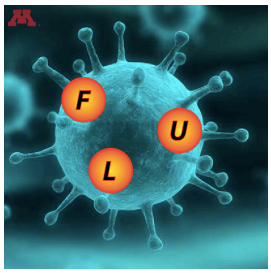

HumNat 2015-16 Influenza Forecast Model

Y. Liu[1], M. Convertino[1,2,3,4]

1. HumNat Lab, Division of Environmental Health Sciences and Public Health Informatics Program, School of Public Health, University of Minnesota, Twin Cities, MN, USA
2. Institute on the Environment, University of Minnesota, Twin Cities, MN, USA
3. Institute for Engineering in Medicine, University of Minnesota, Twin Cities, MN, USA
4. Biomedical Informatics and Computational Biology Program, University of Minnesota, Twin Cities, MN, USA

Correspondence: matteoc@umn.edu, liux3204@umn.edu

Summary

Influenza incidence demonstrates strong seasonality in the form of inter-annual variability at both local/ state and national scale. Environmental factors constituting the suitability for influenza virus transmission from people to people are likely to strongly contribute to such epidemiological pattern. The information theoretical approach that we develop aims to further explore the connection between environmental factors and epidemiological outcomes for producing influenza forecasts. The approach (i) identifies the most important predictors and (ii) uses a stochastic generalized linear model (GLM) to generate decoupled local and national forecasts for the upcoming influenza season (2015-16).

Data

The developed model relies on time series of influenza incidence and of relevant environmental factors. Information on confirmed influenza incidence is not available for a variety of reasons (e.g. high costs to confirm suspected influenza cases). Thus, weighted Influenza-like illness (ILI) percentage is used as a statistically significant estimator of influenza incidence in the context of this forecasting challenge. This information has been obtained through the R package 'cdcfluview' [1] that is updated weekly. The ILI percentage (i.e., ILI over the number of outpatients) is the dependent variable of the model. This variable is available between the 40th week of 1997 and the 41st week of 2015.

The source of environmental factors is the United States Historical Climatology Network (USHCN) created by the National Climatic Data Center (NCDC) and the National Oceanic and Atmospheric Administration (NOAA) [2]. Environmental factors considered for the model are precipitation (*PRCP*), snowfall (*SNOW*), minimum (*TMIN*) and maximum air temperature (*TMAX*). These factors are converted from daily to weekly scale in order to be at the same temporal scale of epidemiological data. Environmental data is available from Jan. 1st 1926 to Dec. 31st 2014. These environmental data do not cover Hawaii, Puerto Rico, U.S. Virgin Island, and Washington D.C. The model assumes that these states have similar environmental conditions of the states belonging to the same Health and Human Service Regions defined by CDC for influenza.

In the calibration of the model we only use data between the 40th week of 1997 and the 52nd week of 2014 because it this is the only temporal window in which both environmental and epidemiological data are available. Missing data within this temporal window are replaced by value estimated by a linear interpolation.

Forecasting Model

The selected model for forecasting influenza is a stochastic model based on the principle of decomposition of variance of the outcome variables. In this particular context we adopt the mutual information as a model to discern predictor importance and interaction explaining the variance of ILI. Considering model factor variance as partial information to reconstruct the outcome variable, a functional information network can be built for different temporal and spatial window. Yet, the model can be considered as a network-based model where variance is apportioned suitably to a maximum information principle model over space and time to maximize model accuracy. Reaction-diffusion models can be considered in analogy when information of interaction of variables is transferred to model output or when different regions are made intercorrelated after proper data analysis.

The signal observed in a preliminary data analysis shows significant lagged effects of environmental factors on ILI percentage. Results from analyses on cross correlation functions verified this observation. Furthermore, auto-correlation functions (ACF) of ILI percentage time series identified a finite correlation function for all macro-regions. The calibration part of the model consisted in identifying the optimal combination of lagged factors from the past that can best explain the ILI patterns. For this purpose we use a mutual information model [3-4] that explores all metamodels built on thresholded mutual information among environmental factors. The mutual information is given by:

$$I(X;Y) = \sum_{x,y} P_{XY}(x,y) log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} \quad (1)$$

Where P_{XY} is the joint probability between any environmental factor X and Y. Y is here taken as ILI but Y can be any other environmental factor when the interdependency of model factors needs to be assessed. P_X and P_Y are the marginal probabilities of X and Y. Lagged environmental factors with the highest mutual information are chosen as the independent variables of the model. These factors explain most of the variance of ILI. The adopted stochastic GLM assumes that the probability distribution of ILI percentage is be a beta distribution as verified by data. The beta distribution is as follows:

$$f(y:\alpha,\beta) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)} \quad (2)$$

where α and β are the shape factors of the distribution in which y is ILI. B is the incomplete beta function. This model has been chosen because the dependent variable considered is a proportion variable bounded in the range of [0,1]. To verify the beta distribution of ILI we plotted the histogram of the ILI percentage and calculated the square of the skew-ness and the kurtosis of that distribution. The value of these moments identified the beta distribution as the best fitting distribution. The forecasting model structure after variable selection and calibration is

$$\log\left(y_t^{HHSi}\right) \sim PRCP_{t-j}^{HHSi} + SNOW_{t-k}^{HHSi} + TMIN_{t-m}^{HHSi} + TMAX_{t-n}^{HHSi} + Season_t^{HHSi} + Trend_t^{HHSi} + AR_{t-p}^{HHSi} \quad (3)$$

where *j , k, m ,n* and *p* are lag terms determined after variable selection (maximizing the mutual information in equation (1)). These lags range between 0 and 12 weeks. *HHS* stands for Health and Human Services Regions, and *i* ranges from 1 to 10. *Season* and *Trend* were decomposed from original time series of ILI percentage using locally weighted smoothing (LOESS) [10], that is a non-parametric regression model. AR represents the autoregressive term. The final model selection was based on the value of pseudo R squareds as well as the Akaike Information Criterion (AIC) for model forecast fitting.

Nation-wide ILI percentage was then estimated through the following model that is also a beta distribution based GLM:

$$\log\left(y_t^{nation}\right) \sim \sum_{i=1}^{10} y_t^{HHSi} \tag{4}$$

Then, the forecasting of ILI percentage during the upcoming season is based on the aforementioned fully calibrated model. This model hopes to achieve this goal by generating local/ state level forecasts of explanatory variables and then plugged them back into the relationship from Equation (3). Autoregressive integrated moving average (ARIMA) model and LOESS are both used in this forecast exercise. Probability distributions of the weekly forecasts were obtained through Monte Carlo simulation (n=500). Local/ state level forecasts are again upscaled to national level using Equation (4).

Further refinement of the model will consider potential spatial dependencies, scaling and universality of mutual information networks. Furthermore a full global sensitivity and uncertainty analyses will be run for the influenza model.

Computational Details

The model is developed in R. Packages used include: '**betareg**' [11], '**cdcfluview**' [1], '**fitdistrplus**' [12], '**forecast**' [13, 14], '**infotheo**' [15],'**PERregress**' [16], '**reshape2**' [17],'**zoo**' [18].

References

[1] Bob Rudis (2015). cdcfluview: Retrieve U.S. Flu Season Data from the CDC FluView Portal. R package version 0.4.0. <http://CRAN.R-project.org/package=cdcfluview>

[2] M. J. Menne, C. N. Williams, Jr., and R. S. Vose, 2015. United States Historical Climatology Network Daily Temperature, Precipitation, and Snow Data. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee. <http://cdiac.ornl.gov/epubs/ndp/ushcn/daily_doc.html>

[3] Meyer, P. E. (2008). Information-Theoretic Variable Selection and Network Inference from Microarray Data. PhD thesis of the Universite Libre de Bruxelles.

[4] Cover, T. M. and Thomas, J. A. (1990). Elements of Information Theory. John Wiley, New York.

[5] Cribari-Neto, F., and Zeileis, A. (2010). Beta Regression in R. Journal of Statistical Software, 34(2), 1–24. <http://www.jstatsoft.org/v34/i02/>.

[6] Ferrari, S.L.P., and Cribari-Neto, F. (2004). Beta Regression for Modeling Rates and Proportions. Journal of Applied Statistics, 31(7), 799–815.

[7] Grün, B., Kosmidis, I., and Zeileis, A. (2012). Extended Beta Regression in R: Shaken, Stirred, Mixed, and Partitioned. Journal of Statistical Software, 48(11), 1–25. http://www.jstatsoft. org/v48/i11/.

[8] Kosmidis, I., and Firth, D. (2010). A Generic Algorithm for Reducing Bias in Parametric Estimation. Electronic Journal of Statistics, 4, 1097–1112.

[9] Simas, A.B., Barreto-Souza, W., and Rocha, A.V. (2010). Improved Estimators for a General Class of Beta Regression Models. Computational Statistics & Data Analysis, 54(2), 348–366.

[10] R. B. Cleveland, W. S. Cleveland, J.E. McRae, and I. Terpenning (1990) STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics*, **6**, 3–73.

[11] Francisco Cribari-Neto, Achim Zeileis (2010). Beta Regression in R. Journal of Statistical Software 34(2), 1-24. < http://www.jstatsoft.org/v34/i02/>.

[12] Marie Laure Delignette-Muller, Christophe Dutang (2015). fitdistrplus: An R Package for Fitting Distributions. Journal of Statistical Software, 64(4), 1-34. < http://www.jstatsoft.org/v64/i04/>.

[13] Hyndman RJ (2015). "forecast: Forecasting functions for time series and linear models". R package version 6.1, <http://github.com/robjhyndman/forecast>.

[14] Hyndman RJ and Khandakar Y (2008). "Automatic time series forecasting: the forecast package for R." Journal of Statistical Software, *26*(3), pp. 1-22. <http://ideas.repec.org/a/jss/jstsof/27i03.html>.

[15] Patrick E. Meyer (2014). infotheo: Information-Theoretic Measures. R package version 1.2.0. <http://CRAN.R-project.org/package=infotheo>

[16] Peter Rossi. (2013). PERregress: Regression Functions and Datasets. R package version 1.0-8. <http://CRAN.R-project.org/package=PERregress>

[17] Hadley Wickham (2007). Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12), 1-20. <http://www.jstatsoft.org/v21/i12/>.

[18] Achim Zeileis and Gabor Grothendieck (2005). zoo: S3 Infrastructure for Regular and Irregular Time Series. Journal of Statistical Software, 14(6), 1-27. <http://www.jstatsoft.org/v14/i06/>