Infectious disease prediction with kernel conditional density estimation

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

[1] *Department of Biostatistics and Epidemiology, School of Public Health and Health Sciences, University of Massachusetts, Amherst*
415 Arnold House, 715 N. Pleasant Street, Amherst, MA 01003, USA

[2] *Dengue Branch, Division of Vector-Borne Infectious Diseases, Centers for Disease Control and Prevention, San Juan, Puerto Rico, USA*

Abstract

Creating statistical models that generate accurate predictions of infectious disease incidence over multiple time points is a challenging problem whose solution could benefit public health decision makers. We develop a new approach to this problem using kernel conditional density estimation (KCDE) and copulas. We obtain predictive distributions for incidence in individual weeks using KCDE and tie those distributions together into joint distributions using copulas. This strategy enables us to create predictions for the timing of and incidence in the peak week of the season. Our implementation of KCDE incorporates two novel kernel components: a periodic component that captures seasonality in disease incidence, and a component that allows for a full parameterization of the bandwidth matrix with discrete variables. We demonstrate via simulation that a fully parameterized bandwidth matrix can be beneficial for estimating conditional densities. We apply the method to predicting dengue fever and influenza, and compare to a seasonal autoregressive integrated moving average (SARIMA) model and a previously published generalized linear model for infectious disease incidence known as HHH4. KCDE outperforms the baseline methods for predictions of dengue incidence in individual weeks. KCDE also offers more consistent performance than the baseline models for predictions of incidence in the peak week, and is comparable to the baseline models on the other prediction targets. Using the periodic kernel function led to better predictions of incidence. Our approach and extensions of it could yield improved predictions for public health decision makers, particularly in diseases with heterogeneous seasonal dynamics such as dengue fever.

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

1 Introduction

With the maturation of digital disease surveillance systems in recent years, accurate and real-time infectious disease prediction has become an achieveable goal in many contexts. These predictions provide valuable information to public health officials planning disease prevention and control measures (Manheim et al. [2016]). For example, interventions designed to reduce person-to-person transmission of disease have been associated with diminished outbreak intensity (Hatchett et al. [2007]). Accurate predictions can help target such interventions more effectively.

Recent efforts by government officials and academic researchers have identified features of outbreaks that can be used by public health decision makers. In this article we focus on three of

these features: weekly incidence, the timing of the season peak, and incidence in the peak week. These quantities have emerged as being targets of particular utility in making planning decisions (Pandemic Prediction and Forecasting Science and Technology Interagency Working Group [2015], Epidemic Prediction Initiative [2016]).

In this work, we use a semi-parametric approach that combines a non-parametric method for conditional density estimation referred to as kernel conditional density estimation (KCDE) with a parametric method for modeling joint dependence structures known as copulas. Using data available up through a given time point, we employ KCDE to obtain separate predictive distributions for disease incidence in each subsequent week of the season. We then combine those marginal distributions using copulas to obtain joint predictive distributions for the trajectory of incidence over the following weeks. Predictive distributions relating to the timing of and incidence at the peak week can be obtained from this joint predictive distribution. Methods combining non-parametric estimates of marginal densities with copulas have been considered previously for other applications such as economic time series (Patton [2012]).

In addition to the novel application of these methods to predicting disease incidence, our contributions include the 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. Previous implementations of kernel methods involving discrete variables have employed a kernel function that is a product of univariate kernel functions (Aitchison and Aitken [1976], Bowman [1980], Grund [1993], Hall et al. [2004, 2007], Li and Racine [2003, 2008], Ouyang et al. [2006], Racine et al. [2004]). This approach forces the kernel function to be oriented in line with the coordinate axes. Motivated by results showing that multivariate kernel functions with a bandwidth parameterization allowing for flexible orientations can result in improved continuous density estimates (Duong and Hazelton [2005]), we introduce an approach that allows for flexible orientation of discrete kernels by discretizing an underlying continuous kernel function.

In a time-series context, KCDE is a local method in the sense that the conditional density estimate for future observations given conditioning variables is a weighted combination of contributions from previous observations with similar conditioning values. Using such local methods is a natural idea in predicting nonlinear systems because it imposes little structure on the assumed relationship between conditioning and outcome variables. Applications range from similar infectious disease settings where nearest neighbors regression has been used to make point predictions for incidence of measles (Sugihara and May [1990]) and influenza (Viboud et al. [2003]) to sports analytics where a version of nearest neighbors regression predicts the career trajectories of current NBA players (Silver [2015]). We note that KCDE can be seen as a distribution-based counterpart of nearest neighbors regression. For example, 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 (e.g., Hastie et al. [2009] discuss the connection between nearest neighbors and kernel methods for regression).

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 (Hall et al. [2004]), female labor force participation (Hall et al. [2004]), bond yields and value at risk in financial markets (Fan and Yim [2004]), and wind power (Jeon and Taylor [2012]), among others. Similar methods can also be formulated in the Bayesian framework. For example, Zhou et al. [2015] 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 for infectious disease prediction, including agent-based models, compartmental models and more generic regression-based time series models such as seasonal autoregressive integrated moving average (SARIMA) models among others. Brown *et al.*[in preparation] and Unkel et al. [2012] are recent reviews of work on forecasting infectious disease, and describe these alternative approaches in more detail.

Little research has been done comparing the predictive performance of more detailed and disease-mechanistic modeling approaches (agent-based or compartmental models) to more generic models

(regression or SARIMA). One difficulty in making comparisons to agent-based models is that these models are often highly parameterized and difficult to independently reproduce or replicate. An additional challenge with both agent-based and compartmental models is that expert knowledge is required to tailor them to the specific disease being modeled, and details of the assumed model specification can have a large impact on the quality of predictions (e.g., Grad et al. [2012]).

One of the most well-developed modern statistical frameworks for infectious disease prediction is the "HHH4" model (Held et al. [2005], Paul et al. [2008], Held and Paul [2012]), a specific variation of a generalized linear model developed for infectious disease. Another commonly used and widely studied approach is the seasonal autoregressive integrated moving average (SARIMA) model. However, both of these approaches have limitations that also hamper generalizability. The HHH4 model specifies a discrete distribution for the observed incidence measure, an appropriate assumption for some data sets, but not for others. The standard SARIMA specification is based on continuous distributions which means that it cannot be directly applied to modeling discrete case count data if low case counts are observed (Unkel et al. [2012]).

Several key features distinguish our approach from existing methods commonly used for predicting infectious disease incidence. First, we generate full predictive distributions to fully characterize uncertainty in the predictions. Compared to point predictions, this gives decision makers additional information in situations where the predictive distribution is skewed or has multiple modes. Second, unlike many methods common in the infectious disease literature, KCDE makes minimal assumptions about the underlying system governing disease dynamics. This flexibility makes KCDE suitable for application to a wide variety of time series, including diseases with different latent dynamics. Third, the method can easily be used with either discrete or continuous data by substituting one kernel function specification for another.

The remainder of this article is organized as follows. First, we describe our approach to prediction using KCDE and copulas. Next, we present the results of a simulation study comparing the performance of KCDE for estimating discrete conditional 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 discrete measure of weekly incidence of dengue fever in San Juan, Puerto Rico and a second with a continuous measure of weekly incidence of influenza in the United States. We conclude with a discussion of these results.

2 Method Description

Suppose we observe a measure z_t of disease incidence at evenly spaced times indexed by $t = 1, \ldots, T$. Our goal is to obtain predictions relating to incidence after time 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 a (conditional) probability measure with respect to an appropriately defined reference measure. We will use a colon notation to specify vectors: for example, $\mathbf{z}_{s:t} = (z_s, \ldots, z_t)$. The variable $t^* \in \{1, \ldots, T\}$ will be used to represent a time at which we desire to form a predictive distribution, using observed data up through t^* to predict incidence after t^*. When we apply the method to perform prediction for incidence after time T, t^* is equal to T; however, t^* takes other values in the estimation procedure we describe below. Let W denote the number of time points in a disease season (e.g., $W = 52$ 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, so that the times in the same season as t^* are indexed by $S_{t^*} + 1, \ldots, S_{t^*} + W$. Finally, let $H_{t^*} = W - (t^* - S_{t^*})$ denote the number of time points after t^* that are in the same season as t^*. H_{t^*} gives the largest prediction horizon for which we need to make a prediction in order to obtain predictions for all remaining time points in the season.

We obtain predictive distributions for each of three prediction targets. We will model the first of these prediction targets directly and frame the second and third 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 with prediction horizon $h \in \{1, ..., W\}$:

$$f(z_{t^*+h}|t^*, \mathbf{z}_{1:t^*})$$

2. Timing of the peak week of the current season, $w^* \in \{1, \ldots, W\}$:

$$\begin{aligned} P(\text{Peak Week} = w^*) &= P(Z_{S_{t^*}+w^*} = \max_w Z_{S_{t^*}+w}|t^*, \mathbf{z}_{1:t^*}) \\ &= \int_{\{\mathbf{z}_{(t^*+1):(t^*+H_{t^*})}: z_{S_{t^*}+w^*} = \max_w z_{S_{t^*}+w}\}} f(\mathbf{z}_{(t^*+1):(t^*+H_{t^*})}|t^*, \mathbf{z}_{1:t^*})\, d\mathbf{z}_{(t^*+1):(t^*+H_{t^*})}. \quad (1) \end{aligned}$$

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

$$\begin{aligned} P(\text{Incidence in Peak Week} \in [a, b)) &= P(a \leq \max_w Z_{S_{t^*}+w} < b|t^*, \mathbf{z}_{1:t^*}) \\ &= \int_{\{(\mathbf{z}_{(t^*+1):(t^*+H_{t^*})}): a \leq \max_w z_{S_{t^*}+w} < b\}} f(\mathbf{z}_{(t^*+1):(t^*+H_{t^*})}|t^*, \mathbf{z}_{1:t^*})\, d\mathbf{z}_{(t^*+1):(t^*+H_{t^*})}. \quad (2) \end{aligned}$$

In practice, we use Monte Carlo integration to evaluate the integrals in Equations (1) and (2) by sampling incidence trajectories from the joint predictive distribution.

We introduce the overall structure of our model here and describe its components and parameter estimation in more detail in the following Subsections. 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 $l_1, \ldots, l_M$: $f(\mathbf{z}_{(t^*+1):(t^*+H_{t^*})}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M})$. For notational simplicity, we take l_M to be the largest of these lags. The model represents this density as follows:

$$\begin{aligned} &f(z_{(t^*+1):(t^*+H_{t^*})}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M}) = \\ &\quad c^{H_{t^*}}\{f^1(z_{t^*+1}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M}; \boldsymbol{\theta}^1), \ldots, f^{H_{t^*}}(z_{t^*+H_{t^*}}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M}; \boldsymbol{\theta}^{H_{t^*}}); \boldsymbol{\xi}^{H_{t^*}}\}. \end{aligned} \quad (3)$$

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.

2.1 KCDE for Predictive Densities at Individual Prediction Horizons

We now discuss the use of KCDE to obtain $f^h(z_{t^*+h}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M}; \boldsymbol{\theta}^h)$, the predictive density for disease incidence at a particular horizon h after time t^*. 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 $f^h(y_{t^*}^h|\mathbf{x}_{t^*}; \boldsymbol{\theta}^h)$.

To estimate this distribution, we use the observed data to form the pairs $(\mathbf{x}_t, y_t^h)$ for all $t = 1 + l_M, \ldots, T - h$ (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:

$$\hat{f}^h(y^h_{t^*}|\mathbf{x}_{t^*}) = \frac{\sum_{t\in\boldsymbol{\tau}} K^{\mathbf{X},Y}\left\{(\mathbf{x}_{t^*}, y^h_{t^*}), (\mathbf{x}_t, y^h_t); \boldsymbol{\theta}^h\right\}}{\sum_{t\in\boldsymbol{\tau}} K^{\mathbf{X}}(\mathbf{x}_{t^*}, \mathbf{x}_t; \boldsymbol{\theta}^h)} \tag{4}$$

$$= \sum_{t\in\boldsymbol{\tau}} \zeta^h_{t^*,t} K^{Y|\mathbf{X}}(y^h_{t^*}, y^h_t|\mathbf{x}_{t^*}, \mathbf{x}_t; \boldsymbol{\theta}^h), \text{ where} \tag{5}$$

$$\zeta^h_{t^*,t} = \frac{K^{\mathbf{X}}(\mathbf{x}_{t^*}, \mathbf{x}_t; \boldsymbol{\theta}^h)}{\sum_{s\in\boldsymbol{\tau}} K^{\mathbf{X}}(\mathbf{x}_{t^*}, \mathbf{x}_s; \boldsymbol{\theta}^h)}. \tag{6}$$

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 $K^{Y|\mathbf{X}}$. With this restriction, we can interpret $K^{\mathbf{X}}$ as a weighting function determining how much each observation $(\mathbf{x}_t, y^h_t)$ 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. These weights are the $\zeta^h_{t^*,t}$ in Equations (5) and (6). $K^{Y|\mathbf{X}}$ is a density function that contributes mass to the final density estimate near y^h_t. 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 (4) through (6), $\boldsymbol{\tau} \subseteq \{(1 + l_M), \ldots, (T - h)\}$ 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, which are measures of disease incidence:

$$K^{\mathbf{X},Y}\left\{(\mathbf{x}_{t^*}, y^h_{t^*}), (\mathbf{x}_t, y^h_t); \boldsymbol{\theta}^h\right\}$$
$$= K^{\text{per}}(t^*, t; \boldsymbol{\theta}^h_{\text{per}}) K^{\text{inc}}\{(z_{t^*-l_1}, \ldots, z_{t^*-l_M}, z_{t^*+h}), (z_{t-l_1}, \ldots, z_{t-l_M}, z_{t+h}); \boldsymbol{\theta}^h_{\text{inc}}\}.$$

Here we have set $\boldsymbol{\theta}^h = (\boldsymbol{\theta}^h_{\text{per}}, \boldsymbol{\theta}^h_{\text{inc}})$.

The periodic kernel function was originally developed in the literature on Gaussian Processes (MacKay [1998]), and is defined by

$$K^{\text{per}}(t^*, t; \rho^h, \eta^h) = \exp\left[-\frac{\sin^2\{\rho^h(t^* - t)\}}{2(\eta^h)^2}\right]. \tag{7}$$

We illustrate this kernel function in Figure 1. It has two parameters: $\boldsymbol{\theta}^h_{\text{per}} = (\rho^h, \eta^h)$, where ρ^h determines the length of the periodicity and η^h determines the strength and locality of this periodic component in computing the observation weights $\zeta^h_{t^*,t}$. In our applications, we have fixed $\rho^h = \pi/52$, 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^h_t. In our applications, these variables are measures of incidence; for brevity of notation, we collect them in the column 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 parameterized in terms of its mode rather than its mean (Figure 1). Using the mode ensures that the contribution to the conditional density is largest near z_{t+h}. 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:

$$K^{\text{inc}}_{\text{cont}}(\tilde{\mathbf{z}}_{t^*}, \tilde{\mathbf{z}}_t; \mathbf{B}) = \frac{\exp\left[-\frac{1}{2}\{\log(\tilde{\mathbf{z}}_{t^*}) - \log(\tilde{\mathbf{z}}_t) - \mathbf{B}\underline{1}\}'\mathbf{B}^{-1}\{\log(\tilde{\mathbf{z}}_{t^*}) - \log(\tilde{\mathbf{z}}_t) - \mathbf{B}\underline{1}\}\right]}{(2\pi)^{\frac{M+1}{2}}|\mathbf{B}|^{\frac{1}{2}} z_{t^*+h} \prod_{m=1}^{M} z_{t^*-l_m}} \tag{8}$$

In this expression, $\underline{1}$ is a column vector of ones. The matrix $\mathbf{B}$ is a bandwidth matrix that controls the orientation and scale of the kernel function. This bandwidth matrix is parameterized by $\boldsymbol{\theta}^h_{\text{inc}}$. In this work we have considered two parameterizations: a diagonal bandwidth matrix, and a fully parameterized bandwidth based on the Cholesky decomposition. To obtain the discrete kernel (Figure 1), we integrate an underlying continuous kernel function over hyper-rectangles containing the points in the range of the discrete random variable (see supplement for details).

We estimate the bandwidth parameters $\boldsymbol{\theta}^h$ by numerically maximizing the cross-validated log score of the predictive distributions for the observations in the training data. For a random variable Y with observed value y the log score of the predictive distribution f_Y is $log\{f_Y(y)\}$. A larger log score indicates better model performance. In obtaining the cross-validated log score for the predictive distribution at time t^*, we leave the year of training data before and after the time t^* out of the set $\boldsymbol{\tau}$ in Equations (4) through (6). 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 (Pandemic Prediction and Forecasting Science and Technology Interagency Working Group [2015], Epidemic Prediction Initiative [2016]). We apply our method to the data sets from those competitions in the applications section below, and report log scores to facilitate comparisons with other results from those competitions that may be published in the future. In general, the log score is a strictly proper scoring rule; i.e., its expectation is uniquely maximized by the true predictive distribution (Gneiting and Raftery [2007]). However, its use as an optimization criterion has been criticised for being sensitive to outliers (Gneiting and Raftery [2007]). In the kernel density estimation literature, this approach to estimation is referred to as likelihood cross-validation, and similar criticisms have been made regarding its performance in handling outliers and estimating heavy-tailed distributions (Schuster and Gregory [1981], Scott and Factor [1981]).

2.2 Combining Marginal Predictive Distributions with Copulas

We use copulas (Nelsen [2007]) 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. The copula is a parametric function that captures the dependence relations among a collection of random variables and allows us to compute the joint distribution from the marginal distributions. Figure 16 in the supplement shows that the copula induces positive correlation in the predictive distributions for incidence in nearby weeks, so that high incidence in one week is more likely to be followed by high incidence in weeks soon after.

To describe our methods for both continuous and discrete distributions, it is most convenient to frame the discussion in this Subsection in terms of cumulative distribution functions (CDF) instead of density functions. We will use a capital C to denote the copula function for CDFs and a lower case c to denote the copula function for densities. Similarly, the predictive densities $f^h(y^h_{t^*}|\mathbf{x}_{t^*};\boldsymbol{\theta}^h)$ we obtained in the previous Subsection naturally yield corresponding predictive CDFs $F^h(y^h_{t^*}|\mathbf{x}_{t^*};\boldsymbol{\theta}^h)$.

Our model specifies the joint CDF for $(Y^1_{t^*},\ldots,Y^{H_{t^*}}_{t^*})$ as follows:

$$\begin{aligned} &F^{H_{t^*}}(y^1_{t^*},\ldots,y^{H_{t^*}}_{t^*}|\mathbf{x}_{t^*};\boldsymbol{\theta}^1,\ldots,\boldsymbol{\theta}^{H_{t^*}},\boldsymbol{\xi}^{H_{t^*}}) = \\ &\quad C\{F^1(y^1_{t^*}|\mathbf{x}_{t^*};\boldsymbol{\theta}^1),\ldots,F^{H_{t^*}}(y^{H_{t^*}}_{t^*}|\mathbf{x}_{t^*};\boldsymbol{\theta}^{H_{t^*}});\boldsymbol{\xi}^{H_{t^*}}\} \end{aligned} \tag{9}$$

The copula function C maps the marginal CDF values to the joint CDF value. We use the isotropic normal copula implemented in the `R` (R Core Team [2016]) package `copula` (Hofert et al. [2015]). The copula function is given by

$$C(u_1,\ldots,u_H;\boldsymbol{\xi}^H) = \Phi_{\Sigma^H}(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_H)), \tag{10}$$

where Φ^{-1} is the inverse CDF of a univariate normal distribution with mean 0 and variance 1 and Φ_{Σ^H} is the CDF of a multivariate normal distribution with mean $\underline{0}$ and covariance matrix Σ^H. The

isotropic specification sets $\Sigma^H = [\sigma^H_{i,j}]$, where

$$\sigma^H_{i,j} = \begin{cases} 1 \text{ if } i = j, \\ \xi^H_d \text{ if } |i-j| = d \end{cases} \tag{11}$$

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 $H = 1$). Broadly, estimation for the model parameters proceeds in two stages: first we estimate the parameters for KCDE separately for each prediction horizon $h = 1, \ldots, H$ as described in the previous Section, and second we estimate the copula parameters while holding the KCDE parameters fixed. We give a more detailed description of this estimation procedure in the supplement. 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 (Joe [2005]). Also, it results in a large reduction in the computational cost of parameter estimation.

3 Simulation Study

We conducted a simulation study to examine the utility of using a non-diagonal bandwidth matrix specification when estimating discrete conditional distributions with KCDE. The simulation study is motivated by the simplest case of predicting incidence in a single week using KCDE as outlined in Subsection 2.1: predicting incidence at time $t + h$ given incidence at time t. A central characteristic of the disease incidence data we analyze in the next Section is the presence of positive correlation between incidence in nearby time points (Supplemental Figure 2). In this simulation study we demonstrate that in the presence of such correlation, using fully parameterized bandwidth matrices can improve conditional density estimates over using a diagonal bandwidth.

There are many factors that determine the relative performance of KCDE estimators with different bandwidth parameterizations. In this simulation study, we vary just one of these factors: the sample size ($N = 100$ or $N = 1000$).

We conducted 500 simulation trials for each sample size. In each trial, we simulated N observations of a discretized bivariate normal random variable $\mathbf{X}$ with mean $\underline{0}$ and covariance matrix Σ where Σ has 1 on the diagonal and 0.9 off of the diagonal (see Supplement for further detail). Using these observations as a training data set, we estimated the bandwidth parameters for two variations on a KCDE model for the conditional distribution of $X_1|X_2$: one with a diagonal bandwidth matrix specification and one with a fully parameterized bandwidth matrix. In this simulation study, the kernel function was obtained by discretizing a multivariate normal kernel function rather than a log-normal kernel function as in the applications below. Otherwise, the method is as described previously.

We evaluated the conditional density estimates by an importance sampling approximation of the Hellinger distance of the conditional density estimate from the true conditional density, integrated over the range of the conditioning variables (see supplement). The Hellinger distance lies between 0 and 1, with smaller values indicating that the density estimate is better. It has been argued that the Hellinger distance is preferred to other measures of the quality of kernel density estimates such as integrated squared error (Kanazawa [1993]). For each combination of the training set sample size, dimension, and simulation trial, we compute the difference between the Hellinger distance from the true conditional distribution achieved with a diagonal bandwidth matrix and with a fully parameterized bandwidth matrix.

The results indicate that in the presence of correlation between the conditioning variable and the density estimation target, using a fully parameterized bandwidth matrix instead of a diagonal bandwidth generally yields improved density estimates as measured by the integrated Hellinger distance (Figure 2). The average improvement from using a fully parameterized bandwidth matrix is larger with a sample size of $N = 100$ instead of $N = 1000$, but there is also more variation in

performance with the smaller sample size. This suggests that using a fully parameterized bandwidth may be helpful in applications similar to infectious disease prediction where there is correlation between the quantity being predicted (e.g., future incidence) and the quantities that we condition on in order to make the predictions.

4 Applications

In this Section, we illustrate our methods through applications to prediction of infectious disease incidence in two examples with real disease incidence data sets (Figure 3). The first provides a weekly integer number of reported cases of dengue fever in San Juan, Puerto Rico. The second provides a weekly weighted proportion of doctor visits with influenza-like illness in the United States. These data sets were used in two recent prediction competitions sponsored by the United States federal government (Pandemic Prediction and Forecasting Science and Technology Interagency Working Group [2015], Epidemic Prediction Initiative [2016]).

We 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, as was chosen by the competition administrators. All predictions are made as though in real time, assuming that once cases are reported they are never revised and that there are no delays in reporting. Specifically, we use only data up through a given week to make predictions for incidence after that week.

We use the three prediction targets described in Section 2 (Supplemental Figure 3). Following the precedent set in the competitions, we make predictions for *binned* incidence in the peak week. For the dengue data set, the bins are $[0, 50), [50, 100), \ldots, [500, \infty)$. For the influenza data set, the bins are $[0, 0.5), [0.5, 1), \ldots, [13, \infty)$. Our predictions for incidence in individual weeks are for the raw, unbinned, incidence measure.

Our applications evaluate four variations on KCDE model specifications:

1. The "Null KCDE" model omits the periodic component of the kernel function and uses a diagonal bandwidth matrix specification for the incidence kernel.

2. The "Full Bandwidth KCDE" model omits the periodic component of the kernel function and uses a fully parameterized bandwidth matrix specification for the incidence kernel.

3. The "Periodic KCDE" model includes the periodic component of the kernel function and uses a diagonal bandwidth matrix specification for the incidence kernel.

4. The "Periodic, Full Bandwidth KCDE" model includes the periodic component of the kernel function and uses a fully parameterized bandwidth matrix specification for the incidence kernel.

We include two baseline models for comparison to our methods. The first is a seasonal autoregressive integrated moving average (SARIMA) model. 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 SARIMA 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 (Hyndman [2015]); this function uses a stepwise procedure to determine the terms to include in the model. This procedure resulted in a SARIMA$(2,0,0)(2,1,0)_{52}$ model for the influenza data and a SARIMA$(3,0,2)(1,1,0)_{52}$ model for the dengue data. In applying this model to the dengue data, we have discretized the predictive distributions obtained from SARIMA using the same methods that we used for KCDE. This discretization was not used in model estimation since it is not available in the standard estimation software.

The second baseline model is the "HHH4" model for infectious disease incidence (Held et al. [2005], Paul et al. [2008], Held and Paul [2012]), available in the `surveillance` (Höhle et al. [2016]) package in R. This is a generalized linear model with either a Poisson or Negative Binomial family. The mean is a linear combination of autoregressive and sinusoidal components. We followed the

model selection and estimation procedures outlined in Held and Paul [2012] (see Supplement for details). Because the HHH4 model specifies a Poisson or Negative Binomial distribution for incidence, it cannot be used with continuous data; therefore, we have only applied this model with the dengue data.

The predictions showed similar performance across the methods in many scenarios, although KCDE showed more stable and calibrated performance across long horizons, especially in years that had unusually high incidence (Figure 4 and Supplemental Figure 4). For predictions of dengue fever incidence in individual weeks, at a prediction horizon of one week the point predictions from all three methods are similar, although the intervals from SARIMA are quite a bit wider than those from KCDE and HHH4. All three methods struggle at larger prediction horizons, but it appears that the SARIMA and HHH4 models have more difficulty with aligning the predictive distribution with the season's peak, particularly in the two seasons with higher incidence. For the more regularly seasonal influenza data, there is less of a noticeable distinction between the predictions given by the KCDE and SARIMA models for incidence in individual weeks.

We evaluated the performance of all KCDE specifications relative to the baseline model(s) for each dataset using log scores. Across both data sets, including the periodic kernel component in the KCDE specification led to consistent improvements in predictions of incidence in individual weeks. (Figure 5 and Supplemental Figure 5). The periodic kernel also generally improved predictions of incidence in the peak week, but it did not have consistent benefits in predicting peak week timing (Supplemental Figures 10 and 11). Using the fully parameterized bandwidth matrix generally had little impact on the quality of the predictive distributions (Figure 5 and Supplemental Figures 5, 10, and 11).

For predictions of incidence in individual weeks in the application to dengue, the KCDE specifications performed similarly to the baseline models near the season nadir but outperformed the baseline models in predictions for times of high incidence near the season peaks (Figure 6 and Supplemental Figures 6 through 8). For example, in weeks with fewer than 93 reported cases (roughly one third of the maximum weekly case count in the testing period), the median log score difference between the predictions from the Periodic, Full Bandwidth KCDE model and SARIMA was about 0.11 [first quartile (Q_1) = -0.14, third quartile (Q_3) = 0.47], where values greater than zero show KCDE making more accurate predictions than SARIMA. The median log score difference in comparison to HHH4 in this incidence range was about -0.09 (Q_1 = -0.30, Q_3 = 0.22). However, in weeks with more than 184 reported cases (the upper tertile) the median difference in log scores was about 1.48 (Q_1 = 0.25, Q_3 = 2.72) relative to SARIMA, and about 0.94 (Q_1 = 0.41, Q_3 = 1.95) relative to HHH4. Translating to a probability scale, in these periods of high incidence this KCDE specification assigned about 5 times higher probability to the observed outcome as SARIMA on average and about **1.25** times higher probability as HHH4 on average. Moreover, there were cases where the KCDE model assigned up to about 450 times as much probability to the realized outcome as SARIMA, and over 1300 times as much probability as HHH4. **Across all weeks in the test period and all prediction horizons, neither baseline model ever outperformed this KCDE specification by a factor of more than 9.** Similar patterns also hold with the other KCDE specifications for dengue. These results for predictions at times of high incidence are visible in Figure 5 as the long upper tails of cases when KCDE outperformed the baseline models. For the influenza application, the KCDE specifications including the periodic kernel component performed about as well as SARIMA throughout the season.

KCDE offered more consistent performance than the baseline methods for predictions of the timing of the peak week and incidence in the peak week that were made before the season peak; all methods did well in predictions for these quantities made after the peak had passed (Figure 7 and Supplemental Figures 12 through 15). For predictions of peak incidence, the baseline methods occasionally struggled in the seasons with high peak incidence (Figure 7 and Supplemental Figure 10). In the two dengue seasons with relatively high incidence, predictions of peak incidence made by the HHH4 model before the season peak had much lower log scores than both KCDE and a naive model that assigns equal probability to each incidence bin. Similarly, SARIMA underperformed

relative to both KCDE and using equal bin probabilities in the influenza season with the highest peak. Meanwhile, pre-peak predictions of peak incidence from the Periodic, Full Bandwidth KCDE model had higher average log scores than the equal bin probabilities model in three of the dengue seasons, and only did slightly worse than the naive model in the 2010/2011 season – when both of the baseline models also did worse than the equal bin probabilities model on average. In every influenza season in the test period, the average performance of each KCDE specification was better than using equal bin probabilities for predicting peak incidence. Across both data sets, the worst performance of any of the KCDE specifications for predicting peak incidence was with the Null KCDE model in the 2012/2013 dengue season – and that performance was much better than the worst-case performance of either of the baseline models. For the most part, the methods were comparable to each other in the remaining seasons, and offered considerable improvements over equal bin probabilities. For predictions of peak week timing, SARIMA consistently underperformed relative to using equal bin probabilities in two of the dengue seasons (Supplemental Figures 9 and 11). KCDE and HHH4 rarely did much worse than using equal bin probabilities.

5 Conclusions

Prediction of infectious disease incidence at horizons of more than a few weeks is a challenging task. We have presented a semi-parametric approach to doing this based on KCDE and copulas and found that it is a viable method that can yield improved predictions relative to commonly employed methods in this field. In predicting incidence of dengue fever in individual weeks, our approach offered consistent and substantial performance gains relative to a SARIMA model and the HHH4 model. These improvements were particularly concentrated in the times that are of most interest to public health decision makers: periods of high incidence near the season peak. In the application to influenza, our method did about as well as SARIMA when predicting incidence in individual weeks.

Overall, across both data sets our method offered more consistency than the baseline models in predictions for incidence in the peak week. Both baseline models suffered in one or more seasons with high incidence where they made substantially worse predictions than a naive model assigning equal probability to each incidence bin, whereas KCDE never did much worse than this naive model. For pre-peak predictions of peak week timing, there were multiple seasons where the SARIMA model consistently underperformed relative to the naive approach of assigning equal probability to each week of the year; KCDE and HHH4 were more consistently at or above the level of this naive approach. These improvements in the reliability of early-season predictions are valuable to public health officials planning interventions several weeks or months before the peak of the disease season. However, since year-to-year variation is substantial, continued evaluation of these methods on datasets with longer prospective testing phases could provide better information about long-run performance of all of these methods.

Our implementation of KCDE offers two main methodological contributions. Most importantly in the context of modeling infectious disease, we have introduced the use of a periodic kernel component that captures seasonality. In both of our applications, including this periodic kernel component in the KCDE specification led to substantial improvements in the predictive distributions for incidence in individual weeks. We also introduced a method for obtaining kernel functions that are appropriate for use with discrete data while allowing for a fully parameterized bandwidth matrix. In our applications, using a fully parameterized bandwidth matrix did not lead to consistent improvements in predictions. However, we have demonstrated through a simulation study that the fully parameterized bandwidth can be helpful in some conditional density estimation tasks. This general method for obtaining discrete kernel functions may be beneficial in other applications of KCDE.

We believe that the difference in relative performance of KCDE and the baseline models for prediction in the dengue and influenza data sets can be explained to a great extent by differences in the underlying disease processes and how they relate to the model specifications. The most salient difference between the two time series is the much greater season-to-season variablility in the dengue data set relative to the influenza data set (Figure 3). For dengue, the peak incidence in the largest

season is about 30 times larger than the peak incidence in the smallest season; this ratio is only about 3 for influenza. It may be the case that the restrictive structure of the SARIMA and HHH4 models means that they are not able to capture the dynamics of dengue incidence accurately. For example, Held and Paul [2012] discuss the fact that the seasonal structure in the HHH4 model does not explicitly allow for different amplitudes in different seasons. Relaxing that structure by using a non-parametric approach such as KCDE may yield improved capability to represent the disease dynamics. This is less of an issue in predicting influenza where there is much more consistency across different seasons.

A major advantage of the approach we have outlined is its flexibility in terms of cleanly handling both discrete and continuous data and a variety of underlying disease mechanisms. Our method consistently yielded reasonable predictions for all three prediction targets in both applications. As we have seen, the HHH4 model is formulated in terms of discrete case counts and so could not be directly applied to the influenza data where the disease measure was continuous. Even in the data set where it could be used, the HHH4 model underperformed relative to KCDE in predictions for incidence in individual weeks and incidence in the peak week. Similarly, the standard SARIMA model is formulated in terms of continuous distributions. The resulting continuous predictive distributions can be discretized as we have done in this article, but without extra coding effort the method is not appropriate for use with case count data when small integer numbers of cases are reported. Furthermore, our approach consistently equalled or exceeded the performance SARIMA across the applications to dengue and influenza.

There is a great deal of room for extensions and improvements to the methods we have outlined in this article. One major limitation of our work lies in the selection of conditioning variables for the predictive model. We have simply used incidence at the two most recent time points, and possibly the observation time, as conditioning variables. We considered using a stepwise variable selection approach to select the model specification, but we found this to be too computationally expensive to be practical; the full grid search suggested by De Gooijer and Gannoun [2000] would be far too slow for our methods.

Another possibility for addressing this problem would be to replace variable selection with shrinkage. Hall et al. [2004] 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. We conjecture that by introducing an appropriate penalty on the elements bandwidth matrix, we could include more (possibly irrelevant) conditioning variables in the model without requiring a dramatically larger sample size. If successful, this would also enable further exploration of using other predictive variables such as weather or incidence measures from neighboring locations in the model.

Another aspect of our method that should be explored further is the use of log score in estimation. We used log scores in this work to match the use of log scores in evaluating and comparing the performance of different models. The log score has the advantage of defining a proper scoring rule, but it has the disadvantage of being sensitive to outlying values. Previous authors have suggested the use of other loss functions in estimation for kernel-based density estimation methods that reduce these effects, such as variations on integrated squared error (e.g., Fan and Yim [2004]) or the continuous ranked probability score (Jeon and Taylor [2012]).

There is also a long history of using other modeling approaches such as compartmental models for infectious disease prediction. 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. On the positive side, these minimal assumptions are what make KCDE appropriate for use with a wide variety of disease processes with minimal changes to the model specification. On the other hand, we believe that a well-specified mechanistic model might outperform KCDE in certain circumstances. However, rather than selecting one "preferred" modeling framework or model formulation, we believe it may be fruitful to incorporate the methods developed in this paper as components of an ensemble with several different types of models. An appropriately constructed ensemble incorporating predictions from KCDE as well as other methods might perform better than

any of the component models on their own, and would be a valuable approach for maximizing the utility of these predictions to public health decision makers.

6 Software

The estimation methods were implemented in `R` and `C`. All source code and data are available in `R` packages hosted on GitHub (Ray et al. [2016]).

7 Supplementary Material

The reader is referred to the on-line Supplementary Materials for technical details and additional figures with further information about the results.

Acknowledgments

The authors thank the competition administrators for making disease incidence data available. This work was supported by the National Institute of Allergy and Infectious Diseases at the National Institutes of Health (grants R21AI115173 and R01AI102939).

References

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

Adrian W Bowman. A note on consistency of the kernel method for the analysis of categorical data. *Biometrika*, 67(3):682–684, 1980.

Jan G De Gooijer and Ali Gannoun. Nonparametric conditional predictive regions for time series. *Computational Statistics & Data Analysis*, 33(3):259–275, 2000.

Tarn Duong and Martin L Hazelton. Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics*, 32(3):485–506, 2005.

Epidemic Prediction Initiative. FluSight: Seasonal Influenza Forecasting, January 2016. http://dengueforecasting.noaa.gov/.

Jianqing Fan and Tsz Ho Yim. A crossvalidation method for estimating conditional densities. *Biometrika*, 91(4):819–834, 2004.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Yonatan H Grad, Joel C Miller, and Marc Lipsitch. Cholera modeling: challenges to quantitative analysis and predicting the impact of interventions. *Epidemiology (Cambridge, Mass.)*, 23(4):523, 2012.

Birgit Grund. Kernel estimators for cell probabilities. *Journal of Multivariate Analysis*, 46(2): 283–308, 1993.

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.

Peter Hall, Qi Li, and Jeffrey S Racine. Nonparametric estimation of regression functions in the presence of irrelevant regressors. *The Review of Economics and Statistics*, 89(4):784–789, 2007.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Science & Business Media, 2nd edition, 2009.

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(18):7582–7587, 2007.

Leonhard Held and Michaela Paul. Modeling seasonality in space-time infectious disease surveillance data. *Biometrical Journal*, 54(6):824–843, 2012.

Leonhard Held, Michael Höhle, and Mathias Hofmann. A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling*, 5(3):187–199, 2005. ISSN 1471-082X. doi: 10.1191/1471082X05st098oa.

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.

Michael Höhle, Sebastian Meyer, and Michaela Paul. *surveillance: Temporal and Spatio-Temporal Modeling and Monitoring of Epidemic Phenomena*, 2016. URL https://CRAN.R-project.org/package=surveillance. R package version 1.12.1.

Rob J Hyndman. *forecast: Forecasting functions for time series and linear models*, 2015. URL http://github.com/robjhyndman/forecast. R package version 6.2.

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.

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

Yuichiro Kanazawa. Hellinger distance and Kullback-Leibler loss for the kernel density estimator. *Statistics & Probability Letters*, 18(4):315–321, 1993.

Qi Li and Jeff Racine. Nonparametric estimation of distributions with categorical and continuous data. *Journal of Multivariate Analysis*, 86(2):266–292, 2003.

Qi Li and Jeffrey S Racine. Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. *Journal of Business & Economic Statistics*, 26(4): 423–434, 2008.

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

David Manheim, Margaret Chamberlin, Osonde A Osoba, Ragffaele Vardavas, and Melinda Moore. *Improving Decision Support for Infectious Disease Prevention and Control: Aligning Models and Other Tools with Policymakers' Needs*. RAND Corporation, 2016.

Roger B Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

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.

Pandemic Prediction and Forecasting Science and Technology Interagency Working Group. Dengue Forecasting, July 2015. http://dengueforecasting.noaa.gov/.

Andrew J Patton. A review of copula models for economic time series. *Journal of Multivariate Analysis*, 110:4–18, 2012.

Michaela Paul, Leonhard Held, and Andr Michael Toschke. Multivariate modelling of infectious disease surveillance data. *Statistics in Medicine*, 27(29):6250–6267, 2008. ISSN 0277-6715. doi: 10.1002/sim.3440.

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/.

Jeff Racine, Qi Li, and Xi Zhu. Kernel estimation of multivariate conditional distributions. *Annals of Economics and Finance*, 5(2):211–235, 2004.

Evan Ray, Krzysztof Sakrejda, Stephen A. Lauer, and Nicholas G. Reich. The Reich Lab at UMass-Amherst, May 2016. https://github.com/reichlab/article-disease-pred-with-kcde.

Eugene F Schuster and Gavin G Gregory. On the nonconsistency of maximum likelihood nonparametric density estimators. In *Computer Science and Statistics: Proceedings of the 13th Symposium on the interface*, pages 295–298. Springer, 1981.

David W Scott and Lynette E Factor. Monte Carlo study of three data-based nonparametric probability density estimators. *Journal of the American Statistical Association*, 76(373):9–15, 1981.

Nate Silver. We're predicting the career of every NBA player. Here's how., October 2015. http://fivethirtyeight.com/features/how-were-predicting-nba-player-career/.

George Sugihara and Robert M. May. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature*, 344:734–741, April 1990.

Steffen Unkel, C Farrington, Paul H Garthwaite, Chris Robertson, and Nick Andrews. Statistical methods for the prospective detection of infectious disease outbreaks: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175(1):49–82, 2012.

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.

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.

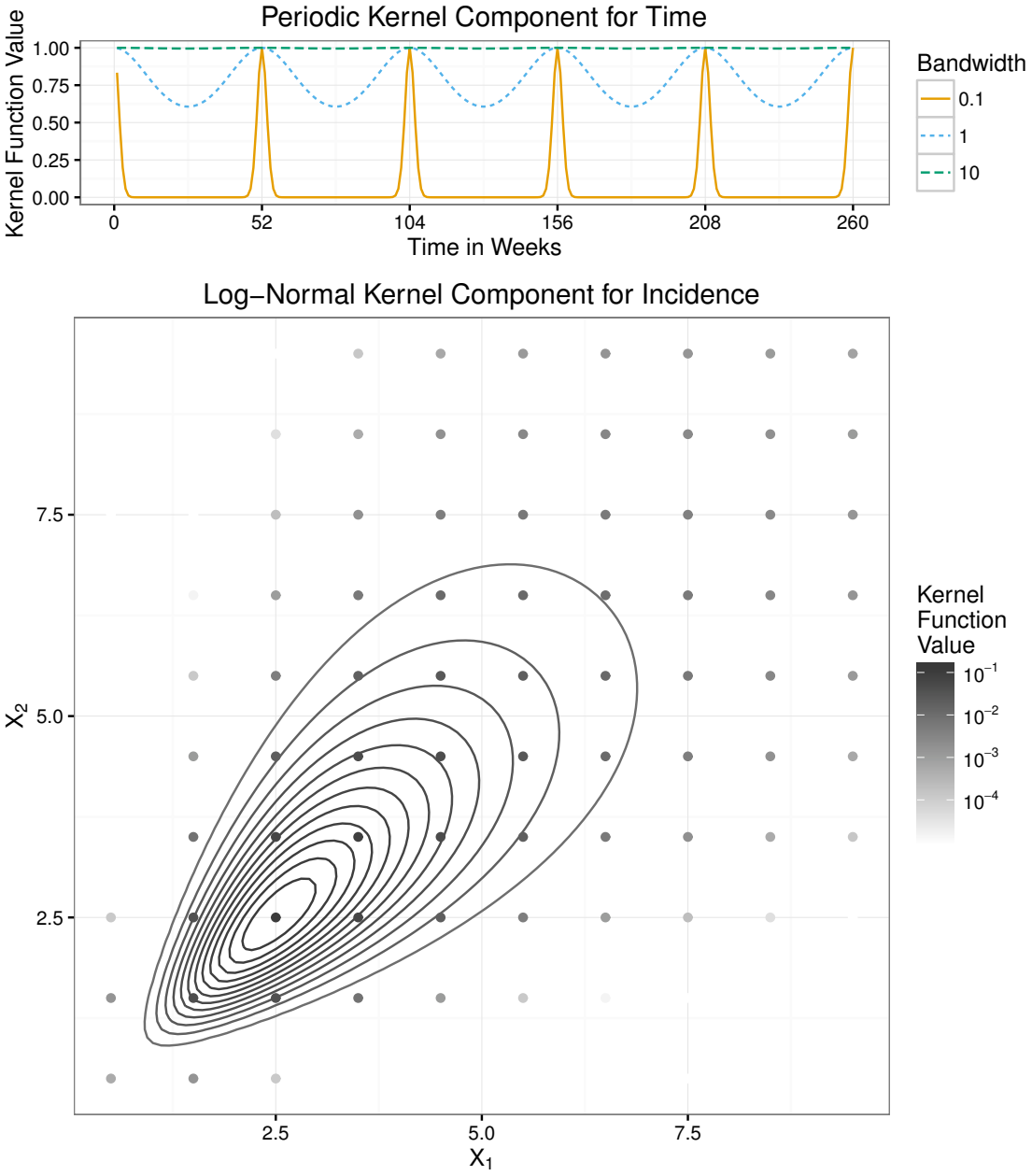

Figure 1: The components of the kernel function. The top panel shows the periodic kernel function illustrated as a function of time in weeks with $\rho = \pi/52$ and three possible values for the bandwidth parameter η. The lower panel shows the log-normal kernel function in the bivariate case. The curves indicate contours of the continuous kernel function and the points indicate the discrete kernel function, which is obtained by integrating the continuous kernel function. The kernel is centered at $(2.5, 2.5)$ and has bandwidth matrix $\begin{bmatrix} 0.2 & 0.15 \\ 0.15 & 0.2 \end{bmatrix}$.

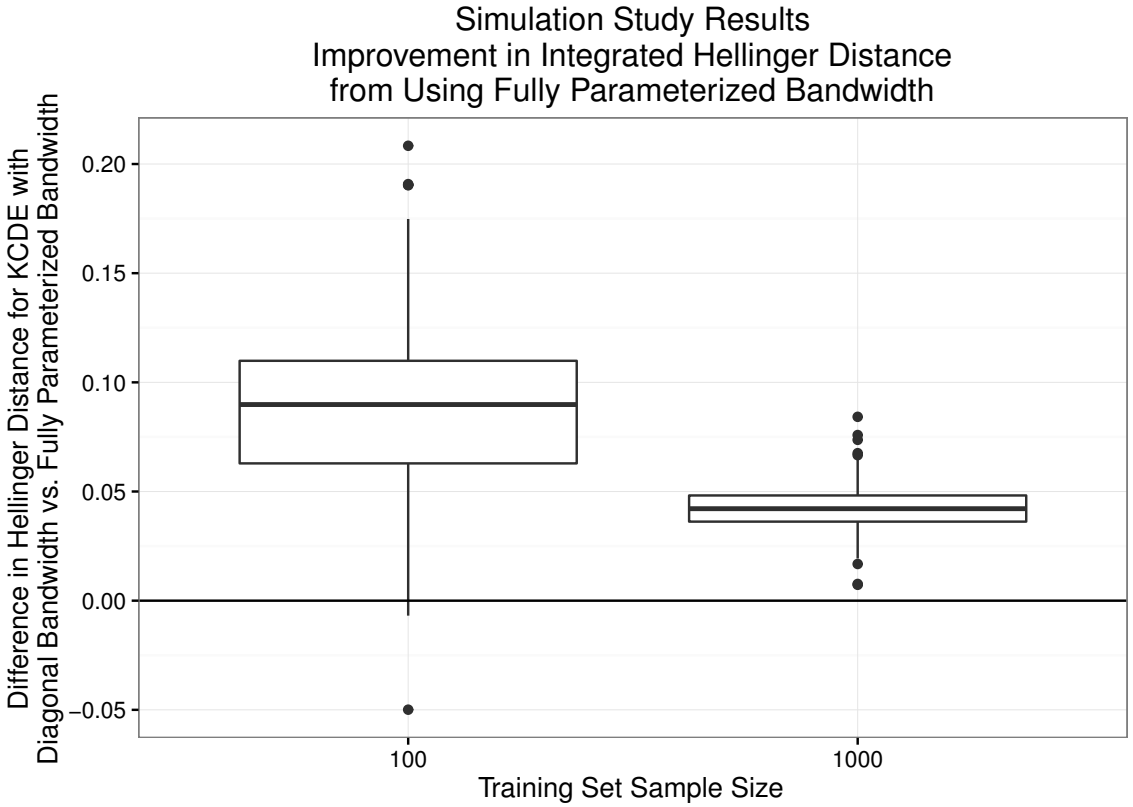

Figure 2: Box plots of results from the simulation study. Positive values indicate simulation trials where the full bandwidth specification outperformed the diagonal bandwidth specification with the same training data set, as measured by Hellinger distance from the target conditional density.

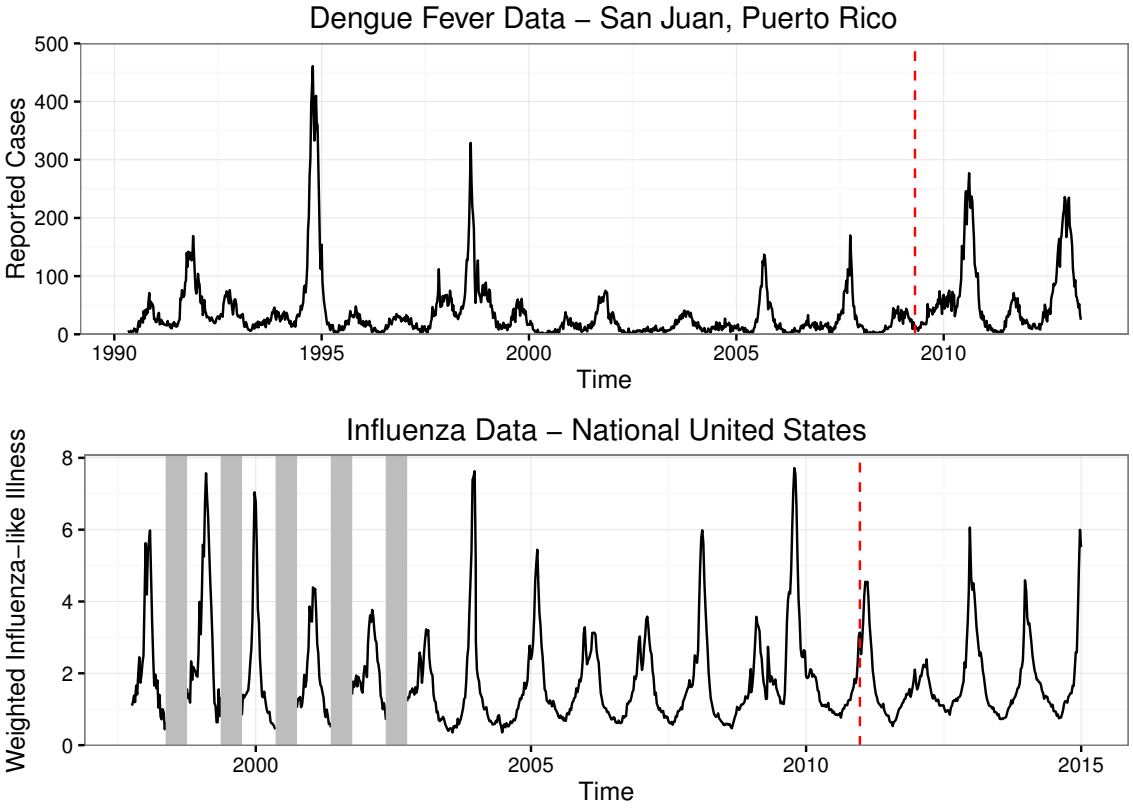

Figure 3: Plots of the data sets we apply our methods to. In each case, the last four years of data are held out as a test data set; this cutoff is indicated with a vertical dashed line. For the flu data set, low-season incidence was not recorded in early years of data collection. These missing data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

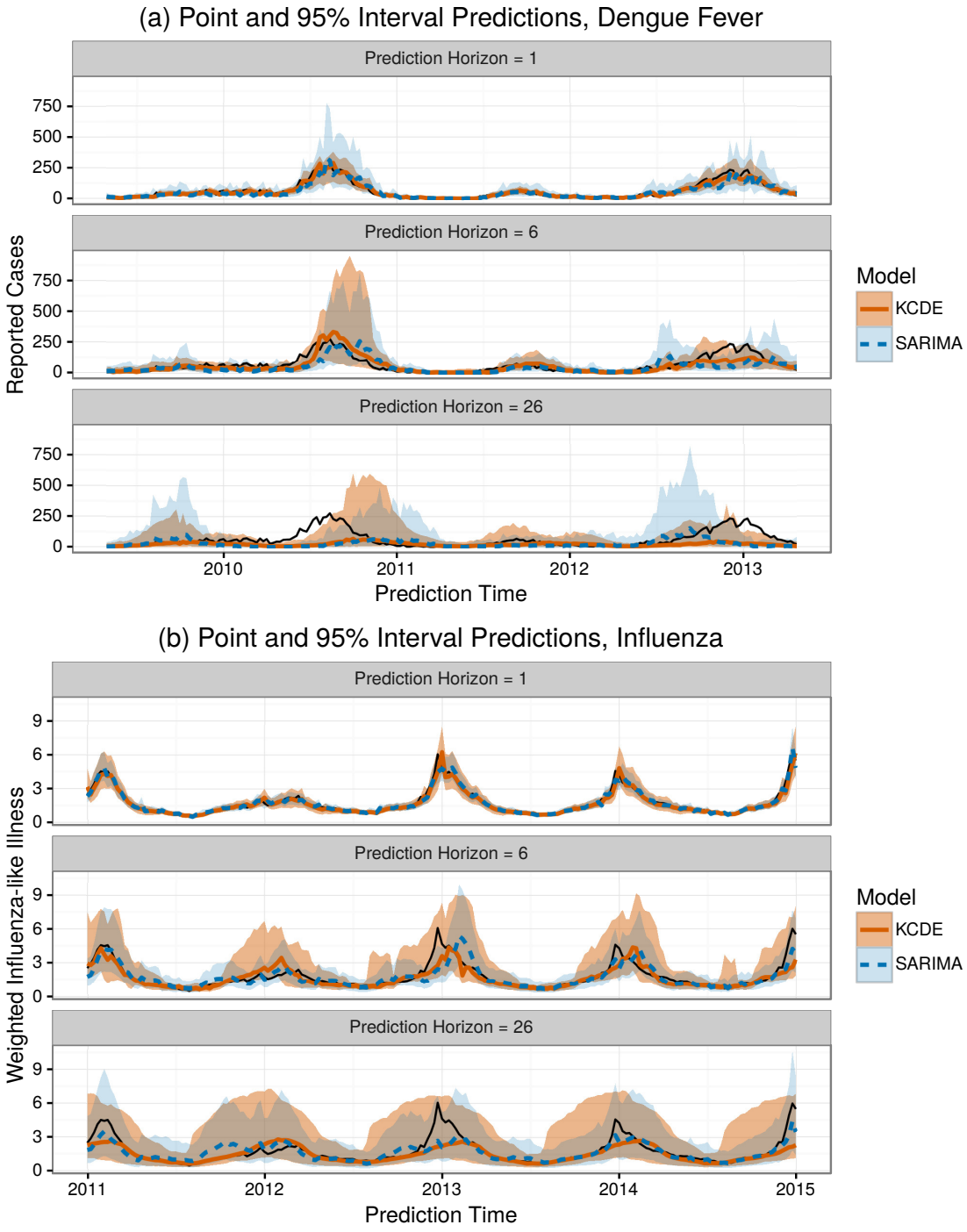

Figure 4: Plots of point and interval predictions from SARIMA and the Periodic, Full Bandwidth KCDE model.

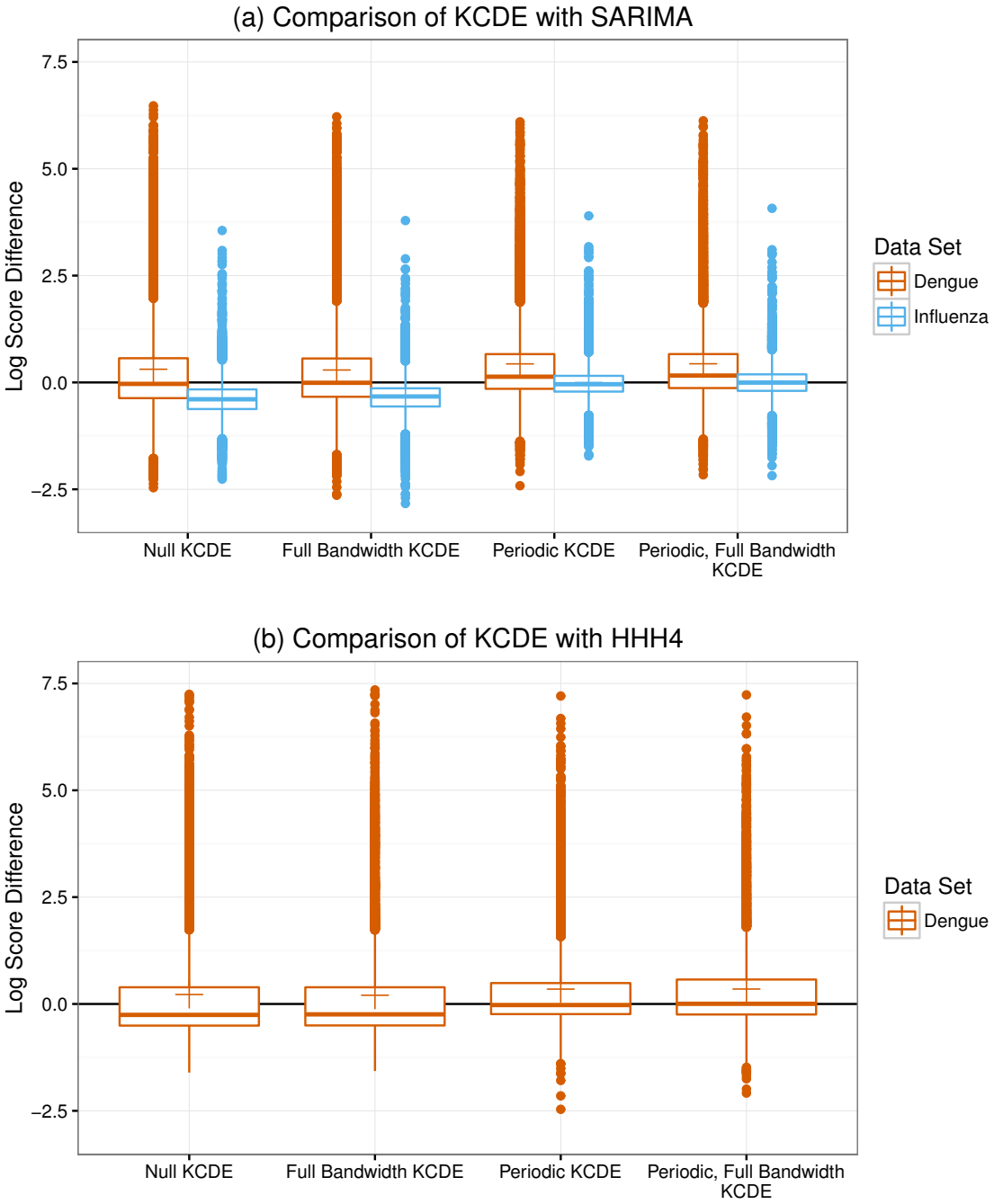

Figure 5: Differences in log scores for the weekly predictive distributions obtained from KCDE specifications and the baseline models. For reference, a log score difference of 2.3 (4.6) indicates that the predictive density from KCDE was about 10 (100) times as large as the predictive density from the baseline model at the realized outcome. The boxplots summarize the results across all combinations of prediction horizon and prediction time in the test period.

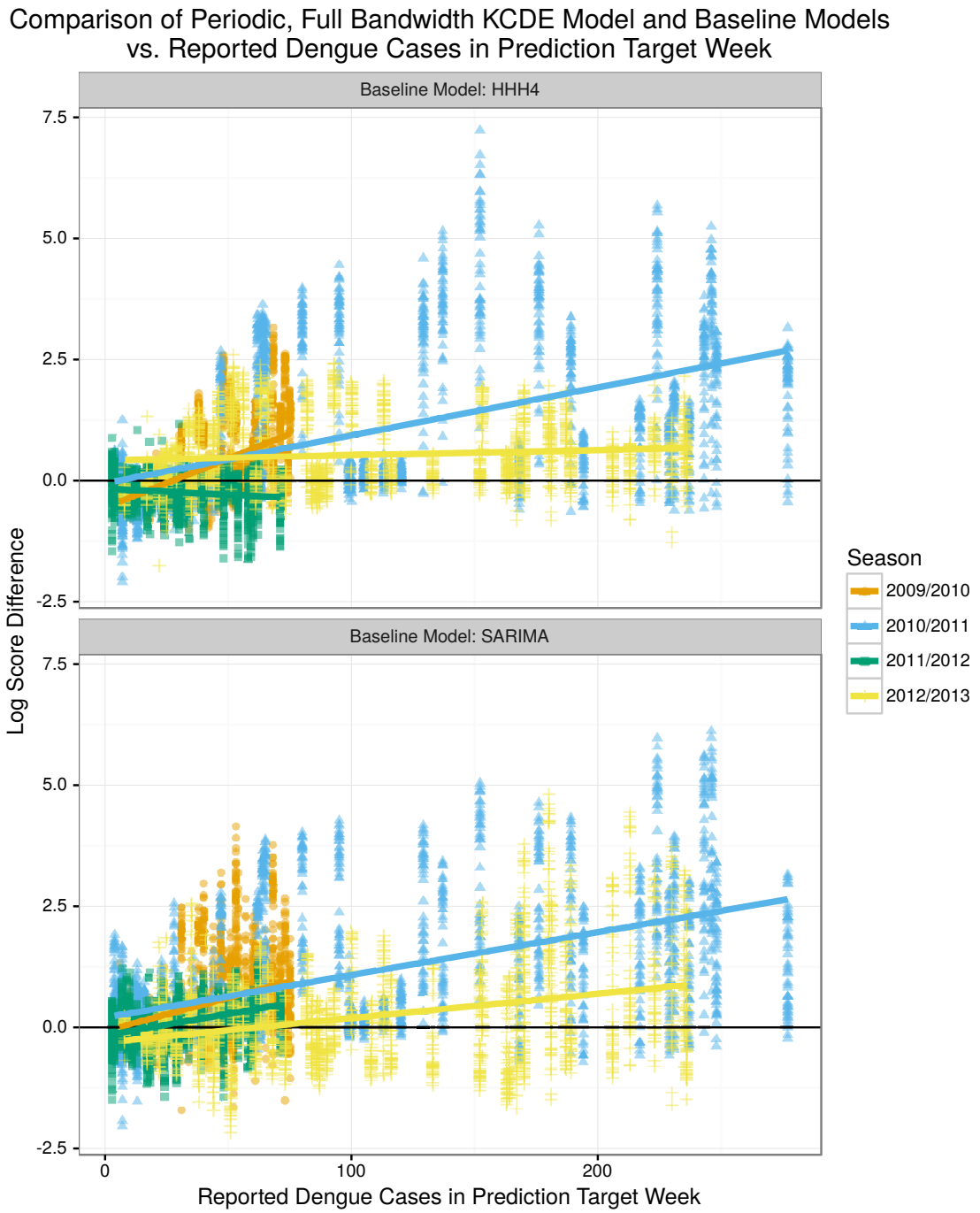

Figure 6: Differences in log scores for the weekly predictive distributions obtained from the Periodic, Full Bandwidth KCDE model and the baseline models, plotted against the observed incidence in the week being predicted. For reference, a log score difference of 2.3 (4.6) indicates that the predictive density from KCDE was about 10 (100) times as large as the predictive density from the baseline model at the realized outcome. Each point corresponds to a unique combination of prediction target week and prediction horizon.

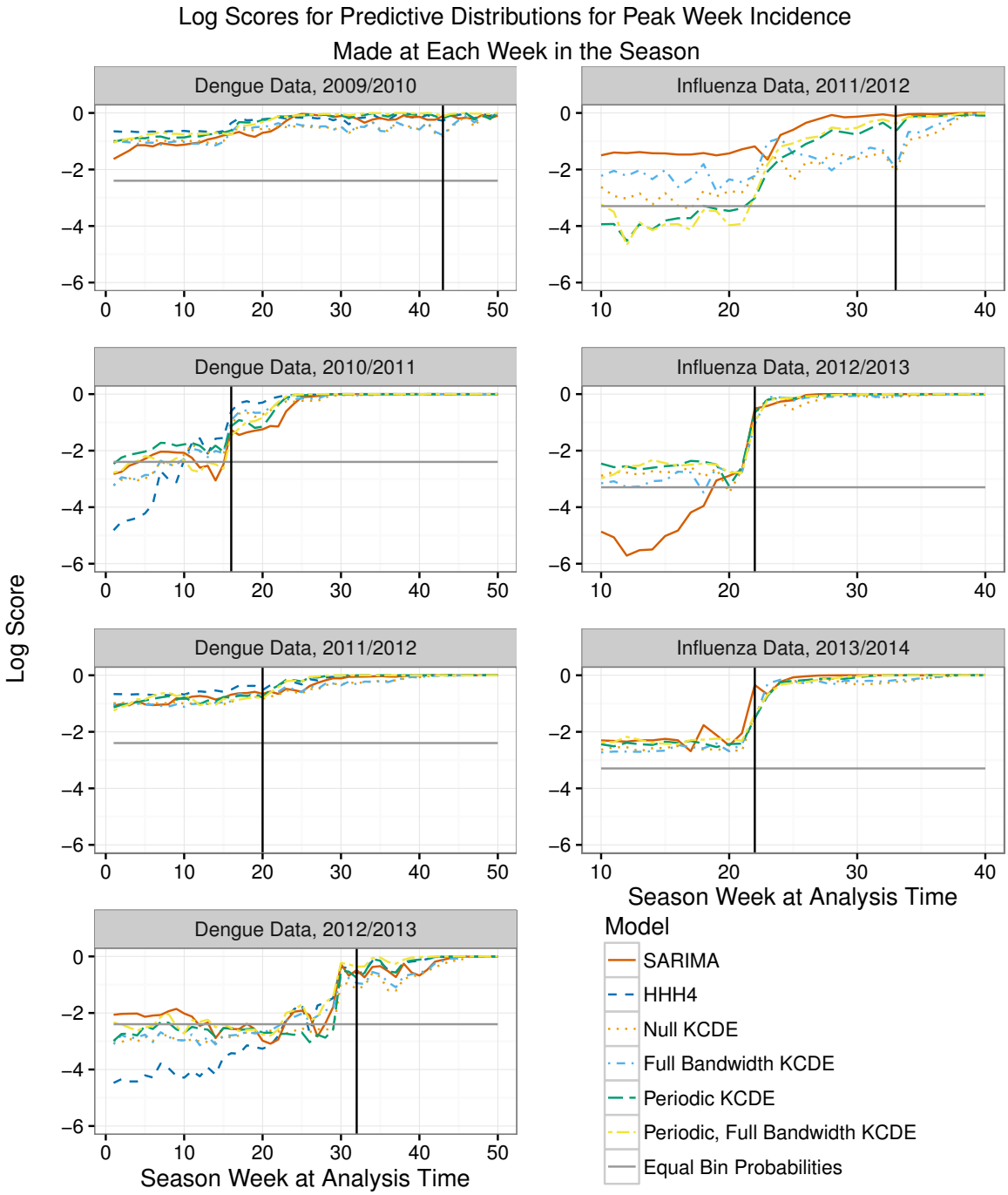

Figure 7: Log scores for predictions of peak week incidence by predictive model and analysis time. The vertical line is placed at the peak week for each season. The log score for "Equal Bin Probabilities" is obtained by assigning equal probability that the peak incidence will be in each of the specified incidence bins. There are 11 incidence bins for dengue and 27 bins for influenza.