2016-17 Influenza Forecast Initiative
HumNat Model-2 the "Modified SARIMA Model"

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

General Motivation

This model is developed using only time-series statistics and public health surveillance data. The original motivation for this model is to develop a baseline for comparison for other models of our team. We are also curious about exploring the value of information (VoI) of different models, that is defined as the value added by having additional knowledge (e.g. driven by environmental information, contact network data, local vaccination rate, etc.) on the predicted outcomes. Such information can be used to motivate future inter-disciplinary disease forecast partnerships.

Model Description

The public health surveillance data used for this model is extracted from the R package **[cdcfluview]**. The dataset is truncated at year 2002, as before this year there were substantial amount of missing data each year. We did not exclude the 2009-2010 H1N1 Pandemic as we believe it is capable of providing crucial information on the population level dynamics of the take-off and decay of epidemics.

The primary approach used for this model is seasonal autoregressive integrated moving average (SARIMA) model:

$$ARIMA(p,d,q)(P,D,Q)_m \tag{1}$$

[*] Email: liux3204@umn.edu

where p is the number of autoregressive terms, d is the differences needed for stationarity, and q isnumber of lagged forecast erros. P,D and Q are the corresponding parameter for the seasonal part of the model. And m captures the length of a unit cycle. In this study particularly, we set m to 52 as the influenza seasonal cycle is defined as one year.

In order to look for the best set-up, we explored the auto-correlation and the partial auto-correlation functions of each time series, as well as the lagged difference and the double lagged difference of that time series. In order to constrain the confidence interval to positive ranges, we transform the ILI% by taking its natural logirthm. We then ran every time series with the function '**auto.arima**' from the packages [**forecast**], using the Akeike Information Criterion (AIC) as key model selection criterion. Although the goodness of fit was good, some of the forecast results using the automatically selected models do not show a typical shape of an epi-curve. Therefore, we continued exploring by looking at the auto-correlation and the partial-correlation functions of the model residuals. In the end, we have decided on the following overall structure based on subjective interpretation:

$$ARIMA(3,1,q)(0,1,Q)_{52} \tag{2}$$

This implies that we are fixing p, d, P and D across the board: they will be the same for all HHS Regions as well as for the national level flu analysis. As for q and Q, we will test all modes with values 0-3 and find the combination that minimizes AIC for each location. Consequently, we have tested 16 (4*4) models for each geographic unit.

To assess the uncertainty of model forecast, we used the function '**simulate**'. However, during the simulation phase, we introduced some additional rejection rules. Reject rules are designed to throw out forecasts that we already know are not realistic. For instance, HHS Regions are not expected to exceed 1.5x the size of their corresponding peaks of the 2009-10 pandemic. HHS regions are also not expected to have more than 3 peaks of similar sizes during a given season. These techniques are similar to the acceptance-rejection methods or the rejection sampling techniques in mathematics, which allows uniform random sampling regulated by N-dimension functions. In our model, simulation is stopped when there are 2000 successful samples, defined as those not meeting any of the rejection rules. Failed samples are not used to deriving the probability distributions of the results.

Computational Requirements

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