Nowcasting influenza activity using Healthmap data

J. Ray[1] and J. Brownstein[2]

[1]Sandia National Laboratories, Livermore, CA and [2]Boston Children's Hospital, Boston, MA

OBJECTIVE

To demonstrate the use of open-source indicators (OSI) of epidemiological activity to nowcast flu

- Data sets: the number of flu-related documents, records, media reports etc. as collected by healthmap.org; also data from sentinel networks (US's ILINet and France)
- Test case # 1: nowcast flu in US (country-wide, regional and for New York City), including during the 2009 swine flu outbreak
- Test case # 2: nowcasting during swine flu outbreak in 2009, in France

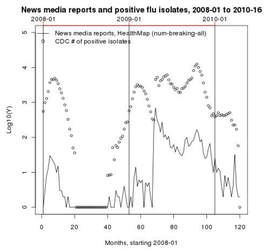
BACKGROUND

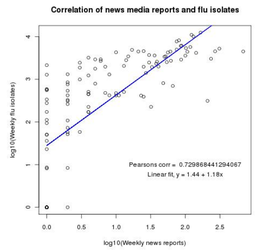
- Disease outbreaks cause changes in our online behavior
  - There are media reports, discussions on social media etc.
  - This data is collected and archived by organizations e.g. Healthmap
- Some of these OSI have been used in the quantitative modeling of disease activity
  - Google searches of flu-related terms are used to nowcast flu activity [Ginsberg et al, 2009]
  - Called Google Flu Trends
- Online OSI are collected in a timely manner
  - In contrast, public health (PH) reporting tends to be delayed by 1-2 weeks
  - Using OSI to compensate for the reporting delay is called *nowcasting*
- While google searches and tweets [Lambos & Cristianini, 2012] can be predictive in a nowcasting setting, there is no prior work investigating the usefulness of Healthmap (HM) data
  - Note : HM data is available even for countries with modest Internet penetration

HEALTHMAP DATA

- Contains news/media reports, Ministry of Health reports, ProMed Mail, "scraped" from the Web
- Tagged by date, location (province or city) and disease
- Data is archived; can be freely accessed at http://healthmap.org
- A data feed for DoD's Biosurveillance Ecosystem

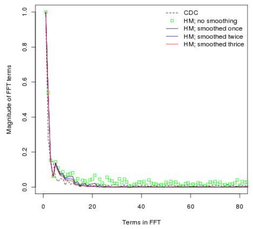
EXPLORING HM DATA



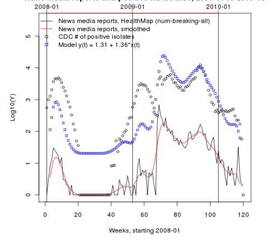


Comparison of ILINet and HM data

- There is a moderate level of correlation (Pearson correlation of 0.7)
- The HM data is very noisy and will require smoothing before being used in modeling



Plot of Fourier amplitude of smoothed HM data



Plot of smoothed HM data and a linear model for ILInet data

Smoothing HM data

- Apply a 5-point smoothing kernel to HM time-series
- Fit a simple linear model to log-transformed ILINet and HM data
- The predictions are in the correct ballpark, but need a far more sophisticated model

MODELING WITH HM DATA

- The low correlation between HM and ILINet data implies that a pure model such as $Y_t = g(X_t)$, where Y_t and X_t are ILINet and HM time-series, will not suffice
- However an autoregressive moving average model with an exogenous input (ARMAX) might work (c is a constant)

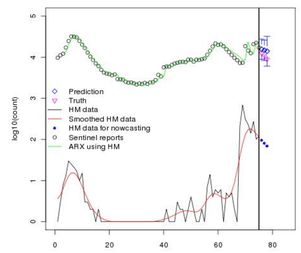
$$Y_t = c + \varepsilon_t + \sum_{l=1}^{L} \alpha_l Y_{t-l} + \sum_{j=0}^{M} \beta_j X_{t-j} + \sum_{k=1}^{K} \gamma_k \varepsilon_{t-k}$$

- The model is fitted to training data by optimizing on (L, M, K) and using AIC to keep the model simple
- Typically we will train the model on N weeks of ILINet data and forecast for weeks (N+1) : (N+3) using HM data.
- We will compare against an autoregressive model too ($\beta_j = \gamma_k = 0$)

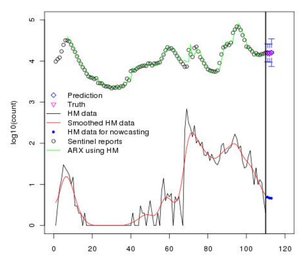
TESTS WITH USA DATA

Country-wide tests

- Use ARMAX models to nowcast morbidity (ILINet data) using HM
- Compute 3σ error bounds on the nowcasts
- Apply during 2008-2010 (the 2009 swine flu outbreak and the big 2009-2010 season)



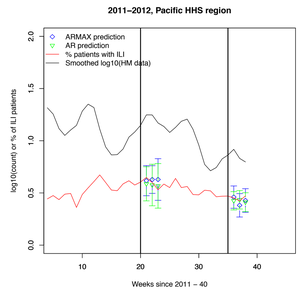
ARIMAX predictions versus ILINet data (training set: 75 weeks)



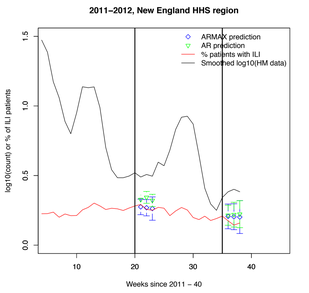
ARIMAX predictions versus ILINet data (training set: 110 weeks)

Regional tests

- Repeat tests for HHS regions where HM data counts are lower (2011-2012 season)
- Compare against AR models (no HM data to assist nowcasting)



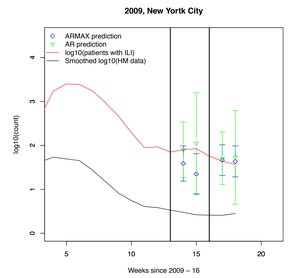
ARIMAX predictions for the pacific region (training set: 35 weeks)



ARIMAX predictions for New England (training set: 35 weeks)

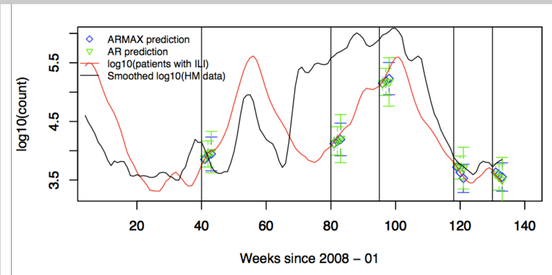
New York City test

- Apply to NYC during the swine flu outbreak
- Start date of analysis : 2009-01-01
- Training period : 16 weeks



In all cases, adding HM data (ARMAX model) provided the same or better accuracy than AR models (no HM data)

WHEN HM DATA MISLEADS



Predicting flu in France

- The spring-summer 2009 swine flu outbreak missed France
  - The media circus did not; HM data was misleading
- The training revealed the weak mutual information in HM data and the ARMAX models gracefully degraded to AR models
- Bottomline: ARMAX models are robust; when HM data misleads, it is ignored

CONCLUSIONS

- HM data can be used for nowcasting
- Its low correlation with flu surveillance data require that that one use more sophisticated time-series models such as ARMAX
- ARMAX models exploit the structural information in historical surveillance data and current HM data to provide accurate nowcasts
- In case the HM data misleads, the model gracefully degrades to a AR model
- HM data is available for countries with low Internet penetration and could be a way of tracking diseases in locations with incomplete PH reporting

Acknowledgements

The project was funded by the US Department of Defense.

References

- [Ginsberg et al, 2009] J. Ginsberg et al, "Detecting influenza epidemics using search engine queries", Nature, 457:1012-1015, 2009.
- [Lambos and Cristianini, 2012] V Lambos and N. Cristianini, "Nowcasting events from the social Web using statistical learning", IEEE Transactions on Intelligent Systems and Technology, 3(2):72, 2012.

For additional information, please contact:

J. Ray, Sandia National Laboratories, jairay@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.

SAND2015-3422C