Biostatistics (2016), **0**, 0, *pp.* 1–31
doi:10.1093/biostatistics/infectious-disease-prediction-with-kcde

Infectious disease prediction with kernel conditional density estimation

EVAN L. RAY*, KRZYSZTOF SAKREJDA, STEPHEN A. LAUER, NICHOLAS G. REICH
Department of Biostatistics and Epidemiology, School of Public Health and Health Sciences,

415 Arnold House, 715 N. Pleasant Street, Amherst, MA 01003, USA

elray@umass.edu

Summary

We develop a novel approach to prediction of infectious disease incidence using kernel conditional density estimation (KCDE). This method obtains predictive distributions for incidence in individual weeks and uses copulas to tie those distributions together into joint distributions in order to make 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 is appropriate for use with discrete variables but also allows for a full parameterization of the bandwidth matrix. A simulation study demonstrates that allowing for a fully parameterized bandwidth matrix can improve conditional density estimates. In applications to predicting dengue fever and influenza, our method outperforms a baseline seasonal autoregressive integrated moving average (SARIMA) model for predictions of dengue incidence in individual weeks, and is comparable to the SARIMA model on the other prediction targets. The periodic kernel function leads to improved predictions of incidence in both applications. Our approach and extensions of it could yield improved predictions

*To whom correspondence should be addressed.

for public health decision makers, particularly in diseases with heterogeneous seasonal dynamics such as dengue fever.

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

1. 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 (Wallinga *and others*, 2010; Hatchett *and others*, 2007). 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 (Pandemic Prediction and Forecasting Science and Technology Interagency Working Group, 2015; Epidemic Prediction Initiative, 2016). 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. We 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 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.

At its heart, KCDE is a local method in the sense that the conditional density estimate for future incidence given conditioning variables is a weighted combination of contributions from previous observations of incidence with similar conditioning values. Using such local methods is a natural idea in predicting nonlinear dynamical systems. For example, in the infectious disease literature nearest neighbors regression has been used to make point predictions for incidence of measles (Sugihara and May, 1990) and influenza (Viboud *and others*, 2003). 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 *and others*, 2009 discuss the connection between nearest neighbors and kernel methods for regression). However, KCDE offers the advantage of providing a complete predictive distribution rather than only a point prediction. 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 *and others*, 2004), female labor force participation (Hall *and others*, 2004), bond yields and value at risk in financial markets (Fan and Yim, 2004), and wind power (Jeon and Taylor, 2012) among others. Methods similar to those we explore in this article can also be formulated in the Bayesian framework. One example along these lines is Zhou *and others* (2015), who model the time to arrival of a disease in amphibian populations using Dirichlet processes and copulas.

To our knowledge, previous implementations of kernel methods for estimating multivariate densities or regression functions 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 *and others*, 2004, 2007; Li and Racine, 2003, 2008; Ouyang *and others*, 2006; Racine *and others*, 2004). 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 (Duong and Hazelton, 2005). We introduce an approach to allowing for discrete kernels with orientation by discretizing an underlying continuous kernel function.

A limitation of local methods such as KCDE is that their performance may not scale well with the dimension of the vector whose distribution is being estimated (Hastie *and others*, 2009). This is particularly relevant in our application, where we wish 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 KCDE. Methods combining nonparametric estimates of marginal densities with copulas have been considered for other applications before. For example, Patton (2012) is a recent review of copula methods for economic and financial time series including several articles that have used nonparametric estimates for the marginal distributions. Our approach differs from these in the details of how the nonparametric marginal distribution estimates are obtained.

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 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.

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, $w^* \in \{1, \ldots, W\}$:

$$\begin{aligned} P(\text{Peak Week} = w^*) &= P(Z_{S_{t^*}+w^*} = \max_w Z_{S_{t^*}+w}|t^*, \mathbf{z}_{1:t^*}) \\ &= \int_{\{\mathbf{z}_{(t^*+1):(t^*+H_{t^*})}: z_{S_{t^*}+w^*} = \max_w z_{S_{t^*}+w}\}} f(\mathbf{z}_{(t^*+1):(t^*+H_{t^*})}|t^*, \mathbf{z}_{1:t^*})\, d\mathbf{z}_{(t^*+1):(t^*+H_{t^*})}. \end{aligned} \tag{2.1}$$

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

$$\begin{aligned} P(\text{Incidence in Peak Week} \in [a, b]) &= P(a \leqslant \max_w Z_{S_{t^*}+w} \leqslant b|t^*, \mathbf{z}_{1:t^*}) \\ &= \int_{\{(\mathbf{z}_{(t^*+1):(t^*+H_{t^*})}): a \leqslant \max_w z_{S_{t^*}+w} \leqslant b\}} f(\mathbf{z}_{(t^*+1):(t^*+H_{t^*})}|t^*, \mathbf{z}_{1:t^*})\, d\mathbf{z}_{(t^*+1):(t^*+H_{t^*})}. \end{aligned} \tag{2.2}$$

In practice, we use Monte Carlo integration to evaluate the integrals in Equations (2.1) and (2.2) by sampling incidence trajectories from the joint predictive distribution.

We will introduce the overall structure of our model here and describe its components 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:

$$\begin{aligned} &f(z_{(t^*+1):(t^*+H_{t^*})}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M}) = \\ &\quad c^{H_{t^*}}\{f^1(z_{t^*+1}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M}; \boldsymbol{\theta}^1), \ldots, f^{H_{t^*}}(z_{t^*+H_{t^*}}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M}; \boldsymbol{\theta}^{H_{t^*}}); \boldsymbol{\xi}^{H_{t^*}}\}. \end{aligned} \tag{2.3}$$

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.

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. 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). 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.

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^*. 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 $t = 1 + l_M, \ldots, T - h$ (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:

$$\hat{f}^h(y^h_{t^*}|\mathbf{x}_{t^*}) = \frac{\sum_{t\in\boldsymbol{\tau}} K^{\mathbf{X},Y}\left\{(\mathbf{x}_{t^*}, y^h_{t^*}), (\mathbf{x}_t, y^h_t); \boldsymbol{\theta}^h\right\}}{\sum_{t\in\boldsymbol{\tau}} K^{\mathbf{X}}(\mathbf{x}_{t^*}, \mathbf{x}_t; \boldsymbol{\theta}^h)} \tag{2.4}$$

$$= \sum_{t\in\boldsymbol{\tau}} \zeta^h_{t^*,t} K^{Y|\mathbf{X}}(y^h_{t^*}, y^h_t|\mathbf{x}_{t^*}, \mathbf{x}_t; \boldsymbol{\theta}^h), \text{ where} \tag{2.5}$$

$$\zeta^h_{t^*,t} = \frac{K^{\mathbf{X}}(\mathbf{x}_{t^*}, \mathbf{x}_t; \boldsymbol{\theta}^h)}{\sum_{s\in\boldsymbol{\tau}} K^{\mathbf{X}}(\mathbf{x}_{t^*}, \mathbf{x}_s; \boldsymbol{\theta}^h)}. \tag{2.6}$$

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. For each y^h_t, $K^{Y|\mathbf{X}}$ 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 (2.4) through (2.6), $\boldsymbol{\tau} \subseteq \{(1+l_M), \ldots, (T-h)\}$ 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

$$K^{\text{per}}(t^*, t; \rho^h, \eta^h) = \exp\left[-\frac{\sin^2\{\rho^h(t^*-t)\}}{2(\eta^h)^2}\right]. \tag{2.7}$$

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 contribution 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:

$$K^{\text{inc}}_{\text{cont}}(\tilde{\mathbf{z}}_{t^*}, \tilde{\mathbf{z}}_t; \mathbf{B}) = \frac{\exp\left[-\frac{1}{2}\{\log(\tilde{\mathbf{z}}_{t^*}) - \log(\tilde{\mathbf{z}}_t) - \mathbf{B}\underline{1}\}'\mathbf{B}^{-1}\{\log(\tilde{\mathbf{z}}_{t^*}) - \log(\tilde{\mathbf{z}}_t) - \mathbf{B}\underline{1}\}\right]}{(2\pi)^{\frac{M+1}{2}}|\mathbf{B}|^{\frac{1}{2}} z_{t^*+h} \prod_{m=1}^{M} z_{t^*-l_m}} \tag{2.8}$$

In this expression, the log operator applied to a vector takes the log of each component of that vector and $\underline{1}$ 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. In order 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 (2.4) through (2.6). 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 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. 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.

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 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 Section 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:

$$F^{H_{t^*}}(y^1_{t^*},\ldots,y^{H_{t^*}}_{t^*}|\mathbf{x}_{t^*};\boldsymbol{\theta}^1,\ldots,\boldsymbol{\theta}^{H_{t^*}},\boldsymbol{\xi}^{H_{t^*}}) = \\ C\{F^1(y^1_{t^*}|\mathbf{x}_{t^*};\boldsymbol{\theta}^1),\ldots,F^{H_{t^*}}(y^{H_{t^*}}_{t^*}|\mathbf{x}_{t^*};\boldsymbol{\theta}^{H_{t^*}});\boldsymbol{\xi}^{H_{t^*}}\} \tag{2.9}$$

The copula function C maps the marginal CDF values to the joint CDF value. We use the isotropic normal copula implemented in the `R` package `copula` (Hofert *and others*, 2015). The copula function is given by

$$C(u_1,\ldots,u_H;\boldsymbol{\xi}^H) = \Phi_{\Sigma^H}(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_H)), \tag{2.10}$$

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

$$\sigma^H_{i,j} = \begin{cases} 1 \text{ if } i = j, \\ \xi^H_d \text{ if } |i-j| = d \end{cases} \tag{2.11}$$

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 $H = 1$). In order to do this, we follow a two-stage estimation strategy (Joe, 2005):

1. Estimate the parameters for marginal predictive distributions using the procedures described in the previous subsection.

2. Estimate the copula parameters, holding the parameters for the marginal predictive distributions fixed:

   (a) 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:

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

   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 steps before the end of the season.

   (b) Estimate the copula parameters $\boldsymbol{\xi}^H$ by maximizing the likelihood of the pseudo-observations.

3. Simulation Study

We conducted a simulation study to examine the utility of using a non-diagonal bandwidth matrix specification when estimating conditional distributions with KCDE. 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 hold other factors that may be related to the relative performance of different bandwidth specifications fixed.

The distributions that we simulate from are discretized bivariate normal distributions. To define this distribution, let $\mathbf{U} \sim MVN(\underline{0}, \Sigma)$ where Σ is a 2×2 matrix with 1 on the diagonal

and 0.9 off of the diagonal. We treat $\mathbf{U}$ as a latent variable and discretize it to obtain the random variable $\mathbf{X}$ using the approach described in the supplement.

We conducted 500 simulation trials for each sample size. In each trial, we simulated N observations of the discretized bivariate normal random variable $\mathbf{X}$. 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 our 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 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.

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: one with a weekly measure of incidence of dengue fever in San Juan, Puerto Rico, and a second with a weekly measure of incidence of 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). In the dengue data set, the incidence measure is an integer number of reported cases in the given week. In the influenza data set the incidence measure is continuous, a weighted proportion of doctor visits with influenza-like illness.

Figure 3 displays each time series. 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. The length of the testing set in the dengue data was set by the competition administrators, and we have adopted that convention of four years for use with the influenza data as well. All predictions are made as though in real time assuming that once cases are reported, they are never revised. Specifically, we use only data up through a given week in order to make predictions for incidence after that week.

There are three prediction targets for each data set, based closely on the prediction targets that were used in the original 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 week of the test data set, we make predictions for the timing of the peak week of the corresponding season. Third, in each week of the test data set we predict incidence in the peak week for the corresponding season. 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. These prediction targets are illustrated in the supplement.

Our applications include 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 use a seasonal autoregressive integrated moving average (SARIMA) model for the log-transformed incidence measure as a baseline to compare our methods to. We performed first-order seasonal differencing to obtain a roughly stationary time series and then 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. Further details of the SARIMA model specification and estimation procedure are in the supplement.

We begin with a discussion of predictive distributions for incidence at individual time points. Figure 4 displays the median and 95% interval limits for the predictive distributions obtained at prediction horizons of 1, 6, or 26 weeks from SARIMA and from the KCDE specification with a

fully parameterized bandwidth matrix and a periodic kernel component. For predictions of dengue fever incidence, at a prediction horizon of one week the point predictions from SARIMA and KCDE are similar, but the distribution from KCDE is much more concentrated around its center. Both methods struggle with larger prediction horizons, but it appears that the SARIMA model has more difficulty with aligning the predictive distribution with the season's peak, particularly in the two seasons with higher incidence. For the predictions of influenza incidence, which exhibits more regular seasonality, there is much less of a noticeable distinction between the predictions given by the two methods. Throughout, the point predictions and intervals are similar.

Figure 5 offers a more quantitative summary of these results in terms of log scores. In the application to predicting dengue fever, the KCDE specifications including a periodic kernel component had slightly larger log scores than the SARIMA model on average, with many outlying cases where KCDE did much better than SARIMA. For the specifications without the periodic kernel component, the median performance was similar to the SARIMA model, but again there were outlying cases where KCDE did much better than SARIMA. Figure 5 (b) shows that KCDE particularly outperformed SARIMA in its predictions for times of high incidence near the season peaks. In weeks with fewer than 93 reported cases (roughly one third of the maximum weekly case count in the testing period), the average log score difference between the predictions from the Periodic, Full Bandwidth KCDE model and the SARIMA model was about 0.25 with a standard deviation of 0.73. In weeks with more than 184 reported cases, in the upper third, the mean difference in log scores was about 1.59 with a standard deviation of 1.6. Translating to a probability scale, in these periods of high incidence the predictive distributions from KCDE assigned about 5 times as much probability to the observed outcome as the predictive distributions from SARIMA on average, with outlying cases where KCDE assigned up to about 450 times as much probability to the realized outcome.

In the application to predicting influenza, the KCDE specifications including a periodic kernel

did about as well as the SARIMA model, while the median performance of the specifications without a periodic kernel was slightly worse than SARIMA. In both applications, including the periodic kernel component led to improved predictions; this can be seen directly in Figure *** of the supplement. Using the fully parameterized bandwidth matrix generally had little impact on the quality of the predictive distributions as measured by the log score.

Figure 6 displays the log score of the predictive distributions for incidence in the peak week obtained from SARIMA and KCDE models over the course of each season in the test data sets, and Figure *** in the supplement displays log scores for predictions of peak week timing. For both of these prediction targets, there is no consistent pattern of KCDE either outperforming or underperforming relative to SARIMA.

5. Conclusions

Prediction of infectious disease incidence at horizons of more than a few weeks is a challenging task. We have presented a semiparametric 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, we saw that our approach offered consistent and substantial performance gains relative to a SARIMA 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. For predicting influenza-like illness, our method did about as well as SARIMA when predicting incidence in individual weeks.

We believe that the difference in relative performance of KCDE and SARIMA 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 differing model specifications. The most salient difference between the two time series depicted in Figure 3 is the much greater season-to-

season variablility in the dengue data set relative to the influenza data set. 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 model means that it is not able to capture the dynamics of dengue incidence accurately. Relaxing that structure by using a nonparametric 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.

Another more subtle effect is present in the influenza data: there is a consistent short-term peak in influenza incidence on Christmas week. This is visible in Figure 3, and is highlighted in Figure *** in the supplement. This "Christmas effect" sometimes coincides with the season peak, but sometimes occurs before the season peak. We have observed evidence that the seasonal structure of the SARIMA model picks up on this structure, and SARIMA tended to outperform KCDE on Christmas week and the weeks immediately thereafter on the influenza data set. We believe that it would be possible to construct a variation on KCDE that captures this effect, for example by including indicator variables for the weeks around Christmas as conditioning variables. However, we have not explored that avenue in this work.

We have also demonstrated that it is feasible to use KCDE in combination with copulas to obtain predictions for the timing of and incidence in the week of the season with the highest incidence. For those prediction targets, our method was competitive with SARIMA; we did better in some of the test seasons and worse in others, with no clear indication that either model was better than the other.

One explanation for the difference in relative performance of the methods on these different prediction tasks may lie in the connection between the objective function used in parameter estimation and the prediction task. We estimated the bandwidth parameters for KCDE by optimizing the log score of predictive distributions for incidence in individual weeks in the training data set.

This is the prediction target where KCDE outperformed SARIMA on the dengue data set. It may be the case that performance on the other two prediction tasks could be improved by implementing a combined one-stage estimation strategy for both the KCDE and copula parameters that optimizes a measure of performance on the specific prediction task at hand.

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.

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 *and others* (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. In particular, we hypothesize that a penalty on the inverse of the bandwidth matrix encouraging it to have small eigenvalues could be helpful. If successful, this would also enable further exploration of using other predictive variables (such as weather) 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 in order 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. A full discussion of those methods is beyond the scope of this article; see Brown *and others* for a 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.

However, rather than selecting one "preferred" modeling framework or model formulation, we believe it may be fruitful to incorporate the models developed in this paper as components of

an ensemble with several different types of models. For example, in our application to influenza, we saw that the SARIMA model captured some features of the data generating process, such as the Christmas-week effect, that KCDE did not capture. On the other hand, the KCDE approach was more flexible and yielded better predictions than SARIMA at other times – most notably, in periods of high incidence in the application to dengue. An appropriately constructed ensemble incorporating predictions from both SARIMA and KCDE as well as other methods such as mechanistic models might perform better than any of these models on its 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` (R Core Team, 2016) and `C`. All source code as well as the data we used in the applications are available in `R` packages hosted on GitHub (Ray *and others*, 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 Michael Johansson for helpful commentary on a draft of the paper and the competition administrators for making disease incidence data available. This work was funded in part by grants *******.

References

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

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

Brown, Alexandria, Lauer, Stephen A., Ray, Evan L., Meng, Xi and Reich, Nicholas G. A systematic review of prediction for infectious disease. In preparation.

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

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

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

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

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

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

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

Hall, Peter, Racine, Jeff and Li, Qi. (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association* **99**(468), 1015–1026.

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

Hatchett, Richard J, Mecher, Carter E and Lipsitch, Marc. (2007). Public health interventions and epidemic intensity during the 1918 influenza pandemic. *Proceedings of the National Academy of Sciences* **104**(18), 7582–7587.

Hofert, Marius, Kojadinovic, Ivan, Maechler, Martin and Yan, Jun. (2015). *copula: Multivariate Dependence with Copulas*. R package version 0.999-14.

Hyndman, Rob J. (2015). *forecast: Forecasting functions for time series and linear models*. R package version 6.2.

Jeon, Jooyoung and Taylor, James W. (2012). Using conditional kernel density estimation for wind power density forecasting. *Journal of the American Statistical Association* **107**(497), 66–79.

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

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

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

Li, Qi and Racine, Jeffrey S. (2008). Nonparametric estimation of conditional CDF and

quantile functions with mixed categorical and continuous data. *Journal of Business & Economic Statistics* **26**(4), 423–434.

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

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

Ouyang, Desheng, Li, Qi and Racine, Jeffrey. (2006). Cross-validation and the estimation of probability distributions with categorical data. *Journal of Nonparametric Statistics* **18**(1), 69–100.

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

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

R Core Team. (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

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

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

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

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

Viboud, Cécile, Boëlle, Pierre-Yves, Carrat, Fabrice, Valleron, Alain-Jacques and Flahault, Antoine. (2003). Prediction of the spread of influenza epidemics by the method of analogues. *American Journal of Epidemiology* **158**(10), 996–1006.

Wallinga, Jacco, van Boven, Michiel and Lipsitch, Marc. (2010). Optimizing infectious disease interventions during an emerging epidemic. *Proceedings of the National Academy of Sciences* **107**(2), 923–928.

Zhou, Haiming, Hanson, Timothy and Knapp, Roland. (2015). Marginal Bayesian nonparametric model for time to disease arrival of threatened amphibian populations. *Biometrics* **71**(4), 1101–1110.

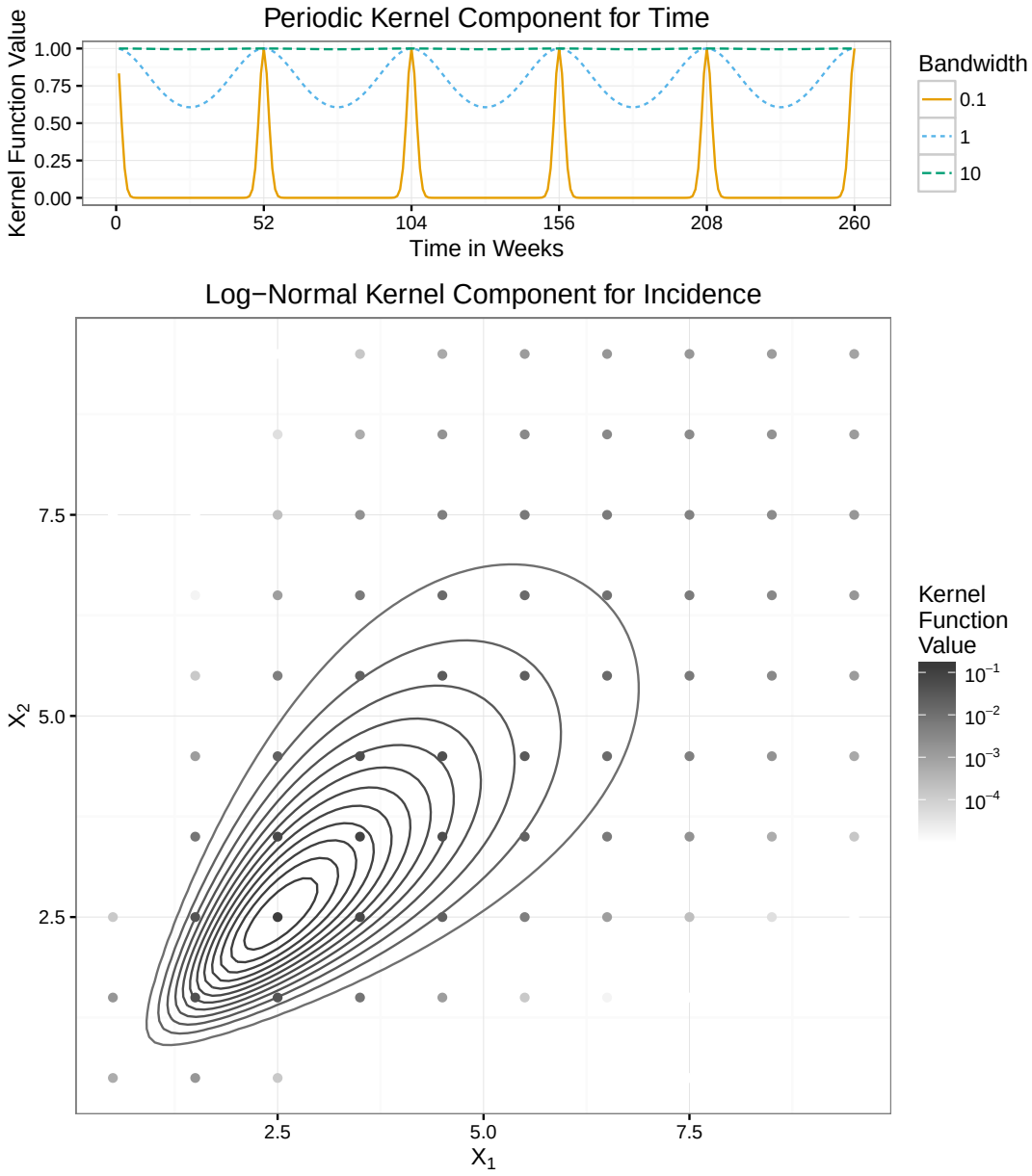

Fig. 1: The components of the kernel function. The top panel shows the periodic kernel function illustrated as a function of time in weeks with $\rho = \pi/52$ and three possible values for the bandwidth parameter η. The lower panel shows the log-normal kernel function in the bivariate case. The curves 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 $\begin{bmatrix} 0.2 & 0.15 \\ 0.15 & 0.2 \end{bmatrix}$.

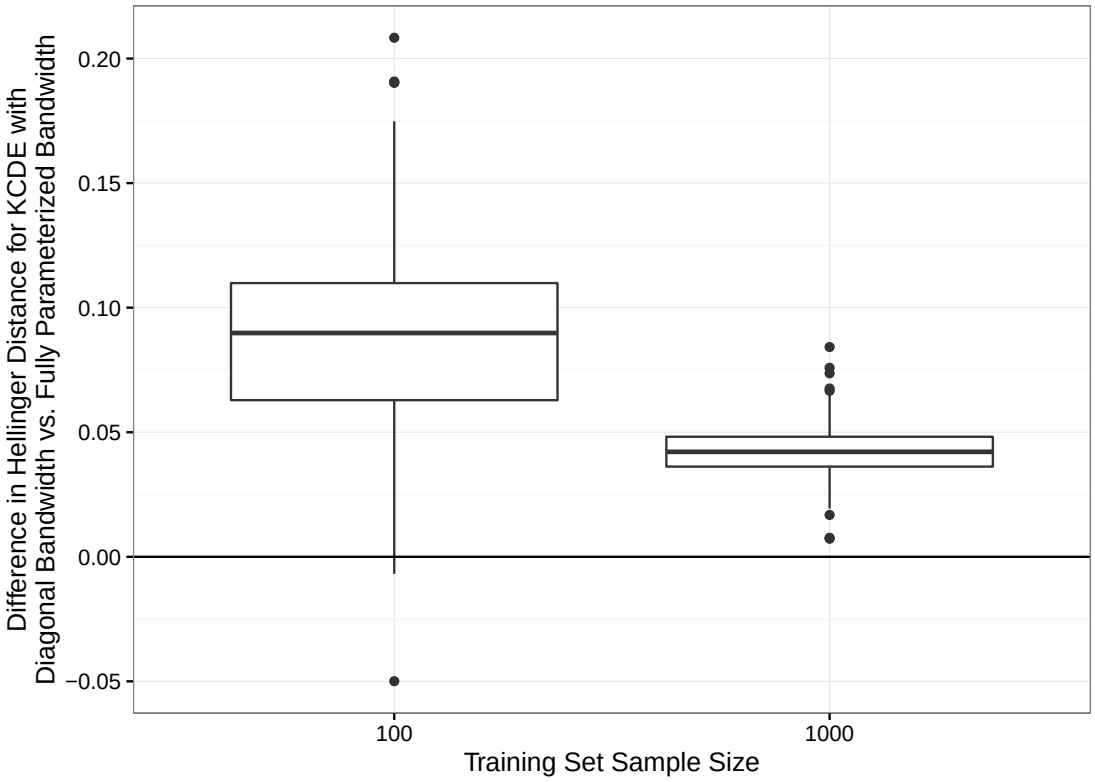

Fig. 2: Results from the simulation study. Positive values indicate simulation trials where the full bandwidth specification outperformed the diagonal bandwidth specification with the same training data set, as measured by Hellinger distance from the target conditional density.

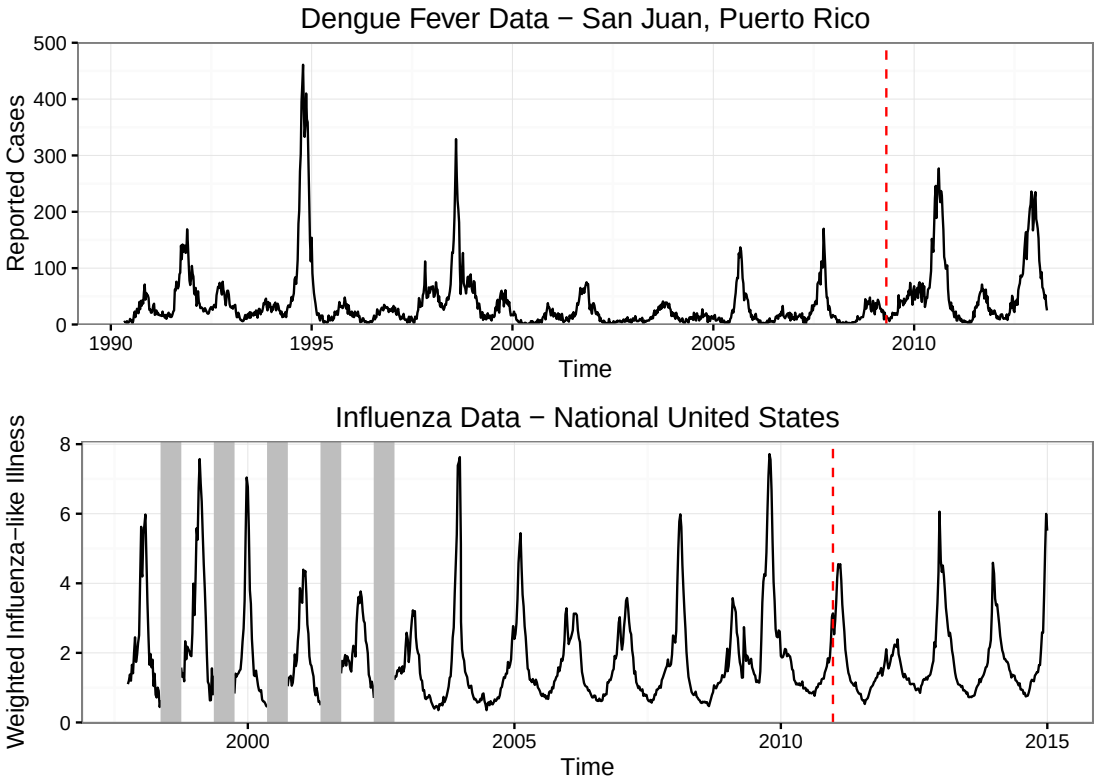

Fig. 3: Plots of the data sets we apply our methods to. In each case, the last four years of data are held out as a test data set; this cutoff is indicated with a vertical dashed line. For the flu data set, low-season incidence was not recorded in early years of data collection. These missing data are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

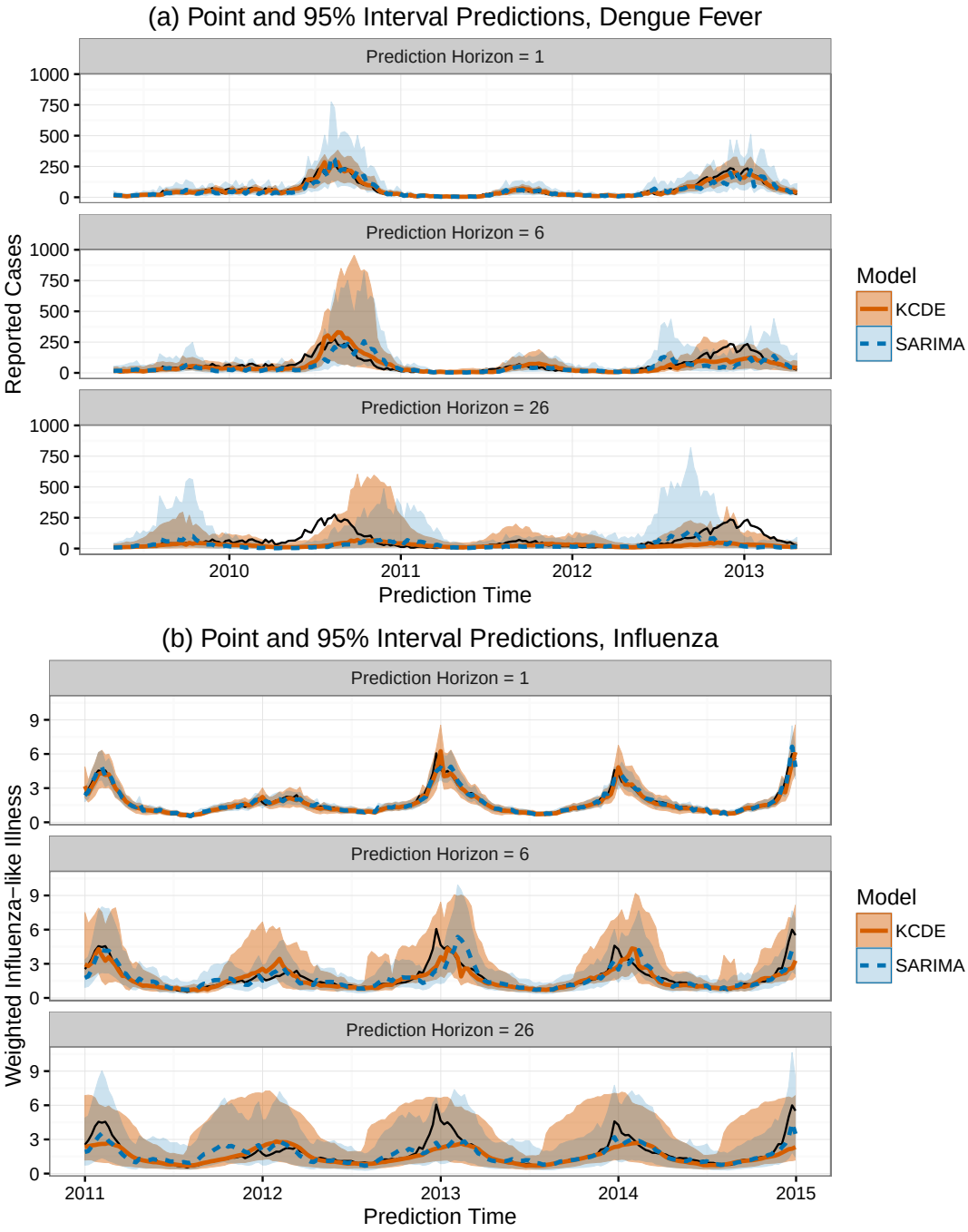

Fig. 4: Plots of point and interval predictions from SARIMA and the Periodic, Full Bandwidth KCDE model.

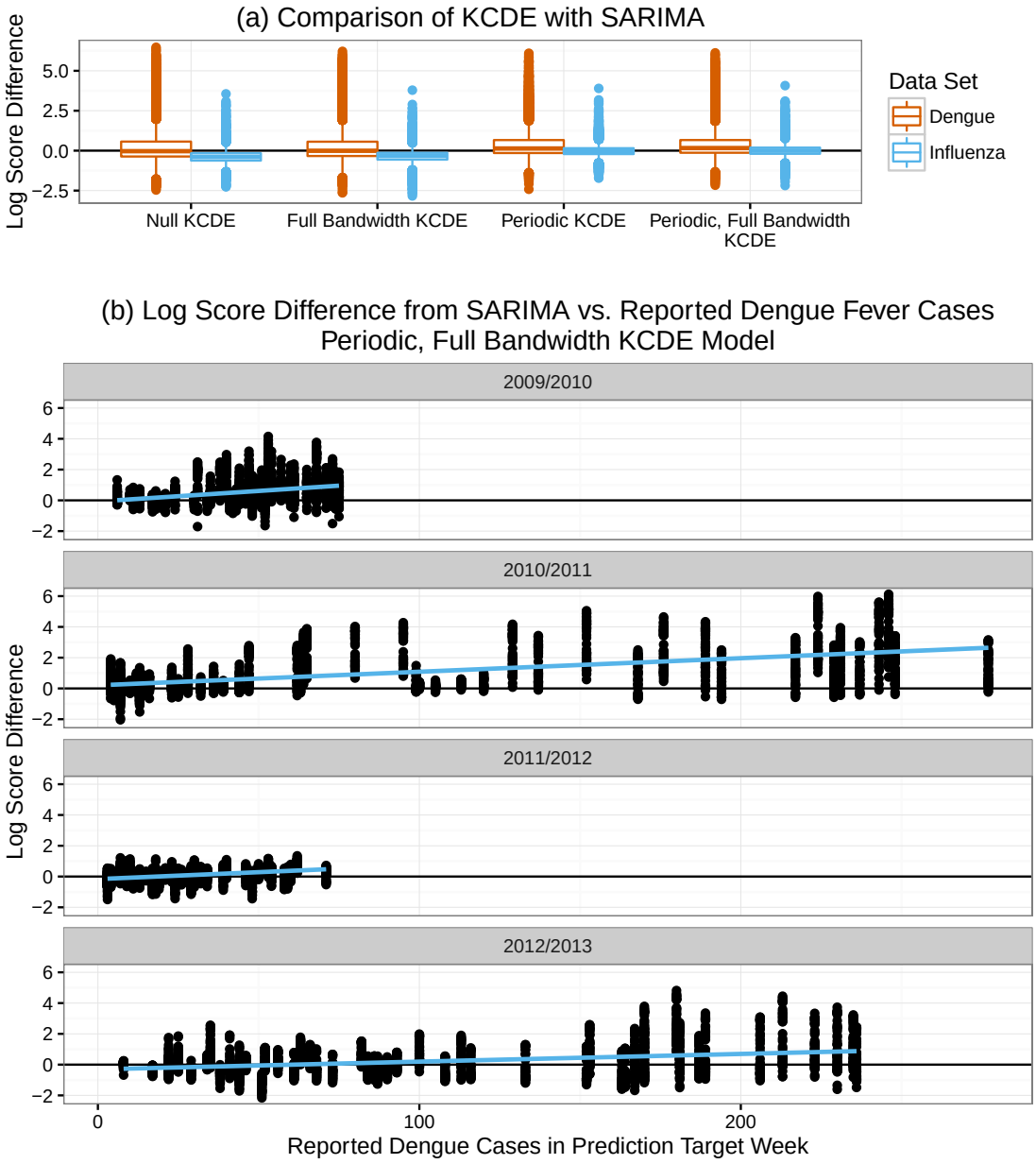

Fig. 5: Differences in log scores for the weekly predictive distributions obtained from KCDE specifications and SARIMA. For reference, a log score difference of 2.3 (4.6) indicates that the predictive density from KCDE was about 10 (100) times as large as the predictive density from SARIMA at the realized outcome. In Panel (a) we summarize the results across all combinations of prediction horizon and prediction time in the test period. In Panel (b) we display more detailed results for just the application to predicting dengue fever and the Periodic, Full Bandwidth KCDE specification. Each point corresponds to a unique combination of prediction target week and prediction horizon.

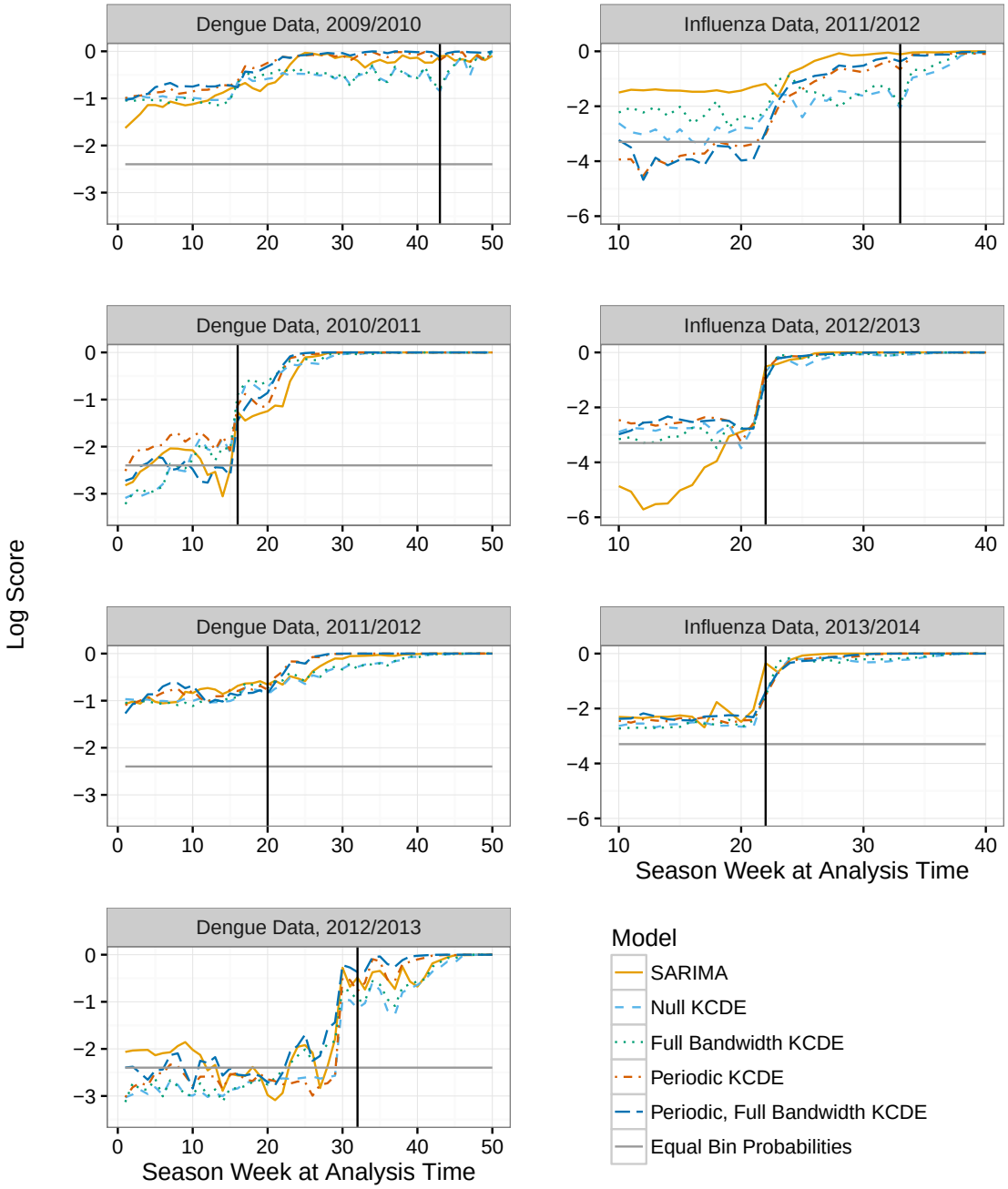

Fig. 6: Log scores for predictions of peak week incidence by predictive model and analysis time. The vertical line is placed at the peak week for each season. The log score for "Equal Bin Probabilities" is obtained by assigning equal probability that the peak incidence will be in each of the specified incidence bins. There are 11 incidence bins for dengue and 27 bins for influenza.