Exceptional service in the national interest

SAND2015-3692C

Sandia National Laboratories



Forecasting Dengue Incidence from Social Media using Meteorological Covariates

Gregory Lambert, Jaideep Ray, Sophie Lefantzi,
Patrick Finley, Halley Smith

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

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. SAND NO. 2014-XXXXP

Problem and Motivation

Sandia National Laboratories

- Aim
  - Develop an early warning system for predicting dengue epidemics in India using easily accessible open-source data
    - Data sources: Google Dengue Trends, Healthmap.org data & meteorology (reanalysis products from NASA)
  - Use information from surrounding states in India, to build better forecasts at the state level
- Motivation
  - Many countries have poor public health reporting systems
    - Using easily accessible open-source info may be the only way to track diseases
  - Develop methods to forecast outbreaks under sparse data conditions leveraging correlated predictors

Background

Sandia National Laboratories

- Disease outbreaks cause changes in our online behavior (e.g. web search on symptoms, cures etc.)
- Such searches have proven to be predictive of flu activity [Ginsberg et al, 2009]
- Open-source indicators (OSI) are timely and collected by many organizations (e.g. Google Dengue Trends, GDT)
  - In contrast, public health reporting tends to be delayed with uneven spatial coverage
  - Meteorological data is easily available at high spatiotemporal resolutions (e.g. reanalysis data products)

Overview

Sandia National Laboratories

- Data collection
  - Google Dengue Trends (GDT): used to measure the severity of dengue activity in India (response)
  - GDT data was averaged by monthly totals for creation of regression components of the SARIMA/SARIMAX models
  - HealthMap articles: used as predictor for GDT
- Modeling approach
  - Hypothesis: If GDT and Healthmap (HM) data (online articles) are representative of dengue activity, then GDT should be modeled by HM data, precipitation & temperature (that govern mosquito lifecycles)
  - Build time series models for forecasting dengue incidence in India
    - Determine if online articles, temperature, and precipitation are predictive of dengue incidence
  - Use findings to construct space filling algorithms for conducting state level forecasting of dengue activity

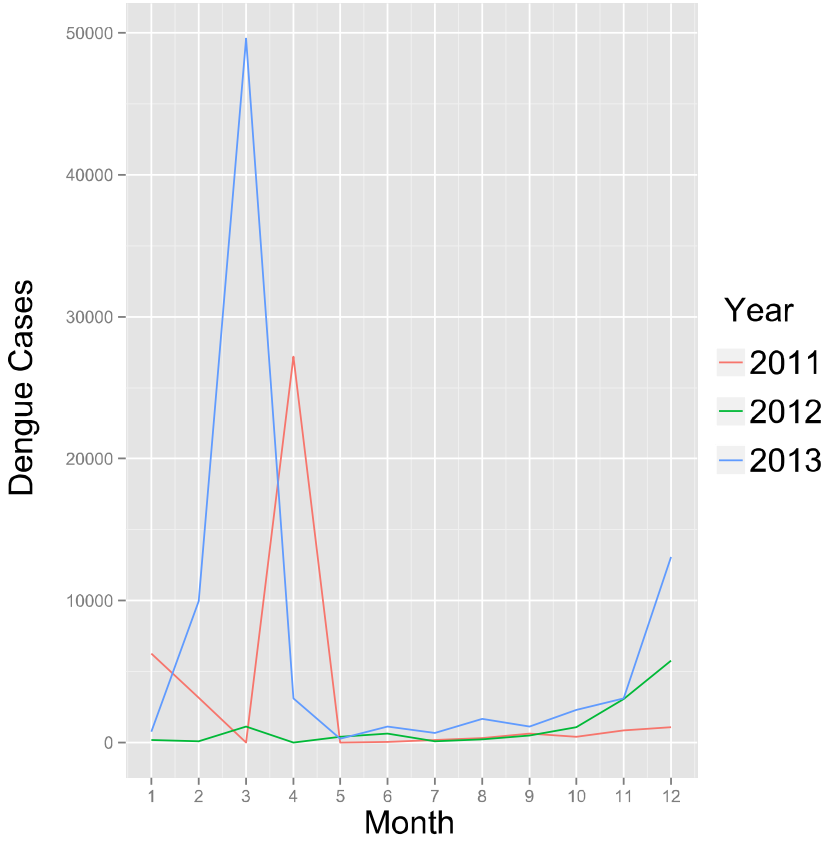
Forecasting dengue incidence in India

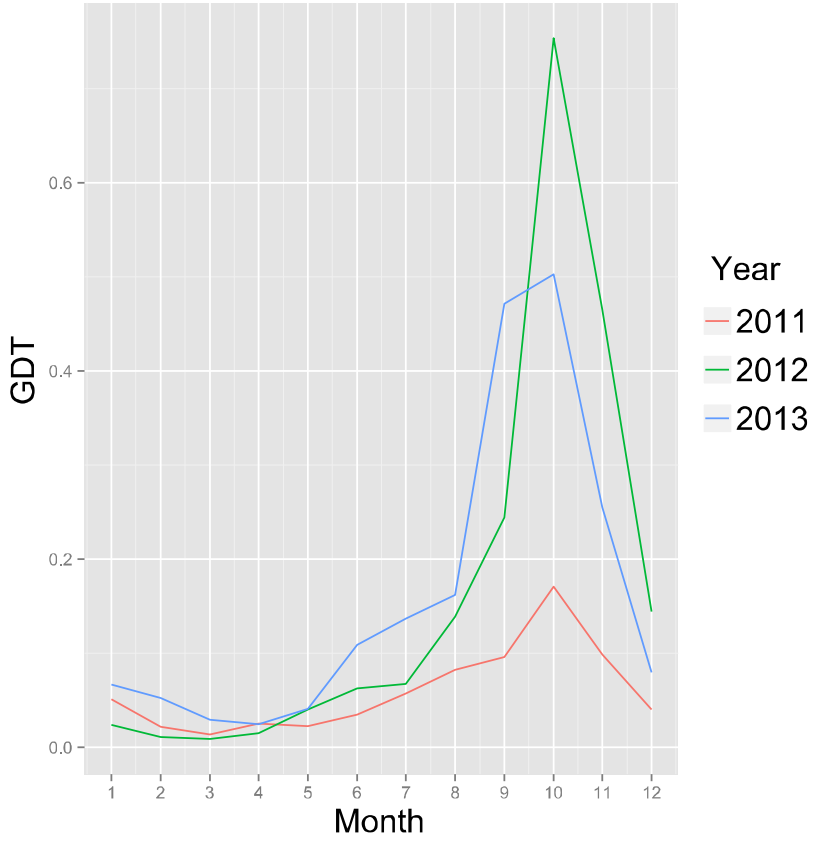
Sandia National Laboratories

- SARIMAX
  - S: adds seasonal component (periodicity)
  - AR(p)
    - Linear model that predicts present value of a time series using immediately prior value in time
    - Ex: AR(1): $x_t = \theta x_{t-1} + \varepsilon_t$
  - MA(q)
    - Term in a time series to capture current and previous (unobserved) white noise error
    - Ex: $w_t \sim N(0, \sigma^2)$ then 1st order (MA(1)) is $x_t = \mu + w_t + \theta_1 w_{t-1}$
  - I: stands for integrated, and allows the model to have a tendency (increasing or decreasing)
  - X: allows for external variables to be considered in the model

Exploratory analysis

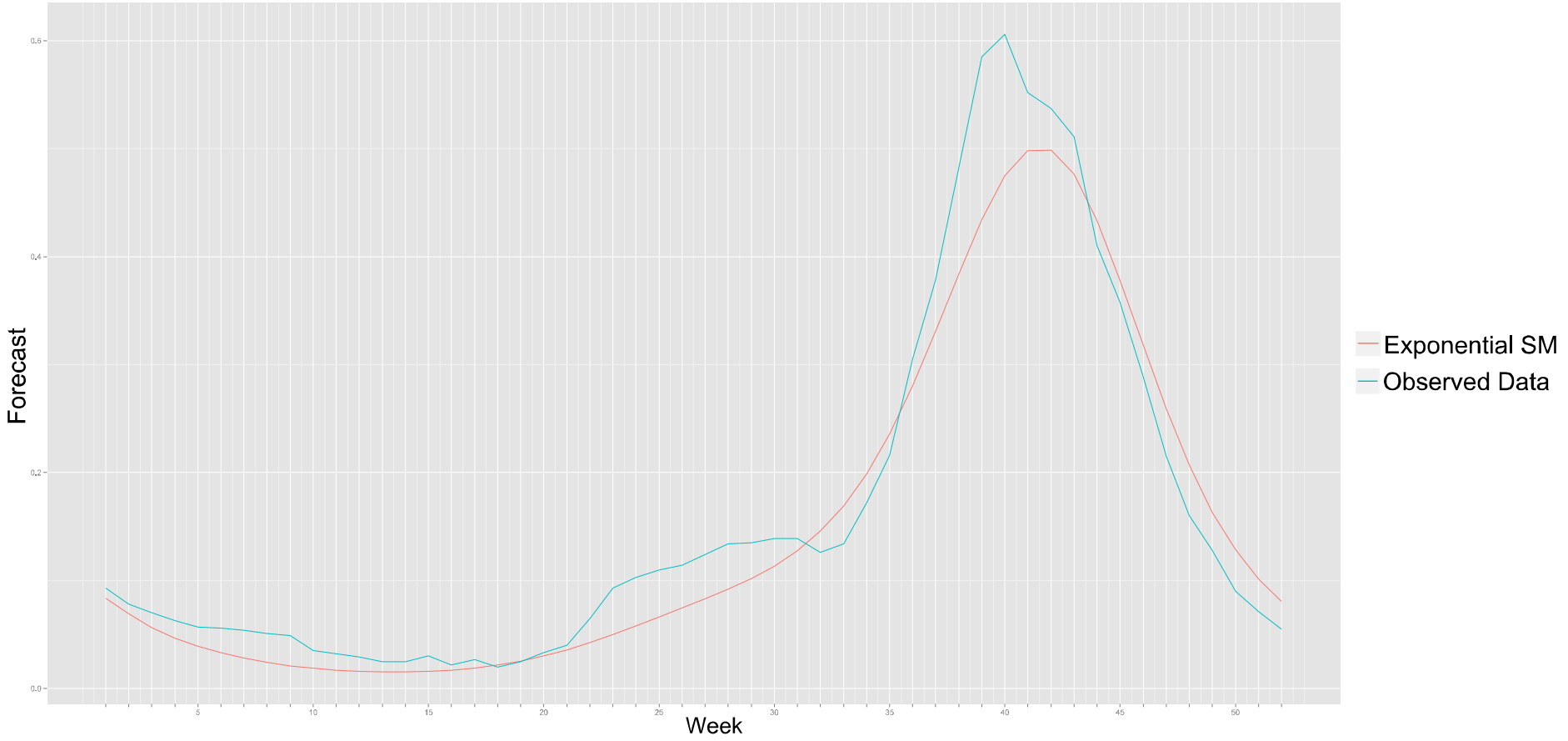
- Systematic effect in dengue time series data
  - Seasonality effect observed both in HM and GDT data





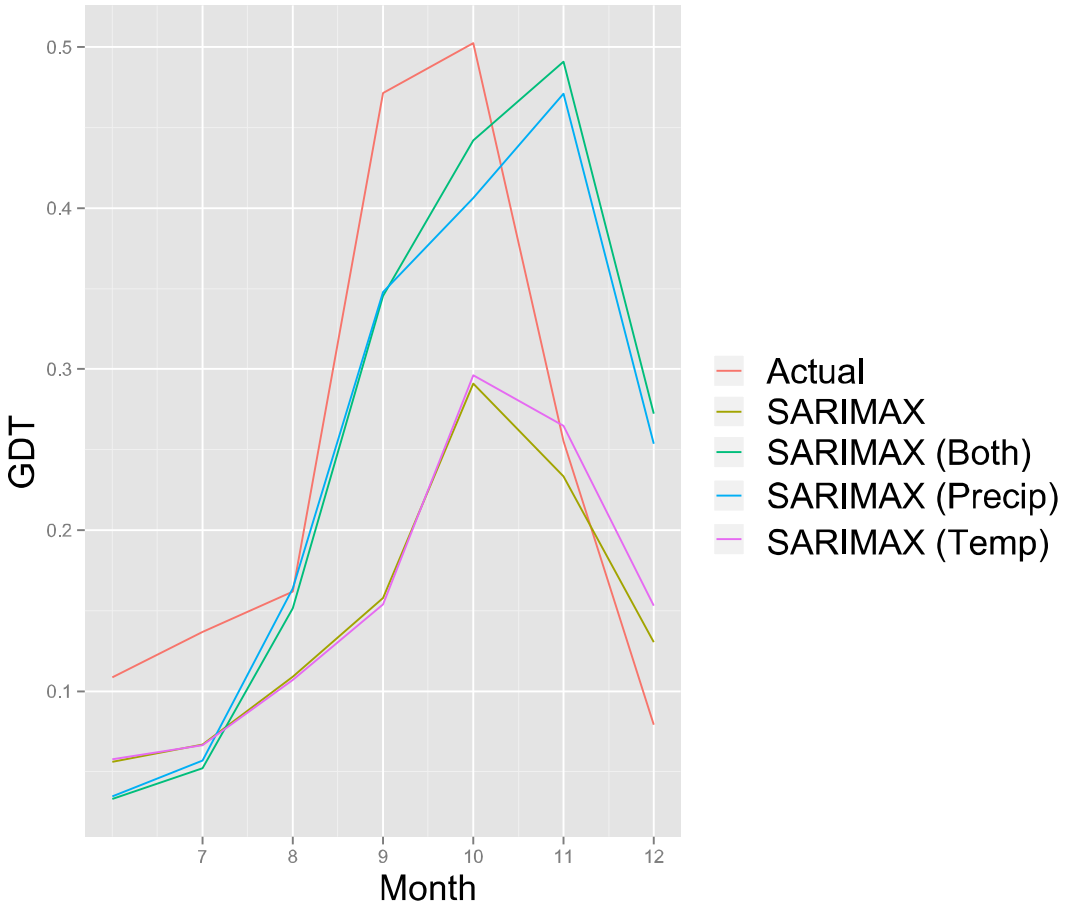
Exponential smoothing model

- Do a good job forecasting out-of-sample GDT data for the entire country
- Problem with this approach, state (province) specific data is not available.



Out-of-sample forecasting

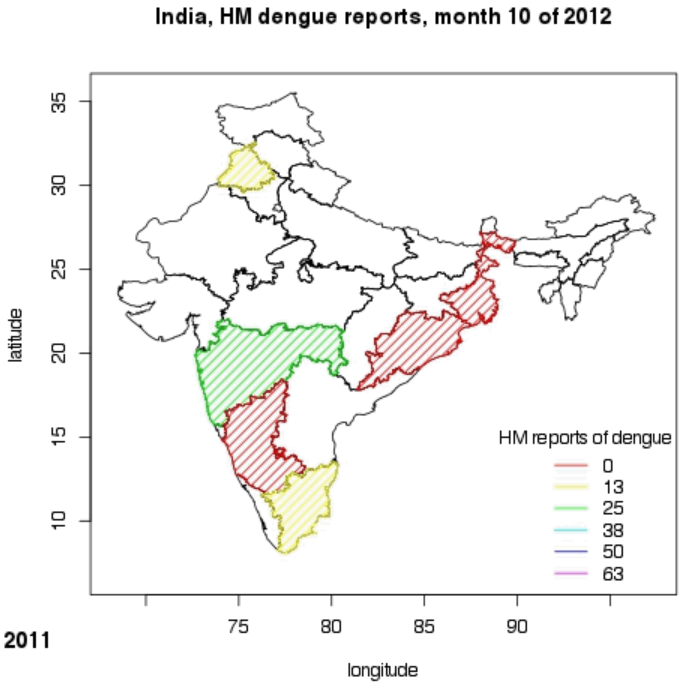
- Meteorological estimates strengthen forecasting of seasonal dengue outbreaks
- $ARIMAX(p,d,q)(P,D,Q)_s$
  - $ARIMAX(2,0,3)(0,0,0)_{12}$



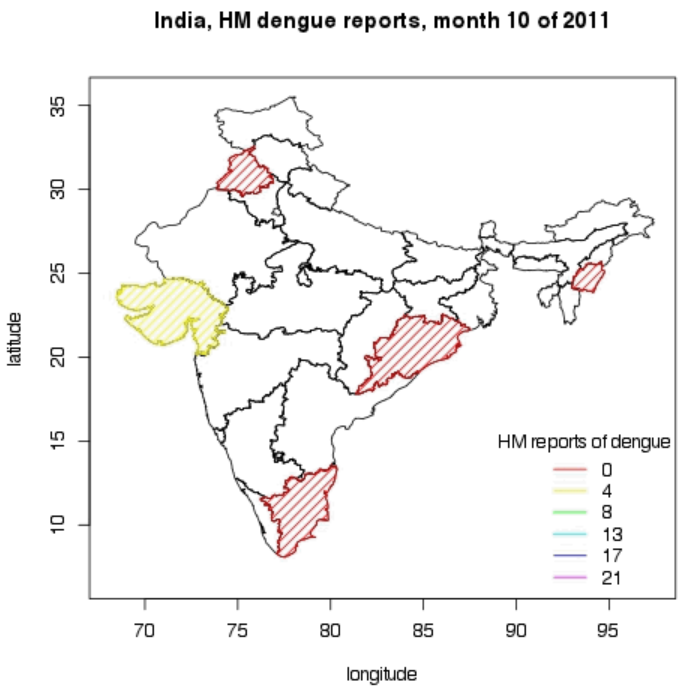
HM data

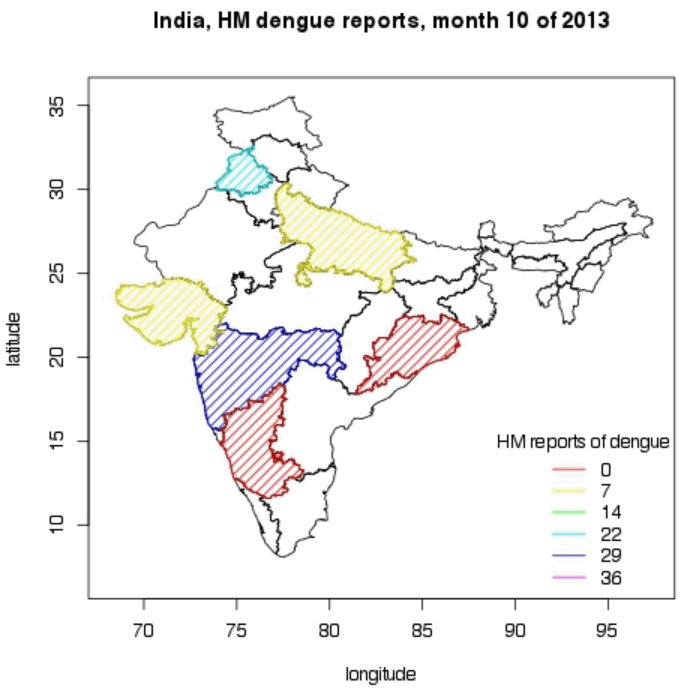
October

2011, 2012, 2013



- The data is gappy
- There are gaps in time too





Dengue space filling approach

Sandia National Laboratories

- Fill in missing meteorology and HM data for specified states in India taking into account both the time and spatial dependency structure of the data
- Algorithm
  - Select a target state with insufficient HM data
  - Define a spatial cluster around a target state with missing dengue data for a specified date range using a threshold from calculated Haversine distances
  - Subset data, taking spatial cluster from previous step including target state
  - Run a missing at random (MAR) multiple imputation to fill in missing data points
  - Cross-validate results with hold-out data

Dengue space filling approach

Sandia National Laboratories

- Continued
  - Square root transformation of the response variable (count data) employed to put less emphasis on outliers and to treat the response as continuous, this allows more information regarding the distribution to be captured under sparse data conditions
  - Transformation of Haversine distances to ordinal ranks to emphasize weighted importance of states closer to target state
  - Add a time predictor to the GLM to aid in predicting missing values

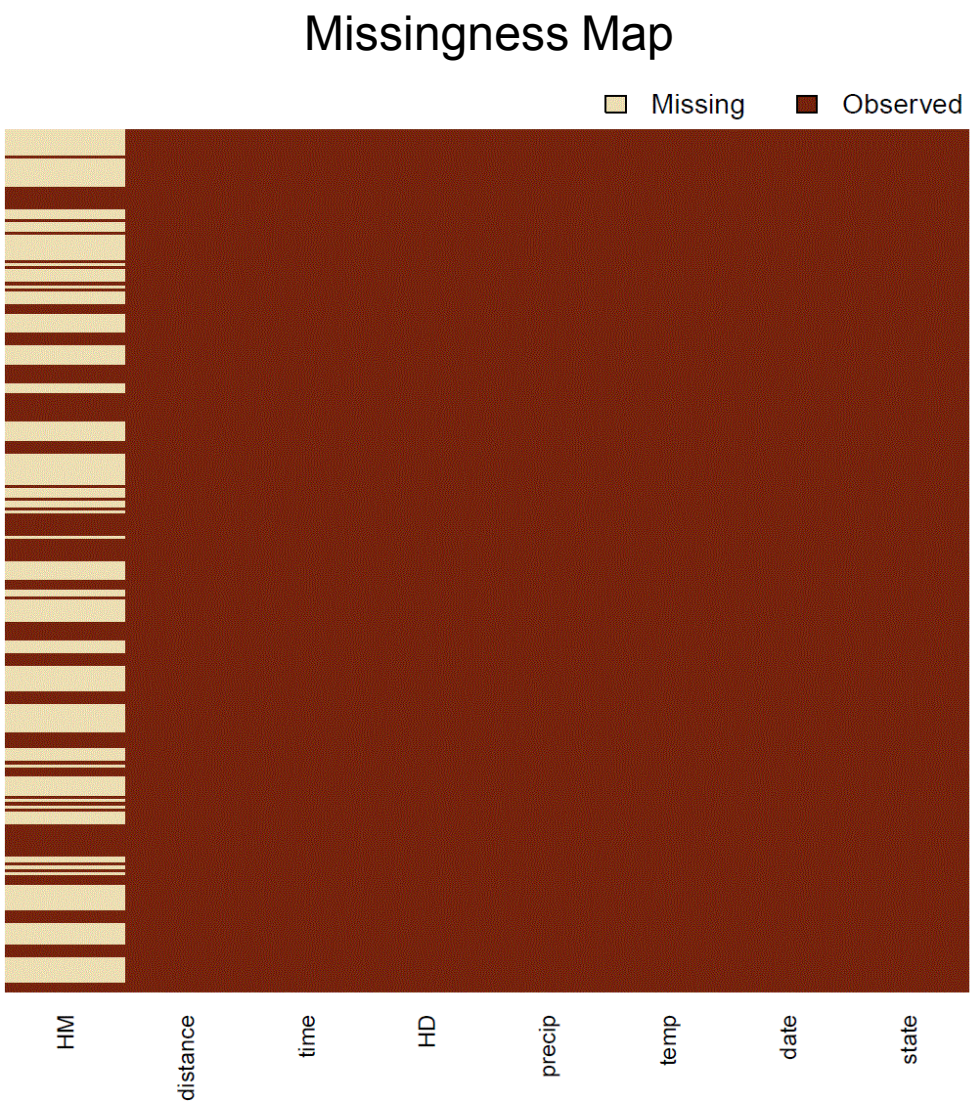
Imputation Algorithm

Sandia National Laboratories

- Algorithm
  1. Missing values in the data set are replaced with mean/median placeholder values
  2. Randomly selected variable (V) with missing values having placeholder values that have not been imputed are set back to missing (NA)
  3. Observed values from variable V from step 2 are regressed on the other variables in the input space including variables with placeholder observations
  4. Missing values from V are then replaced for predictions from the glm results derived from step 3
  5. Steps 2-4 are then repeated for each variable that has missing observations left in the input space, resulting in a complete dataset
- Number of cycles performed (steps 2-5) is determined by assessing the stability of convergence of the parameter estimates derived from the step 3

Imputing Haryana dengue activity

- $D_i = \beta_0 + \beta_1 + B_2 + \beta_3$
  - GLM for predicting missing dengue activity from distance, precipitation, and temperature
- $D_{it} = \beta_0 + \beta_1 t^1 + \beta_1 t^2 + \cdots + \beta_3 t^1 + \beta_3 t^2$
  - Allows for capturing the temporal dependency in the imputation
  - Assumption is that dengue activity varies smoothly over time and that we can describe cross-sectional observation i at time t



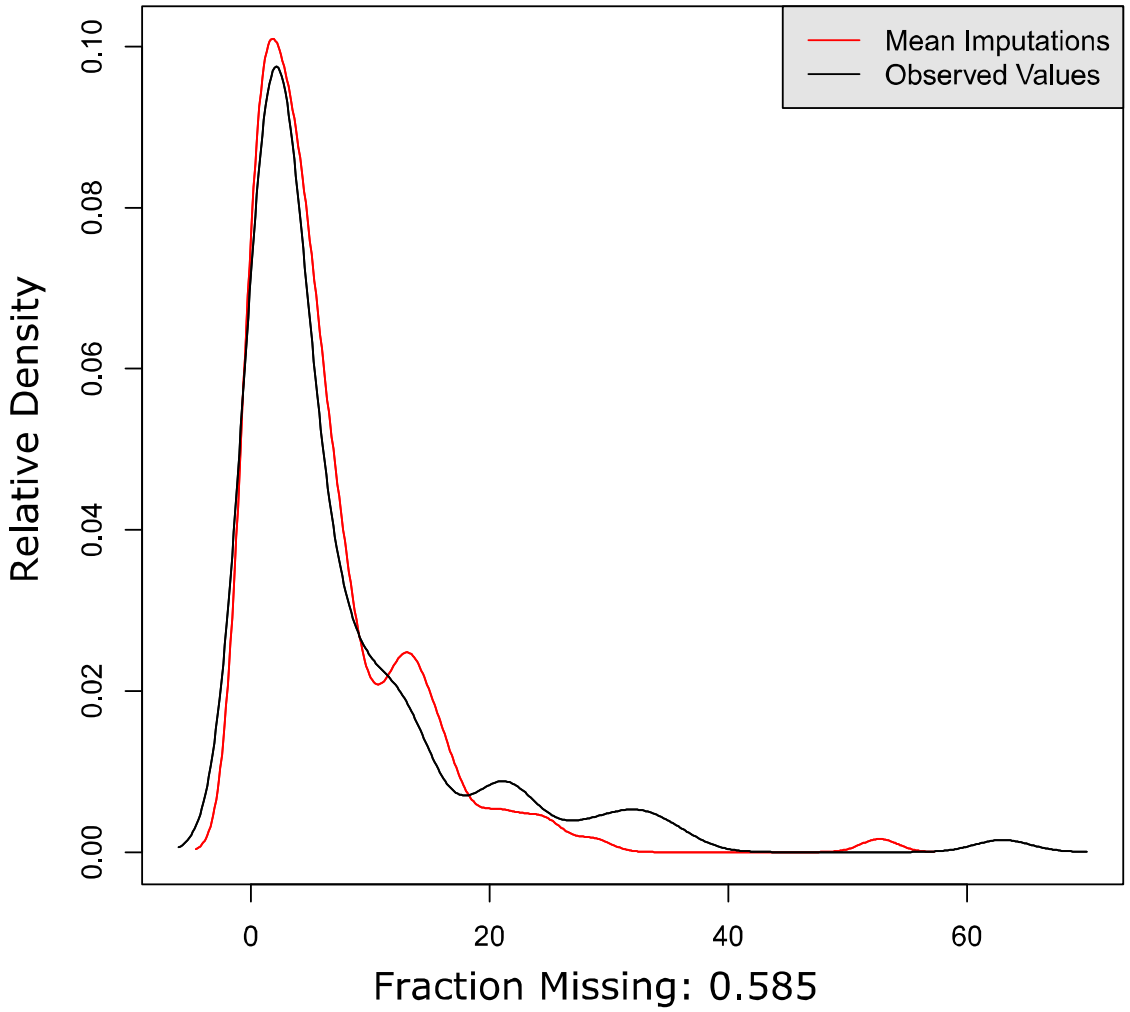
Dengue modeling and space filling

Sandia National Laboratories

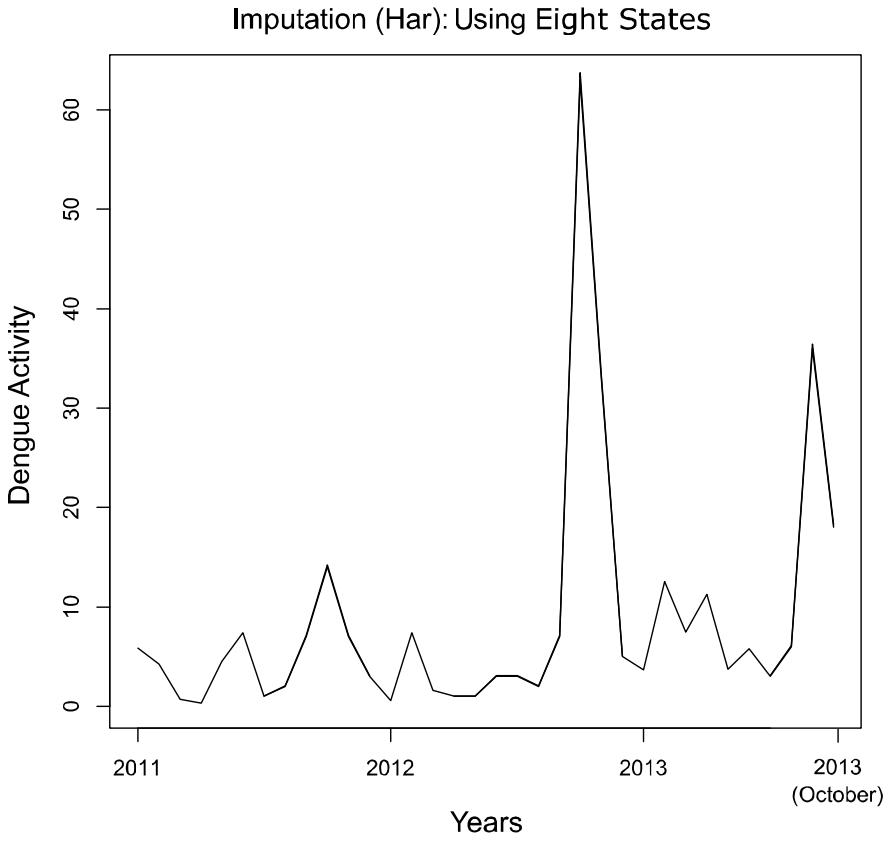
- Hypothesis: dengue in Indian state S in month *m* is related to dengue level in neighboring states $S_k(m + i)$, $-n < I < n$, as well as the precipitations $P_k(m+i)$ and temperatures $T_k(m+i)$
  - k in an index over the neighborhood of state S
  - Neighborhood includes the state itself
  - When confronted with a state with missing data, “borrow” information and meteorological dependence from neighbors within specified threshold

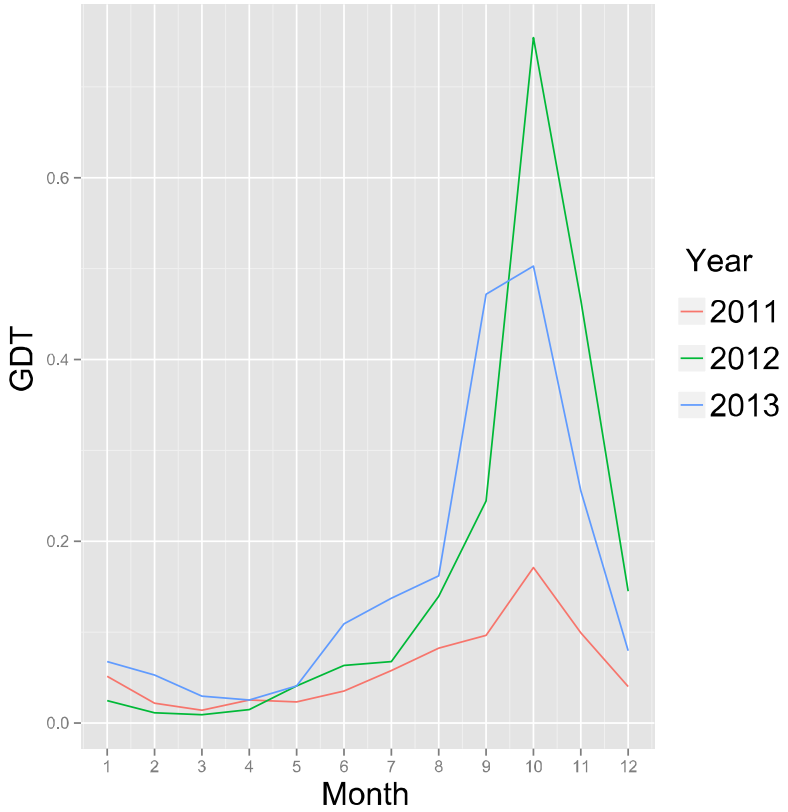
Imputing: Haryana dengue activity

- Expectation-maximization algorithm used for MLE
- Simulate from GLM
- Comparing densities: obviously we cannot expect identical distributions of imputed and observed data but strange comparisons can be signs of problems in an imputation model



Comparison of imputed Haryana dengue media data and GDT for all of India

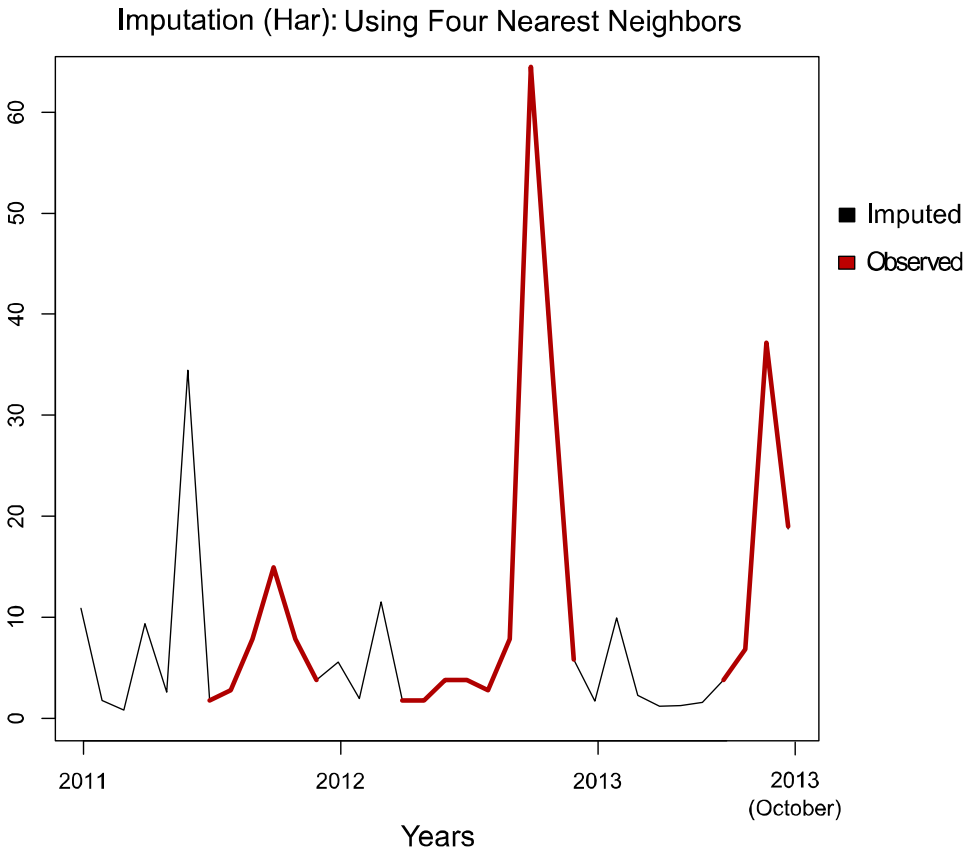




Out-of-sample forecasting – Haryana

- Time series comparison
  - Red lines are the true data
  - Black lines missing data
- To do
  - what are the uncertainty bounds on the predicted/imputed data?
  - How many other states can be gap-filled like this?

Months (2013)	Computed (8 neighbors)	Computed (4 neighbors)	Truth
November	8.2	12.5	8
December	4.9	3.96	3



Conclusions & future work

- OSI seem to be able to spatially and temporally predict dengue activity
  - Can be used for gap-filling i.e., infer dengue information where some OSI (e.g., HM data) are unavailable
  - Exploits meteorological covariates which are always available everywhere
- Future work
  - Improve missing data prediction through viewing the data in a cross-sectional time-series perspective
  - Detailed imputation analysis, to improve out-of-sample predictions
  - Quantify uncertainty in missing-data imputations
  - Compare our areal interpolation method against an (established) kriging method

References

- Azur M, Stuart E, Frangakis C, Leaf PJ. Multiple Imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, 2011
- Blackwell M, Honaker J, King G. In Press. A Unified Approach To Measurement Error And Missing Data: Overview. *Sociological Methods And Research.*
- Brockwell P, David R. Time Series: Theory and Methods. *Springer Series in Statistics*, 2009
- Ginsberg J, Mohebbi M, Patel R, Brammer L, Smolinski M, Brillant L. Detecting influenza epidemics using search engine query data. *Nature*, 2009
- Honaker J and King G. What to do About Missing Values in Time Series Cross-Section Data. *American Journal of Political Science,* 2010
- Livera A, Hyndman R. Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing. *Journal of the American Statistical Association*, 2014