Delphi ArcheFilter: Methodology

http://delphi.midas.cs.cmu.edu

2015-11-02

Overview

The ArcheFilter (Archetype-driven Unscented Kalman Filter (UKF)) is a new system from our group this season. It combines an empirical model of flu (the archetype) with several digital surveillance signals (via the UKF). Our signals currently consist of data from twitter, wikipedia, and CDC. In the near future we hope to add a new signal based on Google search query volume. This system is still largely experimental, and we expect that it, like flu, will rapidly evolve over the coming weeks.

Components

The main components are the Archetype (a model of flu trajectories) and the UKF (a data-assimilation technique). Additionally, we have pre-processing and post-processing strategies to minimize surprise. Essentially we allow for small changes in previously reported values, we assume that nothing is impossible, and we attempt to account for systematic variations in wILI due to holiday effects.

Accounting for Systematic Bais: Modelling Holidays

We have observed a striking, recurring increase in wILI in all regions around weeks 50-01, presumably due to changes in behavior and reporting around holidays rather than to changes in actual flu incidence. To model this effect, we find a multiplier separately for weeks 50, 51, 52, and 01 that minimizes the second derivative of wILI across historical seasons. We estimate the removal or addition of holiday effects by multiplying or dividing wILI on these weeks by these constants.

Estimating the Present: Digital Surveillance

We collect data from Twitter (via http://healthtweets.org), Wikipedia (via http://dumps.wikimedia.org), and CDC (via http://gis.cdc.gov/) and use linear regression to fit these time series to wILI. We then use the regression models to provide an estimate, separately for each data source, of what the following week's wILI value will be. We assimilate these estimates using the UKF, which results in both a point estimate and a distribution of the 1-ahead wILI target.

Predicting the Future: The Archetype

To model how wILI evolves over time we attempt to describe a quintessentail, canonical flu trajectory for each region; we call this the flu "archetype". These archetypes consist of two parts: the expected (i.e. mean) wILI at each week and an amount of variance at each week. To build these archetypes, we gather wILI curves from past seasons (excluding the 2009 pandemic) and:

1. Reduce wILI at weeks 50-01 (using the Holiday model)
2. Smooth the curves (using a Gaussian kernel)
3. Align the curves by peak week
4. Measure the mean and unbiased variance across weeks

Understanding the Past: Modelling Backfill

We have observed that it often takes many weeks to months for a reported wILI value for a given region on a given week to stabilize; we call this process "backfill". Based on data provided by CDC for the 2010-2013 seasons and on data collected automatically starting on 2014w01, we estimate for each region the wILI variance due to backfill. This backfill variance is calculated as a function of the number of weeks that have passed since the first value was published. Essentially we try to describe how wILI is likely to vary over time within each region.

Expecting the Unexpected: Uniform Blending

Even though we may strongly believe that some outcomes are impossible, we assign all outcomes non-zero probability. There are a variety of reasons why an "impossible" outcome can happen (e.g. onset week is changed due to backfill, or peak week is extremely late due to pandemic wave), and we currently enforce a minimum probability of 0.5% in weekly bins and 1% in wILI bins.

Summary

This system gathers digital surveillance, projects it forward using an archetype trajectory, inflates wILI on holiday weeks, injects noise on past weeks proportional to backfill variance, measures target values on each sampled trajectory, finds the median and variance of each target, and blends final distributions with a small uniform probability.