,s,w,\infty} \geq$ onset	$argmax_w Y_{r,s,w,\infty}$	$max_w Y$
...	...	...	...	...	...	...	...
r	s	20	$Y_{r,s,1:20,19:0}$	$Y_{r,s,20+k,\infty}$	$w\ s.t. Y_{r,s,w,\infty} \geq$ onset	$argmax_w Y_{r,s,w,\infty}$	$max_w Y$

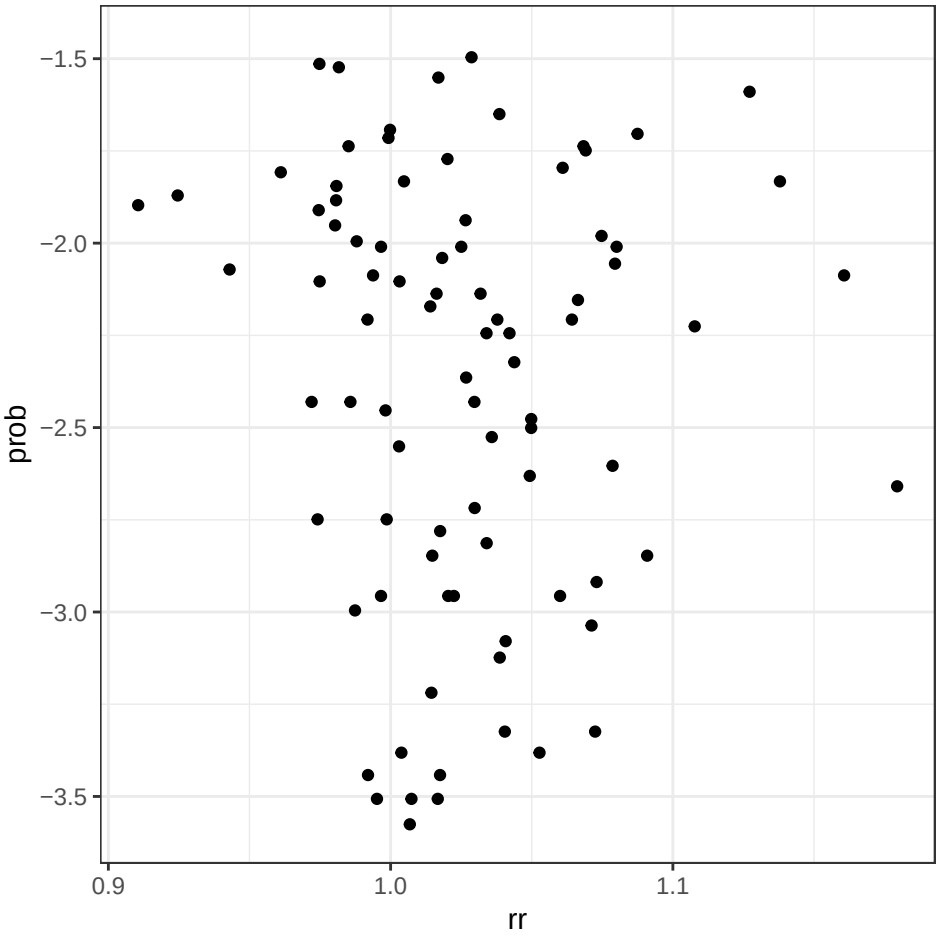
Table 1: Forecast template for year t

Results

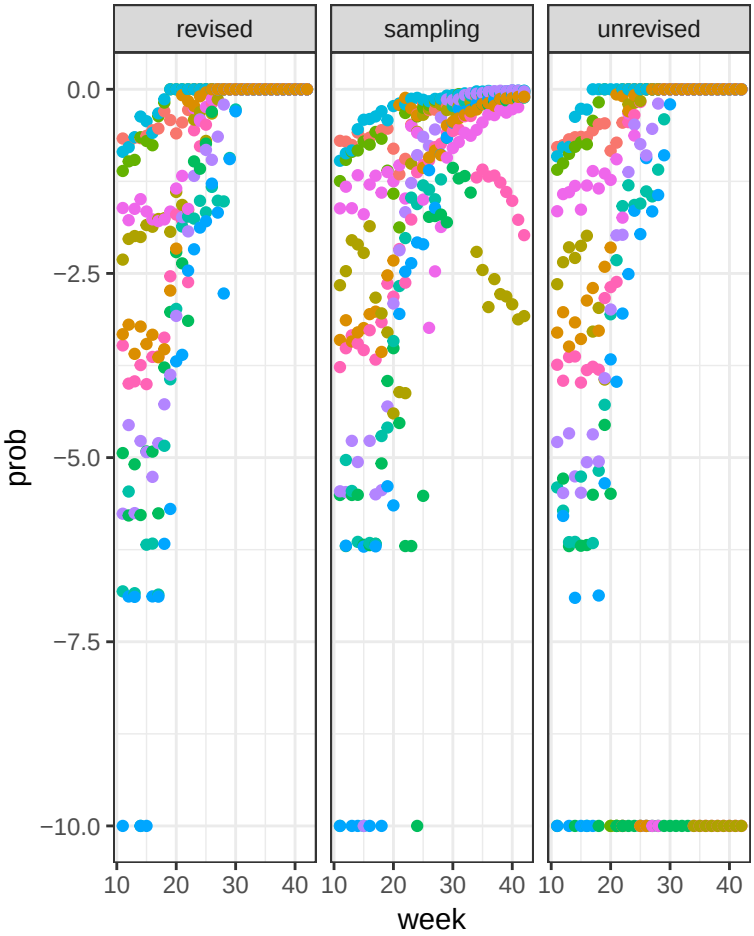


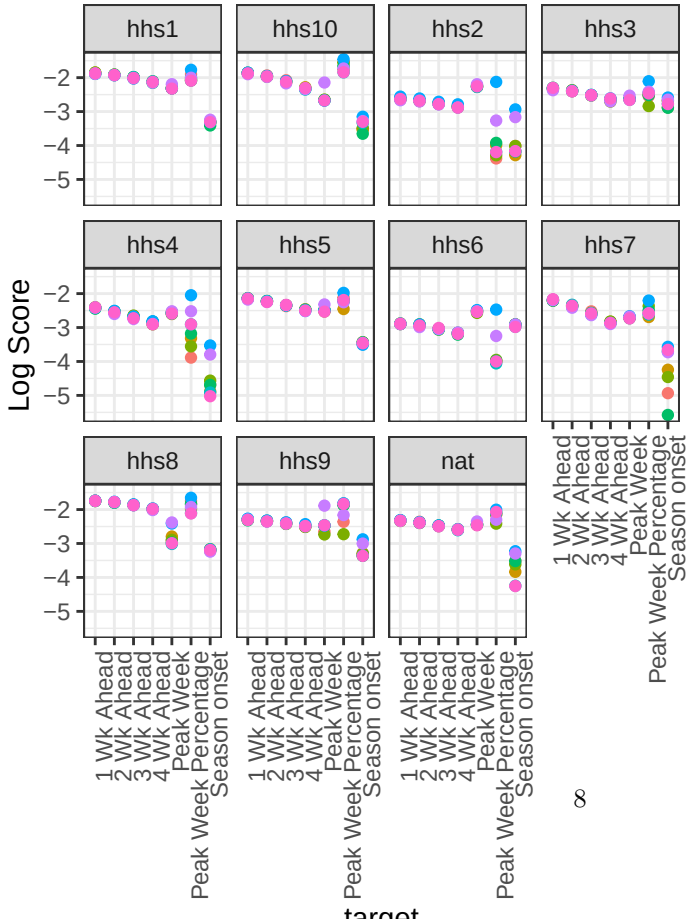
1-4 Week Ahead is largely unaffected

```
## Warning: Removed 7 rows containing missing values (geom_point).
```



Seasonal targets are mostly improved by sampling





High Regional variabil-

Conclusion

Ref

@article{osthus2019even, title={Even a good influenza forecasting model can benefit from internet-based nowcasts, but those benefits are limited}, author={Osthus, Dave and Daughton, Ashlynn R and Priedhorsky, Reid}, journal={PLoS computational biology}, volume={15}, number={2}, pages={e1006599}, year={2019}, publisher={Public Library of Science} }

@article{ray2017infectious, title={Infectious disease prediction with kernel conditional density estimation}, author={Ray, Evan L and Sakrejda, Krzysztof and Lauer, Stephen A and Johansson, Michael A and Reich, Nicholas G}, journal={Statistics in medicine}, volume={36}, number={30}, pages={4908–4929}, year={2017}, publisher={Wiley Online Library} }

@article{ title={https://github.com/reichlab/sarimaTD} }