UMNSPH 2016-17 Influenza Forecast Exercise Model Description

Yang Liu*, Joseph Servadio, and Matteo Convertino

HumNat Lab, Division of Environmental Health Sciences, School of Public Health,
University of Minnesota-Twin Cities

November, 2016

Influenza demonstrates strong seasonality and inter-annual variability at the HHS level as well as on the national level. Existing literature has verified the links between influenza incidence and environmental factors such as temperature and humidity (Lowen et al., 2007; Shaman and Kohn, 2009; Tamerius et al., 2013). Utilizing these links, a generalized linear model (GLM) was developed to forecast for influenza incidence last season. This year, substantial improvements have been made to this model in order to increase its forecast accuracy. Specific procedures used to accomplish improvements will be discussed below.

Data

The primary data source for environmental information is NASA Global Modeling and Assimilation Office (GMAO).The specific data set we used is a long-term global reanalysis named MERRA-2 (Ostrenga, 2015). The product we used, **tavg1_2d_slv_Nx**, is updated every month mid-month and is vailable hourly from 1980. The biggest temporal gap between data availability and forecast target is 2-6 weeks (compared to last year, which had a temporal gap of 9 months.) We extracted data on 2-meter air temperature as the temperature variable and 2-meter specific humidity as the humidity variable. Last year, humidity was not available. We were forced to use rainfall and snowfall as proxies.

In order to match with the temporal scale to public health data, we up-scaled hourly data to weekly. During this process, we examined two statistical characteristics - mean and variation. Weekly mean temperature (T)/ humidity (H) level is calculated by taking the arithmetic mean of daily mean temperature. Weekly temperature variation (TV) or humidity variation (HV) is calculated by taking the standard deviation of daily maximum and minimum temperature and humidity levels. Exploratory analyses showed that there are significant positive correlations between same-week mean temperature/ humidity levels and influenza occurrences, and there are negative correlations between same-week temperature/ humidity variations and influenza occurrences. However, these relationships have different magnitudes when different HHS regions are under consideration.

Model Development

The following steps are repeated for each HHS region:

> **Step 1:** Use mutual information (MI) (Cover and Thomas, 2005) to explore the dependencies between independent variable and its own history and with the history of contributing factors. MI is a concept in information theory that is frequently compared to correlation. However, without assumptions on linearity or continuity, MI is considered a more general measurement of dependency and can be defined as:

$$I(X_t; Y_{t-\tau}) = \sum_{x_t y_{t-\tau}} P_{X_t Y_{t-\tau}}(x_t, y_{t-\tau}) \log \frac{P_{X_t Y_{t-\tau}}(x_t, y_{t-\tau})}{P_{X_t}(x_t) P_{Y_{t-\tau}}(y_{t-\tau})} \tag{1}$$

*Email: liux3204@umn.edu

where τ is the time lags being examined between the two time series. For the purpose of this study, τ from 1 to 52 are examined. The purpose for this step is to set up the dependent variable space for further consideration.

Step 2: Use variance inflation factor (VIF) to eliminate candidates from the dependent variable space that can already be well represented others. For dependent variable i,

$$VIF_i = \frac{1}{1 - R_i^2} \tag{2}$$

where R_i^2 is from the linear model where all other dependent variables were used to describe dependent variable i. Threshold (θ) for VIF_i used in this study is 5. Only dependent variables are considered in this step. The purpose of this step is reducing data-based multicllinearity before dependent variable selection.

Step 3: Use Breiman's random forest algorithm (Breiman, 2001) to evaluate the remaining candidates in the dependent variable space. No model framework has been assumed at this stage. But different from Step 2, independent variable is now under consideration. Percent increase in Mean Square Error (MSE) is used as a selection criteria as the final set of dependent variables. The threshold value is set to 5 in order to keep the model relatively concise without losing information.

Step 4: Use lognormal model to run the regression.

Then the relationship between HHS regions and national influenza occurrence is established as:

$$y_{national} = \sum_{h=1}^{10} \beta_h * y_h \tag{3}$$

Forecast

Environmental information that is not yet available are approximated from bootstrapping of historical records. For instance, the weekly T at week 45 is the mean of a random sample of week 45 Ts in known history. Forecast is made one week at a time and is repeated until the entire rest of the season is covered. However, since the forecast window is short, the uncertainty involved in each forecast is artificially suppressed. To overcome this issue, we introduced a multiplier of standard deviation that add uncertainty back to the analysis. The underlying assumption is that during early season, there is higher uncertainty - there is still many directions where the season can head towards. As time goes on, we know more and more about the season. Towards the end, we are nearly definite about how the season will turn out. All probability disributions the outcomes are obtained through Monte Carlo simulations (n=1000).

Computational Requirements

All analysis described above have been implemented using R (v3.2.4). Core packages used include [**infotheo**],[**usdm**],[**randomForest**] and [**cdcfluview**].

References

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Cover, T. M. and Thomas, J. A. (2005). *Elements of Information Theory*.

Lowen, A. C., Mubareka, S., Steel, J., and Palese, P. (2007). Influenza virus transmission is dependent on relative humidity and temperature. *PLoS Pathogens*, 3(10):1470–1476.

Ostrenga, D. (2015). inst3_3d_asm_Cp: MERRA-2 3D IAU State, Meteorology Instantaneous 3-hourly (p-coord, 0.625x0.5L42), version 5.12.4.

Shaman, J. and Kohn, M. (2009). Absolute humidity modulates influenza survival, transmission, and seasonality. *Proceedings of the National Academy of Sciences of the United States of America*, 106(9):3243–3248.

Tamerius, J. D., Shaman, J., Alonso, W. J., Bloom-Feshbach, K., Uejio, C. K., Comrie, A., and Viboud, C. (2013). Environmental Predictors of Seasonal Influenza Epidemics across Temperate and Tropical Climates. *PLoS Pathogens*, 9(3).