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

# 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 proved valuable for researchers to develop models to forecast the spread of flu epidemics. Besides the CDC data, there are many disparate data of human behavior patterns 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 includes transportation pattern, weather data, social network data, vaccination statistics and flu medication sale. Those data have different characteristics (e.g. counts for CDC data, percentage for vaccination data and connection graph for air transportation) and different spatial or temporal resolutions. Aggregating the data into a forecasting model is challenging but, if successful, can provide much improved forecasting accuracy over 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 use all of the data listed in the next section retrospectively to estimate weekly Influenza-like Illness (ILI) cases in the United States. The spatial resolution of the estimate should be in the level of counties or higher. The results will be compared to select states where ILI cases of individual counties have also been collected. The development in the first phase also helps to identify important covariates and their contributions to spatio-temporal interpolation and prediction.

## Phase 2 Problem

During Phase 2, the goal is for the models developed by the performers to forecast weekly ILI cases in the county level. Hence, the models can only depend on the data sources that will be available prospectively. The performers may use the Phase 1 models for training in conjunction with the data. In the early stage of Phase 2, the performers may use historical data to verify accuracy of their predictions. Later, the forecasting results will be compared weekly to the released data from CDC and select states.

# Data Sources

|  |
| --- |
| CDC Seasonal Flu Data |
| 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 of delay. This is considered a highly authoritative source of influenza related information in the medical field. The publicly available dataset contains the numbers of weekly ILI cases 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.  Source: <http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html>  Availability: Now  Besides ILI, laboratories located in all 50 states, Puerto Rico, and the District of Columbia also report to CDC the number of respiratory specimens tested for influenza and the number of positive cases by influenza virus type and influenza A virus subtype. This data is even more definitive than ILI for flu, since ILI considers cases with “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.”  Source: <http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html>  Availability: Now |
| Twitter Data |
| The basic dataset contains the number of flu related tweets without re-tweets and not from the same user within syndrome elapsed time of 1 week. The flu related tweets are defined as tweets with keywords “flu,” “H1N1,” and “swine flu.” Twitter data provides observations with better granularity in a timely manner but noisier than CDC data. The data is available from 2009-2011.  Availability: Now  More extensive data is to be acquired from a Twitter data reseller such as GNIP (now part of Twitter), which has direct access to the Twitter fire hose. Those resellers can provide a certain percentage (e.g. 1%, 10%) of the Twitter data related to keywords of choice.  Availability: Identify a reseller and pricing by July or earlier; data available at start of contract  To account for the demographic difference of Twitter users from the general population, a study by Pew Research, which summarizes the demographics of Twitter users in terms of gender, race, education and income level, can be used. The demographic information can be used to correct some Twitter sampling bias by normalizing the tweet counts based on the demographics of their corresponding local populations.  Source: <http://www.pewinternet.org/2015/01/09/demographics-of-key-social-networking-platforms-2/>  Availability: Now  US Census Bureau has the demographic information of each county, including gender, race, education and income level.  Source: <http://quickfacts.census.gov/qfd/download_data.html>  Availability: Now |
| Transportation Data |
| Connecting flights between airports are an indicator of connectedness between regions and indirectly influence how fast an epidemic may spread among regions. Airline routes and airport locations are available for incorporating such information.  Source: <http://openflights.org/data.html>  Availability: Now |
| Weather Data |
| Local weather data such as temperature and humidity are related to flu transmission and hence ILI cases. Weather data is available from NOAA’s National Climatic Data Center (NCDC). NOAA aggregates such data from thousands of stations around the country with records going back for decades. Upon request, it will take 24-36 hours for NOAA to prepare the data for download.  Source: <https://www.ncdc.noaa.gov/cdo-web/datatools/selectlocation>  Availability: Most recent years are available now; Earlier years can be available within a few days after request. |
| Flu Vaccination Data |
| People receiving flu vaccines lower the percentage of population susceptible to flu, or have milder symptoms when they are infected with flu viruses, which in turn reduce reported ILI cases. Hence, the flu vaccination data may have strong predictive power for future ILI cases.  HHS maintains flu vaccination claims data (2012-2015) “for those covered by Medicare Fee-for-Service” with about 2 weeks of delay “for every state, county and zip code in the United States.” Since the HHS data contains only Medicare claims, the majority of the claims are for Americans 65 and over.  Source: <http://www.hhs.gov/nvpo/flu-vaccination-map/>  Availability: State level data available now; individual counties over time upon request  CDC has flu vaccination data for the general population divided by state and age group. It is conducted by surveys at the end of a flu season.  Source: <http://www.cdc.gov/flu/fluvaxview/reports/reporti1314/trends/index.htm>  Availability: Now |
| Retail Data |
| Retail data of people buying flu-related medicine or receiving vaccination from national chains such as CVS or Walgreens reflects current status and provides indication for future cases. Walgreens publishes Walgreens Flu Index every week “by tallying the retail prescription data for the pharmacy’s antiviral medications” of the previous week.  Availability: Determine by August if such data can be available for research |

# Spatio-temporal Model

Our current Bayesian hierarchical model captures various effects commonly present in temporally and geographically referenced data. The model can describe spatial clustering and temporal correlations, as well as periodic temporal fluctuation often present in temporal data. For effects that are caused by known factors, we also incorporate 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:

|  |  |
| --- | --- |
| Observations |  |
| Link Function |  |
| Fixed Effects |  |
| Spatio-Temporal Effects |  |
| Periodic Effects |  |
| Parameter Priors |  |

**Distribution**(**): probability distribution of the observations with mean **

**Link**: link function

*s*, *t*: space and time indices respectively

, : random variables in GMRF for spatio-temporal and periodic (e.g. seasonal) effects respectively

*n*, *T*: dimensions of , respectively

*Q*, *P*: sparse precision matrices for spatio-temporal and periodic effects respectively

*l*: dimension of the null space of *Q*

*m*: period of the periodic effect; *m*-1 is dimension of the null space of *P*

*N*(.,.): Gaussian distribution

*G*(.,.): Gamma distribution

*, a*, *b, c*, *d*: hyper-parameters; they are chosen to be relatively uninformative. For example, InvGamma(1,1), *a*=*c*=1 and *b*=*d*=0.1.

We also plan to further improve the model by explicitly incorporating the evolution of an epidemic. An approach is to incorporate an additional term to capture the different phases that a population goes through over time in a flu epidemic – from susceptible to infectious to recovery. More accurate description of the dynamics of the flu epidemics should further improve the accuracy in forecasting future cases.

# Computer Code

The model is implemented in Python. The current code can handle real-valued, count and point process data, and can be easily extended to include other data types in the exponential family.

For inference from observed data, we implemented the Markov chain Monte Carlo (MCMC) method to draw samples from the posterior probability, which provide estimates for the random variables of concern. We leveraged the Python PyMC module for the MCMC computation.

We can also use the model as a generative process to simulate the spatio-temporal evolution of a flu epidemic. The Python code also supports this option. Both the simulated data and the model itself can be used by the performers to develop and test their methods.