Methods Description for Dengue Prediction
Reich Lab Team

Evan L. Ray, Krzysztof Sakrejda, Nicholas G. Reich[*]and Stephen A. Lauer

Department of Biostatistics and Epidemiology, University of Massachusetts – Amherst, Amherst, MA

1 Agreement

By submitting these forecasts, we indicate our full and unconditional agreement to abide by the project's official rules and data use agreements.

2 Introduction

In this document, we describe the method we use for predicting Dengue fever incidence in the Presidential Competition. Briefly, we employ a non-parametric approach to prediction in dynamical systems referred to as state space reconstruction (SSR). Our implementation of SSR is similar in spirit to the simplex projection method of Sugihara and May [1990]; however, our approach is based on kernel regression and kernel density estimation (CITE CITE). We describe the general SSR approach used to obtain predictive densities for counts in a given week in Section **??**. We then discuss post-processing used to obtain predictions for the target quantities in the challenge.

3 State Space Reconstruction

The spread of an infectious disease within a population is often represented as a dynamical process. For example, the SIR model and its variants partition the population into disjoint groups of people according to their infection status, and describe changes in the size of each group over time with a system of differential equations. These models make several assumptions regarding the number of states (i.e., disease categories) and the functional forms describing the rates at which people move through these categories over time.

[*] Email: `nick@schoolph.umass.edu`; Corresponding author

A central challenge in the study of infectious disease is that we typically observe many fewer covariates than are important to the evolution of the process. For example, even the most basic formulation of the SIR model involves the number of Susceptible and Infected individuals in the population at each time point. Most infectious disease data sets include observations of the number of infected individuals in the population, but do not include measures of the number of susceptible individuals. Furthermore, the SIR model is a very simplified representation of the evolution of the disease. In reality, there are likely to be many other important factors affecting disease dynamics such as weather, immigration, and the presence of other diseases. Many of these other factors may also be unobserved.

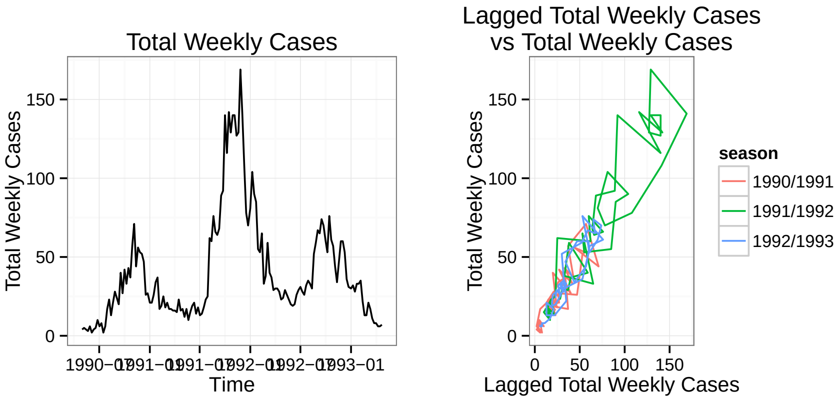
SSR is an alternate, non-parametric aproach to learning about dynamical systems that can be applied when we are uncertain of the functional forms describing evolution of the process and we do not have observations of all of the state variables. The method can be motivated by Takens' theorem [Takens, 1981], which establishes a connection between the original dynamical process and the process that is obtained by forming vectors of lagged observations. Roughly, the theorem states that if the dynamical process satisfies certain conditions, there is a smooth, one-to-one and onto mapping with a smooth inverse between the limit sets of the original dynamical process and the limit sets of the lagged observation process. This provides some justification for studying the lagged observation process in order to learn about long-term dynamics of the system. The theorem also tells us that we must include at least $2D + 1$ lags in the lagged observations in order for the limit sets to be equivalent, where D is the number of states in the state space of the underlying dynamical process.

Although this theorem can help motivate the study of the lagged observation process, it does not answer two key questions for a practical implementation:

1. If the dimension D of the underlying process is unknown, we do not know how many or which lags should be included when we form the lagged observation process.

2. It does not provide us with a specific method for studying the lagged observation process or using it to perform prediction.

For this challenge, we use a non-parametric approach based on kernel regression and kernel density estimation to learn the relationship between the lagged observation process and its future values. We use cross-validation to select the lags that are included and estimate the bandwidth parameters for the kernel functions. Before formally stating the method in Subsection **??**, we give an intuitive overview of it using the Dengue data from the challenge. For the purposes of this demonstration of the method, we use a lag of one in forming the lagged observation vector. We plot the weekly case counts in Figure **??** and their transformation into a lagged coordinate space in Figure **??**.



(a) Total Weekly Cases over time

(b) Lagged Total Weekly Cases vs Total Weekly Cases

Figure 1: Total cases by week for San Juan, Puerto Rico.

4 Model Description

Suppose we observe $\tilde{\mathbf{x}}_t = \{\tilde{x}_{t,1}, \ldots, \tilde{x}_{t,D}\} \in \mathbb{R}^D$ at each point in time $t = 1, \ldots, T$. Our goal is to obtain a predictive distribution for one of the observed variables over a range of prediction horizons spanning the remainder of the current season using lagged observations of all of the variables. In order to do this, we employ weighted kernel density estimation.

We denote the index of the variable in $\tilde{\mathbf{x}}_t$ that we are predicting by $d_{pred} \in \{1, \ldots, D\}$, and the prediction horizon by P. Let $\mathbf{l}^{max} = (l_1^{max}, \ldots, l_D^{max})$ specify the maximum number of lags for each observed variable that may be used for prediction. The lags that are actually used in prediction will be estimated by a procedure we describe in Section ***; we capture which lags are used in the vector

$$\mathbf{u} = (u_{1,0}, \ldots, u_{1,l_1^{max}}, \ldots, u_{D,0}, \ldots, u_{D,l_D^{max}}), \text{ where}$$

$$u_{d,l} = \begin{cases} 0 \text{ if lag } l \text{ of variable } d \text{ is not used in forming predictions} \\ 1 \text{ if lag } l \text{ of variable } d \text{ is used in forming predictions.} \end{cases}$$

Finally, it will be convenient to set

$$\mathbf{y}_t^{(P)} = (F^{(1)}\tilde{x}_{t,d_{pred}}, \ldots, F^{(P)}\tilde{x}_{t,d_{pred}}) \text{ and}$$
$$\mathbf{x}_t^{(\mathbf{l}^{max})} = (\tilde{x}_{1,t}, \ldots, B^{(l_1^{max})}\tilde{x}_{1,t}, \ldots, \tilde{x}_{D,t}, \ldots, B^{(l_D^{max})}\tilde{x}_{D,t})$$

Here, $F^{(a)}$ is the forward shift operator defined by $F^{(a)}\tilde{x}_{t,d} = \tilde{x}_{t+a,d}$ and $B^{(a)}$ is the backward shift operator defined by $B^{(a)}\tilde{x}_{t,d} = \tilde{x}_{t-a,d}$. The variable $\mathbf{y}_t^{(P)}$ represents the prediction target when our most recent observation was made at time t: the vector of counts for the following P weeks. The variable $\mathbf{x}_t^{(\mathbf{l}^{max})}$ represents the vector of all lagged covariates that are available for use in performing prediction. We wish to estimate the distribution of $\mathbf{y}_t^{(P)}|\mathbf{x}_t^{(\mathbf{l}^{max})}$.

To make the notation concrete, suppose that $\tilde{\mathbf{x}}_t$ contains the observed case count for week t in San Juan, the observed case count for week t in Iquitos, and the date on Monday of week t, and our goal is to predict the weekly case count in San Juan. Then $D = 3$ and $d_{pred} = 1$. If we want to predict the weekly case counts for the two weeks after the most recently observation, then $p = 2$. If we specify that the model may include the two most recent observations for the case counts in San Juan and Iquitos, but only the time index at the most recent observation then $\mathbf{l}^{max} = (1, 1, 0)$. If our current model uses only the most recently observed case counts for San Juan and Iquitos then $\mathbf{u} = (1, 0, 1, 0, 0)$, where the 1's are in the positions of the $\mathbf{u}$ vector representing lag 0 of the counts for San Juan and lag 0 of the counts for Iquitos. The variable $y_t^{(P)}$ is a vector containing the observed case counts for San Juan in weeks $t+1$ and $t+2$; $\mathbf{x}_t^{(\mathbf{l}^{max})}$ contains the observed case counts for San Juan and Iquitos in weeks t and $t-1$ as well as the time index variable in week t.

We obtain a predictive distribution for $\mathbf{y}_{t^*}^{(P)}$ at some particular time t^* as follows:

$$\widehat{f}(\mathbf{y}_{t^*}^{(P)}|\mathbf{x}_{t^*}) = \sum_{t \in \tau} w_{t,t^*} K^{pred}(\mathbf{y}_t^{(P)}, \mathbf{y}_{t^*}^{(P)}; H^{pred}), \text{ where} \tag{1}$$

$$K^{pred}(\mathbf{y}_t^{(P)}, \mathbf{y}_{t^*}^{(P)}; H^{pred}) = \prod_{p=1}^{P} K_p^{pred}(F^{(p)}\tilde{x}_{t,d_{pred}}, F^{(p)}\tilde{x}_{t^*,d_{pred}}; h_p^{pred}), \tag{2}$$

$$w_{t,t^*} = K^{wt}(\mathbf{x}_t, \mathbf{x}_{t^*}, \mathbf{u}; H^{wt}) / \sum_{t' \in \tau} K^{wt}(\mathbf{x}_{t'}, \mathbf{x}_{t^*}, \mathbf{u}; H^{wt}), \text{ and} \tag{3}$$

$$K^{wt}(\mathbf{x}_t, \mathbf{x}_{t^*}; H^{wt}) = \prod_{d=1}^{D} \prod_{l=1}^{l_d^{max}} K_{d,l}^{wt}(B^{(l)}\tilde{x}_{t,d}, B^{(l)}\tilde{x}_{t^*,d}; h_{d,l}^{wt})^{u_{d,l}} \tag{4}$$

In these expressions, the functions $K^{pred}(\cdot)$ and $K^{wt}(\cdot)$ are product kernel functions. We discuss our choices for the functional forms of these kernel functions below. τ is an index set of time points. In most settings, we take $\tau = \{t : 1 + \max_d \mathbf{l}^{max} \leq t \leq T - P\}$; these are the time points for which we can form the lagged observation vector $\mathbf{x}_t^{(\mathbf{l}^{max})}$ and the prediction target vector $\mathbf{y}_t^{(P)}$. However, we will place additional restrictions on the time points included in τ in the cross-validation procedure discussed in Section 5.

Equations (1) through (4) can be intrepreted as follows. The weighting kernel $K^{wt}(\mathbf{x}_t, \mathbf{x}_{t^*}; H^{wt})$ is used to assign a weight to each time t that reflects how similar the lagged observation vector formed at time t is to the lagged observation vector formed at the time t^*. If $\mathbf{x}_t$ is similar to $\mathbf{x}_{t^*}$, a large weight is assigned to t; if $\mathbf{x}_t$ is different from $\mathbf{x}_{t^*}$, a small weight is assigned to t. These weights are then used to obtain a weighted kernel density estimate in Equation

(1), placing a kernel centered at $\mathbf{y}_t^{(P)}$ with the corresponding weight w_{t,t^*}.

In general it would be possible to allow for some of the component kernel functions in K^{wt} or K^{pred} to incorporate multiple elements of $\mathbf{x}_t$ or $\mathbf{y}_t$; however, in this work we consider univariate component kernels only. As an example, if each K_c^{wt} is a univariate squared exponential kernel function, then the combined K^{wt} is a multivariate squared exponential kernel function with diagonal bandwidth matrix. We will allow each component kernel function to have a different functional form.

We have employed two functional forms for the kernel functions in our work. We use the squared exponential kernel for most of the component kernel functions in both K^{wt} and K^{pred}. This function has the following form:

$$K(x, x^*; h) \propto \exp\left\{-\frac{1}{2}(\frac{x - x^*}{h})^2\right\}$$

The constant of proportionality is irrelevant when the kernel is used as a component in K^{wt} since the constants cancel out when the weights are calculated. However, when this kernel is used as a component in K^{pred}, we require that the kernel function integrates to 1 in order to obtain a proper density estimate. It should be noted that the use of this kernel function in K^{pred} is technically inappropriate, since the prediction target (case counts) is a discrete random variable but the squared exponential kernel is defined for continuous x. We have adopted the squared exponential kernel for this purpose due to time constraints, and we plan to pursue the use of discrete kernels in future work.

The second kernel we have employed in the weighting kernel K^{wt} is the periodic kernel, which is commonly used in the Gaussian Processes literature. This kernel is defined as follows:

$$K(x, x^*; h) \propto \exp\left\{-\frac{1}{2}sin^2\left(\rho\frac{x - x^*}{h}\right)\right\}$$

The parameter ρ determines the period of the kernel function. In our work, we have fixed ρ so that the kernel has a period of 1 year and estimated only the bandwidth h. We have employed the periodic kernel only for the component of the weighting kernel corresponding to the time index variable. Effectively, this means that observations are scored as being similar to each other if they were recorded at a similar time of the year.

There are several details specific to our application of the SSR method to the Dengue challenge data which we now discuss. First, it can be seen from the plots in Figure 1 that the weekly case counts are not smooth over time. The plot in panel (a) of the figure shows that the weekly case counts are jagged, with many small peaks around a larger trend. In panel (b), we see that this induces some "loops" in the lagged covariate space when a lag of one week is used. These loops can cause difficulties for the SSR method, because lagged covariate vectors representing different points in the disease's evolution may be near each other. This problem can be addressed in several ways. One possibility is the use of a different lag; for example, these loops might be reduced if a lag of three weeks were used instead of one week.

Another alternative that we pursue is the use of smoothed observations in forming the lagged observation vectors. This is appropriate because we are most interested in capturing long-term trends. We therefore perform an initial smooth using the LOESS method. We use these smoothed case counts on a log scale for the weighting kernels, and the unsmoothed case counts on the original scale for the prediction kernels.

5 Parameter Estimation

We use cross-validation in order to select the variables that are used in the model and estimate the corresponding bandwidth parameters by (approximately) minimizing a cross-validation estimate of the mean absolute error (MAE) of the point predictions obtained from the model. Formally,

$$
\begin{aligned}
(\widehat{\mathbf{u}}, \widehat{H}^{wt}) &\approx \underset{(\mathbf{u}, H^{wt})}{\operatorname{argmin}} MAE(\mathbf{u}, H^{wt}) \\
&= \underset{(\mathbf{u}, H^{wt})}{\operatorname{argmin}} \frac{1}{P \cdot (1 + \max_d \mathbf{l}^{max})} \sum_{p=1}^{P} \sum_{t^*=1+\max_d \mathbf{l}^{max}}^{T-P} |\widehat{y}_{t^*}^{(p)} - y_{t^*}^{(p)}|
\end{aligned}
$$

The approximation is due to the fact that our optimization procedures may not find a global minimum.

We use a forward/backward stagewise procedure to obtain the set of combinations of variables and lags that are included in the final model (represented by $\mathbf{u}$). For each candidate model, we use the limited memory box constrained optimization procedure of Byrd et al. [1995] to estimate the bandwidth parameters.

In order to obtain the point estimate $\widehat{y}_{t^*}^{(p)}$ for each time point t^* and number of steps ahead p, we first form the predictive density as in Equations (1) through (4). In order to reduce the potential for our parameter estimates to be affected by local correlation in the time series, we eliminate all time points that fall within one year of t^* from the index set τ; we refer to this timepoint-specific index set as τ^*. We then take our point estimate to be the expected value of this estimated predictive distribution:

$$
\begin{aligned}
\widehat{\mathbf{y}}_{t^*}^{(P)} &= \mathbb{E}\left[\sum_{t \in \tau^*} w_{t,t^*} K^{pred}(\mathbf{y}_t^{(P)}, \mathbf{y}_{t^*}^{(P)}; H^{pred})\right] \\
&= \sum_{t \in \tau^*} w_{t,t^*} \mathbb{E}\left[K^{pred}(\mathbf{y}_t^{(P)}, \mathbf{y}_{t^*}^{(P)}; H^{pred})\right] \\
&= \sum_{t \in \tau^*} w_{t,t^*} \mathbf{y}_t^{(P)}
\end{aligned}
$$

Note that for symmetric prediction kernels, this point estimate does not depend on the bandwidth parameters H^{pred}.

For the prediction kernels, we estimated the bandwidth parameters using the methods of Sheather and Jones [1991], as implemented in `R`'s `bw.SJ` function. We estimate the bandwidth parameter for each prediction horizon $p \in \{1, \ldots, P\}$ separately.

6 Predictions for Challenge

7 Variables

8 Computational Resources

We implemented our variation on SSR in the `R` statistical programming language [R Core Team, 2013].

9 Publications

Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org/`.

Simon J Sheather and Michael C Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, 53(3):683–690, 1991.

George Sugihara and Robert M. May. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series, April 1990.

Floris Takens. *Detecting strange attractors in turbulence*. Springer, 1981.