Forecasting influenza activity using meteorological and Google Flu Trends data

S. Lefantzi[1], J. Ray[1] , G. Lambert[2], P. Finley[2] and H. Smith[2],

[1]Sandia National Laboratories, Livermore, CA and [2]Sandia National Laboratories, Albuquerque, NM

OBJECTIVE

Develop a data assimilation system (DAS) to track and forecast outbreaks using Open Source Indicators (OSI) of epidemiological activity

- Test case: Forecast flu activity in CA in a spatially resolved manner
- Data: Use Google Flu Trends (GFT) available at 11 CA cities as data from a "partially observed epidemic"
- Use meteorological data to spatially model flu activity at locations not tracked by GFT
- Validate against public health (PH) data

BACKGROUND

- Disease outbreaks cause changes in our online behavior e.g. web searches on symptoms, cures etc.
- Such searches have proven to be predictive of flu activity [Ginsberg et al, 2009]
- These online OSI are timely and collected by many organizations e.g. GFT
  - In contrast, public health (PH) reporting tends to be delayed by 1-2 weeks and has uneven spatial coverage (reporting is voluntary for most diseases)
- Meteorology can also be a leading indicator of many diseases
  - Rains precede mosquito-borne diseases, low humidity helps flu [Shaman et al, 2010]
  - Further, meteorological data is easily available at high spatiotemporal resolutions (as reanalysis data products)
- Combining OSI and meteorological data could thus be used to forecast flu
  - In [Shaman et al, 2013], the authors developed an ensemble adjustment Kalman filter to assimilate GFT and meteorological data to forecast flu
  - Demonstrated on 100+ US cities tracked by GFT, but no extension to locations outside the cities e.g. suburbs

FUNCTIONAL REQUIREMENTS

- The DAS should be able to provide forecasts of disease activity at locations (called anchors) tracked by GFT
  - For CA, there are 11 cities (anchors) tracked by GFT
- Forecasts should also contain some measure of predictive uncertainty e.g., confidence bounds
- Using nowcasts and forecasts at the anchors, the DAS should (spatially) predict flu activity at locations not directly tracked by GFT
  - High-res meteorological data and the dependence of disease activity on them could be used to construct a spatial model, or parameterization

TECHNICAL APPROACH

Components of a DAS

- A temporal component that sequentially assimilates time-series GFT data and produces forecasts
- A spatial model that predicts flu activity away from the anchors

The temporal component

- Consists of an ensemble Kalman filter (EnKF) driving a SIR model of flu
- Estimates the number of susceptible and infectious people
- Also estimates the mean infectious period and a time-dependent reproductive number
- Assimilation of GFT data produces a "calibrated" SIR model (after sufficient data has been ingested)
- Forecasts using the calibrated model simply means running the ensemble of SIR models forward
  - Produces a mean prediction and (computed) standard deviation σ around it

The spatial model

- Consists of a prior model of flu activity
- Constructed by fitting a linear model to GFT data at each of the anchor locations i, i ε $\mathcal{A}$, with temperature and humidity as predictors

$$GFT^{(i)}(t) = \sum_{k=0}^{K} \alpha_k^{(i)} T^{(i)}(t-k) + \sum_{l=0}^{L} \beta_l^{(i)} Q(t-l) + \varepsilon$$

- $(\alpha^{(i)}_k, \beta^{(i)}_l)$ are interpolated from i, i ε $\mathcal{A}$, to other locations using kernel smoothing
- This leads to a "mean model" ***MM(t)*** that can provide a prior prediction based solely on meteorology, i.e. without GFT data

$$MM^{(j)}(t) = \sum_{k=0}^{K} \alpha_k^{(j)} T^{(j)}(t-k) + \sum_{l=0}^{L} \beta_l^{(j)} Q(t-l) + \varepsilon$$

- The mean model, i.e. $(\alpha^{(i)}_k, \beta^{(i)}_l)$, is trained on historical data
- For any time outside the training period, one computes a discrepancy with respect to the mean model
  - $\Delta I^{(i)}(t) = I^{(i)}(t) - MM^{(i)}(t)$, i ε $\mathcal{A}$
  - $I^{(i)}(t)$ could be GFT data or a temporal forecast at the anchors
  - The discrepancies are interpolated from the anchors to an arbitrary location *j* via kernel smoothing

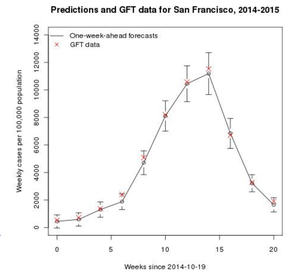
$$I(t;x_j) = \sum_{i,\, i \in \mathcal{A}} K(x_i, x_j)\, I(t, x_i)$$

- This provides a spatial model, or a means or transferring information from the anchors to any other location

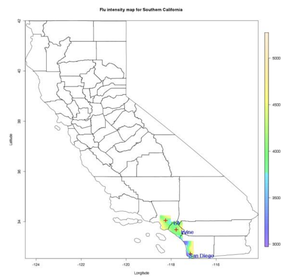
DAS PREDICTIONS

DAS Products

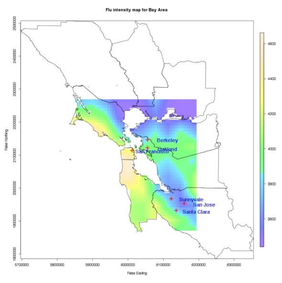
- One-week-ahead flu predictions for a given anchor show predictions and 3σ bounds bracketing the GFT data
  - EnKF is used with 200-member ensembles
  - Data assimilation starts 10 weeks before the start of the flu season
  - Forecasting starts in November
- The spatial model is trained on 2011-2013 GFT and meteorological data
- Together, the DAS produces maps of flu activity
  - Can be nowcasts or forecasts



Forecasts and data for San Francisco, 2014-2015 flu season. Error bars are 3σ bounds

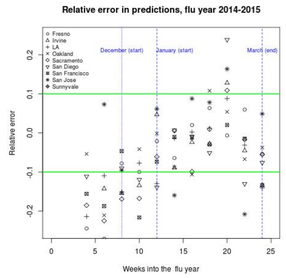


Predictions for Southern California

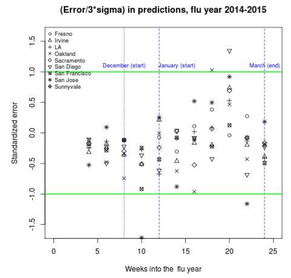


Predictions for San Francisco Bay Area

CHECKING TEMPORAL PREDICTIONS



Relative error in predictions



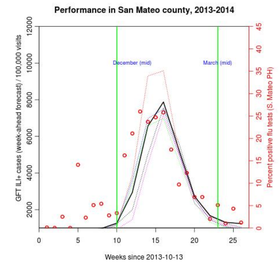
Standardized error

- Temporal predictive skill of the DAS was tested at the anchors
- 1-week-ahead forecasts within 10% of GFT in the January – March period (2014-2015) flu season; 2-week-ahead forecasts are within 20% error
- The 3σ bounds bracket the prediction error during the same period

CHECKING SPATIAL PREDICTIONS

Predicting flu in San Mateo

- Use DAS to provide 1-week ahead forecasts for San Mateo County, CA
- San Mateo is not tracked by GFT
- San Mateo PH department provides a data sheet with percentage of samples testing positive for flu
- Tested with nowcast and forecasts
- Tested on 2013-2014 flu season



CONCLUSIONS

- Open-Source indicators (OSI) of disease outbreaks, such as Google Flu trends, can be used to track and forecast epidemiological activity
- The dependence of epidemiological dynamics on meteorology (temperature and humidity for flu) and the availability of high-resolution meteorological data can be used make a spatial model for epidemiological activity
- A data assimilation system can be set up to assimilate OSI and meteorological data and produce spatiotemporal predictions
- The data streams being used are timely and easily available
- Preliminary tests and comparisons against public health data show that the DAS can be sufficiently accurate to be useful in practice

Acknowledgements

The project is funded by the Laboratory Directed Research and Development program at Sandia National Laboratories, Albuquerque, NM.

References

- [Ginsberg et al, 2009] J. Ginsberg et al, "Detecting influenza epidemics using search engine queries", Nature, 457:1012-1015, 2009.
- [Shaman et al, 2010] J. Shaman et al, "Absolute humidity and the seasonal onset of influenza in the continental United States", PLoS Biology, 8(2):e1000316, 2010.
- [Shaman et al, 2013] J. Shaman et al, "Real-time influenza forecasts during the 2012-2013 season, Nature Communications, 4:2837 doi: 10.1038/ncomms3837, 2013.

For additional information, please contact:

Sophia Lefantzi, Sandia National Laboratories, slefant@sandia.gov

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

SAND2015-3421C