Journal Title
XX(X):1–22
©The Author(s) 0000
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/

Supplementary materials for infectious disease prediction with kernel conditional density estimation

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

Introduction

In this document we collect the supplementary materials for the article, following the organization of the original article.

Method Description

Discretizing the Kernel Function

We obtain the discrete kernel function by discretizing an underlying continuous kernel function:

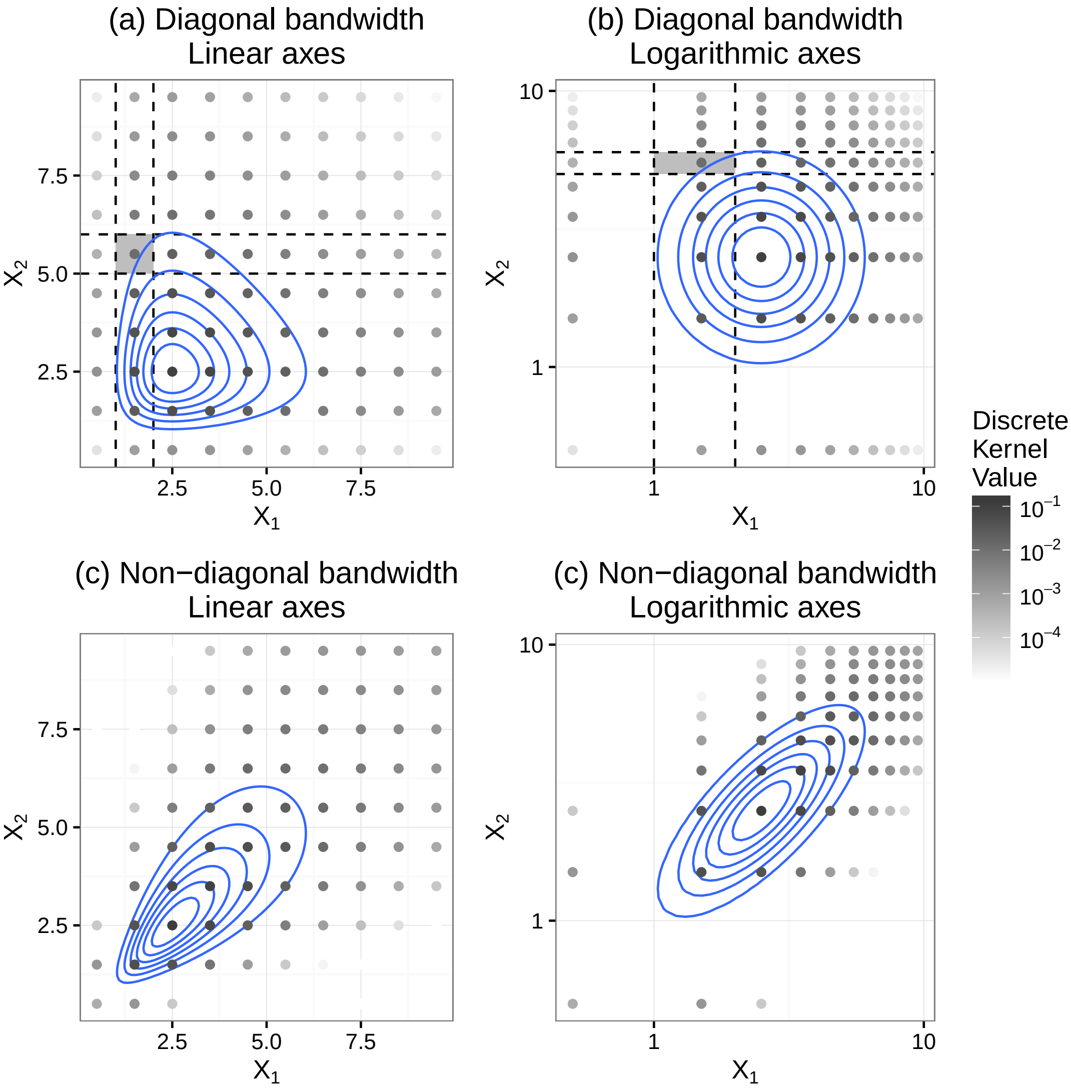
$$K_{disc}^{inc}(\tilde{\mathbf{z}}_{t^*}, \tilde{\mathbf{z}}_t; \mathbf{B}^h) = \int_{a_{z_{t^*-l_1}}}^{b_{z_{t^*-l_1}}} \cdots \int_{a_{z_{t^*+h}}}^{b_{z_{t^*+h}}} K_{cont}^{Incidence}(\tilde{\mathbf{z}}_{t^*}, \tilde{\mathbf{z}}_t; \mathbf{B}^h)\, dz_{t^*-l_1} \cdots dz_{t^*+h}$$

For each component variable in $(z_{t^*-l_1}, \ldots, z_{t^*-l_M}, z_{t^*+h})$, we associate lower and upper bounds of integration a_{z_j} and b_{z_j} with each value in the domain of that random variable. The value of the kernel function is obtained by integrating over the hyper-rectangle specified by these bounds. In our application, the possible values of the random variables are non-negative integer case counts. In order to facilitate use of the log-normal kernel, we add 0.5 to the observed case counts; the corresponding integration bounds are the non-negative integers as illustrated in Figure 1.

[1] Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst

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

Figure 1. Illustrations of $K_{\text{cont}}^{\text{inc}}$ and $K_{\text{disc}}^{\text{inc}}$ in the bivariate case. Solid lines show contours of the continuous kernel function. Grey dots indicate the value of the discrete kernel function. The value of the discrete kernel is obtained by integrating the continuous kernel over regions as illustrated by the dashed lines in panels (a) and (b). In all panels the kernel function is centered at $(2.5, 2.5)$. In panels (a) and (b) the bandwidth matrix is $\begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix}$, and in panels (c) and (d) the bandwidth matrix is $\begin{bmatrix} 0.2 & 0.15 \\ 0.15 & 0.2 \end{bmatrix}$. We illustrate each case with both linear and logarithmic scale axes.



Simulation Study

Simulation Distributions

In the simulation study, we simulate data from discretized multivariate normal distributions. The method for discretizing the underlying multivariate normal is the same as we described above for descritizing the kernel function. As we discussed in the paper, the normal distribution has mean $\underset{\sim}{0}$ and covariance matrix with 1 on the diagonal and 0.9 off of the diagonal. This multivariate normal distribution was used in one of the simulation studies conducted by Duong and Hazelton[4] demonstrating that a fully parameterized bandwidth matrix could yield improved density estimates for joint density estimation with continuous distributions. We discretize this distribution at the half-integers as illustrated for the two-dimensional case in Figure 2.

Hellinger Distance

The Hellinger distance of the estimated density $\widehat{f}(x)$ from the true density $f(x)$ is given by

$$\text{Hellinger}(f, \widehat{f}) = \left[1 - \int \left\{ f(x)\widehat{f}(x) \right\}^{\frac{1}{2}} dx \right]^{\frac{1}{2}}$$

In the simulation study, we measure the quality of a conditional density estimate by integrating the Hellinger distance over the range of the conditioning variables, weighting according to the density of those conditioning variables:

$$\begin{aligned}
&\text{Score}\{\widehat{f}(x_1|x_2, \ldots, x_D)\} \\
&\quad = \int \cdots \int \left[\text{Hellinger}\{f(x_1|x_2, \ldots, x_D), \widehat{f}(x_1|x_2, \ldots, x_D)\}\right] f(x_2, \ldots, x_D) dx_2 \cdots dx_D \\
&\quad = \int \cdots \int \left[1 - \int \left\{ f(x_1|x_2, \ldots, x_D)\widehat{f}(x_1|x_2, \ldots, x_D) \right\}^{\frac{1}{2}} dx_1 \right]^{\frac{1}{2}} f(x_2, \ldots, x_D) dx_2 \cdots dx_D \\
&\quad = \int \cdots \int \left[1 - \int \left\{ \frac{\widehat{f}(x_1|x_2, \ldots, x_D)}{f(x_1|x_2, \ldots, x_D)} \right\}^{\frac{1}{2}} f(x_1|x_2, \ldots, x_D)\, dx_1 \right]^{\frac{1}{2}} f(x_2, \ldots, x_D) dx_2 \cdots dx_D
\end{aligned} \tag{1}$$

We perform Monte Carlo integration to evaluate the integrals in Equation (1) by sampling observations $(x_{i,1}, \ldots, x_{i,D})$ from the joint distribution of $\mathbf{X}$.

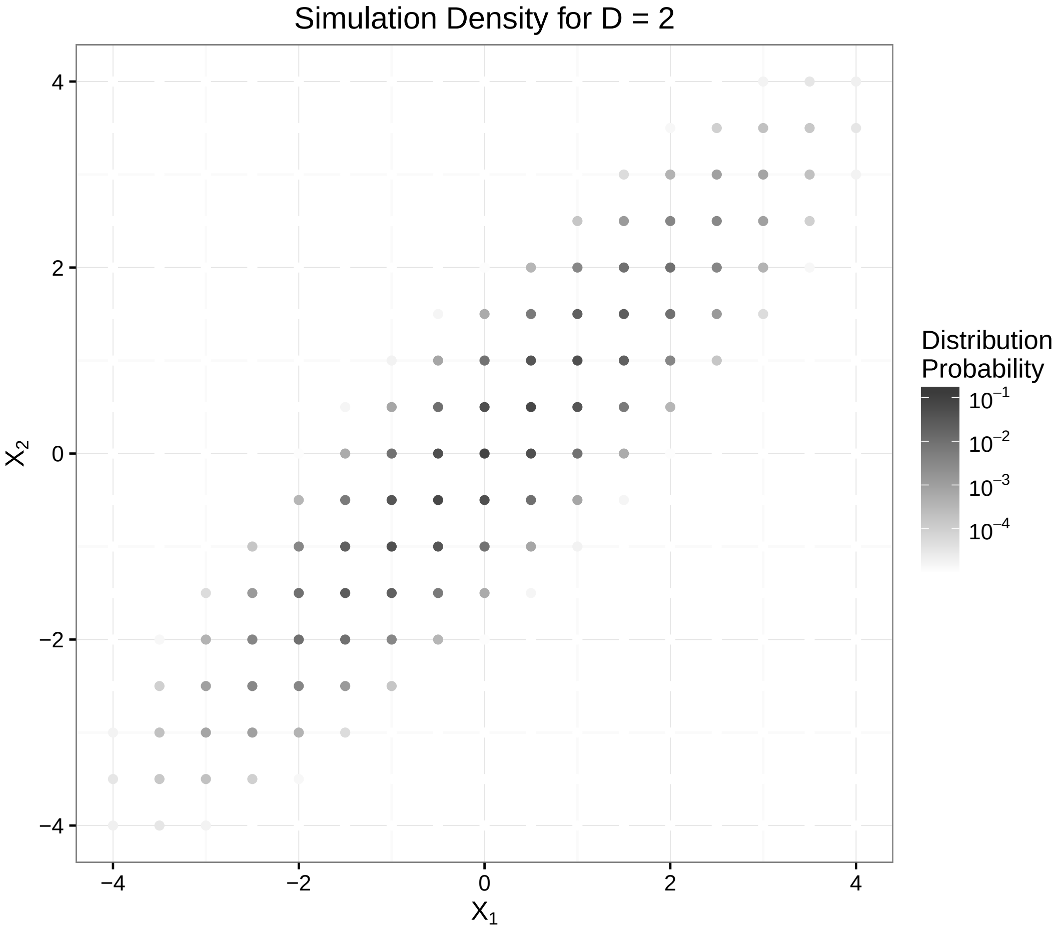
Applications

Prediction Targets

As we discussed in the main article, there are three prediction targets for each data set:

1. 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.

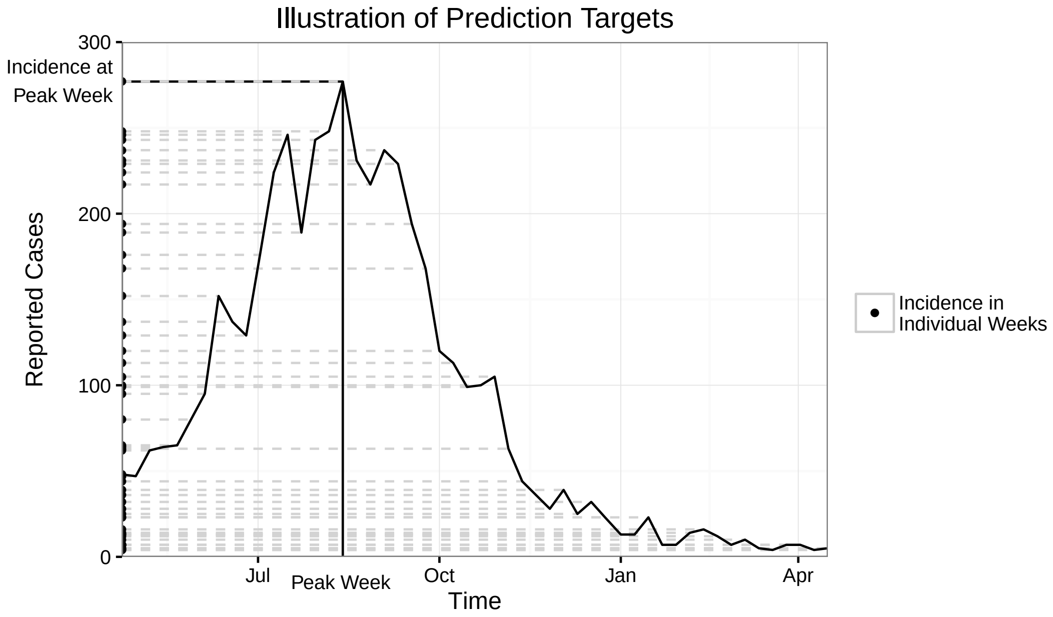
Figure 2. The distribution that we simulate data from in the simulation study, for the case of $D = 2$.



2. In each week of the test data set, we make predictions for the timing of the peak week of the corresponding season.
3. 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.

These prediction targets are illustrated in Figure 3.

Figure 3. Illustration of the prediction targets using one season of the dengue data. The solid vertical line indicates the timing of the peak week. The solid horizontal line indicates the incidence at the peak week. The points along the vertical axis indicate the incidence at every week for the 52 weeks after the time at which predictions are made.



Baseline SARIMA Model

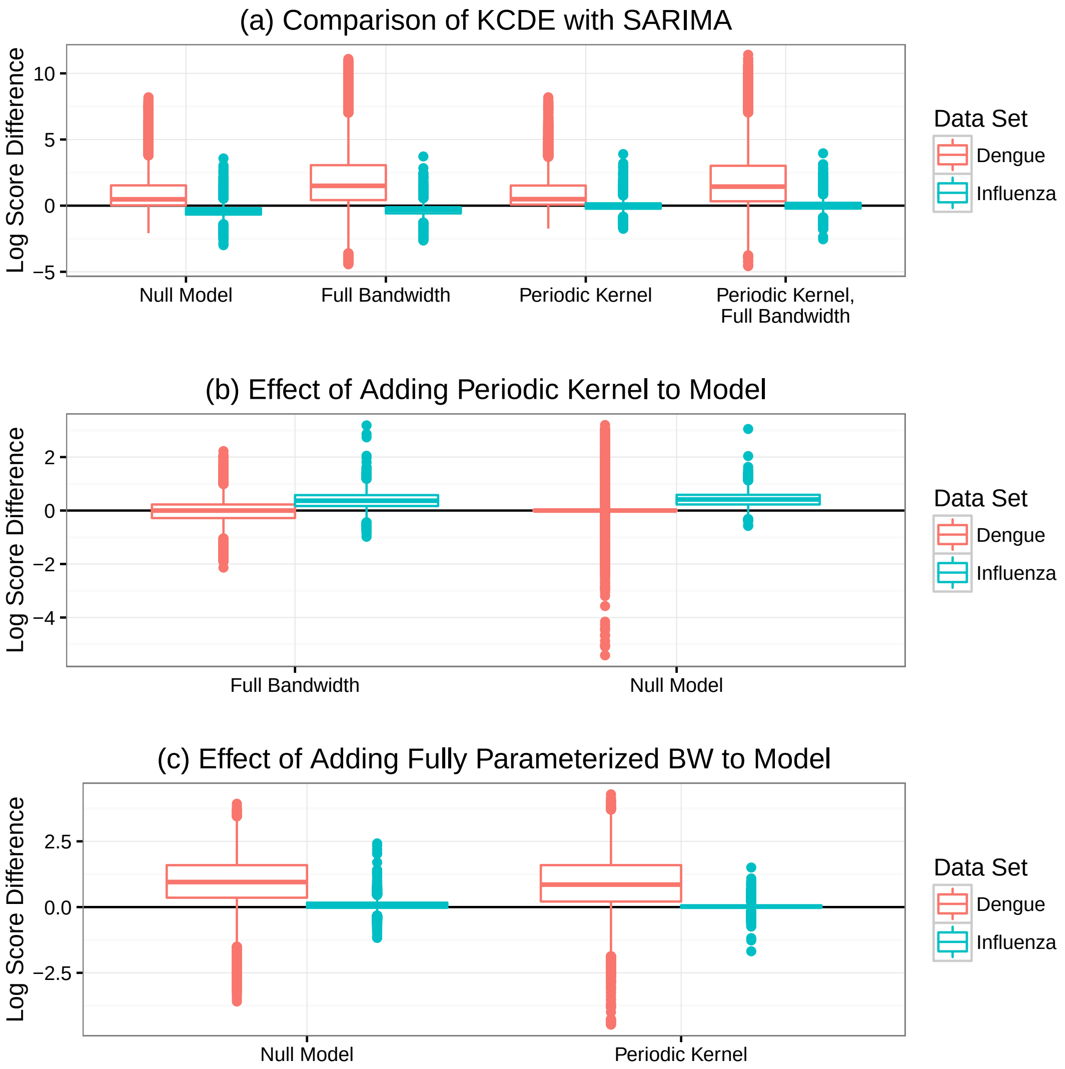
We use a SARIMA model as a baseline to compare our approach to. In fitting this model, we first transformed the observed incidence measure to the log scale (after adding 1 in the Dengue data set, which included some observations of 0 cases); this transformation makes the normality assumptions of the 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[13]; 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. We note that a different SARIMA model was used as a baseline in the Dengue competition.

Predictive Distributions for Individual Weeks

to do: christmas effect plot in supplemental materials

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

```
## Error in ‘$<-.data.frame‘(‘*tmp*‘, "fixed_values", value = structure(c(1L, :
replacement has 21736 rows, data has 43368
```

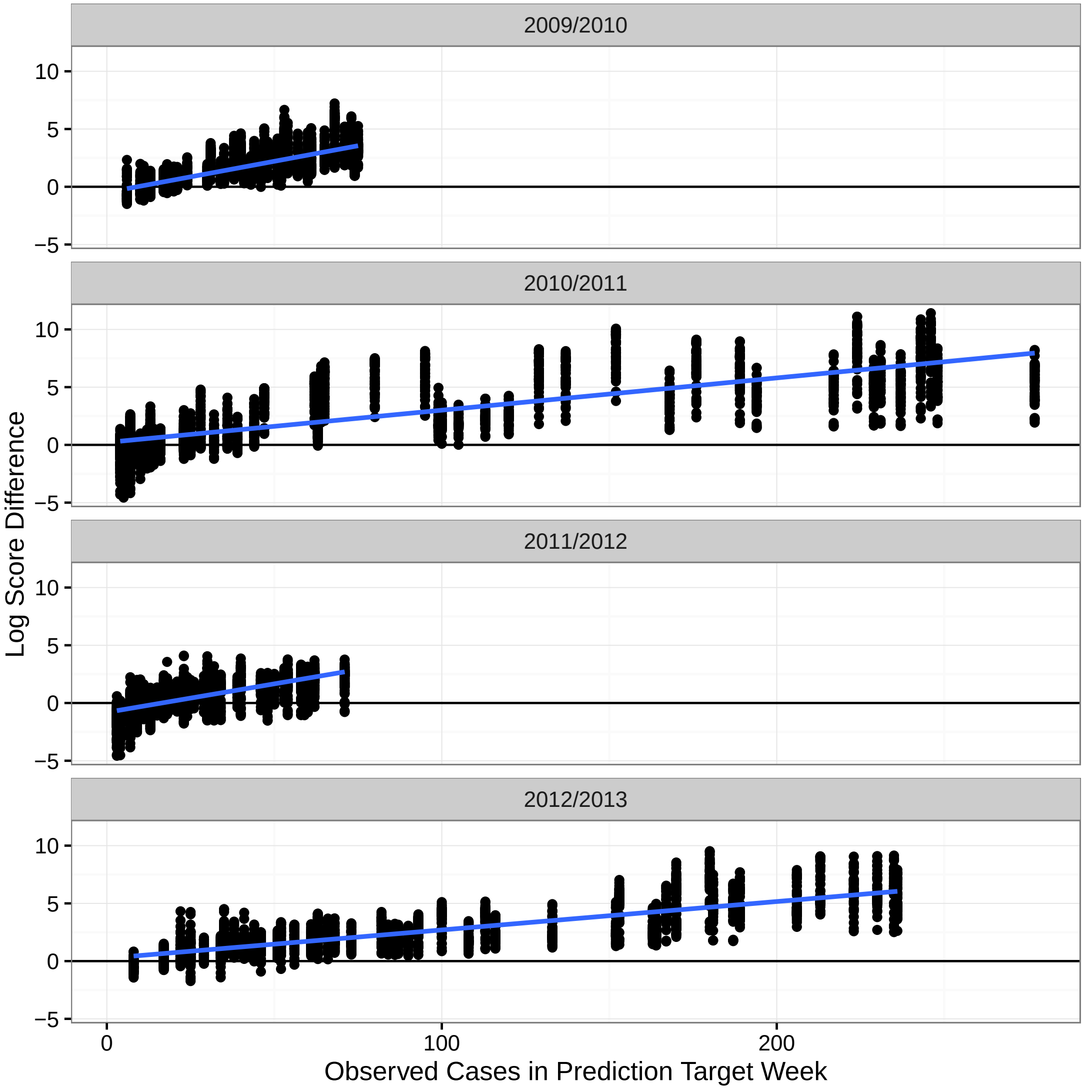


(a) Comparison of KCDE with SARIMA

(b) Effect of Adding Periodic Kernel to Model

(c) Effect of Adding Fully Parameterized BW to Model

Figure 5. Differences in log scores for the weekly predictive distributions from the KCDE specification with a periodic kernel component and fully parameterized bandwidth and the SARIMA model. Each point corresponds to a unique combination of prediction target week and prediction horizon.

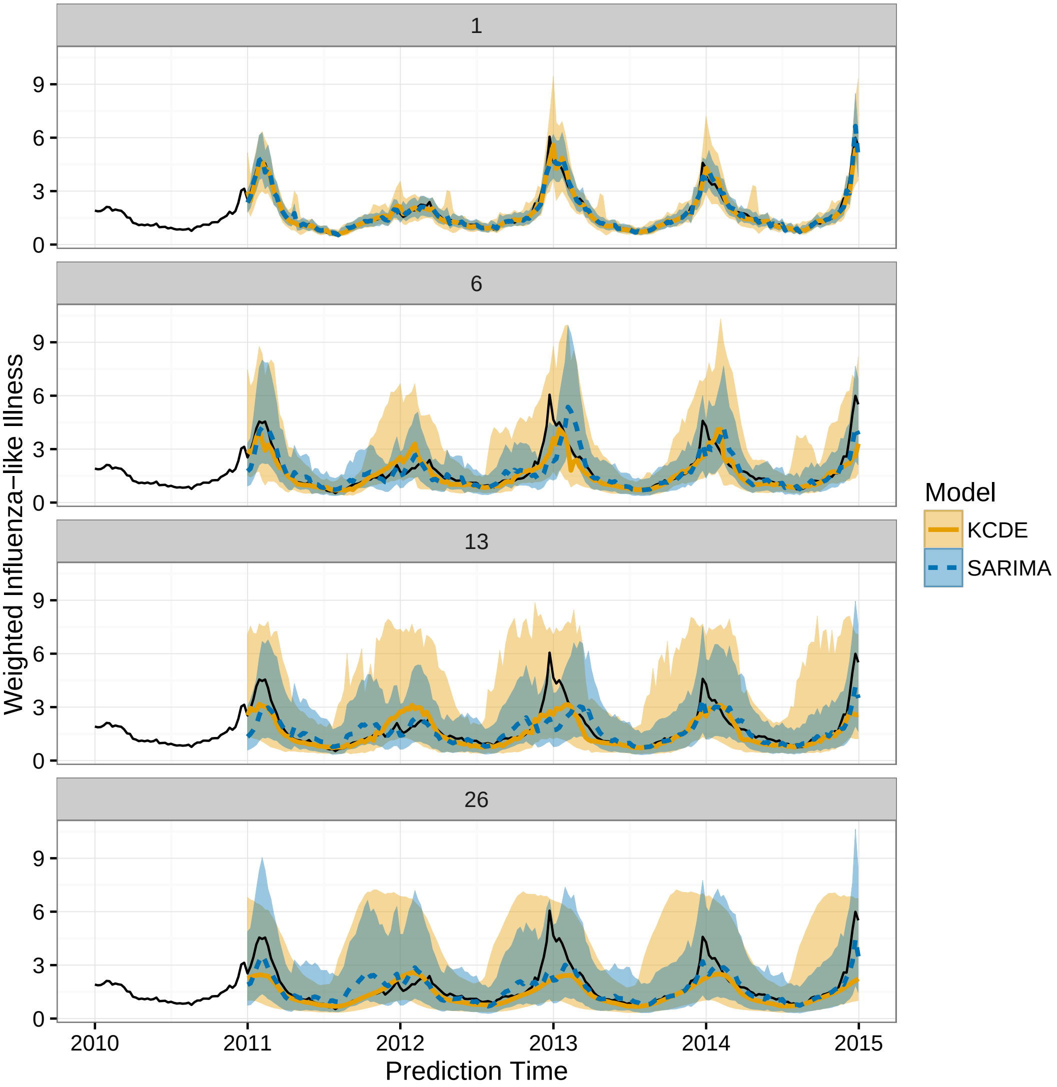


Predictive Distributions for Peak Week and Peak Incidence

Conclusions

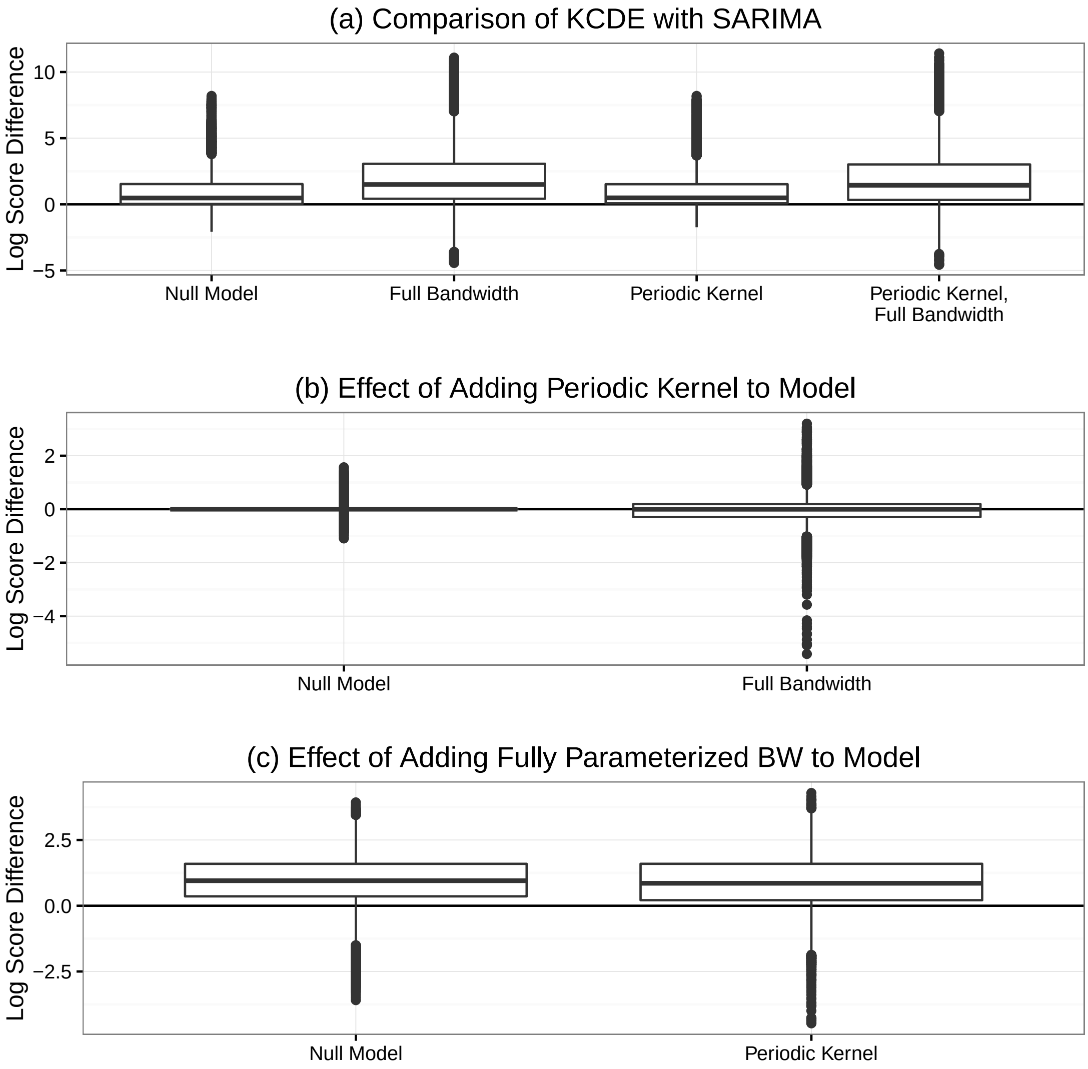
Prediction of infectious disease incidence at horizons of more than a few weeks is a challenging task. We have presented one approach to doing this and found that it is a viable method that

Figure 6. Plots of point and interval predictions from SARIMA and KCDE.



yields improved predictions relative to commonly employed methods in some applications. In an application to predicting Dengue fever, we saw that our approach offered consistent performance gains relative to a SARIMA model in predicting incidence in individual weeks. For predicting influenza-like illness, we saw that our approach did not pick up some features of the data generating process, such as the Christmas-week effect, that a SARIMA model did capture. For

Figure 7. Differences in log scores for the weekly predictive distributions for Dengue among pairs of models across all combinations of prediction horizon and prediction time in the test period. In panel (a) positive values indicate cases when KCDE outperformed SARIMA. In panel (b) positive values indicate cases when the specification of KCDE with the periodic kernel outperformed the corresponding specification without the periodic kernel. In panel (c) positive values indicate cases when the specification of KCDE with a fully parameterized bandwidth outperformed the KCDE specification with a diagonal bandwidth matrix.

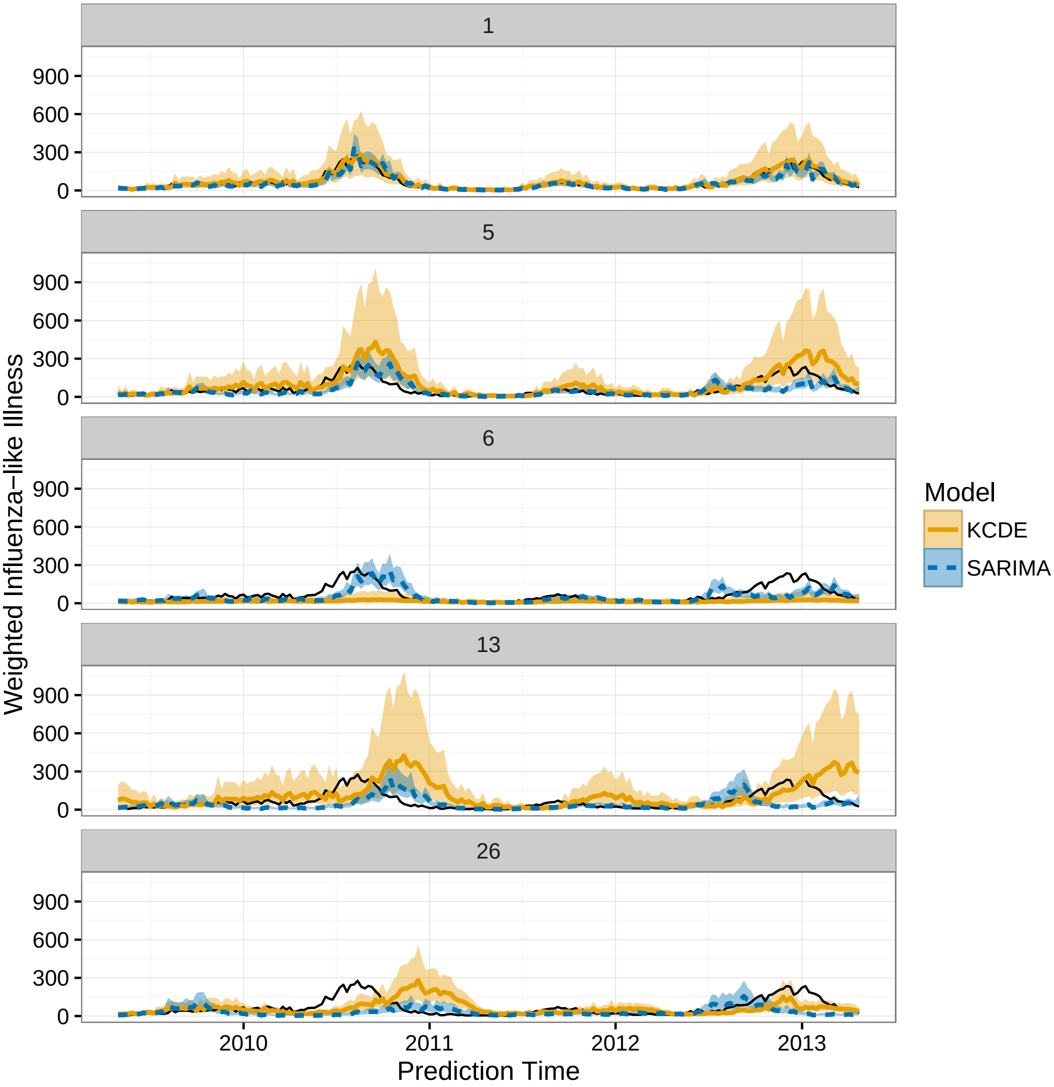


(a) Comparison of KCDE with SARIMA

(b) Effect of Adding Periodic Kernel to Model

(c) Effect of Adding Fully Parameterized BW to Model

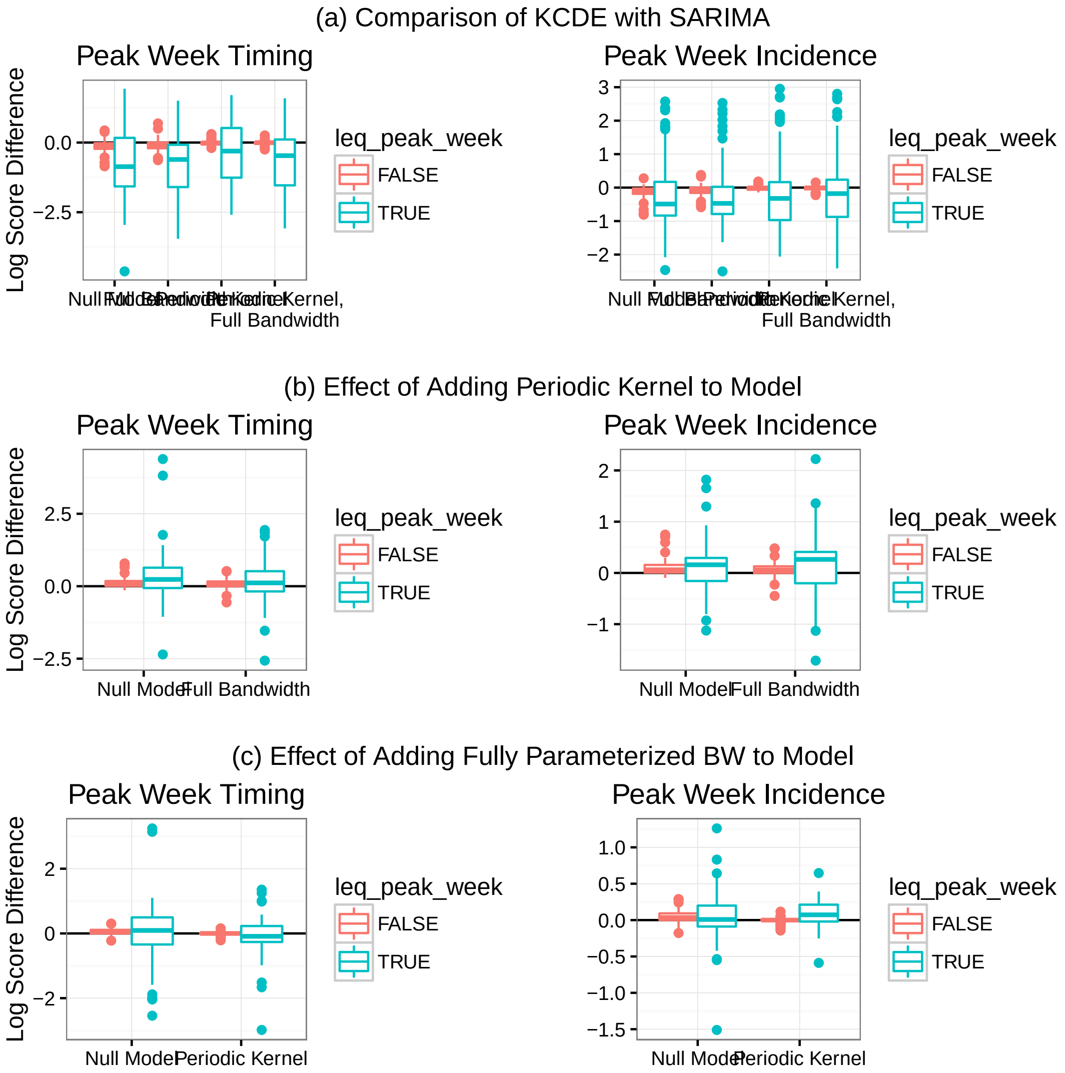
Figure 8. Plots of point and interval predictions from SARIMA and KCDE for Dengue.



some prediction targets, this meant that SARIMA outperformed our method. On the other hand, our method rarely performed worse than a very naive baseline of equal bin probabilities; the same cannot be said for the SARIMA model.

There is much room for extensions and improvements to the methods we have outlined in this article. Hall, Racine, and Li[8] show that when cross-validation is used to select the

Figure 9. Differences in log scores for the predictive distributions for the peak week and incidence at the peak week among pairs of models across all analysis times in the test period. In panel (a) positive values indicate cases when KCDE outperformed SARIMA. In panel (b) positive values indicate cases when the specification of KCDE with the periodic kernel outperformed the corresponding specification without the periodic kernel. In panel (c) positive values indicate cases when the specification of KCDE with a fully parameterized bandwidth outperformed the KCDE specification with a diagonal bandwidth matrix. In the plot for peak week timing in panel (a), the log score differences are not displayed for one analysis time when none of the simulated trajectories from SARIMA peaked at the true peak week. In that case, our monte carlo estimate of the difference in log scores is infinity.



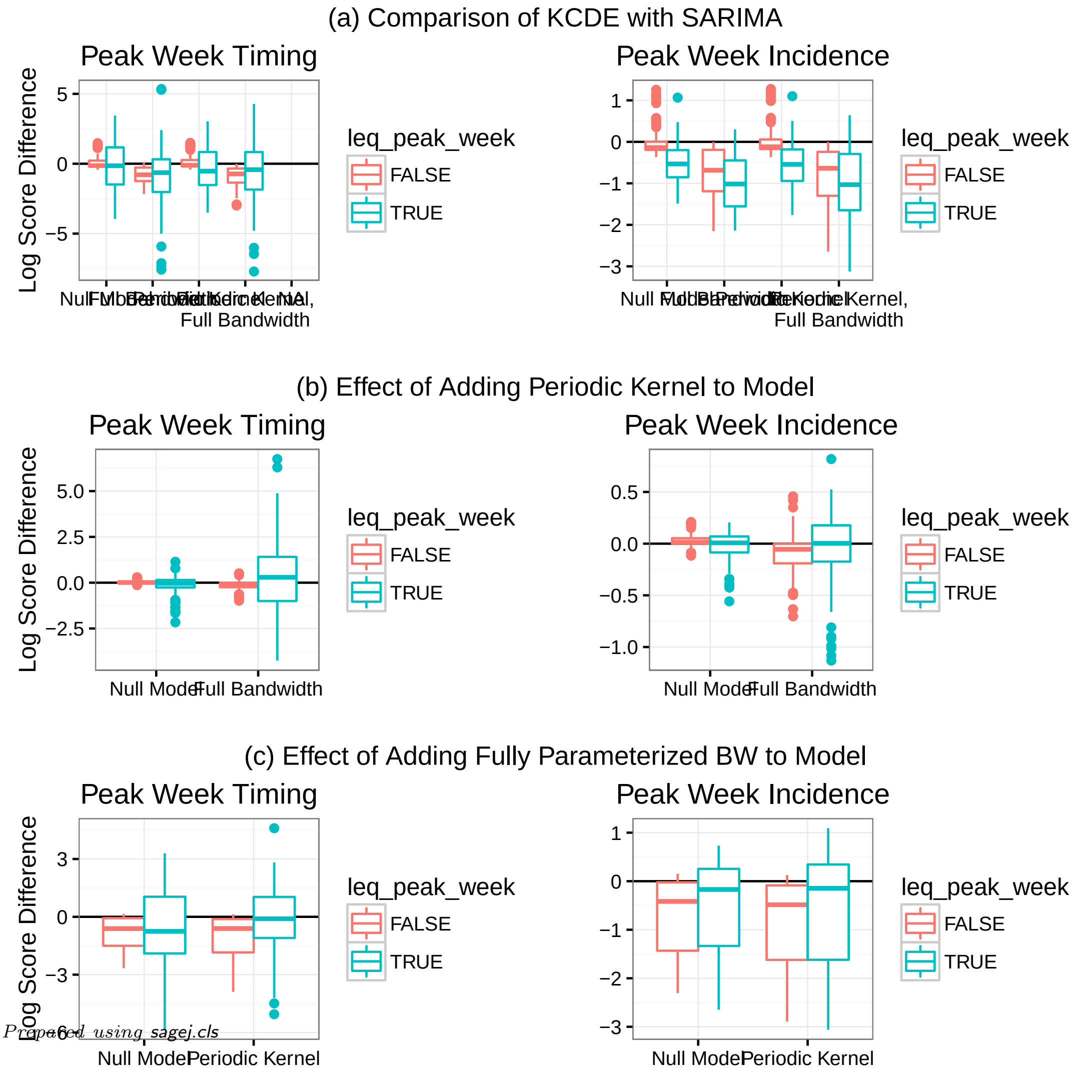
(a) Comparison of KCDE with SARIMA

(b) Effect of Adding Periodic Kernel to Model

(c) Effect of Adding Fully Parameterized BW to Model

Figure 10. Differences in log scores for the predictive distributions for the peak week and incidence at the peak week for Dengue among pairs of models across all analysis times in the test period. In panel (a) positive values indicate cases when KCDE outperformed SARIMA. In panel (b) positive values indicate cases when the specification of KCDE with the periodic kernel outperformed the corresponding specification without the periodic kernel. In panel (c) positive values indicate cases when the specification of KCDE with a fully parameterized bandwidth outperformed the KCDE specification with a diagonal bandwidth matrix. In the plot for peak week timing in panel (a), the log score differences are not displayed for one analysis time when none of the simulated trajectories from SARIMA peaked at the true peak week. In that case, our monte carlo estimate of the difference in log scores is infinity.

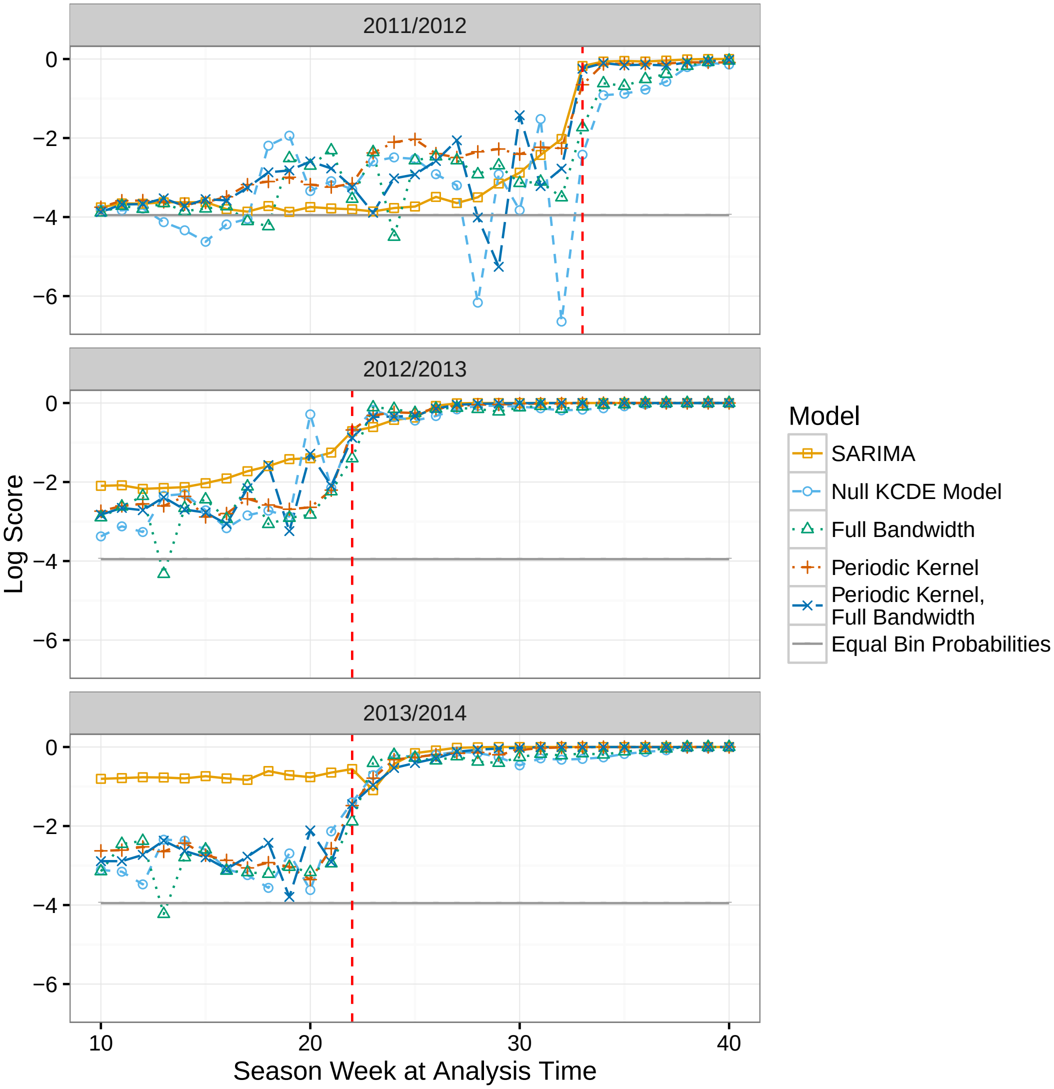
```
## Warning: Removed 11 rows containing non-finite values (stat_boxplot).
## Warning: Removed 9 rows containing non-finite values (stat_boxplot).
## Warning: Removed 9 rows containing non-finite values (stat_boxplot).
```


(a) Comparison of KCDE with SARIMA

(b) Effect of Adding Periodic Kernel to Model

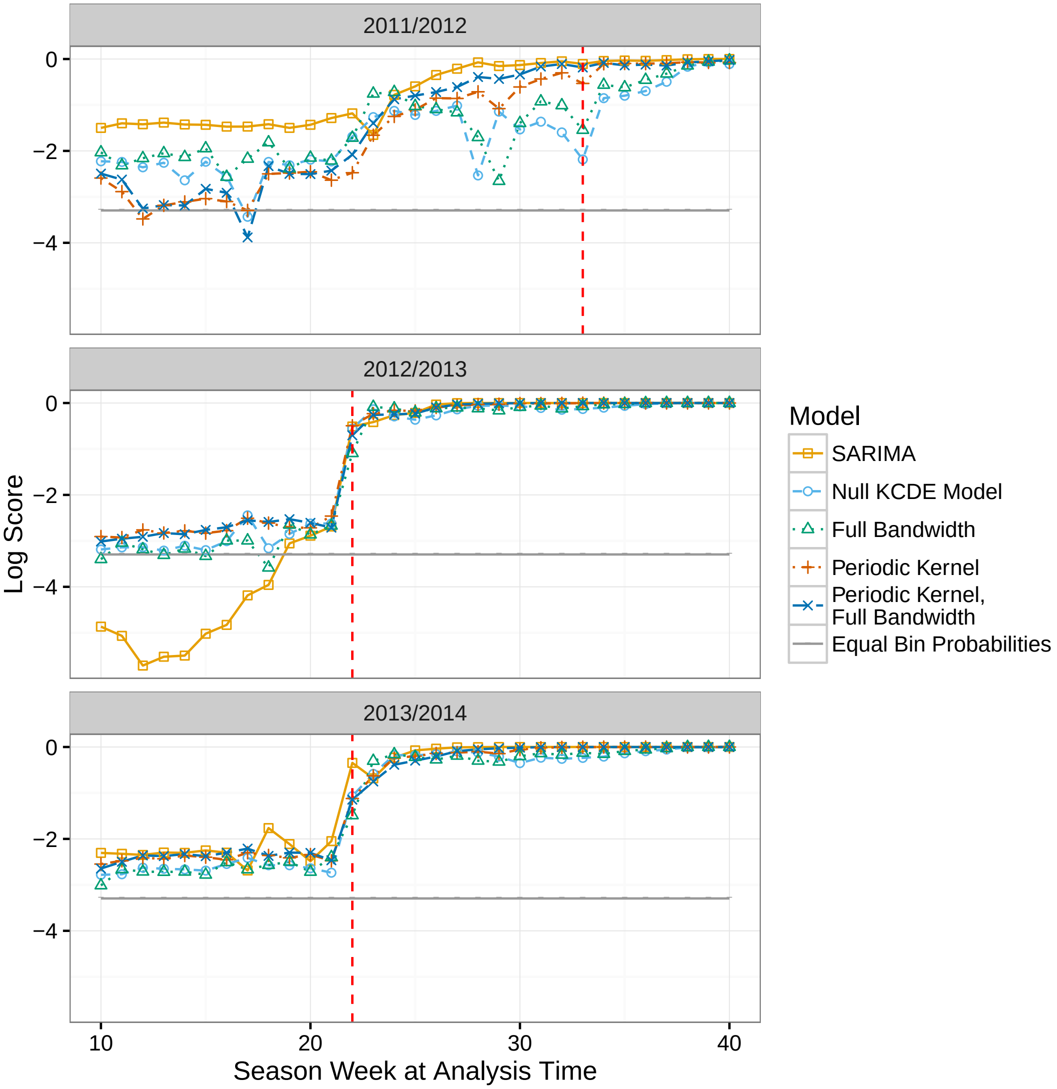
(c) Effect of Adding Fully Parameterized BW to Model

Figure 11. Log scores for predictions of peak week timing by predictive model and analysis time. The vertical gray line is placed at the peak week for each season.



bandwidth parameters in KCDE using product kernels, the estimated bandwidths corresponding to irrelevant conditioning variables tend to infinity asymptotically as the sample size increases. They discuss the fact that similar results could be obtained for linear combinations of continuous variables if a full bandwidth matrix were used. Our approach for obtaining kernels that can be

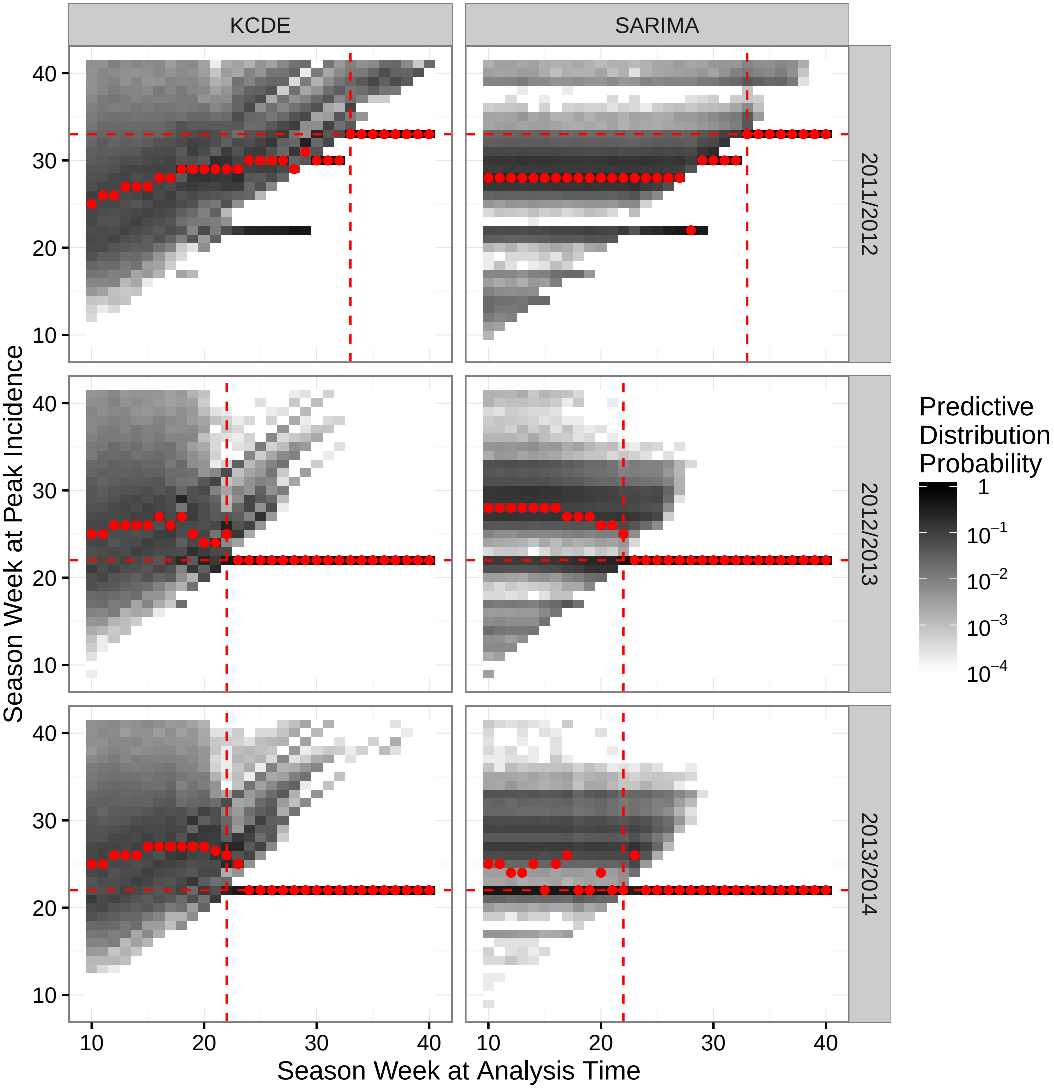
Figure 12. Log scores for predictions of incidence in the peak week by predictive model and analysis time. The vertical gray line is placed at the peak week for each season.



used with mixed discrete and continuous variables opens up an opportunity to extend this analysis to that case; we have not pursued this mathematical analysis here.

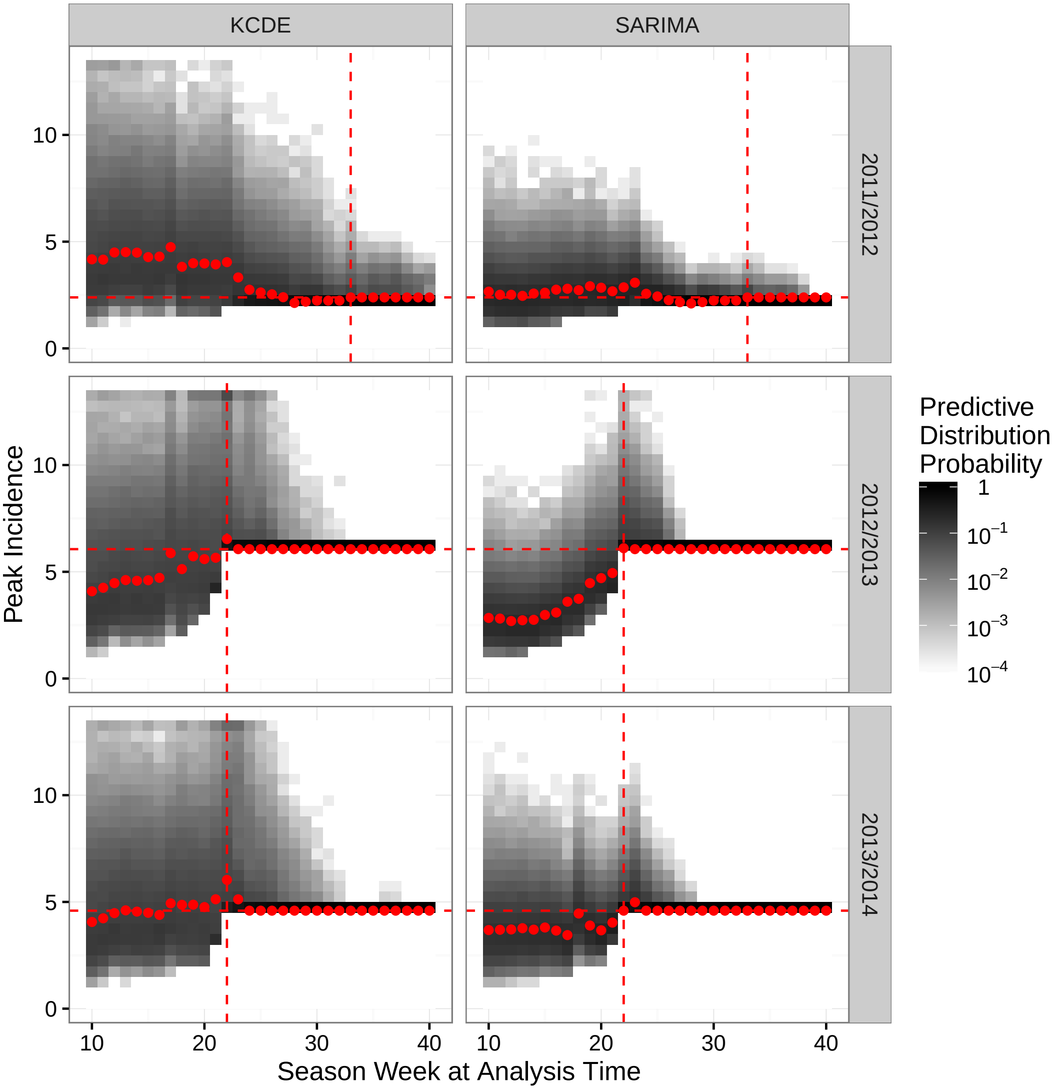
The above results regarding the inclusion of irrelevant conditioning variables hold asymptotically as the sample size increases. However, in practice, data set sizes are often limited.

Figure 13. Predictive distributions for predictions of peak week timing. The horizontal and vertical dashed lines are at the observed peak week for the season.



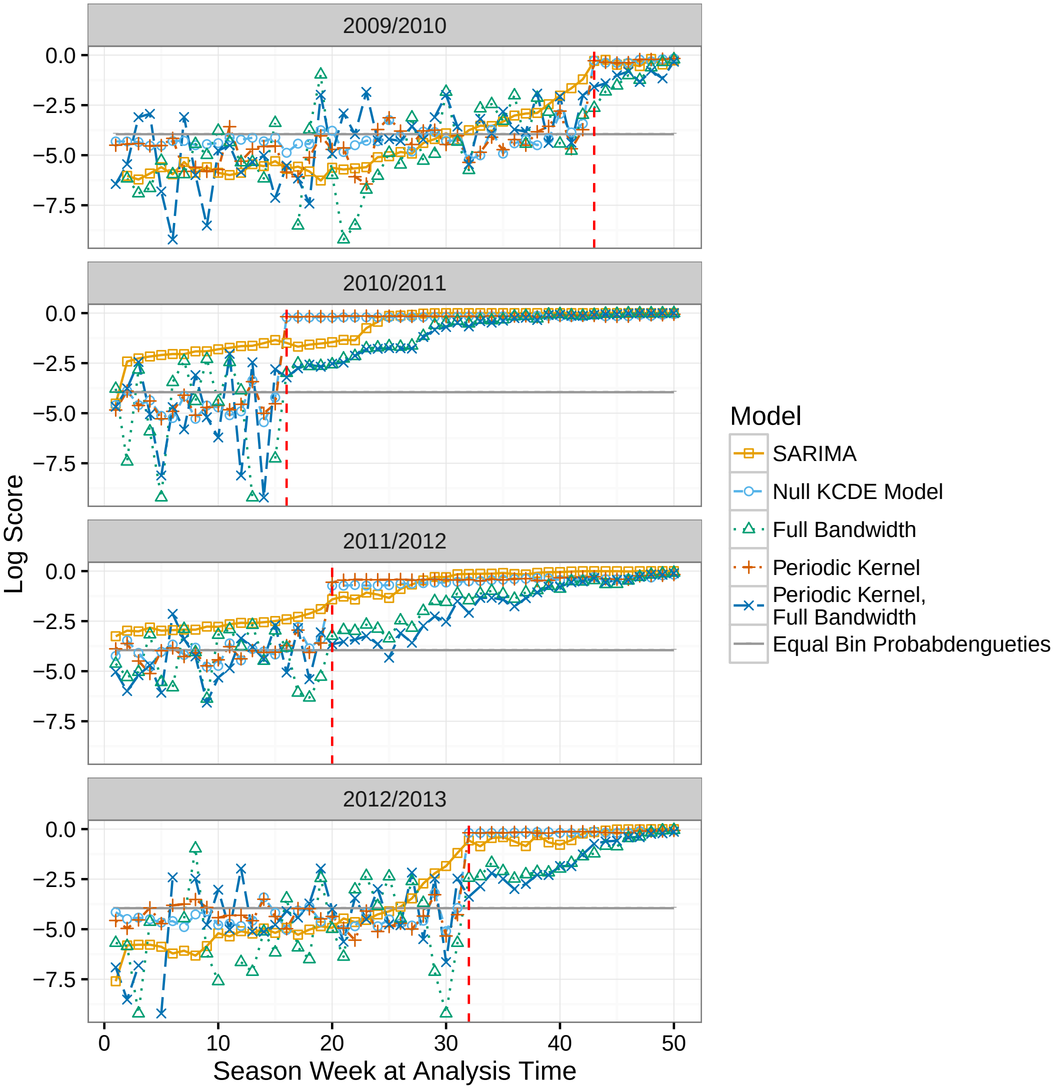
A major limitation of our current implementation of KCDE is that we do not have a good solution to the problem of determining which variables to condition on. In our current application, we allow for a variety of . selection problem. We explored a stepwise variable selection procedure, but found that this was too time-consuming to be practical.

Figure 14. Predictive distributions for predictions of peak week incidence. The horizontal dashed line is at the observed peak incidence for the season. The vertical dashed line is at the observed peak week for the season.



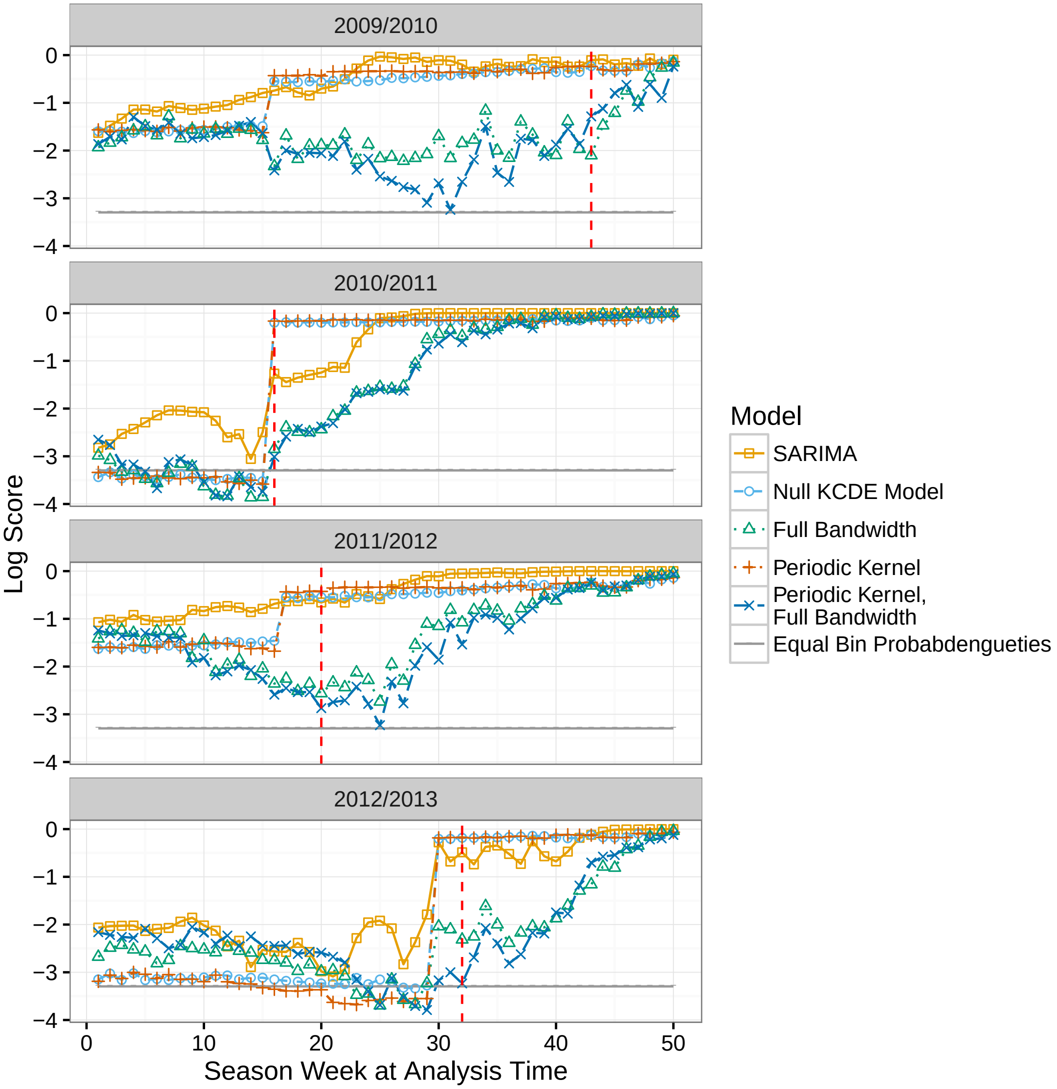
In other modeling settings where some conditioning variables may not be informative, shrinkage methods are often helpful. These methods could be incorporated into a kernel-based approach by imposing a penalty on the elements of the bandwidth matrix; in particular, we

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



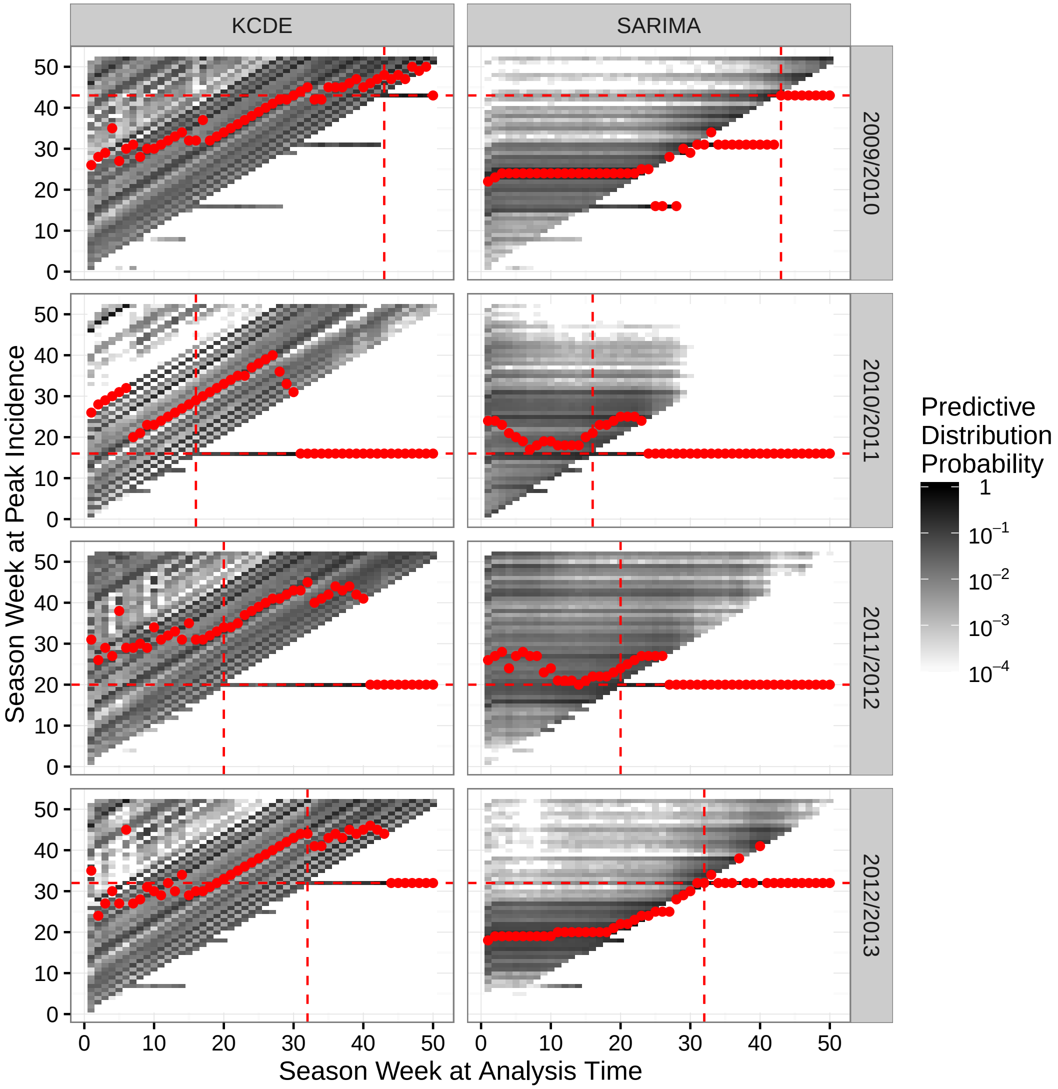
suggest that a penalty on the inverse of the bandwidth matrix encouraging it to have small eigenvalues could be helpful. Another alternative would be to pursue the Bayesian framework, using Dirichlet process mixtures with an informative prior on the mixture component covariance matrices.

Figure 16. Log scores for predictions of incidence in the peak week for Dengue by predictive model and analysis time. The vertical gray line is placed at the peak week for each season.



We could also make some tweaks to our implementation of KCDE. One major limitation of our current implementation is its sensitivity to edge effects. One possibility for addressing this would be to adopt locally linear or polynomial mean functions. Cite Hyndman, Bashtannyk, Grunwald - ”Estimating and Visualizing Conditional Densities”, maybe also Fan and Yim - ”A

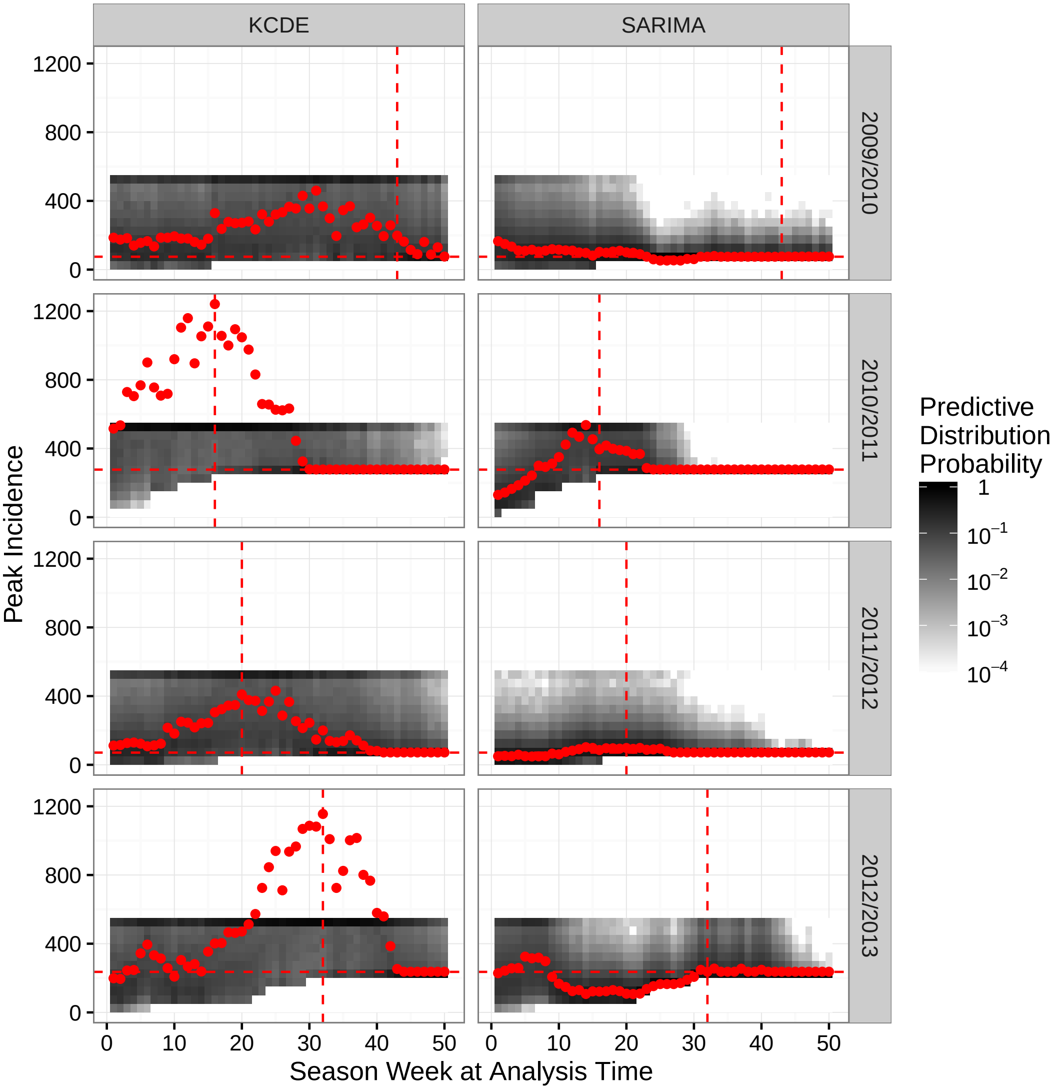
Figure 17. Predictive distributions for predictions of peak week timing for Dengue. The horizontal and vertical dashed lines are at the observed peak week for the season.



crossvaildation method for estimating conditional densities" and Fan et al. 1996 "Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems."

Another limitation to our current approach is in We also plan to include a KCDE-based prediction method as one component in an ensemble-based approach to prediction. There are

Figure 18. Predictive distributions for predictions of peak week incidence for Dengue. The horizontal dashed line is at the observed peak incidence for the season. The vertical dashed line is at the observed peak week for the season.



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

Other covariates

References

1. John Aitchison and Colin GG Aitken. Multivariate binary discrimination by the kernel method. *Biometrika*, 63(3):413–420, 1976.
2. Alexandria Brown, Stephen A. Lauer, Evan L. Ray, Xi Meng, and Nicholas G. Reich. A systematic review of prediction for infectious disease. *Journal*, submitted.
3. Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.
4. Tarn Duong and Martin L Hazelton. Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics*, 32(3):485–506, 2005.
5. Epidemic Prediction Initiative. FluSight: Seasonal Influenza Forecasting, January 2016. URL `http://dengueforecasting.noaa.gov/`.
6. Jianqing Fan and Tsz Ho Yim. A crossvalidation method for estimating conditional densities. *Biometrika*, 91(4):819–834, 2004.
7. Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.
8. Peter Hall, Jeff Racine, and Qi Li. Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99(468):1015–1026, 2004.
9. Jeffrey D Hart and Philippe Vieu. Data-driven bandwidth choice for density estimation based on dependent data. *The Annals of Statistics*, 18(2):873–890, 1990.
10. Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Science + Business Media, 2 edition, 2009.
11. Richard J Hatchett, Carter E Mecher, and Marc Lipsitch. Public health interventions and epidemic intensity during the 1918 influenza pandemic. *Proceedings of the National Academy of Sciences*, 104 (18):7582–7587, 2007.
12. Marius Hofert, Ivan Kojadinovic, Martin Maechler, and Jun Yan. *copula: Multivariate Dependence with Copulas*, 2015. URL `http://CRAN.R-project.org/package=copula`. R package version 0.999-14.
13. Rob J Hyndman. *forecast: Forecasting functions for time series and linear models*, 2015. URL `http://github.com/robjhyndman/forecast`. R package version 6.2.
14. Jooyoung Jeon and James W Taylor. Using conditional kernel density estimation for wind power density forecasting. *Journal of the American Statistical Association*, 107(497):66–79, 2012.
15. Harry Joe. Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94(2):401–419, 2005.
16. Qi Li and Jeff Racine. Nonparametric estimation of distributions with categorical and continuous data. *journal of multivariate analysis*, 86(2):266–292, 2003.
17. David JC MacKay. Introduction to gaussian processes. *NATO ASI Series F Computer and Systems Sciences*, 168:133–166, 1998.
18. Desheng Ouyang, Qi Li, and Jeffrey Racine. Cross-validation and the estimation of probability distributions with categorical data. *Journal of Nonparametric Statistics*, 18(1):69–100, 2006.
19. Pandemic Prediction and Forecasting Science and Technology Interagency Working Group. Dengue Forecasting, July 2015. URL `http://dengueforecasting.noaa.gov/`.
20. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org/`.

21. George Sugihara and Robert M. May. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series, April 1990.
22. Cécile Viboud, Pierre-Yves Boëlle, Fabrice Carrat, Alain-Jacques Valleron, and Antoine Flahault. Prediction of the spread of influenza epidemics by the method of analogues. *American Journal of Epidemiology*, 158(10):996–1006, 2003.
23. Jacco Wallinga, Michiel van Boven, and Marc Lipsitch. Optimizing infectious disease interventions during an emerging epidemic. *Proceedings of the National Academy of Sciences*, 107(2):923–928, 2010.
24. Min-Chiang Wang and John van Ryzin. A class of smooth estimators for discrete distributions. *Biometrika*, 68(1):301–309, 1981.
25. Haiming Zhou, Timothy Hanson, and Roland Knapp. Marginal bayesian nonparametric model for time to disease arrival of threatened amphibian populations. *Biometrics*, 71(4):1101–1110, 2015.