Spatio-temporal prediction of epidemics through fusion of information from diverse sources

Comment [TGD1]: I notice you have removed the seasonal effects from the model. Can you explain why and what difference it might make?

Comment [su2]: I removed the seasonal effects for 2 reasons: (1) Many states and CDC don't collect data outside a flu season, which is between April and September (Alternatively, I could have shorten the period of the seasonal effects). (2) Onset and peaking of a flu season differ from year to year by a couple of months. Hence, I don't think incorporating the seasonal effects would contribute much.

Problem Description

Predicting the spread of epidemics through space and time can help government agencies and organizations better prepare and allocate resources. Seasonal flu epidemics have been closely monitored, and many years of historical data have been collected by the medical community. The data collected and aggregated by CDC has been valuable for researchers trying to develop models to forecast the spread of flu epidemics. In addition to the CDC data, there are many other data collected by different entities for various purposes – many of them unrelated to flu epidemics. When those datasets are considered together with the CDC data, they offer the opportunity to significantly improve our ability to assess and forecast flu epidemics both spatially and temporally. Such data include weather data, social network data, vaccination statistics, and flu medication sales. Those data have different characteristics (e.g. percentages for CDC regional ILI rates and flu vaccination, and quantized flu activity levels for CDC state ILI rates) and different spatial and temporal resolution. Aggregating the data into a forecasting model is challenging but, if successful, can provide much improved forecasting accuracy over a longer time horizon than what current approaches based on limited sources of information can accomplish.

Phase 1 Problem

During Phase 1, the goal is to estimate Influenza-like Illness (ILI) rates at a spatial resolution finer than that of the ILI data from CDC. Performers can use all of the datasets listed in the next section, except for the NREVSS dataset (Phase 2 data), to estimate weekly ILI rates in the 48 contiguous states. The spatial resolution of the estimate should be at the county level. The results will be compared to state ILI rates from selected states (Maryland, North Carolina, Rhode Island and Texas) and district ILI rates from 2 states (Mississippi and Tennessee), where each district consists of multiple counties. The development in the first phase also helps to identify important covariates and their contributions to spatio-temporal interpolation and prediction.

The datasets cover the flu seasons 2012-2013, 2013-2014 and 2014-2015. Performers can use all or some of the datasets for their development. The public set consists of data from the 2012-2013 and 2013-2014 flu seasons. The 2014-2015 flu season data will be used as the private set for evaluation. All of the 2014-2015 Phase 1 data except for the state and district level ILI rates from the selected states are available as input to the models. Fitted models will be evaluated based on their estimated state and district ILI rates against the actual ILI rates of the select states and districts.

Phase 2 Problem

During Phase 2, the goal is to produce estimates ILI rates that are more timely than those published by the CDC data (while maintaining finer spatial resolution finer than the CDC). The ILI data from CDC and states have 1 to 2 weeks of delay. The models developed by the performers will be used to nowcast the weekly state and district ILI rates of the select states. The nowcast results will be compared to the

released data from CDC and select states. Performers can also use the NREVSS dataset in this phase, which may provide additional predictive power.

All of the 2012-2015 data are available to performers in the public set. Performers are to develop models for 2-week-ahead nowcasting of the select states' and districts' ILI rates. Input to the model can include all of the data with 2 weeks of delay except for the Twitter data, where the data from the current week can be used. In Phase 2, the evaluation will be conducted on the current 2015-2016 flu season. Performers are evaluated based on their nowcast state and district ILI rates against the actual ILI rates.

Comment [TGD3]: Isn't this forecasting rather than nowcasting?

Comment [su4]: CDC publishes its data with up to 1-2 weeks of delay. Performers will be using CDC ILI statistics from 2 weeks in the past to estimate current week's ILI statistics.

Data Description

Comment [TGD5]: Can you prepare the data so that all data files have the same basic format? For example, year and week should always be separate columns and always in the same order. I suggest assigning an id number to every county in the US and then defining various kinds of aggregations in a separate file. You can define the regions with states (for those states with regional data), the states (for those states where we have state-level data), and the CDC ILI regions all as aggregations of the counties as the smallest unit of analysis.

We should denote missing values by a standard method, such as the symbol NA.

Comment [su6]: Yes, I'll standardized the formats as you suggested. For a unique ID for each county, I will use the FIPS code, which is a 5-digit integer assigned to every county in the US and is already available in many datasets acquired from the Federal government.

Main Datasets
CDC Seasonal ILI Rate (HHS Regions)
CDC reports the weekly number of Influenza-like Illness (ILI)* cases collected through the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet) with roughly 2 weeks' delay. This is considered a highly authoritative source of influenza related information in the medical field. The publicly available dataset contains the percentages of weekly outpatient visits for ILI since 1997. The cases in each reporting period are also divided into 10 Health and Human Services (HHS) regions, where a region covers multiple states, and is partitioned into 5 age groups.
CDC Seasonal Flu Activity Level (States)
CDC also publishes a weekly measure of flu activity for each state. The flu activity is quantized into 10 levels. "The 10 activity levels correspond to the number of standard deviations below, at or above the mean for the current week compared to the mean of the non-influenza weeks. … An activity level of 1 corresponds to values that are below the mean, level 2 corresponds to an ILI percentage less than 1 standard deviation above the mean, level 3 corresponds to ILI more than 1, but less than 2 standard deviations above the mean, and so on, with an activity level of 10 corresponding to ILI 8 or more standard deviations above the mean."
State and County ILI Rates (Selected states and counties)
This dataset contains percentages of ILI cases from Massachusetts, North Carolina, Rhode Island and Texas for the 2012-2013, 2013-2014, and 2014-2015 flu seasons. For Mississippi and Tennessee, percentages of ILI cases are also broken out by districts/regions. Mississippi is divided into 9 districts and Tennessee into 13 regions, where each district/region consists of multiple counties. However, some districts/regions in some weeks have missing ILI data.
Twitter Data
The dataset contains the number of flu related tweets without re-tweets and not from the same user within the syndrome elapsed time of 1 week. The flu related tweets are defined as tweets with keywords "flu," and "influenza." The locations of the tweets provides observations with finer spatial and temporal resolution than CDC data, but the data are very noisy.
Flu Vaccination Data of Medicare Recipients
As people receive flu vaccines, the percentage of population susceptible to flu is reduced, and even

Deleted: ryland

* "ILI is defined as fever (temperature of 100°F [37.8°C] or greater) and a cough and/or a sore throat without a KNOWN cause other than influenza."

when vaccinated people get the flu, they generally have milder symptoms. This reduces the number of reported ILI cases. Hence, the flu vaccination data may have strong predictive power for future ILI cases. This dataset contains weekly cumulative percentages of Medicare recipients filing flu vaccination claims of each year between 2012 and 2015 for each county in the United States. It is noted that the data covers only Medicare recipients, and the majority of the recipients are age 65 or older.
NREVSS: The National Respiratory and Enteric Virus Surveillance System (Phase 2)
In addition to the ILINet data, CDC also aggregates to HHS regions the weekly percentage and number of respiratory specimens tested positive for influenza and the weekly number of cases for each influenza virus type. Since illness due to Type A and Type B flu virus may peak at different time during the flu season, using this dataset may improve prediction of ILI rates.
Supporting Data
Twitter User Demographics
To account for the demographic differences of Twitter users from the general population, we will use data from a study by Pew Research, that summarizes the demographics of Twitter users by gender, race, education, and income level. The demographic information may help to correct some Twitter sampling bias by normalizing the tweet counts based on the demographics of their corresponding local populations.
US Census
Total population and percentages of population by age group, education and income level of each county in the United States.
County Adjacency Data
The file contains state association of each county and its neighboring counties.

Baseline Spatio-temporal Model

The baseline Bayesian hierarchical model captures various effects commonly present in temporally and geographically referenced data. The baseline model attempts to describe both spatial clustering and temporal correlations. For effects that are caused by known factors, the model also incorporates the fixed-effects term to capture their influence on the observations, in addition to the spatial and temporal random effects. The model has been developed with computational concerns in mind. The use of Gaussian Markov Random Field (GMRF) in modeling the random effects allows this approach to take full advantage of the spatial and temporal correlations while still maintaining manageable memory usage. The specific model structure is below:

County ILI Rate Model	$\log\left(\frac{z_{c,t}+\epsilon_1}{1-z_{c,t}+\epsilon_1}\right) = \beta^T X_{c,t} + y_{c,t} + n_{c,t}$ $\epsilon_1 = 0.0001$ c: county index; t: week index $X_{c,t}$: covariates for each county c and week t $y_{c,t}$: spatio-temporal correlation of ILI $-\infty < y_{c,t} < +\infty$ (see below) $z_{c,t}$: ILI rate (between 0 and 1) of county c in week t

Deleted: latent prevalence

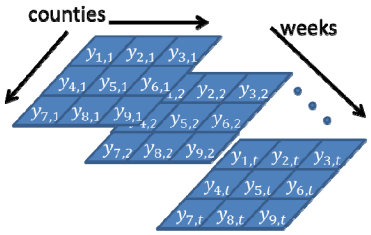
	$\boldsymbol{n_{c,t}}$: zero-mean Gaussian noise (see below) $\boldsymbol{\epsilon_1}$: a small number to ensure numerical stability
Aggregate Observations	$z_{R_i,t} = \sum_{c \in R_i} \left(\frac{N_c}{N_{R_i}}\right) z_{c,t}$ R_i : set of counties in area i; the area can be a HHS Region, a state or a district in a state $z_{R_i,t}$: reported ILI rate of HHS Region i in week t N_c : population of county c N_{R_i} : population of region i
Spatio-temporal correlation	$p(Y) \propto \exp\left(-\frac{1}{2}\tau_1 Y^T (D_w - W)\, Y\right)$ $Y = [y_{1,0} \quad y_{2,0} \quad \dots \quad y_{1,1} \quad y_{2,1} \quad \dots \quad y_{C,T-1} \quad y_{C,T}]^T$ C: total number of counties; T: total number of weeks D_w: diagonal matrix; W: sparse symmetric matrix The elements of W are defined as $\begin{cases} w_{(c,t)(c,j)} = \rho & \text{where } j = t-1 \text{ or } t+1, \\ w_{(c,t)(i,t)} = 1 & \text{if } i \text{ is a neighboring county of } c, \\ w_{(c,t)(i,j)} = 0 & \text{otherwise,} \end{cases}$ where the subscript (c, t) (i, j) denotes the row and column indices of the element of W that corresponds to $y_{c,t}$ and $y_{i,j}$ in Y. The only non-zero entries of W are those whose row and column indices correspond to pairs of elements in Y that represent the effects of neighboring counties from the same week or of the same county from consecutive weeks. $(D_w)_{(c,t)(c,t)} = \Sigma_{(i,j)} w_{(c,t)(i,j)}$ τ_1 : precision (inverse variance) that controls the spatio-temporal correlations (smoothness) of ILI rates
Covariates	$X_{c,t} = \left[\log\left(\frac{S_{c,t} + \epsilon_2}{\widetilde{N}_c}\right), \quad \log\left(\frac{V_{c,t} + \epsilon_3}{1 - V_{c,t} + \epsilon_3}\right)\right]^T$ $S_{c,t}$: number of flu-related tweets from county c in week t. $V_{c,t}$: cumulative percentage of Medicare recipients filing flu vaccination claims from county c in week t. $\widetilde{N}_c = \Sigma_g N_{c,g} U_g$: Twitter user demographics adjusted population of county c $N_{c,g}$: population of county c belonging to age group g. U_g : percentage of Twitter users belonging to age group g. $\epsilon_2 = 0.1$, $\epsilon_3 = 0.001$
County	$n_{c,t} \sim^{iid} N(0, 1/\tau_2)$

Comment [TGD7]: I don't understand what this is doing in the model. I would have expected a random effect n_c.

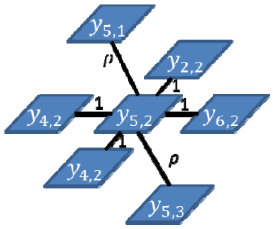
Comment [su8]: This term is essentially a white noise mainly to describe anything that cannot be captured by covariates and spatio-temporal correlations.

heterogeneity	
Hyperprior	$\beta_1 \sim N(0,10)$, $\beta_2 \sim N(0,10)$ $\tau_1 \sim G(3, 0.1)$, Gamma distribution (α (shape), β (rate) definition) $\tau_2 \sim G(10, 0.1)$ $\rho \sim G(1.05, 0.5)$

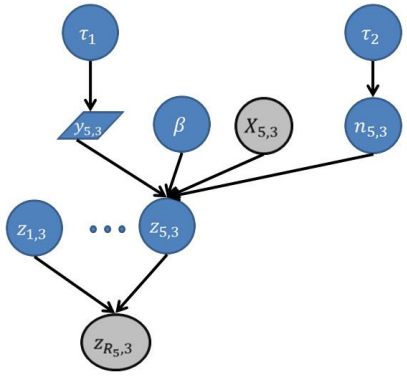
We can visualize the model as follows where we assume that the counties are located on a grid for illustration purpose. The GMRF defines a time series of latent prevalence maps:



Each cell is connected to its neighbors in time and space. The adjacency matrix W contains a 1 for spatial neighbors and ρ for temporal neighbors.



The latent prevalence is combined with covariates and a noise term to generate the observations:



Grey variables are observed at training time. Only X is observed at prediction time. Note that typically several individual county prevalences ($z_{1,3}, \ldots, z_{5,3}$) are aggregated to produce the observed value.

Queries

1. **Reconstruction.** Given weekly $S_{c,t}$ (number of flu related tweets) and $V_{c,t}$ (cumulative percentage of Medicare flu vaccination claims) of every county and weekly $z_{R_i,t}$ (ILI rates) of every HHS region, output a marginal MAP estimate for weekly ILI rates $z_{R_D,t}$ of individual counties in the selected states (Tennessee, Mississippi, Massachusetts, North Carolina, Rhode Island and Texas).
2. **Prediction.** Given weekly $S_{c,t}$ (number of flu related tweets) and $V_{c,t}$ (cumulative percentage of Medicare flu vaccination claims) of every county for weeks $t = 1, \ldots, u$ and weekly $z_{R_i,t}$ (ILI rates) of every HHS region for weeks $t = 1, \ldots, u - 1$, output a marginal MAP estimate for weekly ILI rates $z_{R_D,u}$ of individual counties in the selected states (Tennessee, Mississippi, Massachusetts, North Carolina, Rhode Island and Texas).

Evaluation Metrics

1. **Population-adjusted Squared Error.** Let $\hat{z}_{c,t}$ be the predicted prevalence for county c at time t; let R_i be a region that contains county c; let $z_{R_i,t}$ be the observed prevalence in region R_i; and let N_c be the population in count c. A region in this case is the smallest geographical area where ILI truth data is available. Define $\hat{z}_{R_i,t}$ to be the aggregated prevalence (weighted by relative population):

$$\hat{z}_{R_i,t} = \frac{\sum_{c \in R_i} N_c \hat{z}_{c,t}}{\sum_{c \in R_i} N_c}.$$

Then the loss is

$$\sum_i N_{R_i} \sum_t \left(\hat{z}_{R_i,t} - z_{R_i,t}\right)^2$$

Comment [TGD9]: Is this a joint MAP or marginal MAP for each region?

Deleted: and

Deleted: and for the individual regions in the selected states (Maryland,

Deleted: and

Deleted: and for the individual regions in the selected states (

Deleted: ryland

Comment [TGD10]: Here are some suggestions for the metrics.

2. **Start and Max of the epidemic**. Let $\bar{M}_{R_i}$ be the week in which the highest prevalence was observed during the year. Let S_{R_i} be the week when the prevalence first exceeded 5%. We will compute the absolute difference between the observed and predicted values of these quantities.

Deleted: ,

Deleted: , and Min

Comment [su11]: Lowest prevalence is hard to predict. CDC and many states don't collect off-season ILI data. Also, ILI data during off-season is typically swamped by background noise not related to actual flu.

Deleted: , and $\underline{M}_{R_i}$ be the week in which the lowest prevalence was observed during the year

Deleted: Finally, I

Input and Output Data Formats

Input 1: $S_{c,t}$ (number of flu-related tweets) and $V_{c,t}$ (cumulative percentage of Medicare recipients) in CSV

year	HHS week	county name	state	FIPS county code	cumulative vaccination rate [0,1]	No. of Tweets (integer)
2015	1	ABBEVILLE	SC	45001	0.021815849	323
2015	2	ABBEVILLE	SC	45001	0.021815849	567
2015	3	ABBEVILLE	SC	45001	0.021495027	369
2015	4	ABBEVILLE	SC	45001	0.016361886	34
2015	5	ABBEVILLE	SC	45001	0	3
2015	6	ABBEVILLE	SC	45001	0	765
2015	7	ABBEVILLE	SC	45001	0	27
2015	8	ABBEVILLE	SC	45001	0	32
2015	9	ABBEVILLE	SC	45001	0	565
2015	1	ACADIA	LA	22001	0.02955163	27
2015	2	ACADIA	LA	22001	0.02955163	57
2015	3	ACADIA	LA	22001	0.02955163	17
2015	4	ACADIA	LA	22001	0.018682065	367
2015	5	ACADIA	LA	22001	0.008605072	1222
2015	6	ACADIA	LA	22001	0	232
2015	7	ACADIA	LA	22001	0	343
2015	8	ACADIA	LA	22001	0	733
2015	9	ACADIA	LA	22001	0	546

Input 2: Weekly $z_{R_i,t}$ (ILI rates) of HHS Regions in CSV

YEAR	HHS Week	HHS REGION	ILI Percentage ([0,100])
2015	14	Region 1	1.325439604
2015	14	Region 2	1.943721311
2015	14	Region 3	1.671723983
2015	14	Region 4	1.428197329
2015	14	Region 5	2.138237992
2015	14	Region 6	2.523811924
2015	14	Region 7	1.612231543
2015	14	Region 8	1.458439248
2015	14	Region 9	2.104164378
2015	14	Region 10	0.812929449
2015	15	Region 1	1.09375
2015	15	Region 2	1.564302996
2015	15	Region 3	1.376550268
2015	15	Region 4	1.338711711
2015	15	Region 5	1.92252937
2015	15	Region 6	1.774176794
2015	15	Region 7	1.198237885
2015	15	Region 8	1.305491203
2015	15	Region 9	1.535580524
2015	15	Region 10	0.925834729

Output 1: Weekly $z_{R_{i,t}}$ (ILI rates) of individual counties in CSV

year	HHS week	county name	state	FIPS county code	ILI Percentage ([0,100])
2015	1	ABBEVILLE	SC	45001	1.325439604
2015	2	ABBEVILLE	SC	45001	1.943721311
2015	3	ABBEVILLE	SC	45001	1.671723983
2015	4	ABBEVILLE	SC	45001	1.428197329
2015	5	ABBEVILLE	SC	45001	2.138237992
2015	6	ABBEVILLE	SC	45001	2.523811924
2015	7	ABBEVILLE	SC	45001	1.612231543
2015	8	ABBEVILLE	SC	45001	1.564302996
2015	9	ABBEVILLE	SC	45001	1.376550268
2015	1	ACADIA	LA	22001	1.338711711
2015	2	ACADIA	LA	22001	1.92252937
2015	3	ACADIA	LA	22001	1.774176794
2015	4	ACADIA	LA	22001	1.198237885
2015	5	ACADIA	LA	22001	1.305491203
2015	6	ACADIA	LA	22001	1.535580524
2015	7	ACADIA	LA	22001	0.925834729
2015	8	ACADIA	LA	22001	2.104164378
2015	9	ACADIA	LA	22001	0.812929449