# Methodology for “Forecast the 2016-2017 Influenza Season Collaborative Challenge”

Alessandro Vespignani and Qian Zhang, Northeastern University.

## Summary

This document describes the methodology and models we are using to provide the prediction entries for the “Forecast the 2016-2017 Influenza Season Collaborative Challenge”. The proposed methodology utilizes Twitter data to estimate the initial conditions of a generative epidemic model that can provide geolocalized predictions for the influenza season. The generative model used is Global Epidemic and Mobility model [1, 2] (GLEAM), a data-driven stochastic epidemic modeling tool. Although, the model couples 220 countries worldwide and is able to simulate the global circulation of influenza like illnesses, here we restrict simulations to the USA. GLEAM divides this country in 582 census areas, hereafter called basins, centered on major transportation hubs and/or urban areas. The basins are coupled by mean of human mobility derived from real data. In each census area the population is divided in different compartments according the disease status. Here, we use a variation of the classic SEIR model, that includes asymptomatic transmission, traveling and non-traveling symptomatic individuals. The model is described in detail in the Appendix. The results of the model initialized with Twitter data and calibrated on historical national surveillance data are finally scaled on the ILINet data format and reported according to the formats provided by CDC.

## Methodology

Our forecasting method combines digital surveillance data, historical ILINet data and an epidemic stochastic generative model (GLEAM) to provide weekly probabilistic predictions for the season start week, the peak week and the peak intensity, as well as the ILINet percent for one to four weeks in advance at both national level and HHS regional level. Our methodology, illustrated in Figure 1, consists of three stages.

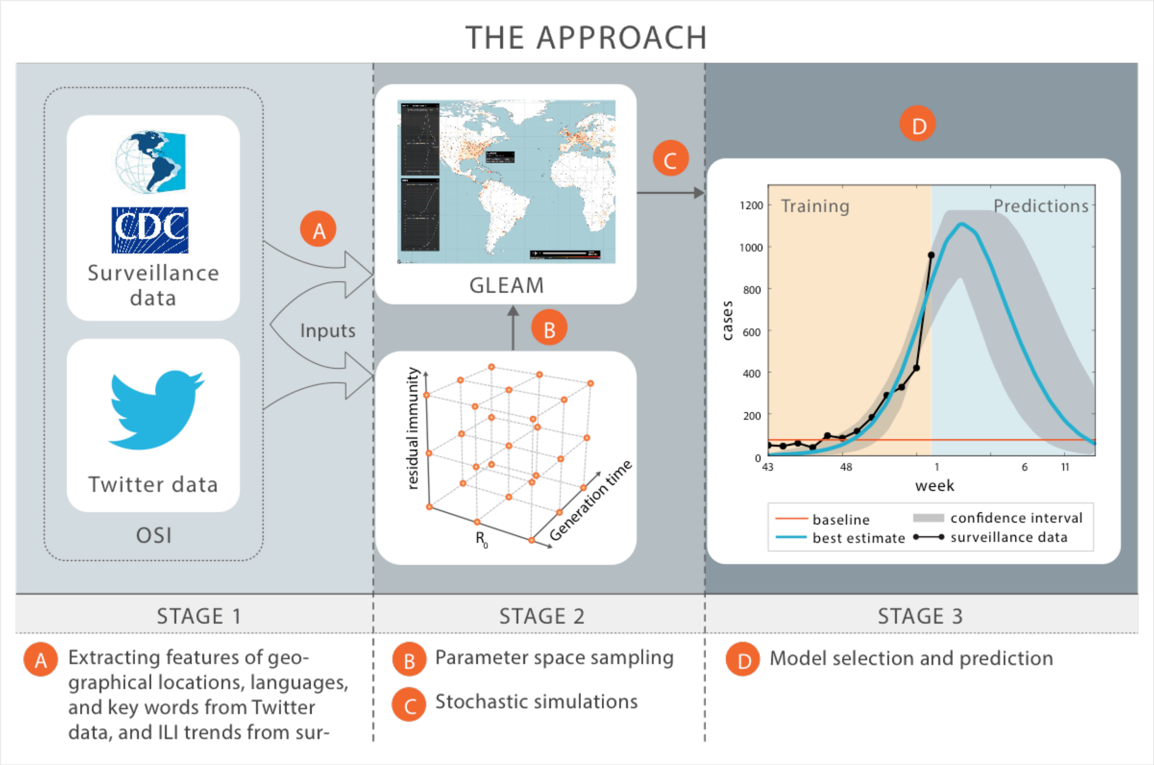

**Figure 1. Schematic representation of the forecasting methodology**

## Stage 1: Twitter and historic ILINet data mining.

In the first stage we mine Twitter's gardernhose (a live stream of about 10% to 20% of the global volume) extracting ILI related tweets in the USA. This task is executed applying three filters to the raw the data:

1) **Geolocalization**. We take advantage of the highest geographical resolution available considering just the subset of tweets containing explicit GPS coordinates.
2) **Language detection**. We analyze the subset of geolocalized tweets filtering out those not written in English. Although, just 4% of tweets written in the USA are not in English, we adopt this strategy to avoid biases in the next filter step.
3) **Keyword match**. Considering the literature on digital surveillance we compiled a list of 48 keywords that have been proved positively correlated with the ILI. We mine the subset of tweets coming out the second filter extracting tweets with a least one match. This allows us to define time series for each keyword.

The final output from Twitter is a set of geolocalized time series for each keyword. In order to match the model data structure, we map each time series to the census area level structure of the GLEAM model by aggregating all the GPS coordinates in the same census area.

In this first stage, we define the seeding window to be the consecutive four weeks from the starting week. We also extract the historic time series of the percentage of

ILI visits provided by the CDC in the same window. For each keyword we measure the correlation of determination $R^2$ between the time series of its tweet volume and the time series of the surveillance data.

## Stage 2: Model initialization and numerical simulations

In the second stage we evaluate the initial conditions necessary to run GLEAM and perform the model simulations exploring the phase space of the model's parameters.

### Initialization of GLEAM with Twitter data:

Though, the minimal time scale of GLEAM is the day and this is used to simulate the spreading, we aggregate the data at the week level to match CDC data. We define the starting week evaluating the first week characterized by an effective reproduction number larger that one.

We estimate the number of infected persons in each census area *k*, during the week *w* as:

$$I_{k,w} = \left(\sum_l \omega_{l,w} R_l^2\right) \cdot \alpha_k \cdot Y \quad .$$

In the above expression we have that;

- $\omega_{l,w}$ is the number of matches for the keyword, ILI related, *l* in the week *w* in the USA,
- $R^2{}_l$ is the coefficient of determination that provides the weight of each keyword in the window considered. The coefficient of determination is evaluated measuring the correlation between the time series of each keyword and the CDC data. This comparison is done in a window of four weeks after the starting week w.
- $\alpha_k$ is the ratio of population size (as reported by census) to the total number of Twitter users that were determined to live in the census area *k*.
- $Y$ is a free parameter to fit that provides a season-dependent relation between the actual number of cases of flu and the buzz generated on Twitter. The value of this parameter is calibrated at stage 3.

### Sampling Parameter Space:

Any epidemic forecasting approach relies on a number of assumptions introduced by the model structure, parameter settings, scale, etc. The initial conditions do not set uniquely the epidemic dynamics as the model output depends also on the basic epidemic parameters such as the basic reproduction number $R_0$ that is determined by the transmissibility and the infectious period, residual immunity of the population etc. Thus it is important to estimate parameter values consistently with the assumptions introduced by the model structure. We take consideration of $R_0$, residual immunity of the population $r$ and the free scaling parameter into parameterization. The basic reproduction number $R_0$ we use is defined as:

$$R_0 = C \cdot \beta \cdot \mu^{-1}$$

where $C$ is a constant determined by parameters describing the asymptomatic infectious individuals (see Appendix for details), $\beta$ is the transmissibility of the disease and $1/\mu$ is the infectious period. We also consider the effective $R_0$,

$$R_0^{eff} = (1 - r)R_0$$

We explore via Latin-Hypercube sampling a four-dimensional parameter space defined as $\boldsymbol{Y} \times \boldsymbol{r} \times \boldsymbol{\beta} \times \boldsymbol{\mu}$. Specifically, we consider $Y \in (10^{-6}, 10^{1})$ in logarithmic scale, the fraction of immunized population $r \in (0.0, 0.6)$ with interval 0.05, the recovery rate $\mu \in (0.2, 0.5)$ with interval 0.05 (the infectious period is from 2 days to 5 days). The value of $\beta$ is determined by $\mu$ and the basic reproduction number $R_0$. We consider $R_0^{eff}$ ranging from 0.8 to 3.0 with step 0.1 for each value of $r$. In total, the parameter space we explore contains 75,300 different combinations of parameters.

### Numerical simulations:

GLEAM is an ab-initio model, and it can simulate the entire influenza season at the geographic resolution of 582 basins. For each point in the model's parameter space, we perform numerically on the computer a large number of stochastic realizations (generally about 1,000 realizations) that project the epidemic behavior for 52 weeks, after the starting week we have considered. The simulations provide, for each census area and week of the year, an ensemble of possible epidemic evolutions from which it is possible to extract median, mean, and reference ranges for epidemic observables, such as newly generated cases, peak time of the epidemic, intensity etc. Each point in the parameters' phase space defines a different stochastic forecast output (SFO) set that can now be compared to the existing historical ILINet data in order to select the parameter's values the best fit the current influenza season so far.

## Stage 3: Stochastic forecast output (SFO) sets selection and rescaling to ILINet values.

In the final stage, we select the set of parameters $[\boldsymbol{Y}, \boldsymbol{r}, \boldsymbol{\beta}, \boldsymbol{\mu}]$ that best describe the historical ILINet data from CDC in the current season. First of all, we notice that ILINet CDC's data and GLEAM's data are provided in different scales. The CDC provides a normalized sample of the number of cases in the population, while GLEAM simulates the spread of the disease in the country providing the total number of symptomatic cases. In the ongoing season the maximum value (peak) is not known yet, implying the impossibility of evaluating the rescaling factor $\delta$

exactly. In order to overcome this problem, we consider the average maximum value (peak) in the past seasons plus minus several standard deviations to include possible fluctuations.

Based on the observation of holiday effects, we have performed statistical analysis on the total number of patient visits and estimated the “holiday scaling factor” to rescale the simulated epidemic profile curves on the weeks of Thanksgiving, Christmas and New Year. Since our model is currently restricted to one virus strains, to include the effects of influenza B on the ILI visits, we also estimated the “flu B scaling factor” to rescale the simulated epidemic profile curves on the weeks in February and March.

We therefore generate a set of rescaled SFO sets that we can statistically evaluate with respect to the historical ILINet data of the current season. To select proper models from this set, we first compare the simulation results of each point in the phase space explored against real data. This step is performed using the Aikake Information Criterion (AIC), which provides an estimation of the information lost when the selected model is used to represent the real data. We first select the model that minimizing AIC in the sliding fitting window of size 6 with updated surveillance data. We choose model $i$ with $\Delta AIC_i = AIC_i - \min AIC < 7$ into the best fitting rescaled SFO set. This set of models is then used to produce a set of estimations of the season start, peak week and intensity indicators, and therefore provide the corresponding probabilistic forecast of these quantities.

The geographical granularity of Twitter data and the geographical resolution of the GLEAM allow us to provide also predictions for the ten HHS regions with an analogous rescaling approach. In this case, for consistency, the best fitting SFO set at the national level is used to restrict the data to each one of the ten separate HHS regions. The rescaling and model selections are done accordingly to the historical ILINet data in each region.

## Appendix: The GLEAM model

We used a data-driven global stochastic epidemic model, which is based on the metapopulation approach. The model has been extensively described previously, and all the technical details and the algorithms underpinning the model results are reported in [1,2]. By integrating real demographic and mobility data, the model divides the world population into geographic census areas that are defined around transportation hubs and connected by mobility fluxes, which then defines a subpopulation network. Within each subpopulation, a compartmental structure models the disease spread between individuals. Individuals can move from one subpopulation to another along the mobility network; in this way, an outbreak originating in a seed subpopulation can lead to a global-scale epidemic. The GLEAM model can simulate the global spread of ILIs, and also allows study of the implementation of a wide range of intervention strategies. The GLEAM model architecture integrates three different data layers: 1) the population layer, 2) the transportation mobility layer, and 3) the epidemic layer.

The population layer is based on the high-resolution population database of the 'Gridded Population of the World' project of the Socioeconomic Data and Application Center at Columbia University (SEDAC). This database provides a population estimate by using a grid of cells covering the whole planet, with a resolution of 15 × 15 minutes of arc. The subpopulations of the metapopulation structure correspond to geographic census areas defined around transportation hubs, which are represented by the world airports, as provided by international databases of air travel. The census areas are obtained using a Voronoi-like tessellation of the Earth's surface by assigning each cell of the grid to the closest airport, taking into account distance constraints (see Figure.2). The resulting network of subpopulations counts 3,362 census areas in 220 different countries. In the USA the model considers 582 census areas.

The mobility layer takes into account the multiscale nature of human mobility. The GLEAM model integrates the mobility by global air travel (obtained from the International Air Transport Association and Official Airline Guide databases) and the short-scale mobility between adjacent subpopulations, which represents the daily commuting patterns of individuals. We obtained the commuting fluxes by collecting and integrating the data of 30 countries in 5 continents across the world [1]. We have fully integrated commuting and mobility data for the USA from the official Census. The model simulates the number of passengers traveling daily worldwide by using the real data obtained from the airline transportation databases, which contain the number of available seats on each airline connection in the world. The commuting short-range couplings between subpopulations are accounted for by defining the effective force of infections in subpopulations connected by commuting flows [2].

The epidemic model within each subpopulation considers a compartmental approach specific for the disease under study. In the present application we assumed that each individual can be in one of the following discrete states: susceptible, latent, symptomatic infectious able to travel, symptomatic infectious unable to travel, asymptomatic infectious, and permanently recovered [2]. The model assumes homogeneous mixing within each subpopulation. The disease transmission rate of symptomatic infectious individuals is $\beta$, and it is assumed to be rescaled by a factor $r_\beta$ = 50% for asymptomatic individuals. After the infection, susceptible individuals enter the latent compartment, where they are infected but not yet contagious. After the latency period, assumed to be equal to the incubation period and of average duration $\varepsilon^{-1}$, exposed individuals become infectious and have a probability $(1-p_a)$ of developing clinical symptoms, with $p_a$ considered to be the probability of becoming asymptomatic equal to 33%. Change in traveling behavior after the onset of symptoms is modeled by setting to 50% the probability $1-p_t$ that individuals would not travel when ill. Eventually, infected individuals recover after the average infectious period $\mu^{-1}$, and they are no longer susceptible. A fraction of the population r is assumed to have residual immunity to influenza from past season or vaccination. The basic reproduction number can be written as $R_0 = (1 - p_a + r_\beta p_a) \cdot \beta \cdot \mu^{-1}$[2]. All stochastic processes, modeling either the transitions of individuals in the different compartments and/or their mobility, are mathematically defined by discrete stochastic chain binomial and multinomial processes in order to preserve the discrete and stochastic nature of the individuals. Individuals are discrete but indistinguishable, because no additional population structure (for example, households or workplaces) is being considered. The other unspecified parameters are set using the methodology detailed in the previous sections.

The spreading rate of the disease at the level of a single subpopulation is governed by the basic reproduction number, $R_0$, which is a function of the parameters defining the natural history of the disease. However, in a metapopulation framework in which space is explicitly considered, the reproductive number is dependent on space and time, and it is more appropriate to define an effective reproduction number R(t). In more detail, to take into account seasonal effects in the transmission of influenza, we considered a seasonal forcing of the reproduction number, dependent on the calendar time. We assumed the US to have the northern hemisphere seasonal forcing. We denoted by $R_0$ the reference value of the reproduction where the model reproduces seasonality by means of a sinusoidal rescaling of $R_0$, by a factor ranging from $\alpha_{min}$ =0.1 (during the summer season) to $\alpha_{max}$ =1 (during the winter season). These are standard values from the literature. It is worth remarking that the seasonality farcing alone is not able to set the peak time of the seasonal influenza as residual immunity, transmissibility and initial conditions in the number of infectious individuals have are all at play in defining the unfolding of the epidemic dynamic during the winter season.

The GLEAM model is implemented in C/C++. Briefly, GLEAM is implemented in a modular manner, with each module performing a specific function. The compartmental model and the epidemic parameters are defined in a configuration text file that is loaded when the program starts. Subsequently, the program loads three data input files: the population database, the short-range mobility network, and the long-range mobility network. During each time step, which represents a full day, the following modules are called into the sequence: air travel, the compartmental transitions (where the force of infection takes into account both the infection dynamics and the short-range movement of individuals), and the partial aggregation of the results at the desired level of geographic resolution. After the last time step, the program generates the final output, which can be further processed for analysis. The model generates a large number of nominally identically initialized numerical stochastic simulations of the progression the epidemic. The simulations provide, for each point in space and time allowed by the resolution of the model, the set of possible epidemic evolution by statistically defining the median, mean, and reference range of a number of epidemic parameters. Each ensemble of identically initialized stochastic simulations corresponds to a stochastic forecast output (SFO) set. In the previous sections we have described the methodology of the initialization and selection of the SFO set used to provide the prediction of the challenge.

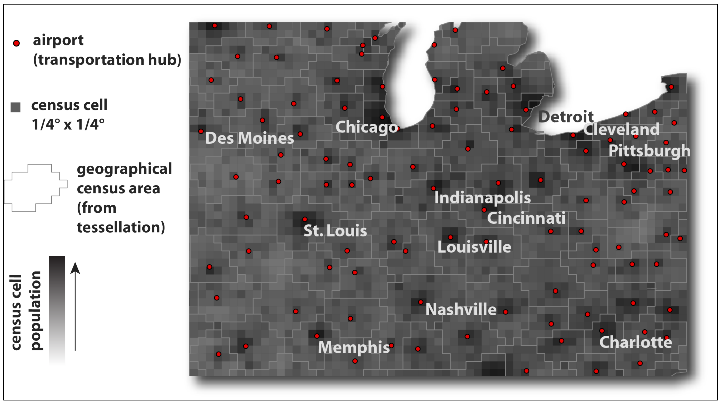

**Figure 2. Census areas, basins in GLEAM**

References

1. Balcan, D., et al., *Multiscale mobility networks and the spatial spreading of infectious diseases.* Proc. Natl. Acad. Sci. U.S.A., 2009. 106: p. 21484.

2. Balcan, D., et al., *Modeling the spatial spread of infectious diseases: The GLobal Epidemic and Mobility computational model.* Journal of Computational Science, 2010. 1: p. 132.