Journal Title
XX(X):1–27

DOI: 10.1177/ToBeAssigned
www.sagepub.com/

Infectious disease prediction with kernel conditional density estimation

Evan L. Ray[1], Krzysztof Sakrejda[1], Stephen A. Lauer[1], Michael Johansen[2] and Nicholas G. Reich[1]

Abstract
Abstract

Keywords
copula, dengue fever, infectious disease, influenza, kernel conditional density estimation, prediction

Introduction

Accurate prediction of infectious disease incidence is important for public health officials planning disease prevention and control measures such as vector control and increased use of personal protective equipment by medical personnel during periods of high disease incidence [11;23]. Several quantities have emerged as being of particular utility in making these planning decisions; in this article we focus on measures of weekly incidence, the timing of the season peak, and incidence in the peak week. Predictive distributions for these quantities are preferred to point predictions because they communicate uncertainty in the predictions and give decision makers more information in cases where the predictive distribution is skewed or has multiple modes. In this work, we employ a non-parametric approach referred to as kernel conditional density estimation (KCDE) to obtain separate predictive distributions for disease incidence in each week of the season, and then combine those marginal distributions using copulas to obtain joint predictive distributions for the trajectory of incidence over the course of multiple weeks. Predictive distributions relating to the timing of and incidence at the peak week can be obtained from this joint predictive distribution for the trajectory of disease incidence. In addition to the novel application of these methods to predicting disease incidence, our contributions include the

[1]Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst
[2]CDC, Puerto Rico

Corresponding author:
Evan Ray, UMass Address Here
Email: elray@umass.edu

Prepared using sagej.cls *[Version: 2015/06/09 v1.01]*

use of a periodic kernel specification to capture seasonality in disease incidence and a method for obtaining multivariate kernel functions that handle discrete data while allowing for a fully parameterized bandwidth matrix.

KCDE has not previously been applied to obtain predictive distributions for infectious disease incidence, but it has been successfully used for prediction in other settings such as survival time of lung cancer patients[8], female labor force participation[8], bond yields and value at risk in financial markets[6], and wind power[14] among others. Although KCDE has not previously been applied to predicting infectious disease, closely related methods for obtaining point predictions have been employed for diseases such as measles[21] and influenza[22]. In the infectious disease literature these methods have been referred to as state space reconstruction and the method of analogues, but they amount to applications of nearest neighbors regression. The point prediction obtained from nearest neighbors regression is equal to the expected value of the predictive distribution obtained from KCDE if a particular kernel function is used in the formulation of KCDE[10]. However, KCDE offers the advantage of providing a complete predictive distribution rather than only a point prediction. Methods similar to those we explore in this article can also be formulated in the Bayesian framework. One example along these lines is Zhou et al.[25], who model the time to arrival of a disease in amphibian populations using Dirichlet processes and copulas.

There is also a long history of using other modeling approaches such as compartmental models for infectious disease prediction. A full discussion of those methods is beyond the scope of this article; see Brown et al.[2] for a recent review. KCDE is distinguished from these approaches in that it makes minimal assumptions about the data generating process. This can be either an advantage or a disadvantage of KCDE. In general, we would expect a well-specified parametric model to outperform KCDE. On the other hand, because non-parametric approaches such as KCDE make fewer assumptions about the data generating process, they may outperform incorrectly specified parametric models. An evaluation of the benefits of an approach such as KCDE is therefore dependent on the particular characteristics of the system being modeled, the data that are available, and the quality of the models that are considered as alternatives. We will return to this point in our conclusions.

FIX CITATION INFO FOR LEXI'S REVIEW PAPER OR FIND AN ALTERNATIVE

As we will describe in more detail below, KCDE estimates the conditional density of a random vector $\mathbf{Y}$ given another vector $\mathbf{X}$ as a weighted sum of contributions from previously observed pairs $(\mathbf{x}_t, \mathbf{y}_t)$. In our work, $\mathbf{Y}$ is a measure of disease incidence at some future date (the prediction target) and $\mathbf{X}$ is a vector of predictive variables that we condition on in order to make our prediction. In our example applications, $\mathbf{X}$ includes observations of incidence over several recent time points and variables indicating the time of year at which we are making a prediction. In general, it would be possible to include other predictive variables such as weather covariates.

The observation weights and the scale of the contribution from each observation to the final density are determined by a kernel function. To our knowledge, all previous authors using kernel methods to estimate multivariate densities involving discrete variables have employed a kernel function that is a product of univariate kernel functions[1;16;18;24]. Using a product kernel simplifies the mathemetical formulation of the kernel function when discrete variables are present, but has the effect of forcing the kernel function to be orientied in line with the coordinate axes. In settings with only continuous variables, asymptotic analysis and experience with applications

have shown that using a multivariate kernel function with a bandwidth parameterization that allows for other orientations can result in improved density estimates in many cases (cite ***). We introduce an approach to allowing for discrete kernels with orientation by discretizing an underlying continuous kernel function.

A limitation of kernel-based density estimation methods is that their performance may not scale well with the dimension of the vector whose distribution is being estimated. This is particularly relevant in our application, where it is desired to obtain joint predictive distributions for disease incidence over the course of many weeks. Copulas present one strategy for estimating the joint distribution of moderate to high dimensional random vectors, and work by specifying a relatively simple parametric model for the dependence relations among those variables. This simple dependence model ties separate marginal distribution estimates together into a joint distribution. In our case, we obtain those marginal distribution estimates through an application of KCDE to each prediction horizon.

The remainder of this article is organized as follows. First, we describe our approach to prediction using KCDE and copulas, including development of the discretized kernel function and periodic kernel function. Next, we present the results of a simulation study comparing the performance of KCDE for estimating discrete distributions using a fully parameterized bandwidth matrix and a diagonal bandwidth matrix. We then illustrate our methods by applying them to predicting disease incidence in two data sets: one with a measure of weekly incidence of influenza in the United States and a second with a measure of weekly incidence of Dengue fever in San Juan, Puerto Rico. We conclude with a discussion of these results.

Method Description

In this Section, we give a detailed discussion of our methods. Suppose we observe a measure z_t of disease incidence at evenly spaced times indexed by $t = 1, \ldots, T$. We allow the incidence measure to be either continuous or discrete and use the term density to refer to the Radon-Nikodym derivative of the (conditional) cumulative distribution function with respect to an appropriately defined measure. We will use a colon notation to specify vectors: for example, $\mathbf{z}_{s:t} = (z_s, \ldots, z_t)$. Let W denote the number of time points in a disease season (typically $W = 52$ if we have weekly data). For each time t^*, let S_{t^*} denote the time index of the last time point in the *previous* season, and let $H_{t^*} = W - (t^* - S_{t^*})$ denote the number of time points remaining in the current season. At time T, we obtain predictive distributions for each of three prediction targets. We frame these quantities as suitable integrals of a predictive distribution $f(\mathbf{z}_{(T+1):(T+H_T)}|T, \mathbf{z}_{1:T})$ for the trajectory of incidence over all remaining weeks in the season:

1. Incidence in a single future week:

$$f(z_{T+h}|T, \mathbf{z}_{1:T}) = \int \cdots \int f(\mathbf{z}_{(T+1):(T+H_T)}|T, \mathbf{z}_{1:T})\, dz_{T+1} \cdots dz_{T+h-1}\, dz_{T+h+1} \cdots dz_{T+H_T} \quad (1)$$

2. Timing of the peak week of the current season:

$$P(\text{Peak Week} = w) = P(Z_{S_T+w} \geq Z_{S_T+w^*} \forall w^* = 1, \ldots, W | T, \mathbf{z}_{1:T}) = \int_{\{\mathbf{z}_{(T+1):(T+H_T)}: z_{S_T+w} \geq z_{S_T+w^*} \forall w^* = 1,\ldots,W\}} f(z_{(T+1):(T+H_T)} | T, \mathbf{z}_{1:T}) \, d\mathbf{z}_{(T+1):(T+H_T)}. \quad (2)$$

3. Binned incidence in the peak week of the current season:

$$P(\text{Incidence in Peak Week} \in [a, b]) = P(a \leq \max w Z_{S_T+w} \leq b | T, \mathbf{z}_{1:T}) = \int_{\{(\mathbf{z}_{(T+1):(T+H_T)}): a \leq \max w Z_{S_T+w} \leq b\}} f(\mathbf{z}_{(T+1):(T+H_T)} | T, \mathbf{z}_{1:T}) \, d\mathbf{z}_{(T+1):(T+H_T)}. \quad (3)$$

Our approach is to specify a model for $f(\mathbf{z}_{(T+1):(T+H_T)}|T, \mathbf{z}_{1:T})$, and then obtain predictive distributions for the desired quantities by computing the integrals above. In practice, we use Monte Carlo integration to evaluate the integrals in Equations (2) and (3) by sampling incidence trajectories from the joint predictive distribution.

At time t^*, our model approximates $f(\mathbf{z}_{(t^*+1):(t^*+H_{t^*})}|t^*, \mathbf{z}_{1:t^*})$ by conditioning only on the time at which we are making the predictions and observed incidence at a few recent time points with lags given by the non-negative integers $l_1, \ldots, l_M$: $f(\mathbf{z}_{(t^*+1):(t^*+H_{t^*})}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M})$. The time t^* is equal to T when we are applying the method to perform prediction, but takes other values in the estimation procedure we describe below. The model represents this density as follows:

$$f(z_{(t^*+1):(t^*+H_{t^*})}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M}) = c^{H_{t^*}}\{f^1(z_{t^*+1}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M}; \boldsymbol{\theta}^1), \ldots, f^{H_{t^*}}(z_{t^*+H_{t^*}}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M}; \boldsymbol{\theta}^H); \boldsymbol{\xi}^{H_{t^*}}\}. \quad (4)$$

Here, each $f^h(z_{t^*+h}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M}; \boldsymbol{\theta}^h)$ is a predictive density for one prediction horizon obtained through KCDE. The distribution for each prediction horizon depends on a separate parameter vector $\boldsymbol{\theta}^h$. The function $c^{H_{t^*}}(\cdot)$ is a copula used to tie these marginal predictive densities together into a joint predictive density, and depends on parameters $\boldsymbol{\xi}^{H_{t^*}}$. In our applications, we will obtain a separate copula fit for each trajectory length H_{t^*} of interest for the prediction task.

Broadly, estimation for the model parameters proceeds in two stages: first we estimate the parameters for KCDE separately for each prediction horizon $h = 1, \ldots, H_{t^*}$, and second we estimate the copula parameters while holding the KCDE parameters fixed. The efficiency of two-stage estimation procedures for copula models has been studied in the literature both theoretically and through simulation studies. In general the two-stage approach may result in some loss of efficiency relative to one-stage methods, but this efficiency loss is small for some model specifications[15]. We pursue the two-stage strategy in this work because it results in a large reduction in the computational cost of parameter estimation.

In the following subsections we describe the formulations of KCDE and the copula in more detail and give our estimation strategy for each set of model parameters.

KCDE for Predictive Densities at Individual Prediction Horizons

We now discuss the methods we use to obtain the predictive density $f^h(z_{t^*+h}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M}; \boldsymbol{\theta}^h)$ for disease incidence at a particular horizon h after time t^*. In order to simplify the notation we define two new variables: $Y_t^h = Z_{t+h}$ represents the prediction target relative to time t, and $\mathbf{X}_t = (t, Z_{t-l_1}, \ldots, Z_{t-l_M})$ represents the vector of predictive variables relative to time t. With this notation, the distribution we wish to estimate is $f^h(y_{t^*}^h|\mathbf{x}_{t^*}; \boldsymbol{\theta}^h)$.

In order to estimate this distribution, we use the observed data to form the pairs $(\mathbf{x}_t, y_t^h)$ for all $t = 1 + \max_m l_m, \ldots, T - h$; for smaller values of t there are not enough observations before t to form $\mathbf{x}_t$ and for larger values of t there are not enough observations after t to form y_t^h. We then regard these pairs as a (dependent) sample from the joint distribution of $(\mathbf{X}, Y^h)$ and estimate the conditional distribution of $Y^h|\mathbf{X}$ via KCDE:

$$\widehat{f^h}(y_{t^*}^h|\mathbf{x}_{t^*}) = \frac{\sum_{t\in\boldsymbol{\tau}} K^{\mathbf{X},Y}\left\{(\mathbf{x}_{t^*}, y_{t^*}^h), (\mathbf{x}_t, y_t^h); \boldsymbol{\theta}^h\right\}}{\sum_{t\in\boldsymbol{\tau}} K^{\mathbf{X}}(\mathbf{x}_{t^*}, \mathbf{x}_t; \boldsymbol{\theta}^h)} \tag{5}$$

$$= \frac{\sum_{t\in\boldsymbol{\tau}} K^{Y|\mathbf{X}}(y_{t^*}^h, y_t^h|\mathbf{x}_{t^*}, \mathbf{x}_t; \boldsymbol{\theta}^h) K^{\mathbf{X}}(\mathbf{x}_{t^*}, \mathbf{x}_t; \boldsymbol{\theta}^h)}{\sum_{t\in\boldsymbol{\tau}} K^{\mathbf{X}}(\mathbf{x}_{t^*}, \mathbf{x}_t; \boldsymbol{\theta}^h)} \tag{6}$$

$$= \sum_{t\in\boldsymbol{\tau}} w_t^h K^{Y|\mathbf{X}}(y_{t^*}^h, y_t^h|\mathbf{x}_{t^*}, \mathbf{x}_t; \boldsymbol{\theta}^h), \text{ where} \tag{7}$$

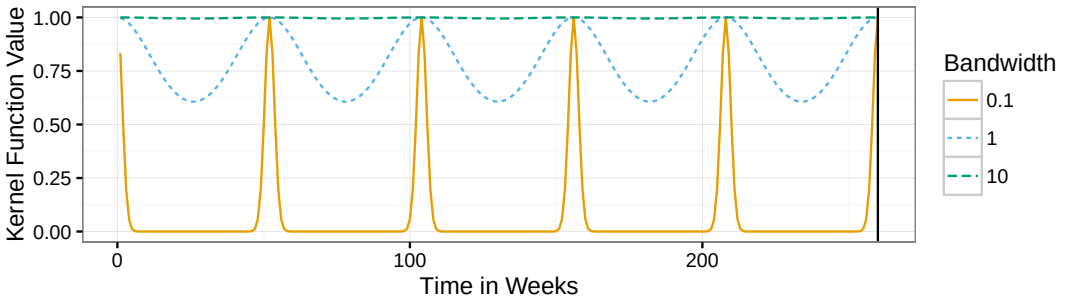
$$w_t^h = \frac{K^{\mathbf{X}}(\mathbf{x}_{t^*}, \mathbf{x}_t; \boldsymbol{\theta}^h)}{\sum_{s\in\boldsymbol{\tau}} K^{\mathbf{X}}(\mathbf{x}_{t^*}, \mathbf{x}_s; \boldsymbol{\theta}^h)} \tag{8}$$

Here we are working with a slightly restricted specification in which the kernel function $K^{\mathbf{X},Y}$ can be written as the product of $K^{\mathbf{X}}$ and a "conditional kernel" $K^{\mathbf{Y}|\mathbf{X}}$. With this restriction, we can interpret $K^{\mathbf{X}}$ as a weighting function determining how much each observation $(\mathbf{x}_t, y_t^h)$ contributes to our final density estimate according to how similar $\mathbf{x}_t$ is to the value $\mathbf{x}_{t^*}$ that we are conditioning on. For each y_t^h, $K^{\mathbf{Y}|\mathbf{X}}$ is a density function that contributes mass to the final density estimate near y_t^h. The parameters $\boldsymbol{\theta}^h$ control the locality and orientation of the weighting function and the contributions to the density estimate from each observation. In Equations (5) through (8), $\boldsymbol{\tau} \subseteq \{1 + \max_m l_m, \ldots, T - h\}$ indexes the subset of observations used in obtaining the conditional density estimate; we return to how this subset of observations is defined in the discussion of estimation below.

We take the kernel function $K^{Y,\mathbf{X}}$ to be a product kernel with one component being a periodic kernel in time and the other component capturing the remaining covariates:

$$\begin{aligned} &K^{\mathbf{X},Y}\left\{(\mathbf{x}_{t^*}, y_{t^*}^h), (\mathbf{x}_t, y_t^h); \boldsymbol{\theta}^h\right\} \\ &\quad = K^{\mathbf{X},Y}\left\{(t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M}, z_{t^*+h}), (t, z_{t-l_1}, \ldots, z_{t-l_M}, z_{t+h}); \boldsymbol{\theta}^h\right\} \\ &\quad = K^{Periodic}(t^*, t; \boldsymbol{\theta}^h) K^{Incidence}\{(z_{t^*-l_1}, \ldots, z_{t^*-l_M}, z_{t^*+h}), (z_{t-l_1}, \ldots, z_{t-l_M}, z_{t+h}); \boldsymbol{\theta}^h\} \end{aligned}$$

Figure 1. The periodic kernel function illustrated as a function of time in weeks with $\rho = \pi/52$ and three possible values for the bandwidth parameter θ.



The periodic kernel function was originally developed in the literature on Gaussian Processes[17], and is defined by

$$K^{Periodic}(t^*, t; \rho, \theta) = \exp\left[-\frac{\sin^2\{\rho(t^* - t)\}}{2\theta^2}\right]. \tag{9}$$

We illustrate this kernel function in Figure 1. It has two parameters: ρ, which determines the length of the periodicity, and θ, which determines the strength and locality of this periodic component in computing the observation weights w_t^h. In our applications, we have fixed $\rho = \pi/52$, so that the kernel has period of length 1 year with weekly data. Using this periodic kernel provides a mechanism to capture seasonality in disease incidence by allowing the observation weights to depend on the similarity of the time of year that an observation was collected and the time of year at which we are making a prediction.

The second component of our kernel is a multivariate kernel incorporating all of the other variables in $\mathbf{x}_t$ and y_t^h. In our applications, these variables are measures of incidence; for brevity of notation, we collect them in the vector $\tilde{\mathbf{z}}_t = (z_{t-l_1}, \ldots, z_{t-l_M}, z_{t+h})$. These incidence measures are continuous in the application to Influenza and discrete case counts in the application to Dengue fever. In the continuous case, we have used a multivariate log-normal kernel function. This kernel specification automatically handles the restriction that counts are non-negative, and approximately captures the long tail in disease incidence that we will illustrate in the applications Section below. This kernel function has the following functional form:

$$K_{cont}^{Incidence}(\tilde{\mathbf{z}}_{t^*}, \tilde{\mathbf{z}}_t; \mathbf{B}^h) = \frac{\exp\left[-\frac{1}{2}\{\log(\tilde{\mathbf{z}}_{t^*}) - \log(\tilde{\mathbf{z}}_t)\}'\mathbf{B}^{-1}\{\log(\tilde{\mathbf{z}}_{t^*}) - \log(\tilde{\mathbf{z}}_t)\}\right]}{(2\pi)^{\frac{M+1}{2}}|\mathbf{B}|^{\frac{1}{2}} z_{t^*+h}\prod_{m=1}^{M} z_{t^*-l_m}} \tag{10}$$

In this expression, the log operator applied to a vector takes the log of each component of that vector. The matrix $\mathbf{B}$ is the bandwidth matrix, controling the orientation and scale of the kernel function as illustrated in Figure 2. This bandwidth matrix is parameterized by $\boldsymbol{\theta}^h$. In this work we have considered two parameterizations: a diagonal bandwidth matrix, and a fully parameterized bandwidth based on the Cholesky decomposition.

In the discrete case, we obtain the kernel function by discretizing an underlying continuous kernel function:

$$K_{disc}^{Incidence}(\tilde{\mathbf{z}}_{t^*}, \tilde{\mathbf{z}}_t; \mathbf{B}^h) = \int_{a_{z_{t^*-l_1}}}^{b_{z_{t^*-l_1}}} \cdots \int_{a_{z_{t^*+h}}}^{b_{z_{t^*+h}}} K_{cont}^{Incidence}(\tilde{\mathbf{z}}_{t^*}, \tilde{\mathbf{z}}_t; \mathbf{B}^h)\, dz_{t^*-l_1} \cdots dz_{t^*+h}$$

For each component variable in $(z_{t^*-l_1}, \ldots, z_{t^*-l_M}, z_{t^*+h})$, we associate lower and upper bounds of integration a_{z_j} and b_{z_j} with each value in the domain of that random variable. The value of the kernel function is obtained by integrating over the hyper-rectangle specified by these bounds. In our application, the possible values of the random variables are non-negative integer case counts. In order to facilitate use of the log-normal kernel, we add 0.5 to the observed case counts; the corresponding integration bounds are the non-negative integers as illustrated in Figure 2.

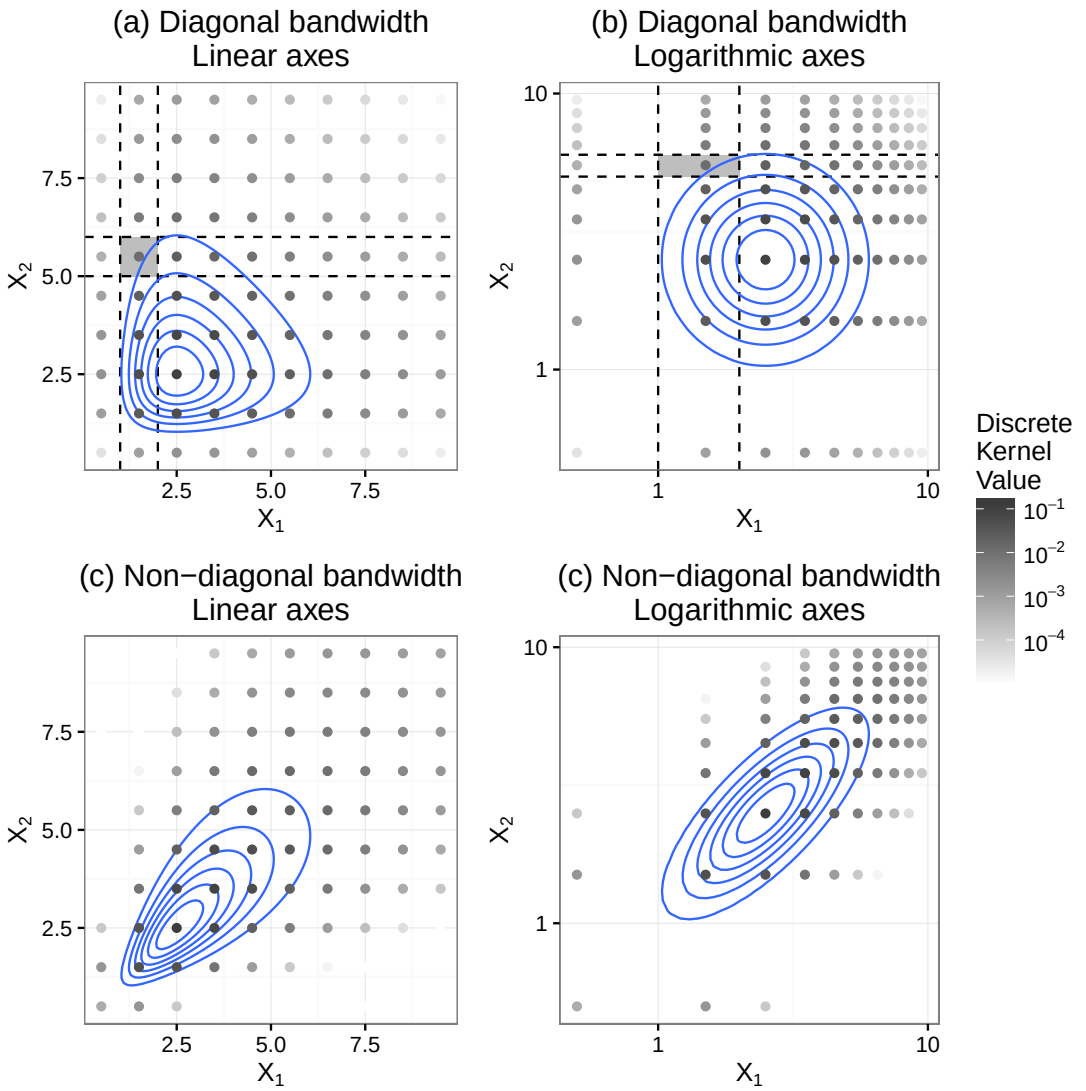
We estimate the bandwidth parameters by numerically maximizing the cross-validated log score of the predictive distributions for the observations in the training data: $\widehat{\boldsymbol{\theta}}^h \approx \text{argmax}_{\boldsymbol{\theta}^h} \sum_{t^*=1}^{T_{\text{train}}} \widehat{f}^h_{-t^*}(y^h_{t^*}|\mathbf{x}_{t^*}; \boldsymbol{\theta}^h)$. Here $\widehat{f}^h_{-t^*}(y^h_{t^*}|\mathbf{x}_{t^*})$ is as in Equation (5). In order to obtain the term corresponding to time t^*, we leave the year of training data before and after the time t^* out of the set $\boldsymbol{\tau}$. Hart and Vieu[9] show that when kernel density estimation is used to estimate a marginal density with dependent observations, leaving out a window of times around the target time point in cross validation can yield small improvements in the integrated squared error of the density estimate under certain assumptions about the form of the dependence. We expect that a similar result holds for the case of conditional density estimation. We perform the optimization using the limited memory box constrained optimization method of Byrd *et al.*[3], implemented by the `optim` function in `R`[20].

Our primary motivation for using the log score as the optimization target during estimation is that this is the criteria that has been used to evaluate and compare prediction methods in two recent government-sponsored infectious disease prediction contests[5;19]. We will apply our method to the data sets from those competitions in the applications Section below, and will report log scores in order to facilitate comparisons with other results from those competitions that may be published in the future. Our intuition is that it is beneficial to align the criteria used in estimation with the criteria used for comparing methods. In general, the log score is a strictly proper scoring rule; i.e., its expectation is uniquely maximized by the true predictive distribution[7]. However, its use as an optimization criterion can be criticised as it may be sensitive to outliers[7].

Combining Marginal Predictive Distributions with Copulas

We use copulas to tie the marginal predictive distributions for individual prediction horizons obtained from KCDE together into a joint predictive distribution for the trajectory of incidence over multiple time points. In this Section, we will provide a brief overview of copulas and our approach to using them in this application. A complete review of copulas is beyond the scope of this article; see (cite cite) for more thorough introductions. In order to describe our methods for both continuous and discrete distributions, it is most convenient to frame the discussion in this Section in terms of c.d.f.s instead of density functions. We will use a capital C to denote the copula function for distributions and a lower case c to denote the copula function for densities.

Figure 2. Illustrations of $K_{cont}^{Incidence}$ and $K_{disc}^{Incidence}$ in the bivariate case. Solid lines show contours of the continuous kernel function. Grey dots indicate the value of the discrete kernel function. The value of the discrete kernel is obtained by integrating the continuous kernel over regions as illustrated by the dashed lines in panels (a) and (b). In all panels the kernel function is centered at $(2.5, 2.5)$. In panels (a) and (b) the bandwidth matrix is $\begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix}$, and in panels (c) and (d) the bandwidth matrix is $\begin{bmatrix} 0.2 & 0.15 \\ 0.15 & 0.2 \end{bmatrix}$. We illustrate each case with both linear and logarithmic scale axes.



Similarly, the predictive densities $f^h(z_{t^*+h}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M}; \boldsymbol{\theta}^h)$ we obtained in the previous section naturally yield corresponding predictive c.d.f.s $F^h(z_{t^*+h}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M}; \boldsymbol{\theta}^h)$.

Our model specifies the joint c.d.f. for $\mathbf{Z}_{(t^*+1):(t^*+H_{t^*})}$ as follows:

$$F^{H_{t^*}}(\mathbf{z}_{(t^*+1):(t^*+H_{t^*})}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M}; \boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^{H_{t^*}}, \boldsymbol{\xi}^{H_{t^*}}) = C\{F^1(z_{t^*+1}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M}; \boldsymbol{\theta}^1), \ldots, F^h(z_{t^*+H_{t^*}}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M}; \boldsymbol{\theta}^{H_{t^*}}); \boldsymbol{\xi}^{H_{t^*}}\} \tag{11}$$

The copula function C maps the marginal c.d.f. values to the joint c.d.f. value. We use the isotropic Gaussian copula implemented in the `R` package `copula`[12]. The copula function is given by

$$C(u_1, \ldots, u_J; \boldsymbol{\xi}^H) = \Phi_{\Sigma^H}(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_J)), \tag{12}$$

where Φ^{-1} is the inverse c.d.f. of a standard univariate Gaussian distribution and Φ_{Σ^H} is the c.d.f. of a multivariate Gaussian distribution with mean $\underline{0}$ and covariance matrix Σ^H. The isotropic specification sets $\Sigma^H = [\sigma^H_{i,j}]$, where

$$\sigma^H_{i,j} = \begin{cases} 1 \text{ if } i = j, \\ \xi^H_d \text{ if } |i-j| = d \end{cases} \tag{13}$$

Intuitively, ξ^H_d captures the amount of dependence between incidence levels at future times that are d weeks apart.

We obtain a separate copula fit for each value of H from 2 to W (note that a copula is not required for "trajectories" of length $H = 1$). In order to do this, we follow the two-stage estimation strategy outlined by Joe[15]. Briefly, this procedure follows three main steps:

1. Estimate the parameters for marginal predictive distributions using the procedures described in the previous Subsection.
2. Form vectors of "pseudo-observations" by passing observed incidence trajectories from previous seasons through the marginal predictive c.d.f.s obtained in step 1:

   $$(u_{k,1}, \ldots, u_{k,H}) = \{F^1(z_{t^*_k+1}|t^*_k, z_{t^*_k-l_1}, \ldots, z_{t^*_k-l_M}; \boldsymbol{\theta}^1), \ldots, F^H(z_{t^*_k+1}|t^*_k, z_{t^*_k-l_1}, \ldots, z_{t^*_k-l_M}; \boldsymbol{\theta}^H)\}$$

   We form one such vector of pseudo-observations for each season in the training data; in the notation here, these seasons are indexed by k. The relevant time points t^*_k are the times in those previous seasons falling H time points before the end of the season.
3. Estimate the copula parameters $\boldsymbol{\xi}^H$ by maximizing the likelihood of the pseudo-observations.

Simulation Study

In this Section, we conduct two sets of simulation studies designed to answer two separate questions:

1. How much does using a kernel function with a non-diagonal bandwidth matrix contribute to the quality of conditional density estimates relative to density estimates obtained through KCDE using diagonal bandwidth matrices?
2. How does our method perform in the context of seasonal time series data? Specifically, how does the method perform relative to common alternatives, and how much do each of our three contributions (non-diagonal bandwidth matrices for discrete data, using a periodic function of time as predictive variable, and use of low band-pass filtered observatiosn as predictive variables) contribute to predictive performance?

Comparison of KCDE approaches

Our first set of simulation studies is based closely on those conducted in [4]; their examples demonstrate the utility of using a fully parameterized bandwidth matrix in kernel density estimation of continuous distributions. We modify their simulation study to examine the benefits of fully parameterized bandwidth matrices in the context of conditional density estimation with discrete variables.

We simulate observations from each of seven distributions. The first five of these are plotted in Figure ***.

Applications

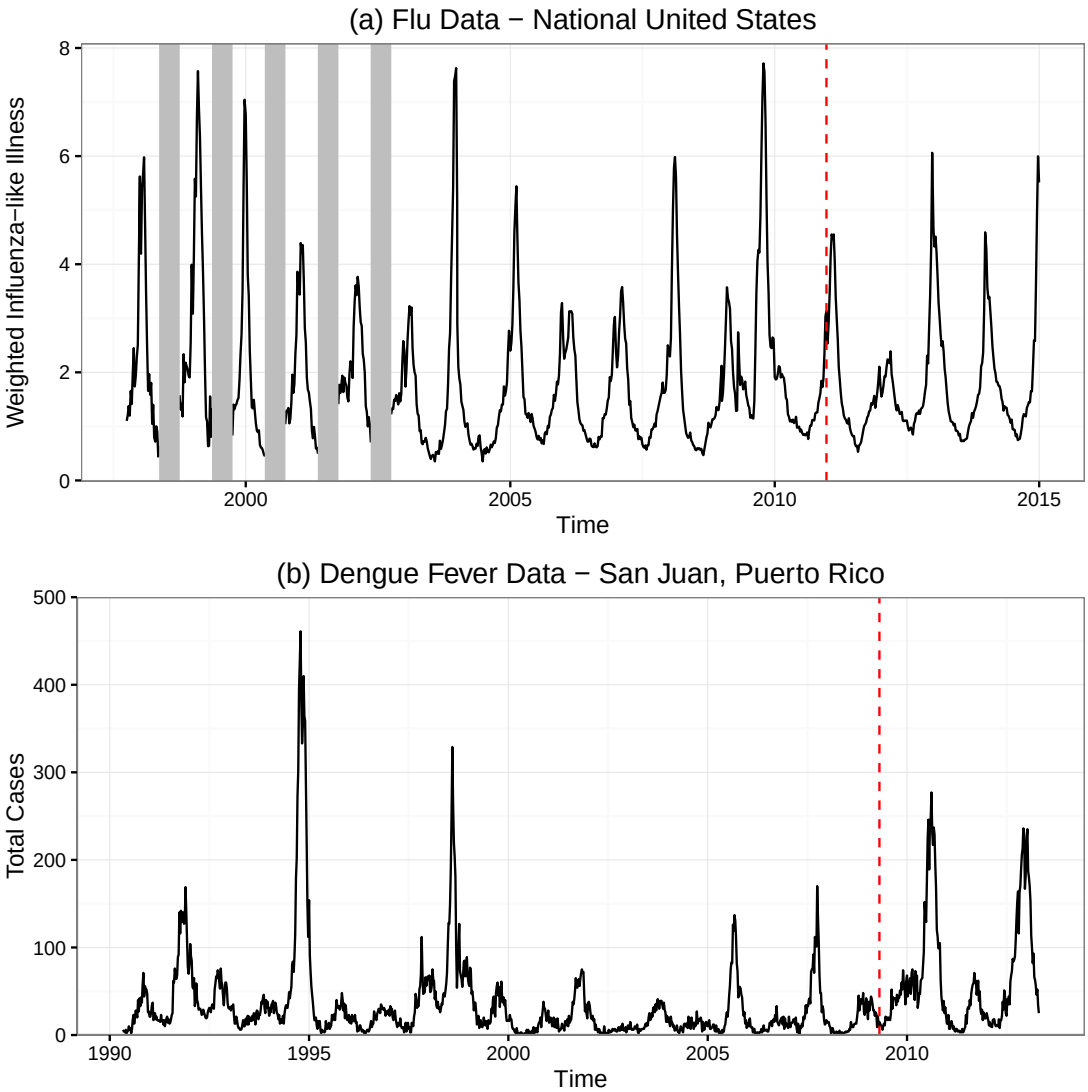
In this Section, we illustrate our methods through applications to prediction of infectious disease in two examples with real disease incidence data sets: one with a weekly measure of incidence of influenza like illness in the United States, and a second with a weekly measure of incidence of Dengue fever in San Juan, Puerto Rico. These data sets were used in two recent prediction competitions sponsored by the United States federal government [5;19].

We plot the data in Figure 3. As indicated in the figure, we have divided each data set into two subsets. The first period is used as a training set in estimating the model parameters. The last four years of each data set are reserved as a test set for evaluating model performance.

There are three prediction targets for each data set, based closely on the prediction targets that were used in those competitions. First, for each week in the test data, we obtain a predictive distribution for the incidence measure in that week at each prediction horizon from 1 to 52 weeks ahead. Second, in each season of the test data set, we make predictions for the timing of the peak week. Third, we predict the incidence measure in the peak week. In all cases, we compare the models using log score.

We use a seasonal ARIMA model as a baseline to compare our approach to. In fitting this model, we first transformed the observed incidence measure to the log scale (after adding 1 in the Dengue data set, which included some observations of 0 cases); this transformation makes the normality assumptions of the ARIMA model more plausible. We then performed first-order seasonal differencing, and obtained the final model fits using the `auto.arima` function in `R`'s `forecast` package [13]; this function uses a stepwise procedure to determine the terms to include in the model. This procedure resulted in a SARIMA(4 NA -)$_{52}$ model for the influenza data and a SARIMA(6 NA -)$_{52}$ model for the Dengue data. We note that a different SARIMA model

Figure 3. Plots of the data sets we apply our methods to. In each case, the last four years of data are held out as a test data set; this cutoff is indicated with a vertical dashed line. For the flu data set, low-season incidence was not recorded in early years of data collection; these missing data are indicated with vertical grey bars.



was used as a baseline in the Dengue competition, but the SARIMA model we obtained using this procedure performed slightly better on the test set than that previous baseline model.

Discuss variations on KCDE models.

Predictive Distributions for Individual Weeks

Predictive Distributions for Peak Week and Peak Incidence

Conclusions

Prediction of infectious disease incidence at horizons of more than a few weeks is a challenging task. We have presented one approach to doing this and found that it is a viable method that may lead to improved predictions relative to commonly employed methods in some applications. In an application to predicting Dengue fever, we saw that our approach offered consistent performance gains relative to a SARIMA model in predicting incidence in individual weeks. For predicting influenza-like illness, we saw that our approach did not pick up some features of the data generating process, such as the Christmas-week effect, that a SARIMA model did capture. For some prediction targets, this meant that SARIMA outperformed our method. On the other hand, our method rarely performed worse than a very naive baseline

Hall, Racine, and Li[8] show that when cross-validation is used to select the bandwidth parameters in KCDE using product kernels, the estimated bandwidths corresponding to irrelevant conditioning variables tend to infinity asymptotically as the sample size increases. They discuss the fact that similar results could be obtained for linear combinations of continuous variables if a full bandwidth matrix were used. Our approach for obtaining kernels that can be used with mixed discrete and continuous variables opens up an opportunity to extend this analysis to that case; we have not pursued this mathematical analysis here.

The above results regarding the inclusion of irrelevant conditioning variables hold asymptotically as the sample size increases. However, in practice, data set sizes are often limited. In other modeling settings where some conditioning variables may not be informative, shrinkage methods are often helpful. These methods could be incorporated into a kernel-based approach by imposing a penalty on the elements of the bandwidth matrix; in particular, we suggest that a penalty on the inverse of the bandwidth matrix encouraging it to have small eigenvalues could be helpful. Another alternative would be to pursue the Bayesian framework, using Dirichlet process mixtures with an informative prior on the mixture component covariance matrices.

We could also make some tweaks to our implementation of KCDE. Locally linear – help address edge effects. Cite Hyndman, Bashtannyk, Grunwald - "Estimating and Visualizing Conditional Densities", maybe also Fan and Yim - "A crossvaildation method for estimating conditional densities" and Fan et al. 1996 "Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems."

Ensembles – either ensembles of KCDE and/or include as a component in an ensemble. Also Bayesian model averaging. Return to discussion of bias/variance trade-off?

Other covariates

References

1. John Aitchison and Colin GG Aitken. Multivariate binary discrimination by the kernel method. *Biometrika*, 63(3):413–420, 1976.

Figure 4. Differences in log scores for the weekly predictive distributions among pairs of models across all combinations of prediction horizon and prediction time in the test period. In panel (a) positive values indicate cases when KCDE outperformed SARIMA. In panel (b) positive values indicate cases when the specification of KCDE with the periodic kernel outperformed the corresponding specification without the periodic kernel. In panel (c) positive values indicate cases when the specification of KCDE with a fully parameterized bandwidth outperformed the KCDE specification with a diagonal bandwidth matrix.

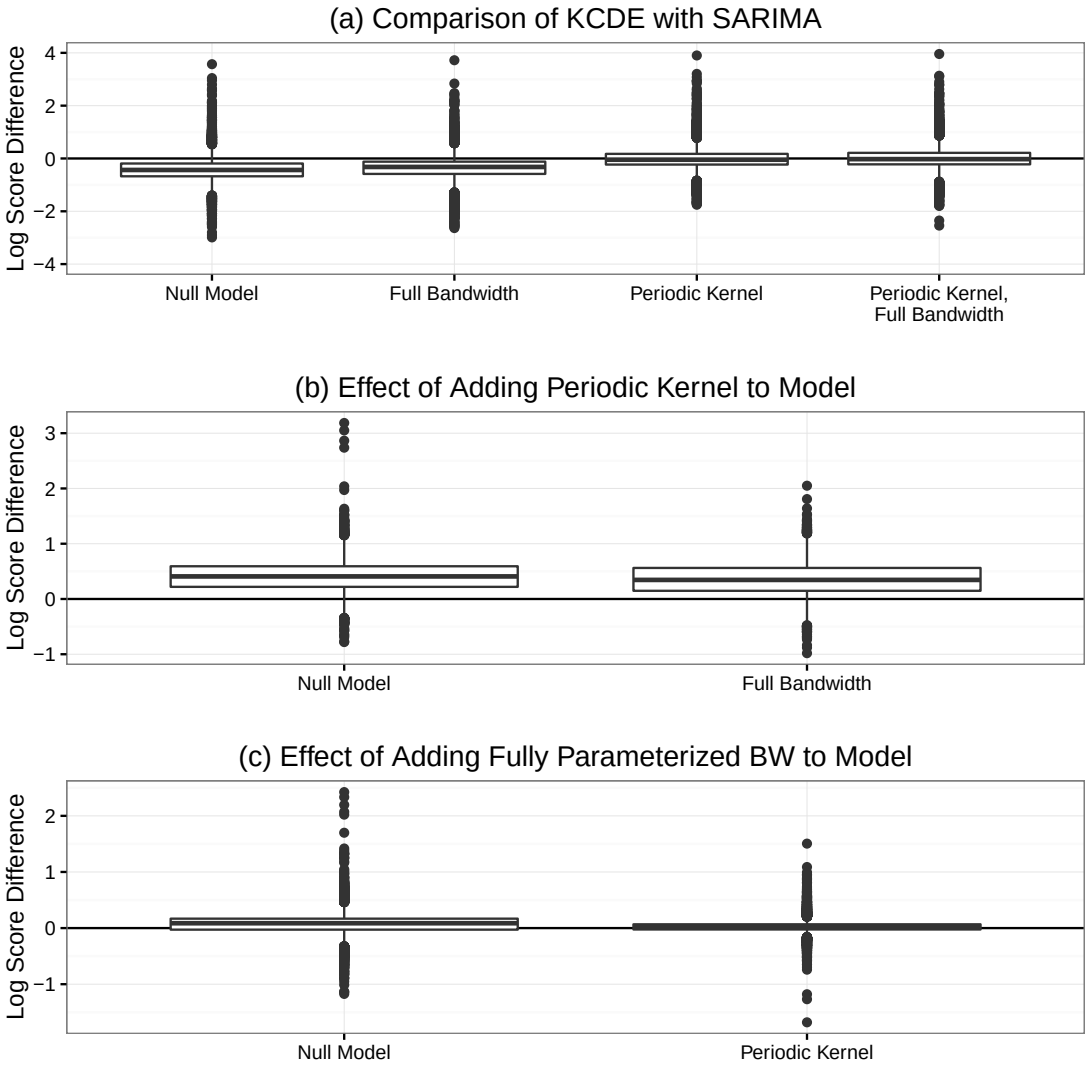
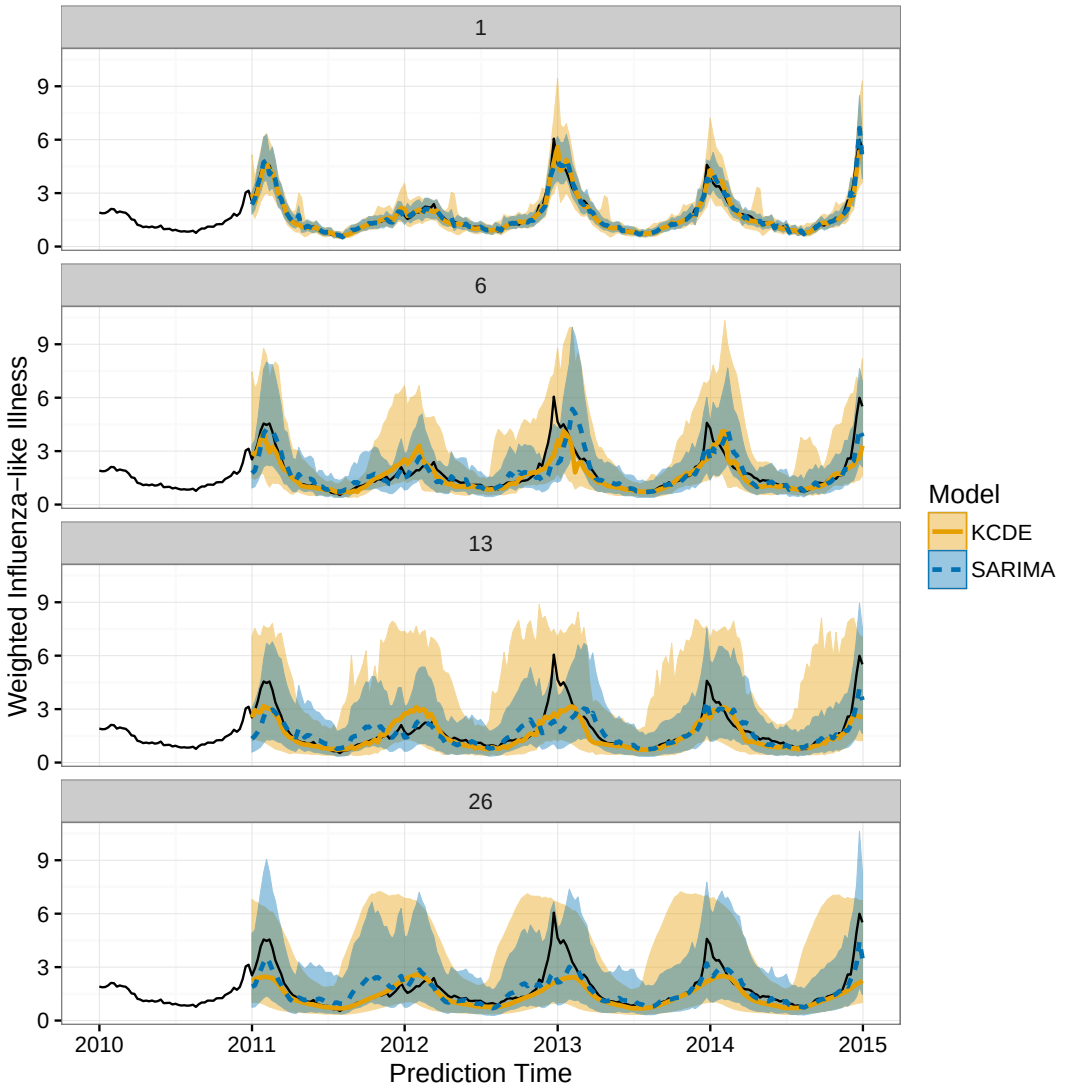


Figure 5. Plots of point and interval predictions from SARIMA and KCDE.



2. Alexandria Brown, Stephen A. Lauer, Evan L. Ray, Xi Meng, and Nicholas G. Reich. A systematic review of prediction for infectious disease. *Journal*, submitted.
3. 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.

Figure 6. Differences in log scores for the weekly predictive distributions for Dengue among pairs of models across all combinations of prediction horizon and prediction time in the test period. In panel (a) positive values indicate cases when KCDE outperformed SARIMA. In panel (b) positive values indicate cases when the specification of KCDE with the periodic kernel outperformed the corresponding specification without the periodic kernel. In panel (c) positive values indicate cases when the specification of KCDE with a fully parameterized bandwidth outperformed the KCDE specification with a diagonal bandwidth matrix.

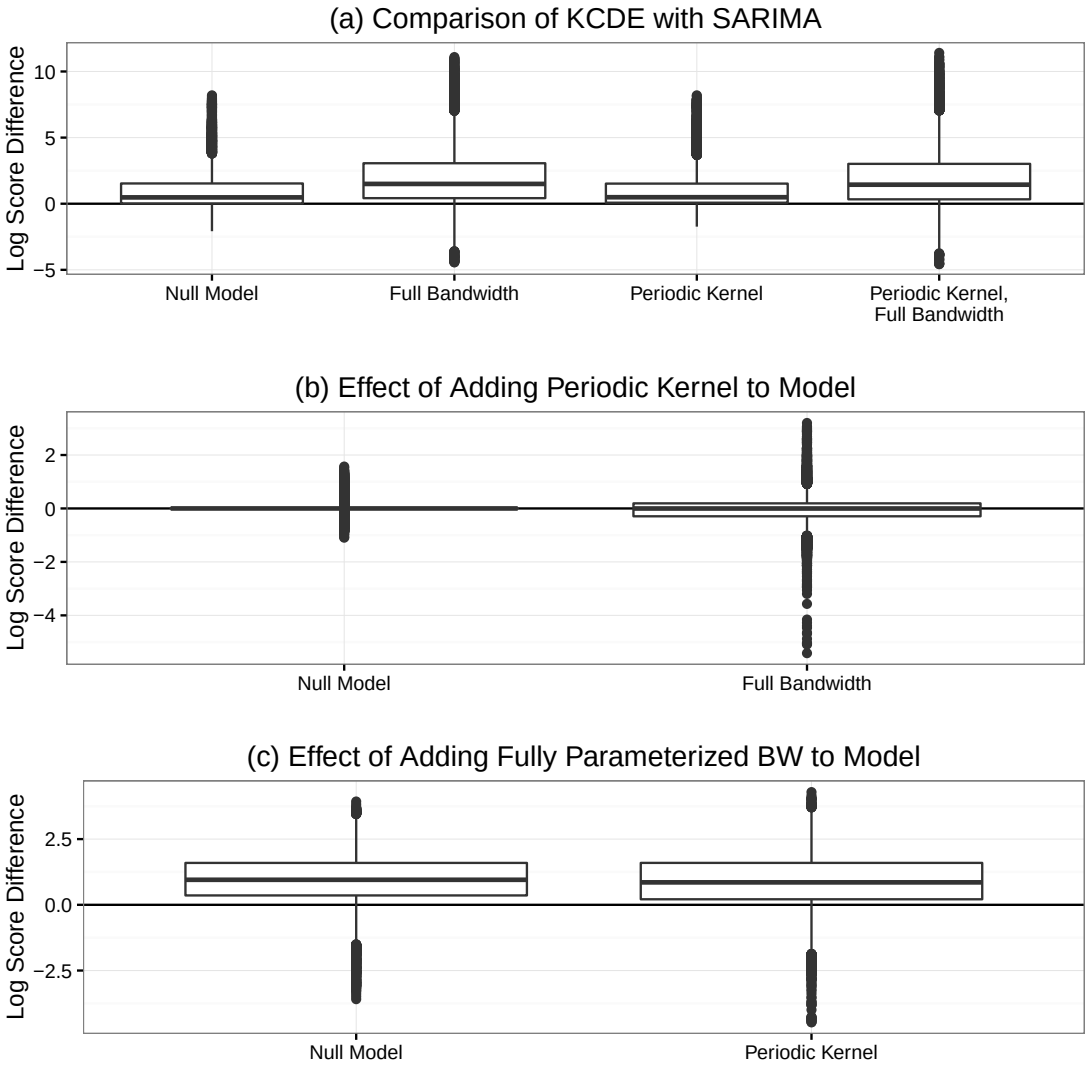
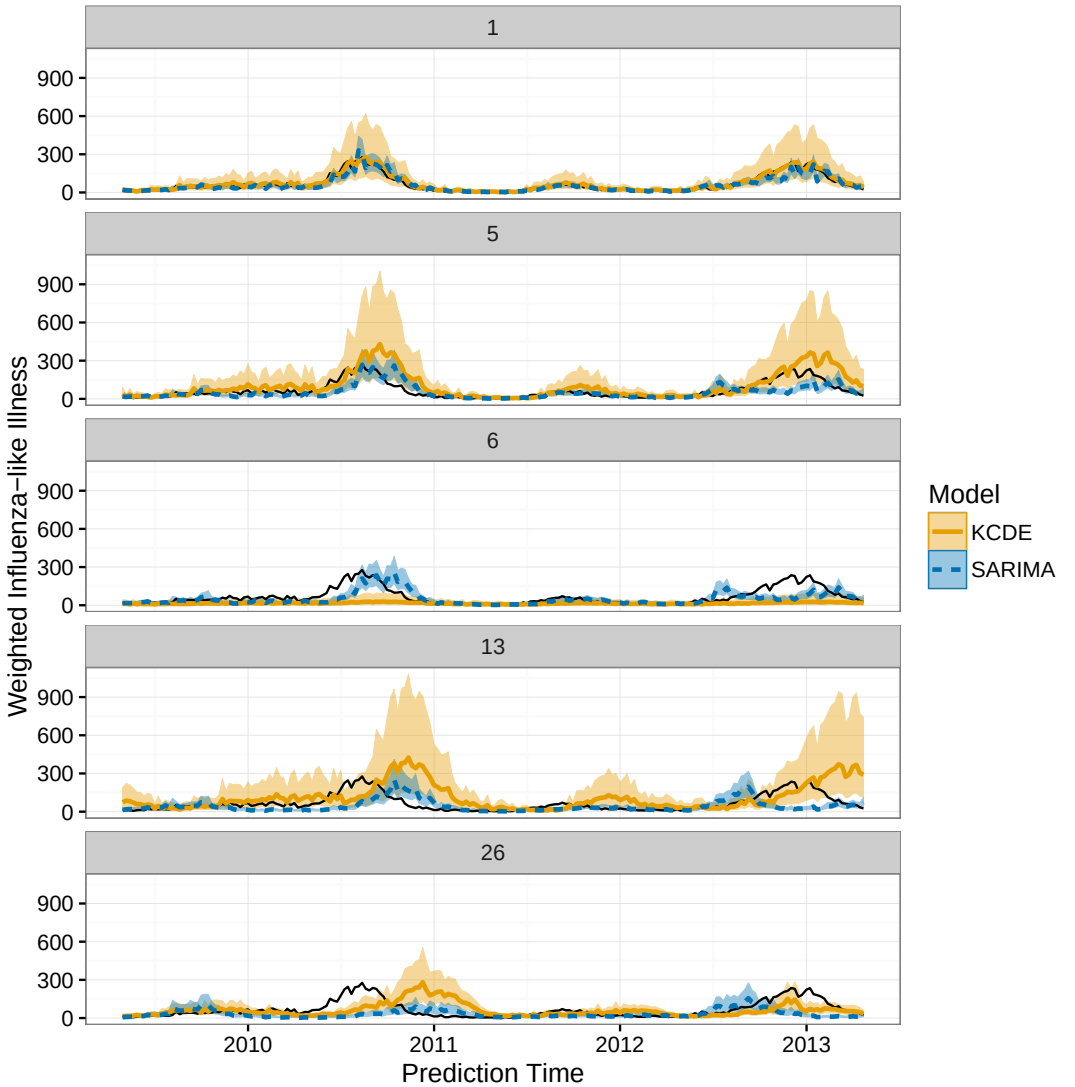


Figure 7. Plots of point and interval predictions from SARIMA and KCDE for Dengue.



4. Tarn Duong and Martin L Hazelton. Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics*, 32(3):485–506, 2005.
5. Epidemic Prediction Initiative. FluSight: Seasonal Influenza Forecasting, January 2016. URL `http://dengueforecasting.noaa.gov/`.

Figure 8. Differences in log scores for the predictive distributions for the peak week and incidence at the peak week among pairs of models across all analysis times in the test period. In panel (a) positive values indicate cases when KCDE outperformed SARIMA. In panel (b) positive values indicate cases when the specification of KCDE with the periodic kernel outperformed the corresponding specification without the periodic kernel. In panel (c) positive values indicate cases when the specification of KCDE with a fully parameterized bandwidth outperformed the KCDE specification with a diagonal bandwidth matrix. In the plot for peak week timing in panel (a), the log score differences are not displayed for one analysis time when none of the simulated trajectories from SARIMA peaked at the true peak week. In that case, our monte carlo estimate of the difference in log scores is infinity.

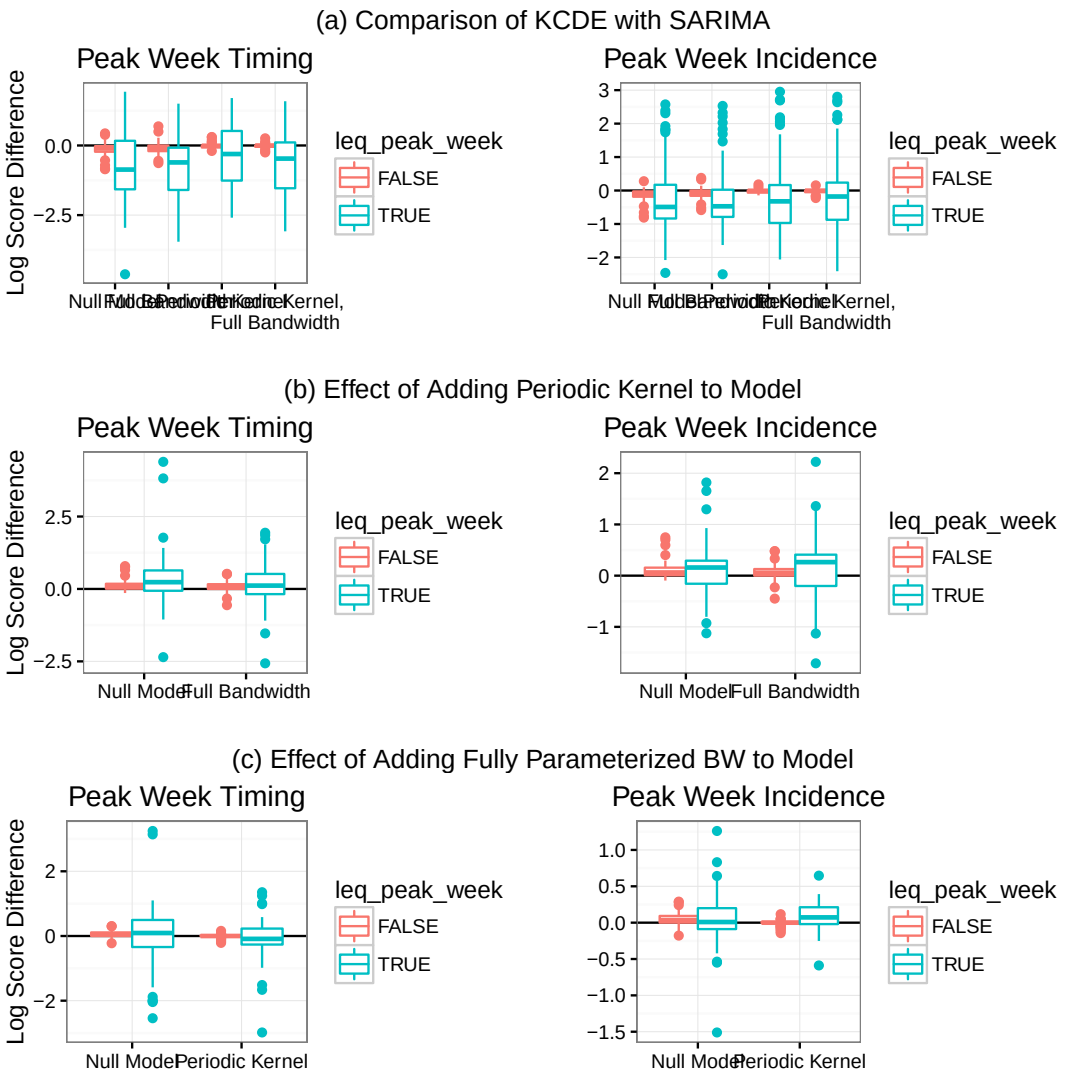
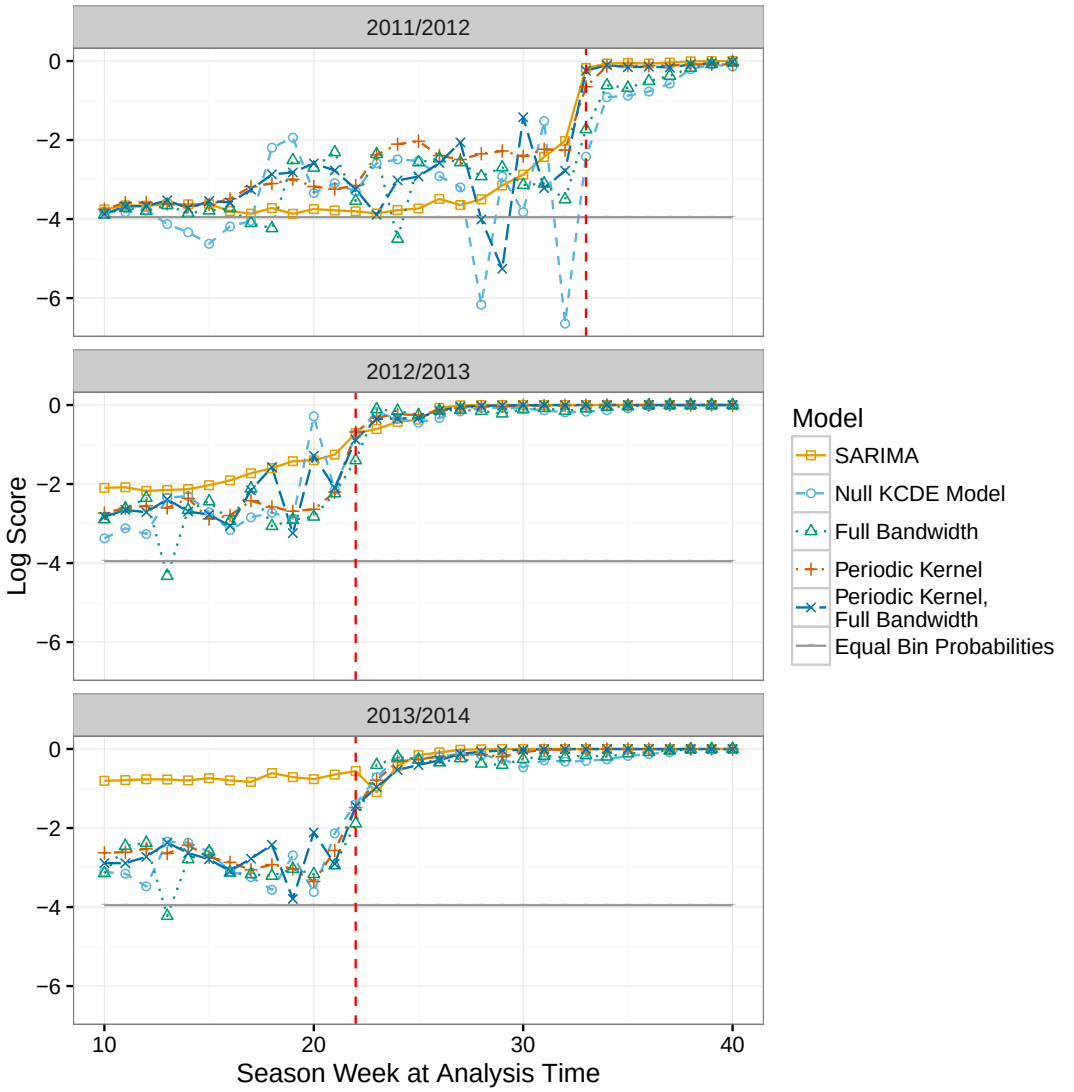
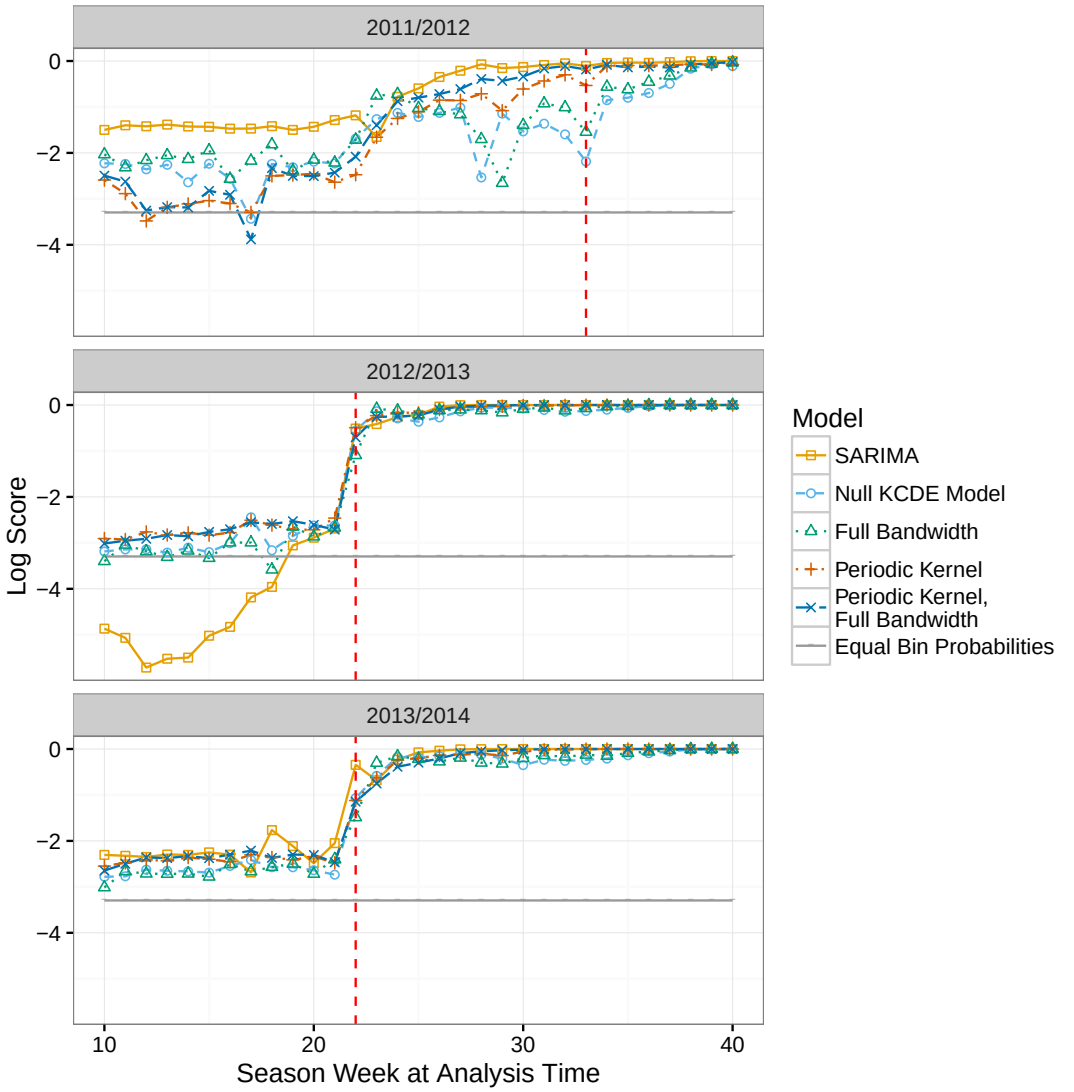


Figure 9. Log scores for predictions of peak week timing by predictive model and analysis time. The vertical gray line is placed at the peak week for each season.



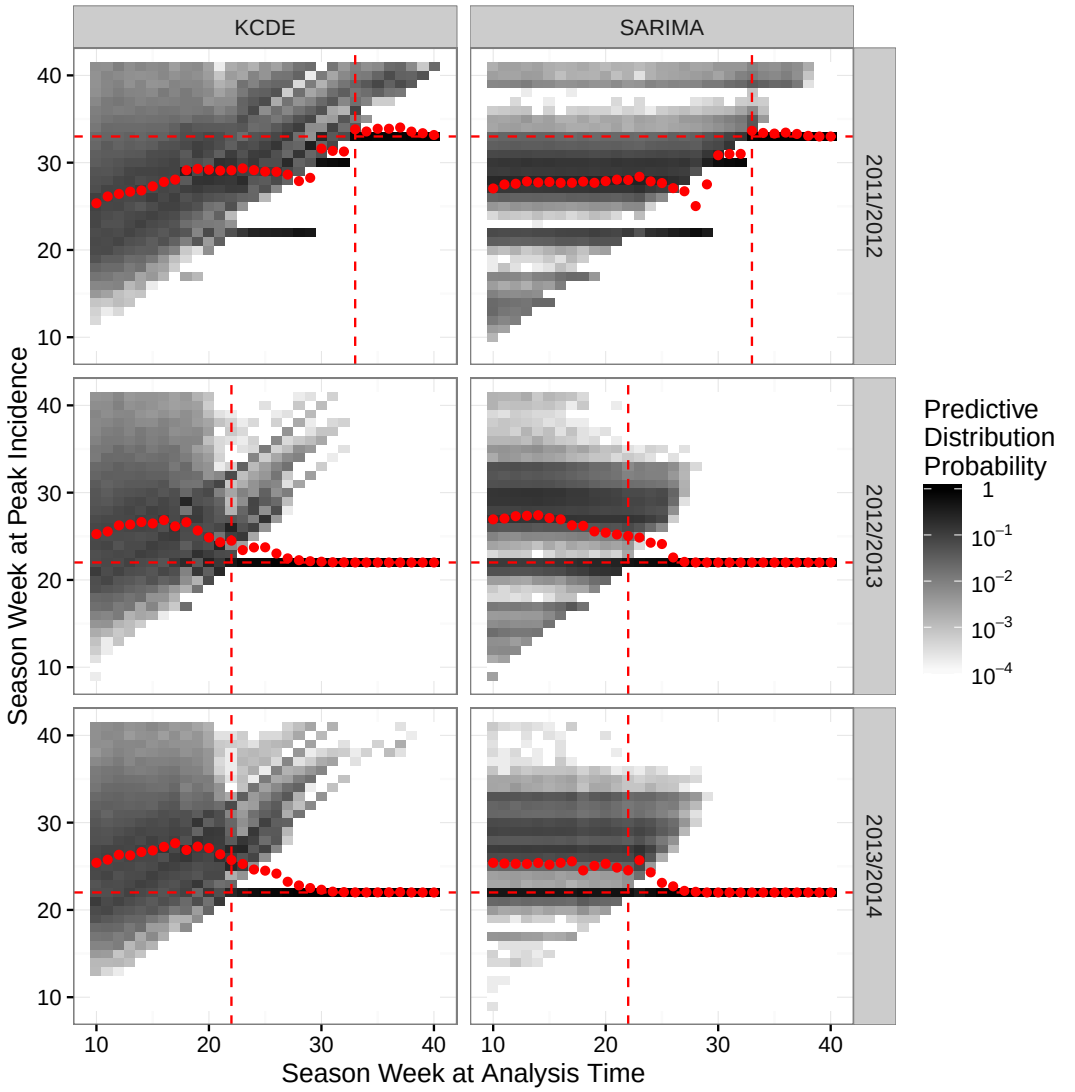
6. Jianqing Fan and Tsz Ho Yim. A crossvalidation method for estimating conditional densities. *Biometrika*, 91(4):819–834, 2004.
7. Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Figure 10. Log scores for predictions of incidence in the peak week by predictive model and analysis time. The vertical gray line is placed at the peak week for each season.



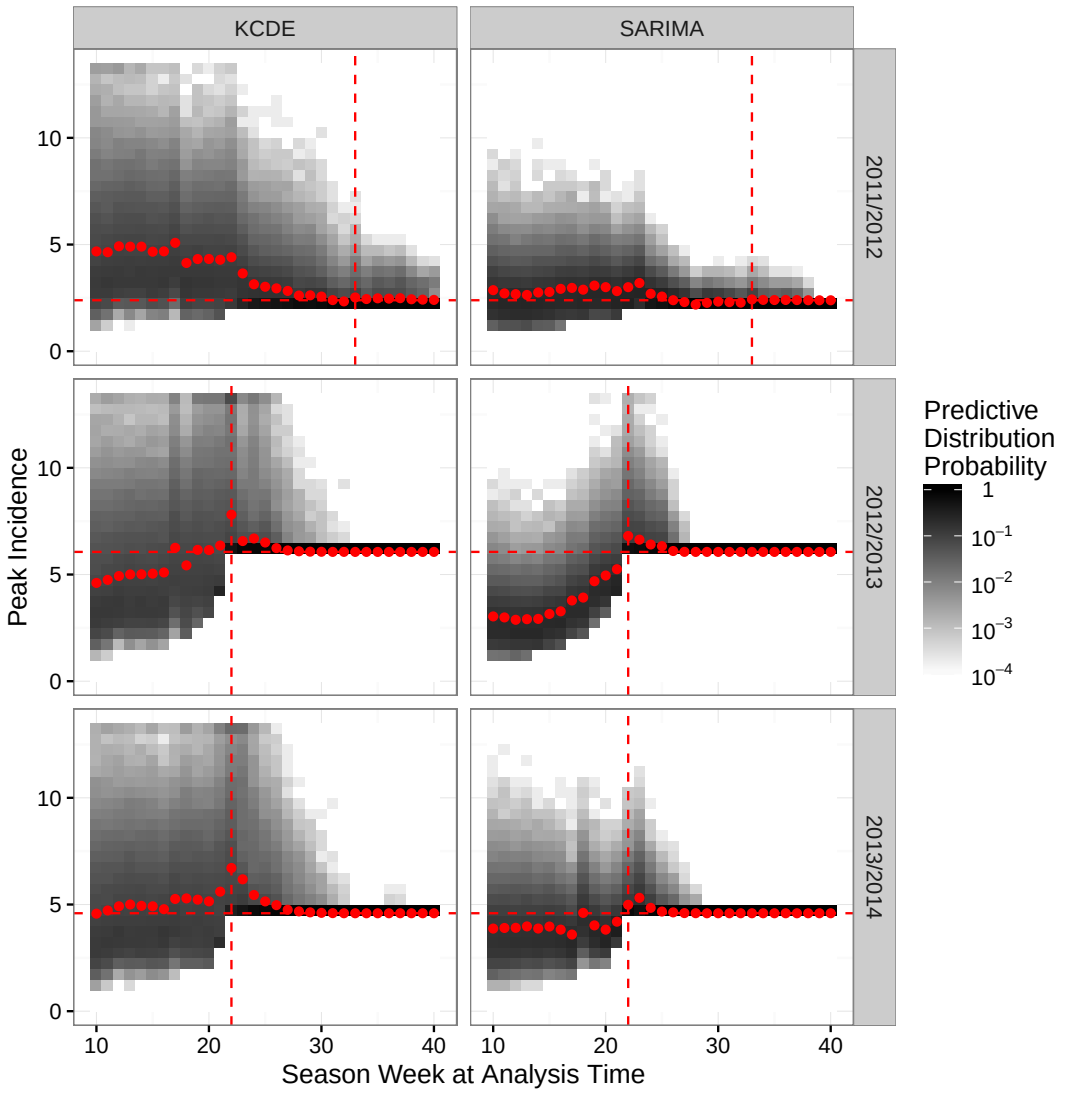
8. Peter Hall, Jeff Racine, and Qi Li. Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99(468):1015–1026, 2004.
9. Jeffrey D Hart and Philippe Vieu. Data-driven bandwidth choice for density estimation based on dependent data. *The Annals of Statistics*, 18(2):873–890, 1990.

Figure 11. Predictive distributions for predictions of peak week timing. The horizontal and vertical dashed lines are at the observed peak week for the season.



10. Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Science + Business Media, 2 edition, 2009.
11. Richard J Hatchett, Carter E Mecher, and Marc Lipsitch. Public health interventions and epidemic intensity during the 1918 influenza pandemic. *Proceedings of the National Academy of Sciences*, 104

Figure 12. Predictive distributions for predictions of peak week incidence. The horizontal dashed line is at the observed peak incidence for the season. The vertical dashed line is at the observed peak week for the season.



(18):7582–7587, 2007.

Figure 13. Differences in log scores for the predictive distributions for the peak week and incidence at the peak week for Dengue among pairs of models across all analysis times in the test period. In panel (a) positive values indicate cases when KCDE outperformed SARIMA. In panel (b) positive values indicate cases when the specification of KCDE with the periodic kernel outperformed the corresponding specification without the periodic kernel. In panel (c) positive values indicate cases when the specification of KCDE with a fully parameterized bandwidth outperformed the KCDE specification with a diagonal bandwidth matrix. In the plot for peak week timing in panel (a), the log score differences are not displayed for one analysis time when none of the simulated trajectories from SARIMA peaked at the true peak week. In that case, our monte carlo estimate of the difference in log scores is infinity.

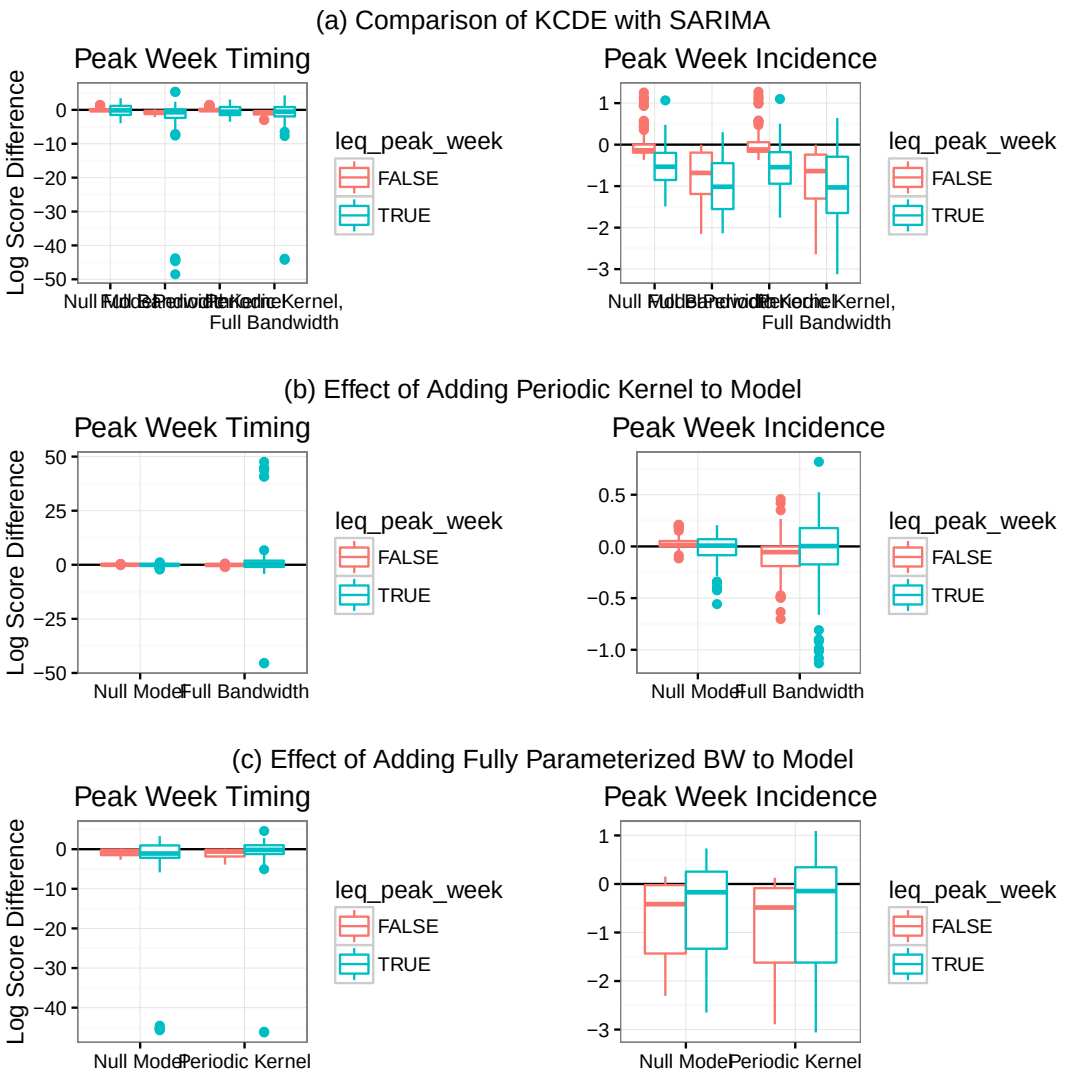
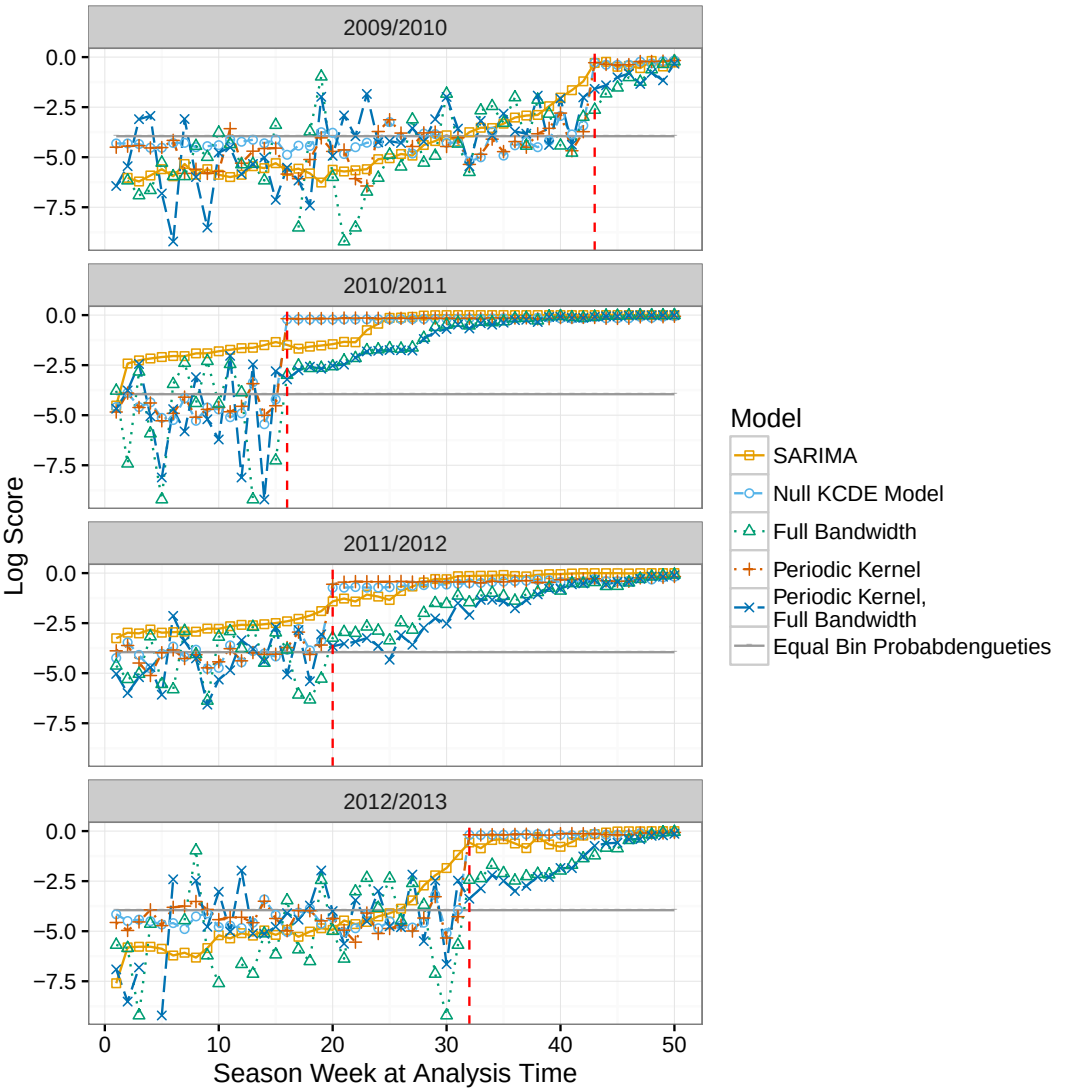
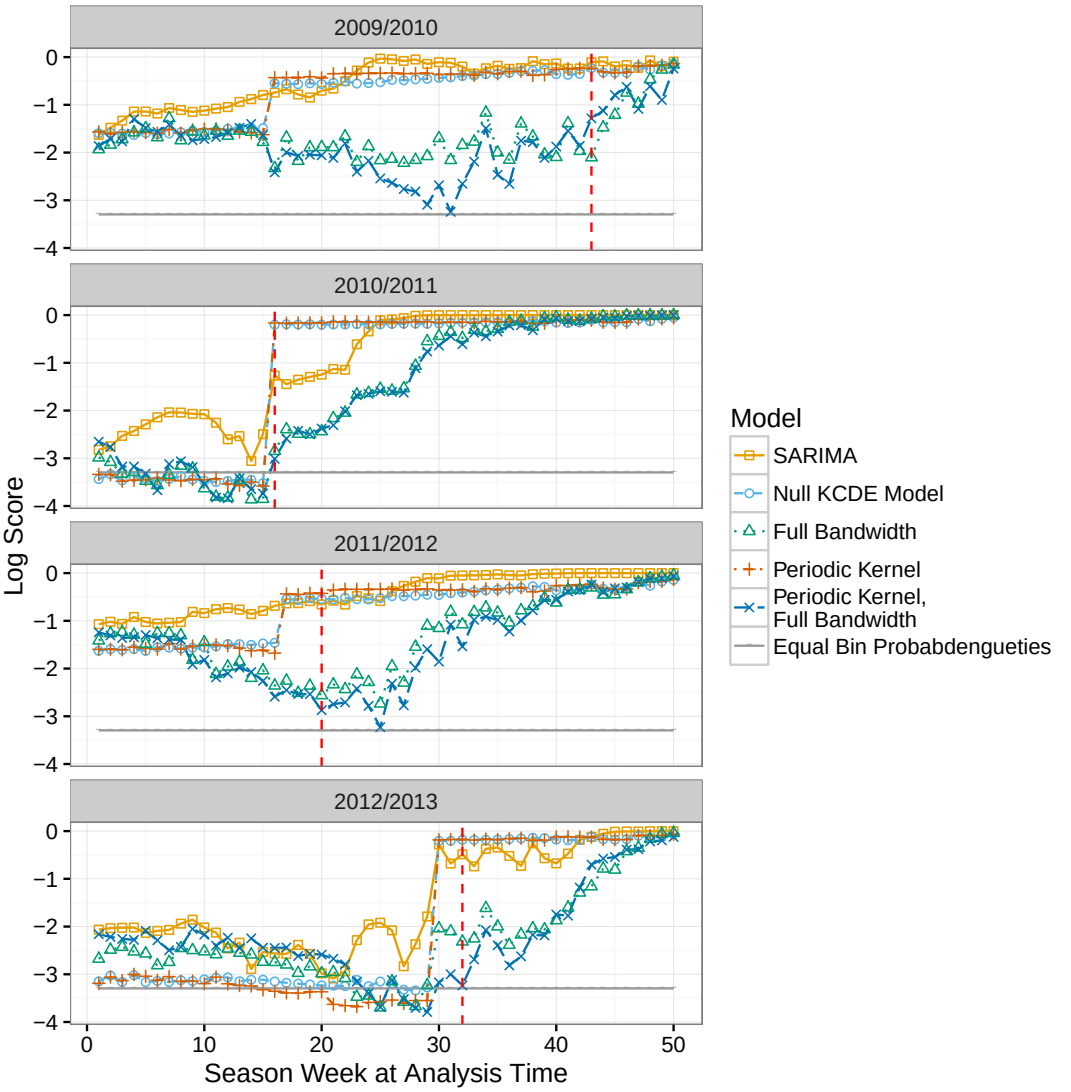


Figure 14. Log scores for predictions of peak week timing by predictive model and analysis time. The vertical gray line is placed at the peak week for each season.



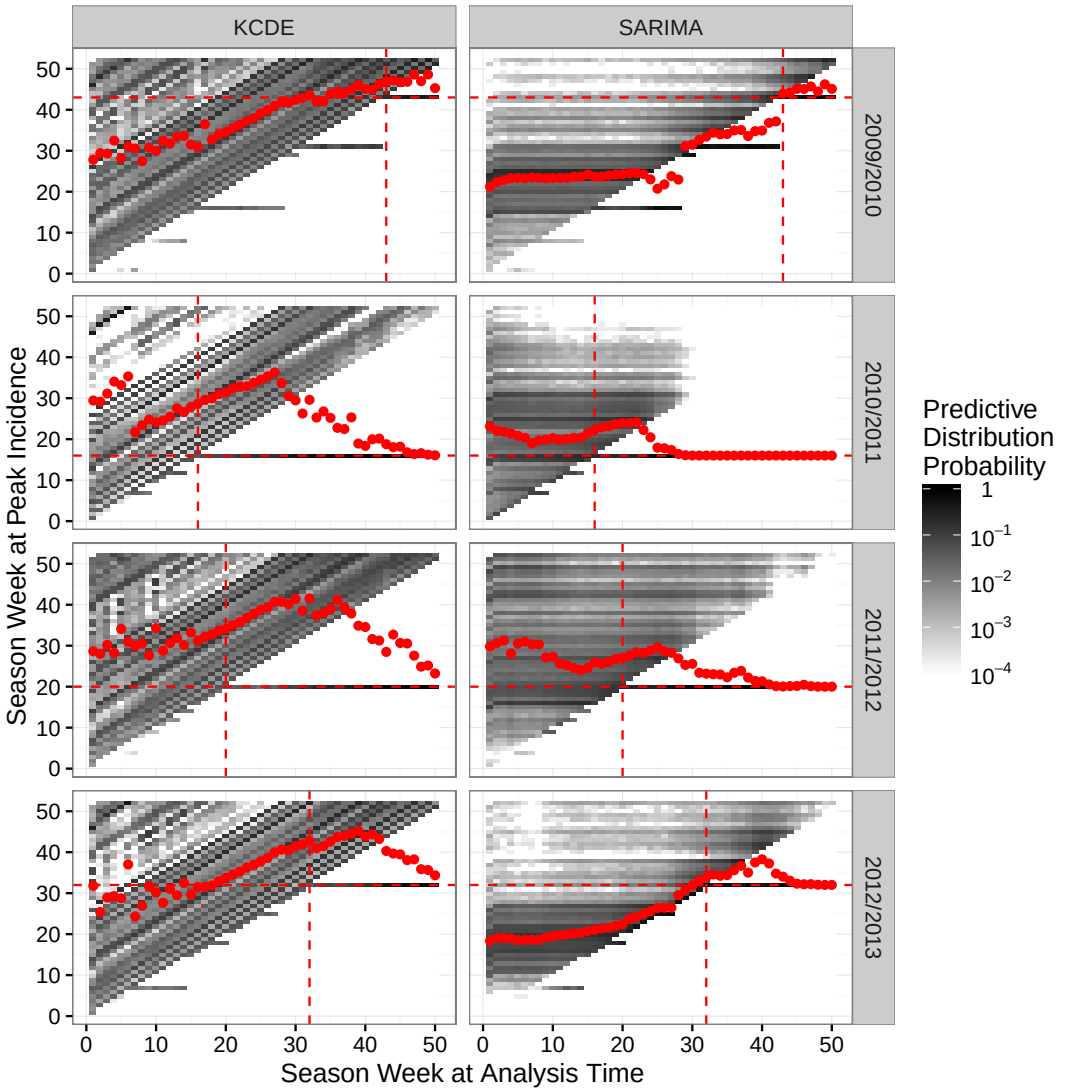
12. Marius Hofert, Ivan Kojadinovic, Martin Maechler, and Jun Yan. *copula: Multivariate Dependence with Copulas*, 2015. URL `http://CRAN.R-project.org/package=copula`. R package version 0.999-14.

Figure 15. Log scores for predictions of incidence in the peak week for Dengue by predictive model and analysis time. The vertical gray line is placed at the peak week for each season.



13. Rob J Hyndman. *forecast: Forecasting functions for time series and linear models*, 2015. URL `http://github.com/robjhyndman/forecast`. R package version 6.2.
14. Jooyoung Jeon and James W Taylor. Using conditional kernel density estimation for wind power density forecasting. *Journal of the American Statistical Association*, 107(497):66–79, 2012.

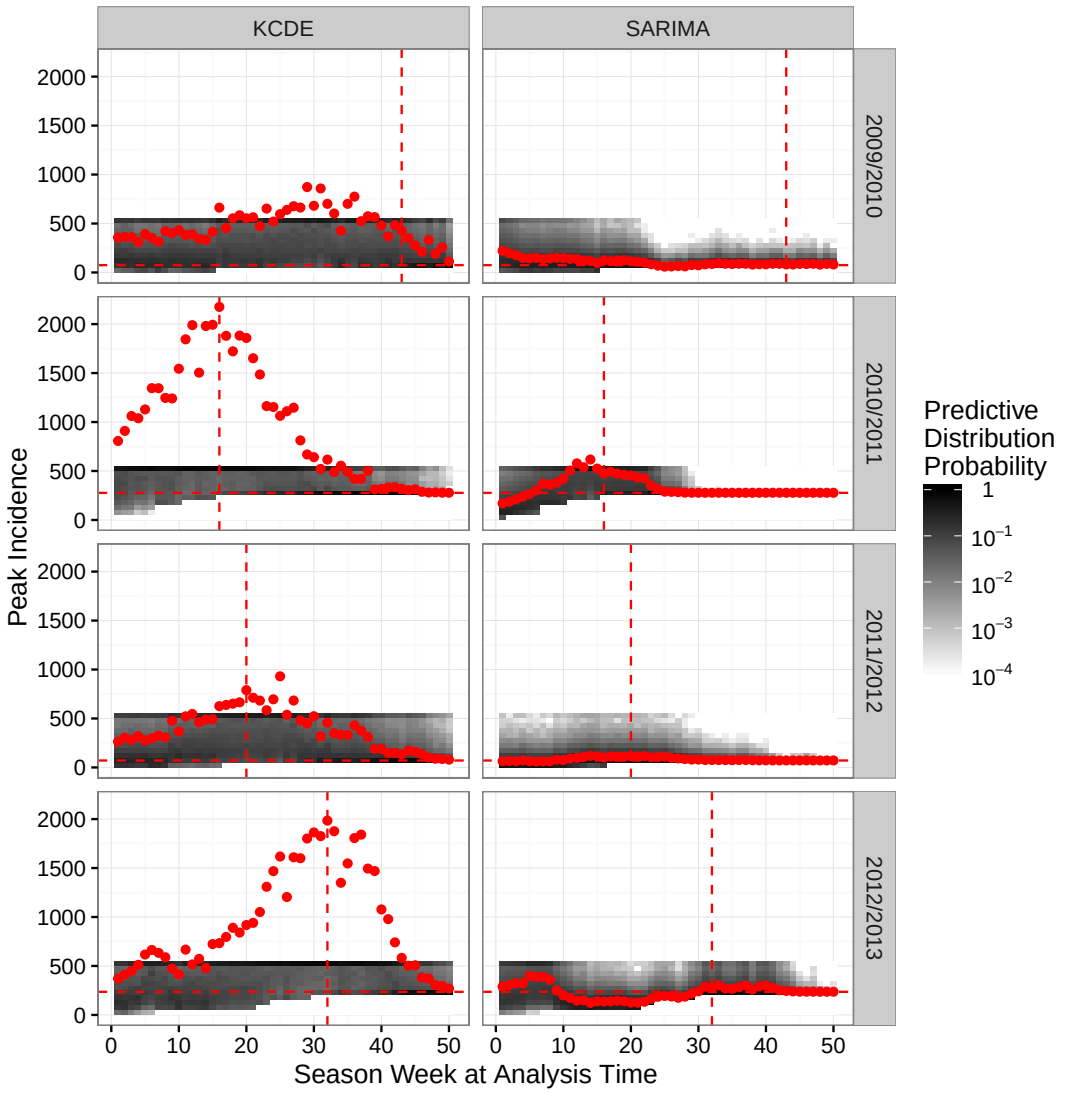
Figure 16. Predictive distributions for predictions of peak week timing for Dengue. The horizontal and vertical dashed lines are at the observed peak week for the season.



15. Harry Joe. Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94(2):401–419, 2005.

16. Qi Li and Jeff Racine. Nonparametric estimation of distributions with categorical and continuous data. *journal of multivariate analysis*, 86(2):266–292, 2003.

Figure 17. Predictive distributions for predictions of peak week incidence for Dengue. The horizontal dashed line is at the observed peak incidence for the season. The vertical dashed line is at the observed peak week for the season.



17. David JC MacKay. Introduction to gaussian processes. *NATO ASI Series F Computer and Systems Sciences*, 168:133–166, 1998.

18. Desheng Ouyang, Qi Li, and Jeffrey Racine. Cross-validation and the estimation of probability distributions with categorical data. *Journal of Nonparametric Statistics*, 18(1):69–100, 2006.
19. Pandemic Prediction and Forecasting Science and Technology Interagency Working Group. Dengue Forecasting, July 2015. URL `http://dengueforecasting.noaa.gov/`.
20. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org/`.
21. George Sugihara and Robert M. May. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series, April 1990.
22. Cécile Viboud, Pierre-Yves Boëlle, Fabrice Carrat, Alain-Jacques Valleron, and Antoine Flahault. Prediction of the spread of influenza epidemics by the method of analogues. *American Journal of Epidemiology*, 158(10):996–1006, 2003.
23. Jacco Wallinga, Michiel van Boven, and Marc Lipsitch. Optimizing infectious disease interventions during an emerging epidemic. *Proceedings of the National Academy of Sciences*, 107(2):923–928, 2010.
24. Min-Chiang Wang and John van Ryzin. A class of smooth estimators for discrete distributions. *Biometrika*, 68(1):301–309, 1981.
25. Haiming Zhou, Timothy Hanson, and Roland Knapp. Marginal bayesian nonparametric model for time to disease arrival of threatened amphibian populations. *Biometrics*, 71(4):1101–1110, 2015.