Ongoing Influenza Activity Inference with Real-time Digital Surveillance Data

Lisheng Gao
Department of Machine Learning
Carnegie Mellon University
lishengg@andrew.cmu.edu

Friday 8th December, 2017

Background. In order to track influenza activity, flu-like patient records collected from health care providers are combined to generate an estimate every week during the flu season. However, this estimate is problematic in terms of timeliness and granularity. A promising alternative is to infer flu activity with fine-grained, timely digital surveillance data. Fitting these extra measurements of flu activities will ideally generate real-time estimate of disease spread, at least one week ahead of official report.

Aim. The goal of this project is to make reliable real-time inference for flu activity with both historical flu data and digital surveillance data. The primary focus is on regional level, which could be extended to state or county level given finer-grained flu activity measurement.

Data. Considered a gold standard, ILINet Surveillance Data (ILINet in short) is flu-like activity record collected by Centers for Disease Control and Prevention (CDC). This data contains wILI - the percent of influenza like illness (ILI) cases out of total cases weighted by state population - that available for nine census regions, the ten Health and Human Services (HHS) regions, and for the US as a whole. As for digital surveillance data sources, this project mainly conducts analysis for flu lab test data (lab data for short) and thermometer measurement data (thermo data for short).

Methods. Each digital surveillance data provides some measurements for flu activity. Regression metrics will be used to produce inferences of flu activity based on historical measurements and wILI. The set of inferences are called sensors. Those sensors generated from different data sources will finally be fused to make a more reliable inference of flu activity for current week.

Results. The outcome of this project contains detailed analysis of new data sources and a better understanding of the end-to-end system for real-time flu activity inference used by CMU Delphi Group.

1 Introduction

Influenza has been a serious threat to human health for long. The highly infectious nature of its viruses resulting in a significant number of patients each year. Also, some subtype of flu viruses may lead to severe respiratory diseases and even death, especially for the elderly. However, there are many protective methods that can largely diminish the intensity of flu activity. Taking vaccine is a very effective way for people to gain the immunity towards influenza. In order to block or impede the spread of viruses, methods including frequent hygiene checks and reduced public gatherings are helpful as well.

All these methods have societal and monetary costs, which presents a challenging task for policy makers to find the appropriate time to take actions. Hence, it is necessary to keep close track of the ongoing flu activity and to be alerted of flu outbreaks. ILINet Surveillance Data (ILINet in short) is flu-like activity record collected by Centers for Disease Control and Prevention (CDC)[3]. CDC issues weekly influenza surveillance report which contains a measurement of flu activity intensity[4]. However, this record is problematic in terms of timeliness, granularity and stability.

An alternative to solve this problem is using multiple digital surveillance data sources to infer the flu activity. For example, more appearance of flu related terms in Google search queries[6] or Twitter posts[1, 5] could be clear evidence of more flu cases. Meanwhile, these data sources are available in real-time with excellent geographical granularity. The focus of this project is to find a systematic approach to combine different data sources and generate reliable ongoing flu activity inference.

2 Problem Definition

In this section, the optimization target for ongoing flu activity estimation will be defined. First, it is important to clarify a few flu related terms used by CDC:

Epiweek. An epidemiological week (epiweek) is defined as beginning on Sunday and ending on Saturday. The first epiweek of the year is the week of the first Wednesday in that year.

Flu season. The official flu season in the US starts on epiweek 40 (early October) and runs through epiweek 20 (mid May) of the following calendar year.

HHS region. The U.S. Department of Health and Human Services (HHS) divides the country into ten administrative regions, which are somehow aligned with climatic conditions. Figure 2 shows the division for the regions.

WILI. The percent of ILI cases out of total cases reported by health care providers on volunteer basis gives a number called %ILI (or, simply ili), which provides a measurement of flu activity in each state. wILI — ili weighted by state population — is available in the nine census regions, the ten Health and Human Services (HHS) regions, and for the US as a whole.

For this project, the estimation target is (HHS) regional wILI for each in-season epiweek. At epiweek ew_t, with historical wILI records for $wILI_1, ..., wILI_{t-1}$ and digital surveillance data $d_1, ..., d_t$ available, a point estimation $w\hat{IL}I_t$ will be generated. This process is called "Nowcast" since it infers the flu activity happening now. This can also be considered a "+1" week wILI prediction as it generates wILI estimation one week ahead of the official report.

3 Data

In this project, the major data sources are wILI records, flu lab test data and thermometer measurement data. There are several auxiliary digital surveillance datasets mentioned for comparing sensors' performance.

3.1 wILI

Although wILI is considered the "gold standard" for flu activity measurement, there are some problems with this data. One of the major caveats is that, the wILI report is subject to revision for several subsequent weeks due to the delays of records collection. This problem is called "backfill". Figure 1 shows significance of this sort of revisions. As a result, we call the wILI reports that are likely to be changed "unstable wILI", while the final wILI values are referred as "stable wILI". After epiweek 30 in each year, the "ground truth" of wILI reports during the past flu season is available. Another issue of wILI data is that it is published with one or two weeks' delay, which motivates our nowcast inference.

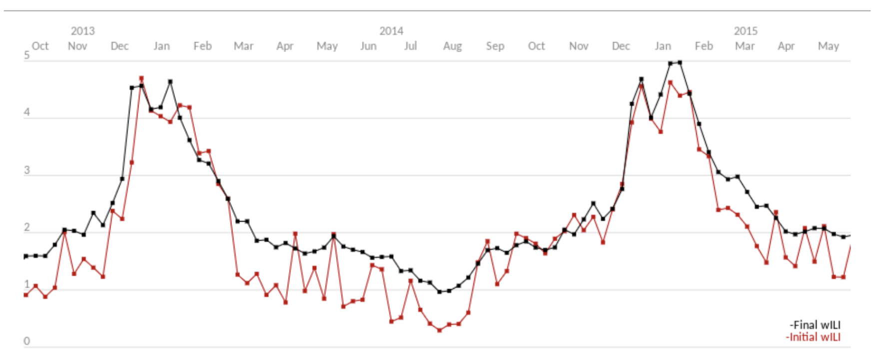

Figure 1: Backfill problem in region HHS9 from David Farrow's Thesis[7]

3.2 Flu lab test data

Flu lab test data (lab data for short) is daily lab test results returned by health facilities since Tuesday, week 35, 2015. Each record contains test result for flu subtype A, flu subtype B and flu in general, together with the patient's age. Some extra information for the facility and device, including the device serial number, facility identification and facility type, is included in the data as well. Geographically, this data is fine grained, down to the level of zip5. It also has fields describing state, city and county that the test taken place.

Spatial distribution of lab data

Since this project focuses on regional inference for real-time flu activity, it is important to look at the spatial distribution of lab data, especially the volume for each region. Figure 2 shows the distribution of flu test devices deployed across the nation in 2015. Comparing with the division of hhs regions, we can conclude that lab test records are more dense in east coast regions. The data volume for northwest regions was smaller at that time.

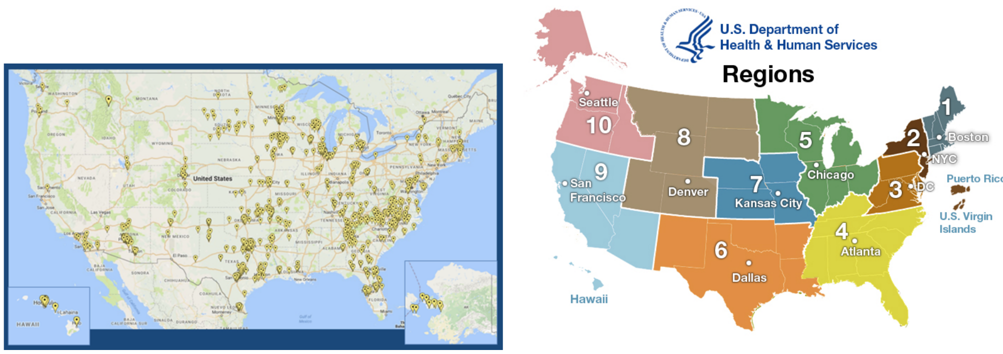

Figure 2: **(left)** Geographical distribution of lab test devices in 2015, provided by the company.**(right)** CDC's Definition of HHS Regions.

Figure 3 showed the detailed story for data volume across regions. The y axis is the total number of test records seen within a week in log scale, while x axis is time axis. We can observe that region 4 and regions 6 have greatest amount of records. For region 2 and 10, missing data points appear from time to time until April 2016. The data volume in region 2, 8 and 10 are small, which can introduce large noise during inference. Hence, it is reasonable to exclude these three regions for the inference task.

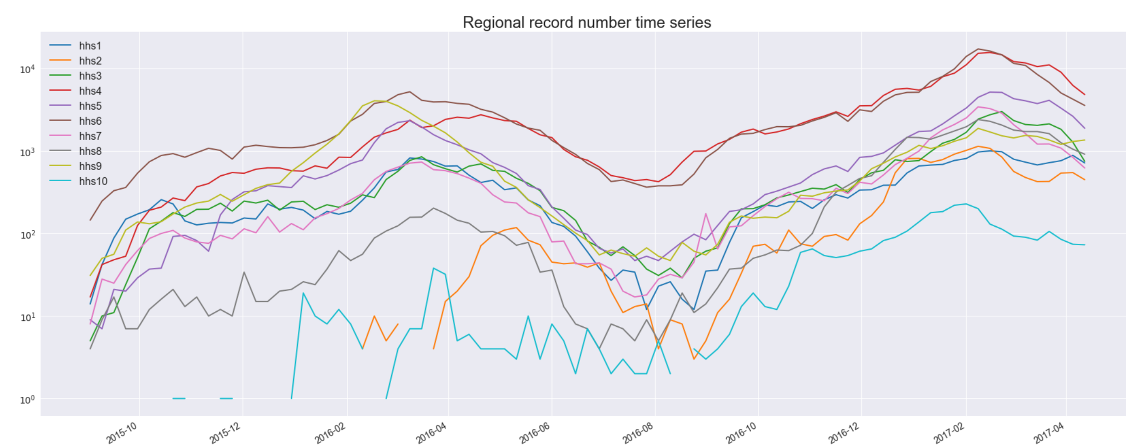

Figure 3: Spatial-temporal record distribution in log scale for lab data.

Temporal distribution of lab data

One important temporal characteristic of this data is the increasing trend in data volume across the years. From Figure 3, it can be seen that for most regions, there is a significant increase in record volume. The exceptions are region 1, in which the total number of records are relatively stable, and region 9 where the data volume decreases. This trend is possibly because of the deployment of more devices in more health facilities. In order to generate more reliable inference based on historical record, we have to find certain ways to de-trend this data.

3.3 Thermometer measurement data

Thermometer measurement data(thermo data for short) consists of body temperature readings. The symptom data is not used for now since it is relatively sparse and needs extra efforts to parse. The thermo data includes an id for user identification together with the age and gender disclosed on volunteer basis. Each measurement record contains the temperature reading, device id, measurement time in UTC(+0) and the measurement location.

Spatial-temporal distribution of thermo data

The thermo data is available since 2015 epiweek 35. For each hhs region, there are at least hundreds of measurements collected each week. So we can expect thermo data to generate reliable inferences for all regions. Meanwhile, there is no significant change in data volume across years, hence de-trending operations may not be necessary at this point.

Potential issues for thermo data

There are several issues to be handled for thermo data. First, a large proportion of data does not have detailed patient information (identification, age, gender). This will be problematic if we would like to distinguish the users. Second, the time stamp is in UTC time, which means that in order to convert the timestamp back to the real time at the measurement location, specific time shifting must be applied.

4 Metrics

Given various digital surveillance input, the final output of this pipeline is a point estimation for ongoing wILI, which could be national, regional or state level inferences for flu activity. This process is called Delphi Nowcast. Online visualization is available for real-time flu activity inference[8]. Comparing with the official wILI report, the nowcast inference has two advantages: (a) the nowcast inference is real-time while official report has an one-week delay; (b) since the digital surveillance data is final once collected, the nowcast inference does not have the backfill problem that offical wILI has.
The whole “Nowcast” procedure is designed by David Farrow[7]. The pipeline contains three steps. First, each data is collected, pre-processed and aggregated to a series of weekly “measurements”. For example, the “measurements” for wiki data is the weekly traffic of several influenza related pages, while the “measurements” for flu lab test data could be the number of positive records during that week. Due to the different level of geographical granularity of datasets, the measurements are in national, regional or state level. Some of the state level measurements are aggregated to regional

level since finer-granularity could lead to much larger noise. Second, the measurements are fit to historical wILI using a weighted linear regression approach to create sensor. Sensors are assumed to be unbiased estimations of stable wILI with IID Gaussian noise. Third, sensors are passed through sensor fusion step, which produced the nowcast inference by linearly combining the sensors based on the estimated precision matrix of each sensor's historical prediction error.

4.1 Sensor Generation

In this step, weighted (potentially multiple) linear regression is applied to each measurement to generate the sensors. Taking weighted multiple linear regression of Wiki as an example, there is a series of measurements available, denoted as $M = [m_1, ..., m_t]$, where m_i is the measurement that derived i weeks ago. In addition, m_0 is the measurement for current week. Each measurement is a vector of form $[m_{i1}, m_{i2}, ..., m_{ip}]$, where m_{ij} is the traffic volume of a certain Wiki page j and p is the total number of selected influenza related pages. The historical wILI series is $V = [v_1, v_2..., v_t]$, with v_0 missing. Before the regression, a weight term $w_1, w_2, ., w_t$ is applied to each data point. The regression weights were designed for the following purposes: (a) samples at or near to the same week of year should be given more weight due to the seasonality of flu, (b) sample weight should fall exponentially as time passes, and (c) very recent samples should be penalized as wILI on these weeks is subject to revision. In mathematical form, the weights are defined as:

$$
\begin{aligned}
year =& 52.2 \\
a =& \frac{1}{20} + \frac{19}{20} \exp\left(-\min(i \bmod year, year - i \bmod year)/2)^2\right) \\
b =& 2^{-(i/year)} \\
c =& 1 - 2^{-i} \\
weight(i) =& a \times b \times c
\end{aligned}
$$

Let W denotes the t-dimensional diagonal matrix of all weights $w_1, w_2, ..., w_t$, we can obtain the model parameters:

$$\hat{\beta} = \arg\min_{\beta} W\|M\beta - V\|_2^2 = (M^T W M)^{-1} M^T W V$$

The inference of %ili for states or wILI for regions during current week is simply $\hat{v}_0 = \hat{\beta}^T m_0$.
It's worth noting that this model is re-trained every week. There are several reasons for this design. The wILI data is subject to "backfill", so the model should be adjusted as the data is corrected. We also suspect that the model could evolve along time due to the seasonality of flu and some characteristics of measurement time series. For example, Wiki page visit counts will get larger if Wiki's user base increases. Meanwhile, some digital surveillance sources are available for only a short period of time, so re-training of model helps take advantage of newly collected data points.
Another issue for both digital surveillance data and wILI report is the quality of off-season data. Hence, only in-season data points are used when prepare the training data in this project, while Delphi Nowcast does consider off-season data for sensor generation.

4.2 Sensor Fusion

Derivation of sensor fusion kernal

The sensor fusion step is performed with a sensor fusion kernel derived from a Kalman Filter[9]. Since the canonical process of flu is unknown, the posterior error covariance matrix $P_{t|t-1}$ becomes infinity and the Kalman Filter is then a one-step update procedure. The remaining problem is how to compute the Kalman Gain K, error covariance matrix $P_{t|t}$ and generate the prediction x.
First, following the original Kalman Filter design, we can express K, P as:

$$
\begin{aligned}
K =& (P_{t|t-1}^{-1} + H^T R^{-1} H)^{-1} H^T R^{-1} \\
P_{t|t} =& (I - KH) P_{t|t-1} (I - KH)^T + KRK^{-1}
\end{aligned}
$$

H is a matrix mapping state space to observation space (we avoid using "measurement" again to reduce confusion), where the observation would be the sensors generated in previous steps. R is the error covariance matrix of all observations. The Kalman Gain term is derived using matrix operations and the error correlation matrix is written in "Joseph form"[2]. From here, we can eliminate the infinite error covariance $P_{t|t-1}$ and get:

$$
\begin{aligned}
K =& (H^T R^{-1} H)^{-1} H^T R^{-1} \\
P_{t|t} =& (H^T R^{-1} H)^{-1}
\end{aligned}
$$

The derivation of K is straight forward while for $P_{t|t}$ the properties of vector covariance is used. David Farrow's thesis includes all mathematical details[7]. Finally, plugging in these two matrix to the update step for state in Kalman Filter, we have:

$$x = (H^T R^{-1} H)^{-1} H^T R^{-1} z = P H^T R^{-1} z$$

Within this specific Nowcast system, the inputs for sensor fusion kernel are national, regional and state-level sensors for all past weeks. x is estimated wILI for all states. So, the error correlation matrix R has dimension $M \times M$ where M is the total number of sensors, and mapping matrix H has dimension $M \times S$ with S being the number of states. Since x is on state level, there is an additional step to linearly combine x to regional and national estimation y.

Estimation of error correlation matrix

One significant issue for sensor fusion kernel is that different sensors are available for different period of time. Some of them has existed for more than 5 years while others are only available for less than 2 years. Also, some of the sensors are only available during flu season. Hence, instead of generating the full error correlation matrix R at one time, the sensor fusion kernel computes pair-wise correlation of sensors and approximate the correlation matrix by adding a shrinked identity matrix to pair-wise correlation matrix to create a positive semi-definite, invertible matrix. This helps the kernel use as much as historical data as possible.

Alternative interpretation of sensor fusion kernel

It is important to know that when the error correlation matrix R is error covariance matrix and mapping matrix $H = \mathbf{1}$, the sensor fusion kernel is equivalent to multiple linear regression. This also explains the potential advantages of sensor fusion kernel over linear regression model. First, it uses more historical data to estimate R, which could reduce the noise. Second, it takes advantage of sensors available for different geographical level to improve inference accuracy.

5 Experiments

5.1 Flu lab test data analysis

For the lab data analysis, we are interested in (a) which combination of covariates can generate the best inference; (b) how does the inference perform comparing to other sensors; (c) by how much it can help the final inference - nowcast - after being passed through the sensor fusion step. Since there is only two years of data available, we are not confident how stable the performance will be in the future.

Covariates analysis

Our purpose for this step is to generate the best inference for the flu activity during current week, which is also a "+1" week point wILI prediction since there is an one-week lag for wILI report. The method used is the weighted linear regression described in previous section. In addition, since we do not have enough training points, we would like to limit the number of covariates used to 1 or 2.

The experiment is designed to use four sets of covariates. For simplicity, let's denote "flu type A (B) positive records count" as "flu A (B)", "flu in general positive records count" as "flu all", "total number of records" as "total" and "the rate of flu positive records against total records" as "flu rate". The four sets of covariates are respectively: [flu A, flu B], [flu all], [flu rate], [total].

To understand the performance of each set of covariates, the correlation between each covariate and wILI truth are plotted in Figure 4. The x axis denotes 10 regions and y axis includes five covariates, namely flu A, flu B, flu all, total, flu rate. Darker color means higher correlation. According to this visualization, flu A, flu all and flu total are more promising covariates, while flu B is the worst one. The bar plot shows the performance of sensors created with each set of covariates from 2015 week 51 to 2017 week 10, where higher bar means higher root mean squared error. Sensor created using total record count beats those using flu A + flu B or flu all and is consistent among regions. "Flu rate" sensor behaves quite well in some region but is very noisy in region 5 and 6. One possible reason is that flu rate is de-trended data which remains stable across years. But at the same time, it is noisy when the sample size is small.

De-trending

The increasing trend of total record number of lab data can affect a lot in inference. In order to do de-trend, we take advantage of the facility identifications the data provides to compute the number of records per facility in each week. The underlying assumption is that the number of facilities that deployed with the device fully captures the volume increase. From Figure 5 we can observe notably improvement of de-trended covariates and sensors in terms of correlation and prediction error. Among all the sensors, the one using de-trended total record number is an easy pick since it behave well in almost every region.

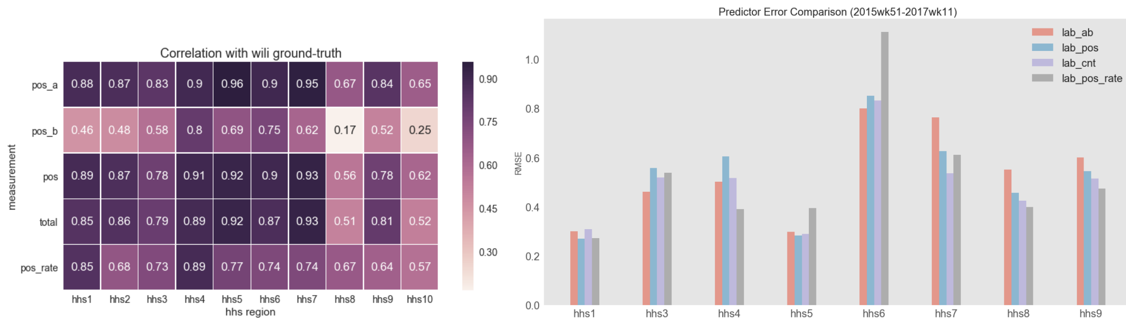

Figure 4: **(left)** In-season correlation heatmap between covariates and wILI across regions **(right)** Sensors' out-of-sample RMSE comparison across regions.

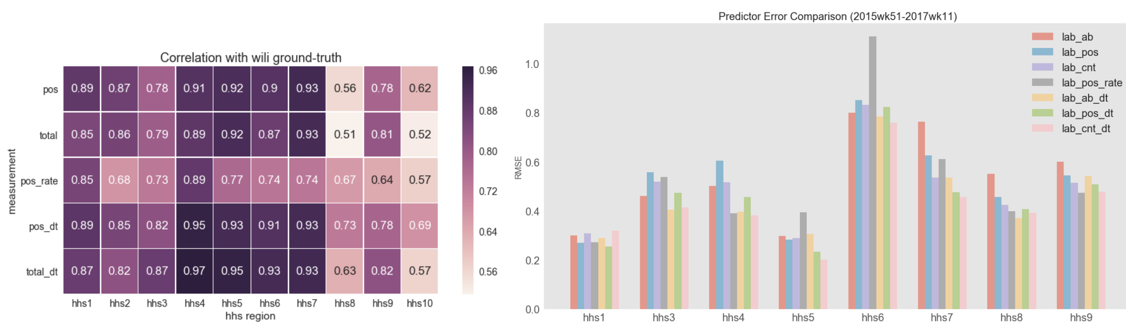

Figure 5: **(left)** In-season correlation heatmap between covariates, de-trended covariates and wILI across regions **(right)** Expanded set of sensors' RMSE comparison across regions.

Sensor performance comparison

In order to evaluate the lab data sensor's performance, we compare its performance to all regional sensors existing in Delphi's system from week 51, 2015 to week 19, 2017 with only in-season data point considered. This interval covers near two whole seasons. From Figure 7, we can conclude that lab data helps create a good sensor in general. Lab data sensor tops in region 3,4 and 5 during last season and behaves relatively well in region 1,6,7,8 and 9. We can also expect that as the data volume increases, the inference accuracy for region 2 and 10 can improve as well.

Sensor fusion performance

Two methods are used to evaluate the benefits of using a new sensor: simple linear regression with seasonal autoregressive (sar3) sensor and fusion into Delphi nowcast inference. Empirically, sar3 is the best performing and the most dominant sensor in Delphi's system. Meanwhile, sar3 merely uses the historical wILI. So combining a new sensor together with sar3 could explain how well extra digital surveillance data can help the pure autoregressive metric.
Figure 8 and 9 show that this sensor does not help sar3 by a linear regression. The underlying reason could be the inconsistency of correlations along time or simply the insufficient amount of

training data. However, lab test sensor does help nowcast inference by a noticeable margin with Delphi's Sensor Fusion metric in region 3, 4, 5 and 7. Meanwhile, it does not hurt the performance of nowcast in other regions. Hence, it seems that lab test sensor can improve the accuracy of nowcast in general.

5.2 Thermometer measurement data analysis

The analysis of thermo data follows the same procedure. We will look at its potential of generating sensor with regressive models and evaluate the benefits it brings to sensor fusion.

Covariates analysis

Similar to lab data, for thermo data we can use the positive measurement count or total measurement count to build sensor. Also, since it is very likely that a user takes multiple measurements in a short while, it may be beneficial to estimate the total number of users rather than records. In order to distinguish the users, we created a user id by joining (using "_") identification of devices and fields of demographic information. We assume that records with the same user id and date were taken by the same user. Sensors are generated with four single covariate: total record count, total user count, feverish record count and feverish user count. Here, "feverish" is defined as body temperature over 37.5 Celcius. According to the correlation heatmap in Figure 6, the count of total or feverish records seem to be more correlated with the wILI value, while the feverish one works better than total one. The plot for sensors' proves this argument. The best sensor generated by thermo data is with the feverish record count.

There are two questions remained unanswered. Since CDC's definition of Influenza-like Illness uses a higher body temperature than 37.5 Celcius, it could be helpful to increase our threshold of "feverish" definition. Another issue is that a user could take measurements in several consecutive days if he/she fells sick. So, a sliding window of a couple days could help improve the accuracy of user count estimation. Both experiments have been conducted. Unfortunately, neither a higher threshold for temperature nor a sliding window to distinguish users helps the prediction.

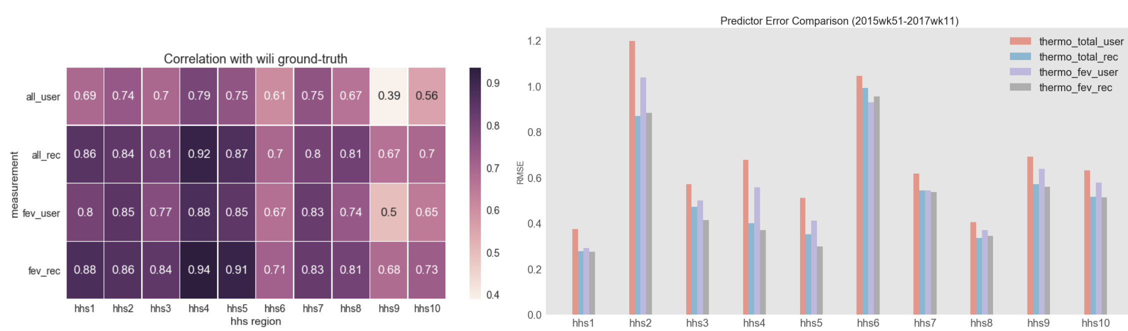

Figure 6: **(left)** In-season correlation heatmap between thermo data covariates and wILI across regions **(right)** Expanded set of thermo sensors' RMSE comparison across regions.

Sensor performance comparison

Thermo data sensor's performance is evaluated for all 10 HHS regions during week 51, 2015 to week 16, 2017, with only in-season data point considered. Figure 7 shows that thermo data sensor behaves well in region 1, 2, 3, 4, 5 and 8, where it is at least a top 3 sensor. However, the inference is not very accurate in region 6, 9 and 10. In general, this sensor is promising as well since it generates good estimates in many regions and is relatively stable across different regions.

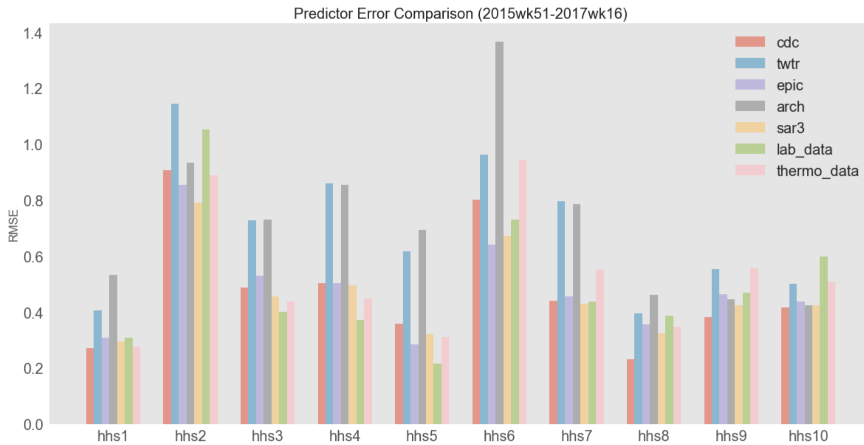

Figure 7: Comparison of lab sensor, thermo sensor with all other regional sensors' RMSE.

Sensor fusion performance

Figure 8 and 9 show the result of linearly combining thermo data sensor and sar3 sensor and fusing thermo data sensor with other sensors. Similar to lab data, linear regression does not work well for estimating wILI with thermo data sensor and sar3. If we fuse thermo data sensor into nowcast with sensor fusion kernel, the inference accuracy for half of the regions will increase. For other regions like hhs6 and hhs7, adding thermo data leads to lower accuracy. Also, if we compare lab + nowcast case with lab + thermo + nowcast case, the latter seems have even worse performance on average. In general, the difference made by adding thermo sensor to nowcast is not significant. Since we only have less than two seasons of data to evaluate, it is hard to conclude whether thermo data will help by nowcast inference at this point.

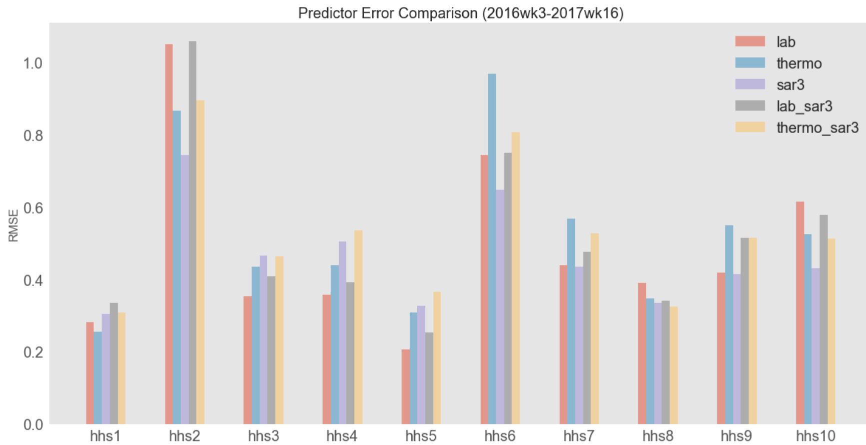

Figure 8: Performance of fusing lab, thermo data sensor with sar3.

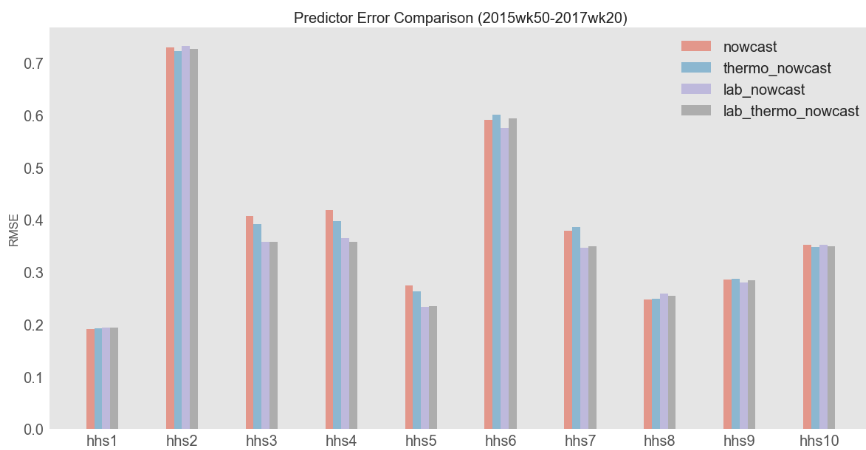

Figure 9: Performance of fusing lab, thermo data sensor into nowcast system.

5.3 Sensor fusion kernel performance analysis

Comparing with multiple linear regression, current sensor fusion kernel has two main advantages. First, it utilizes the sensors of different geographical granularity through using a mapping matrix. Second, it estimates the covariance matrix with pairwise covariances to make better use of sensors that available for different time periods. In this section, we try to verify that the actual performance of sensor fusion kernel agrees with the assumptions above.

Benefits of using different geographical granularity

In this experiment, we try to figure out whether higher or lower geographical level sensors can help with the inference of regional flu activity. The sensor fusion kernal is applied to regional sensors, regional + national sensors and regional + national + state-level sensors. The accuracy is evaluated using RMSE of in-season point predictions from week 40, 2015 to week 20, 2017. The performance is also compared with the arguably best sensor sar3.

Figure 10 shows the experiment result. It is clear that fused sensors achieve better performance than using sar3 alone in every single region. National sensors do not help regional level predictions. However, regional inferences benefit a lot from state-level sensors. So it is confirmed that sensor fusion kernel achieves better performance since finer-grained sensors are utilized.

This result also poses a question that whether sensor fusion kernel is a good choice for state-level forecast since national and regional level sensors may not help much. Figure 11 shows the comparison of sensor fusion result using every available sensor with each individual sensor on ten different states. The reasons of choosing these states are that they all have relatively large population and they belong to different HHS regions. Nowcast achieves very good performance in most of the states, with the exceptions being NY and IL. One hypothesis is that Nowcast benefits from implicitly utilizing the error correlations among different states.

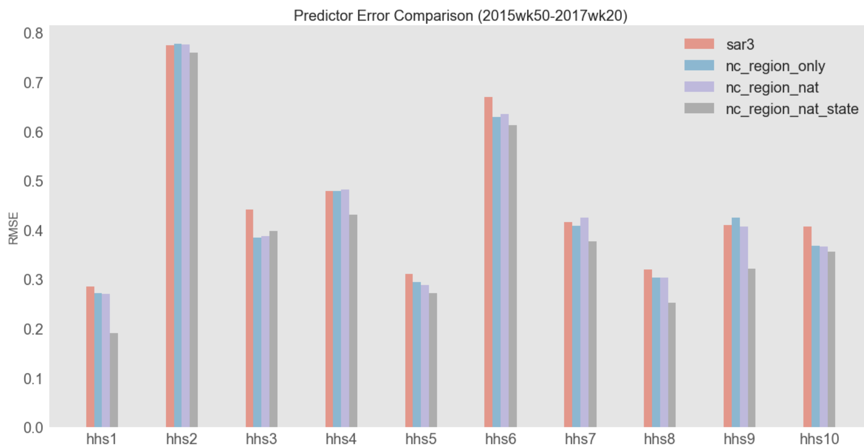

Figure 10: Comparison of sensor fusion kernel performance given sensors of different geographical levels.

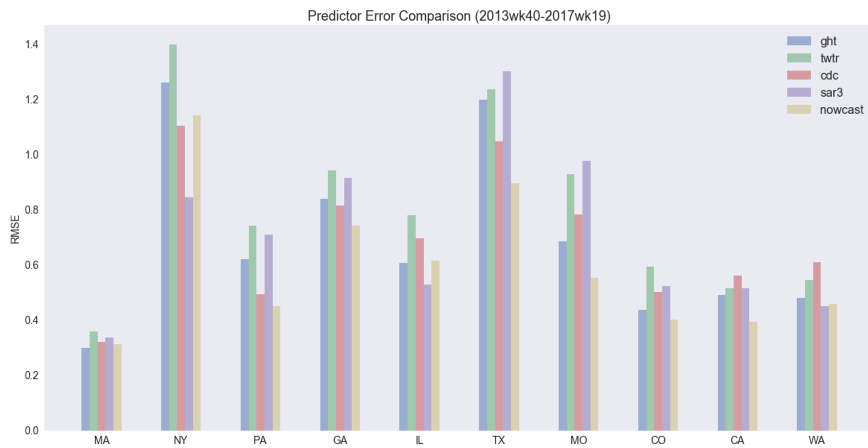

Figure 11: Performance comparison of sensor fusion kernel with individual state-level sensors on ten different states.

Analysis of evolving data generating process

There are two temporal factors that could affect the data generating process of sensor fusion step. First, off-season data generation and collection could be different from in-season data. For example, official flu season is a time of enhanced surveillance. As a result, wILI is expected to be more accurate during this period. Second, there could be long-term trend in different data sources. One example is that the demographics of social network users may change overtime. Also for wILI, the number of volunteering health care providers might change across years as well.
Two experiments are designed to test these hypotheses. In the first experiment, off-season points are excluded from the training data. The result is shown in table 1. In season 2012, 2013 and 2016, excluding the off-season training point helps improve the estimation accuracy in terms of RMSE, while it is slightly worse in the other two years. So, it seems that excluding off-season training points could be potentially useful. However, the different between in-season and off-season should not be considered a major factor affecting sensor fusion step. In the second experiment, a two-year sliding window is used to exclude data points from too long ago. Table 2 shows that the result improves by a notable amount in season 2012 and 2016 in terms of RMSE. A less than 1% decrease in is observed in other three regions. Again, sliding window may help with nowcast inference but more data is needed to make the judgement.

6 Conclusion

This project primarily focuses on two new data sources that are potentially helpful to Nowcast: flu lab test data and thermometer measurement data. The distributional characteristics of both data sources as well as their performance in generating estimators for ongoing flu activity are evaluated. In order to improve the final output of Nowcast, several experiments are designed for better understanding sensor fusion kernel. The varied temporal availability of sensors and the

year	hhs1	hhs2	hhs3	hhs4	hhs5	hhs6	hhs7	hhs8	hhs9	hhs10	avg
2012	**-8.0%**	**-4.9%**	**-0.3%**	**-6.1%**	**-6.1%**	**-3.0%**	**-1.6%**	**-4.7%**	**-1.0%**	**-3.7%**	**-3.9%**
2013	1.0%	3.1%	**-0.2%**	**-1.1%**	**-1.4%**	1.9%	**-1.0%**	1.2%	**-1.4%**	**-8.7%**	**-0.7%**
2014	**-2.9%**	**-2.1%**	**-2.0%**	0.8%	3.1%	0.7%	4.1%	**-3.0%**	3.1%	0.7%	0.3%
2015	**-0.2%**	0.8%	**-1.9%**	1.2%	**-1.5%**	9.5%	1.4%	1.8%	3.3%	2.2%	1.7%
2016	4.9%	**-2.8%**	**-1.5%**	**-3.5%**	**-4.6%**	**-2.6%**	**-1.7%**	5.2%	**-1.4%**	**-3.4%**	**-1.2%**

Table 1: Comparison of sensor fusion kernels using in-season or all training data points. Values shown are the rates of increase in RMSE for the experiment using in-season points only (the lower the better).

year	hhs1	hhs2	hhs3	hhs4	hhs5	hhs6	hhs7	hhs8	hhs9	hhs10	avg
2012	**-19.9%**	11.0%	**-7.0%**	0.7%	**-6.8%**	**-6.0%**	**-8.6%**	**-7.2%**	**-8.8%**	7.2%	**-4.5%**
2013	0.6%	**-0.3%**	4.1%	**-1.1%**	2.8%	0.8%	2.1%	0.4%	**-5.6%**	2.9%	0.7%
2014	9.9%	5.1%	**-7.2%**	**-4.3%**	**-12.2%**	2.6%	6.7%	4.2%	**-5.6%**	1.8%	0.1%
2015	2.0%	3.3%	5.4%	6.3%	4.5%	**-9.1%**	**-9.0%**	5.0%	**-1.4%**	0.4%	0.7%
2016	5.8%	**-0.2%**	**-9.2%**	**-7.3%**	**-11.6%**	4.7%	3.7%	**-5.4%**	**-17.9%**	5.4%	**-3.2%**

Table 2: Comparison of sensor fusion kernels using a sliding window or not. Values shown are the rate of increaments in RMSE for the experiment with sliding window (the lower the better).

insufficient data points impose a great challenge in evaluating the Nowcast result. It is very likely that de-trended flu lab test data will help with the inference significantly. But for other topics covered, it is still too early to draw a conclusion.

References

[1] D. A. Broniatowski, M. J. Paul, and M. Dredze. National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic. *PLoS ONE*, 8(12):e83672, dec 2013.

[2] R. G. Brown and P. Y. Hwang. Introduction to random signals and applied kalman filtering: with matlab exercises and solutions. *Introduction to random signals and applied Kalman filtering: with MATLAB exercises and solutions, by Brown, Robert Grover.; Hwang, Patrick YC New York: Wiley, c1997.*, 1997.

[3] Centers for Disease Control and Prevention. Overview of Influenza Surveillance in the United States, 2017.

[4] Centers for Disease Control and Prevention. Weekly U.S. Influenza Surveillance Report, 2017.

[5] M. Dredze, R. Cheng, M. J. Paul, and D. Broniatowski. HealthTweets.org: A Platform for Public Health Surveillance Using Twitter. 2014.

[6] A. F. Dugas, Y.-H. Hsieh, S. R. Levin, J. M. Pines, D. P. Mareiniss, A. Mohareb, C. A. Gaydos, T. M. Perl, and R. E. Rothman. Google Flu Trends: Correlation With Emergency Department Influenza Rates and Crowding Metrics. *Clinical Infectious Diseases*, 54(4):463–469, feb 2012.

[7] D. Farrow. Modeling the Past, Present, and Future of Influenza. 2016.

[8] D. Farrow and Z. Wang. ILI Nearby, 2016.

[9] R. E. Kalman et al. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.