Semi-parametric Modeling of Temporal Trends for Cholera Outbreak in Haiti

Shaojun Zhang

April 5, 2016

Motivation

Haiti had been cholera-free for at least 100 years. However, there has been a severe outbreak of cholera in Haiti about 10 months after the catastrophic earthquake in 2010 and the reason is still not clear. Statistical modeling can help us understand how the disease is evolved and transmitted, but not much work has been done.

Background

Some research have been done to model the cholera transmission dynamics.

- Gaudart et al. (2013) used generalized additive model (GAM) with environment factors to study the daily cholera cases during the first year of the epidemic at the commune level.
- Kirpich et al. (2015) applied an extended SIR model to study the transmission mechanism in Ouest department.

Data

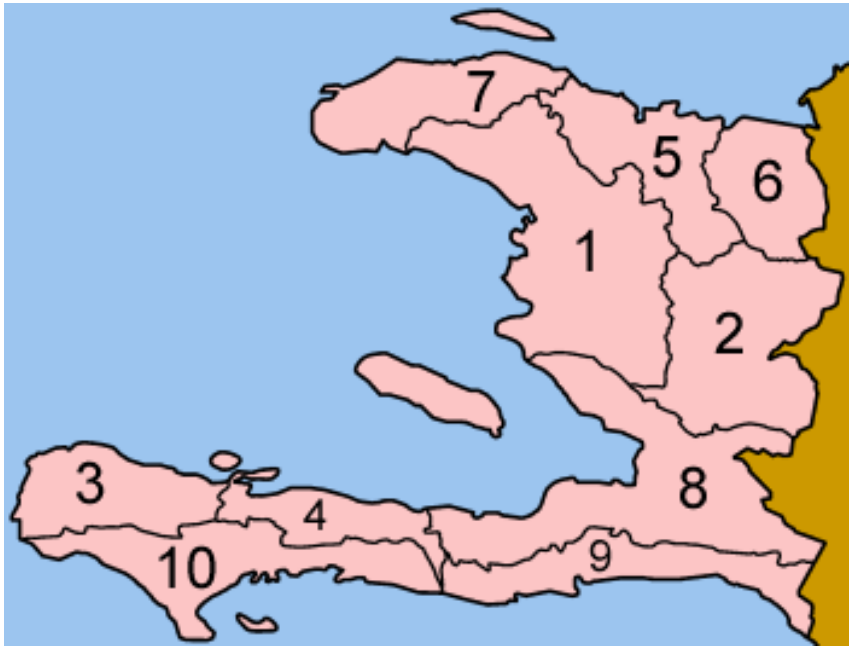

Figure 1: Departments of Haiti

Data

Haiti is divided into $N = 10$ departments. The cholera cases per week from November 21, 2010 to June 1, 2013 for each department are obtained from the website of Pan American Health Organization (PAHO). There are exactly $T = 132$ weeks in this time period.

Notation

- Y_{it}: the number of cholera cases in department i at week t for $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$
- $\{z_i(\cdot)\}_{i=1}^{K}$: the collection of nonconstant cubic spline basis functions where K is the dimension of the basis
- $\mathbf{Z}$: the design matrix consisting of the basis functions evaluated at each time point, that is

$$\mathbf{Z} = \begin{bmatrix} z_1(1) & z_2(1) & \ldots & z_K(1) \\ z_1(2) & z_2(2) & \ldots & z_K(2) \\ \ldots & \ldots & \ldots & \ldots \\ z_1(T) & z_2(T) & \ldots & z_K(T) \end{bmatrix}$$

Methods

Smoothing spline Poisson regression

The model can be written as

$$Y_{it}|\mu_{it} \sim \text{Poisson}(\mu_{it})$$

with log link function

$$\log \mu_{it} = \eta_{it}$$

where η_{it} is captured by a univariate function $f_i(\cdot)$ such that

$$f_i(t) = \eta_{it}.$$

Further, $f_i(\cdot)$ is assumed to lie in the space of all twice continuously differentiable functions and by using the roughness penalty

$$\lambda_i \int |f_i''(t)|^2 dt,$$

the estimate is forced to be a cubic spline.

Methods

Here, $f_i(\cdot)$ is modeled to be a linear combination of the cubic spline basis functions as

$$f_i(t) = \beta_{i0} + \sum_{j=1}^{K} z_j(t)\beta_{ij}.$$

Let $\log \boldsymbol{\mu_i} = (\log \mu_{i1}, \log \mu_{i2}, \ldots, \log \mu_{iT})^T$, $\mathbf{1} = (1, 1, \ldots, 1)^T$ and $\boldsymbol{\beta_i} = (\beta_{i1}, \beta_{i2}, \ldots, \beta_{iK})^T$. We can have the matrix form

$$\log \boldsymbol{\mu_i} = \beta_{i0}\mathbf{1}^T + \mathbf{Z}\boldsymbol{\beta_i}.$$

Note that the offsets of population sizes are ignored since they are absorbed in separate intercepts.

Methods

Cross-validation is used to determine the number of basis K as the following. The time period is divided into 10 time intervals of roughly equal width, denoted by $\{I_j\}_{j=1}^{10}$ in chronological order. The mean square prediction error (MSPE) is estimated as

$$\widehat{\text{MSPE}} = \frac{1}{N(T - |I_1| - |I_{10}|)} \sum_{i=1}^{N} \sum_{j=2}^{9} \sum_{t \in I_j} (Y_{it} - \hat{Y}_{it|-I_j})^2$$

where $|I_j|$ is the number of time points in I_j and $\hat{Y}_{it|-I_j}$ is the fitted value for Y_{it} without using the data in I_j. The $\widehat{\text{MSPE}}$s for K from 10 to 40 are calculated and the K corresponding to the minimum $\widehat{\text{MSPE}}$ is selected.

Methods

Zero-inflated Poisson regression

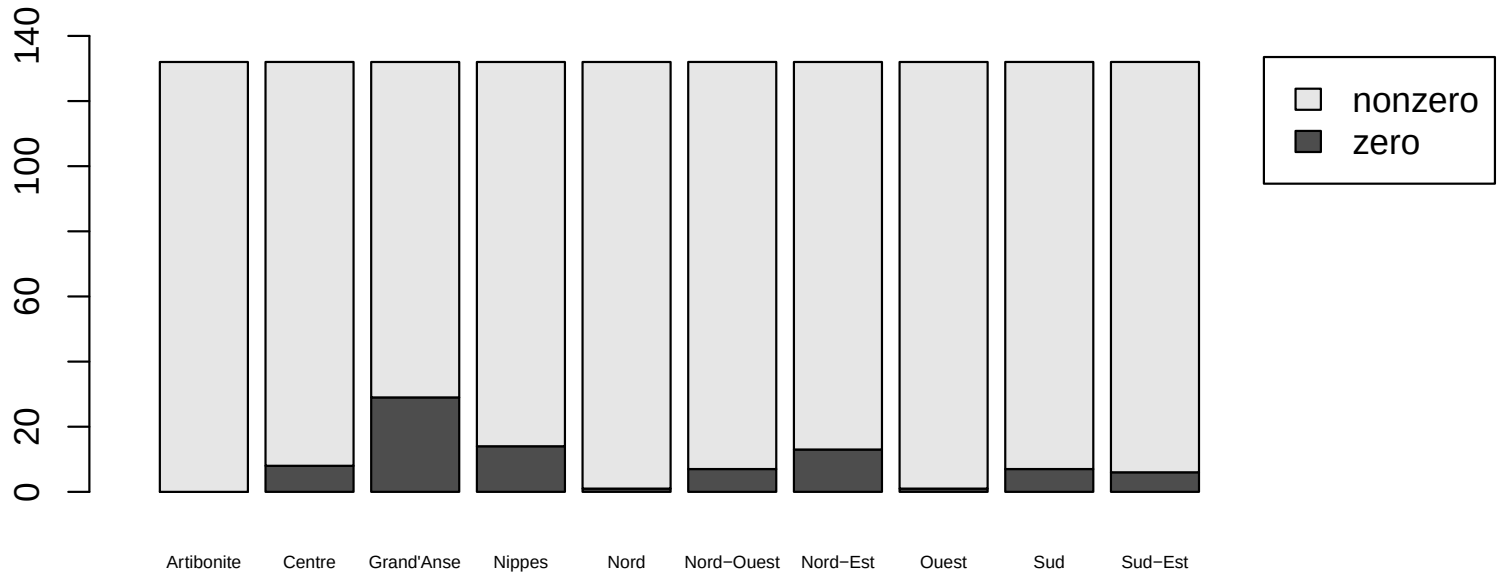

Figure 2: Number of zero cases for each department

Methods

The model can be expressed as

$$\begin{aligned} Y_{it} &\sim 0 \text{ with probability } p_i \\ &\sim \text{Poisson}(\mu_{it}) \text{ with probability } 1 - p_i \end{aligned}$$

with log link function

$$\log \boldsymbol{\mu_i} = \beta_{i0} \mathbf{1}^T + \mathbf{Z}\boldsymbol{\beta_i}.$$

Methods

Latent Gaussian model

The likelihood model is

$$Y_{it}|\mu_{it} \sim \text{Poisson}(\mu_{it})$$

with log link function

$$\log \mu_{it} = \eta_{it}.$$

The model for the latent variable can be written as

$$\eta_{it} = g_i(t),$$

where t is the observed time whose effect is modeled as a smooth function $g_i(\cdot)$.

Methods

Let $g_{it} = g_i(t)$ for $t = 1, 2, \ldots, T$. The random vector $\boldsymbol{g_i} = (g_{i1}, g_{i2}, \ldots, g_{iT})^T$ is assumed to have a random walk of order 2 (RW2) prior. It is constructed assuming independent second-order increments:

$$\Delta^2 g_{ij} = g_{ij} - 2g_{i(j+1)} + g_{i(j+2)} \sim N(0, \tau_i^{-1}).$$

The density for $\boldsymbol{g_i}$ is derived from its $T-2$ second-order increments as

$$\pi(\boldsymbol{g_i}|\tau_i) \propto \tau_i^{(T-2)/2} \exp\left\{-\frac{\tau_i}{2}\sum_{j=1}^{T-2}(\Delta^2 g_{ij})^2\right\}$$

with prior

$$\tau_i \sim \text{Gamma}(1, 0.00005).$$

The RW2 model represents 'smooth curves' with small squared second derivative, which serves the same goal as the penalty term in the smoothing spline Poisson regression.

Methods

Bayesian generalized linear mixed models

Two models are applied to the cholera data. Both assume

$$Y_{it}|\mu_{it} \sim \text{Poisson}(\mu_{it})$$

with log link function

$$\log \boldsymbol{\mu_i} = \beta_{i0}\mathbf{1}^T + \mathbf{Z}\boldsymbol{u_i}$$

and the prior for β_{i0}

$$\beta_{i0} \sim N(0, 1000).$$

Methods

The first model assumes different covariance matrices for different departments,

$$\boldsymbol{u_i}|\sigma^2_{ui} \sim N(0, \sigma^2_{ui}\mathbf{I})$$

with hyperprior

$$\sigma^2_{ui} \sim \text{Gamma}(1, 1).$$

The second model assumes common covariance matrix for all departments,

$$\boldsymbol{u_i}|\sigma^2_{u} \sim N(0, \sigma^2_{u}\mathbf{I})$$

with hyperprior

$$\sigma^2_{u} \sim \text{Gamma}(1, 1).$$

Mathematically, the covariance matrix of $\boldsymbol{u_i}$ also corresponds to the penalty term in the smoothing spline Poisson regression model.

Results

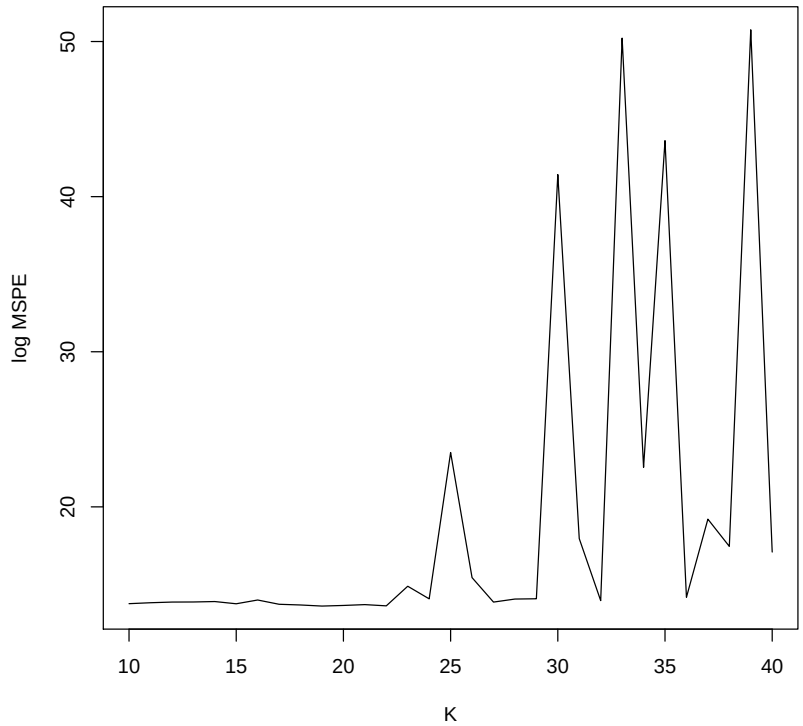

Figure 3: Log $\widehat{\text{MSPE}}$ for different K

Results

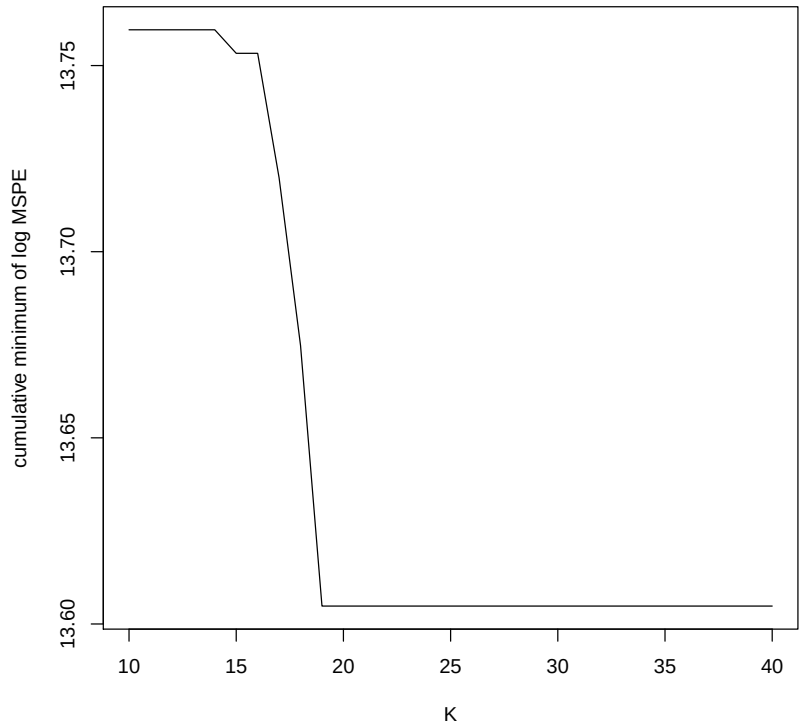

Figure 4: Cumulative minimum of log $\widehat{\text{MSPE}}$

Results

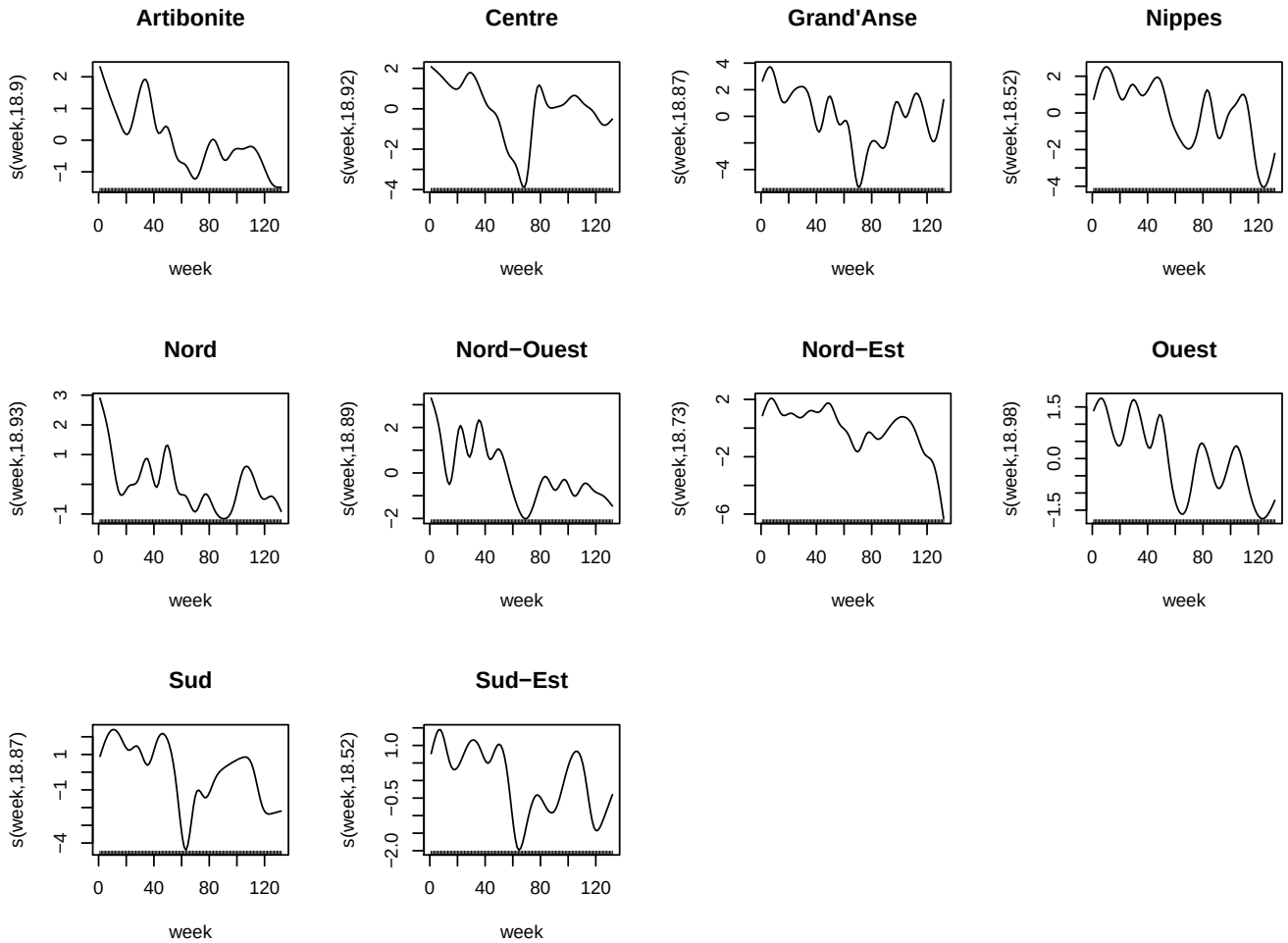

Figure 5: Fitted temporal trends of the smoothing spline Poisson regression model

Results

Table 1: Fitted point masses and p-values of Vuong's nested test

i	1	2	3	4	5	6	7	8	9	10
$\hat{p}_i$	0.000	0.061	0.183	0.037	0.008	0.098	0.053	0.008	0.045	0.019
p-value		0.082	0.001	0.066	0.149	0.004	0.005	0.175	0.010	0.146

Results

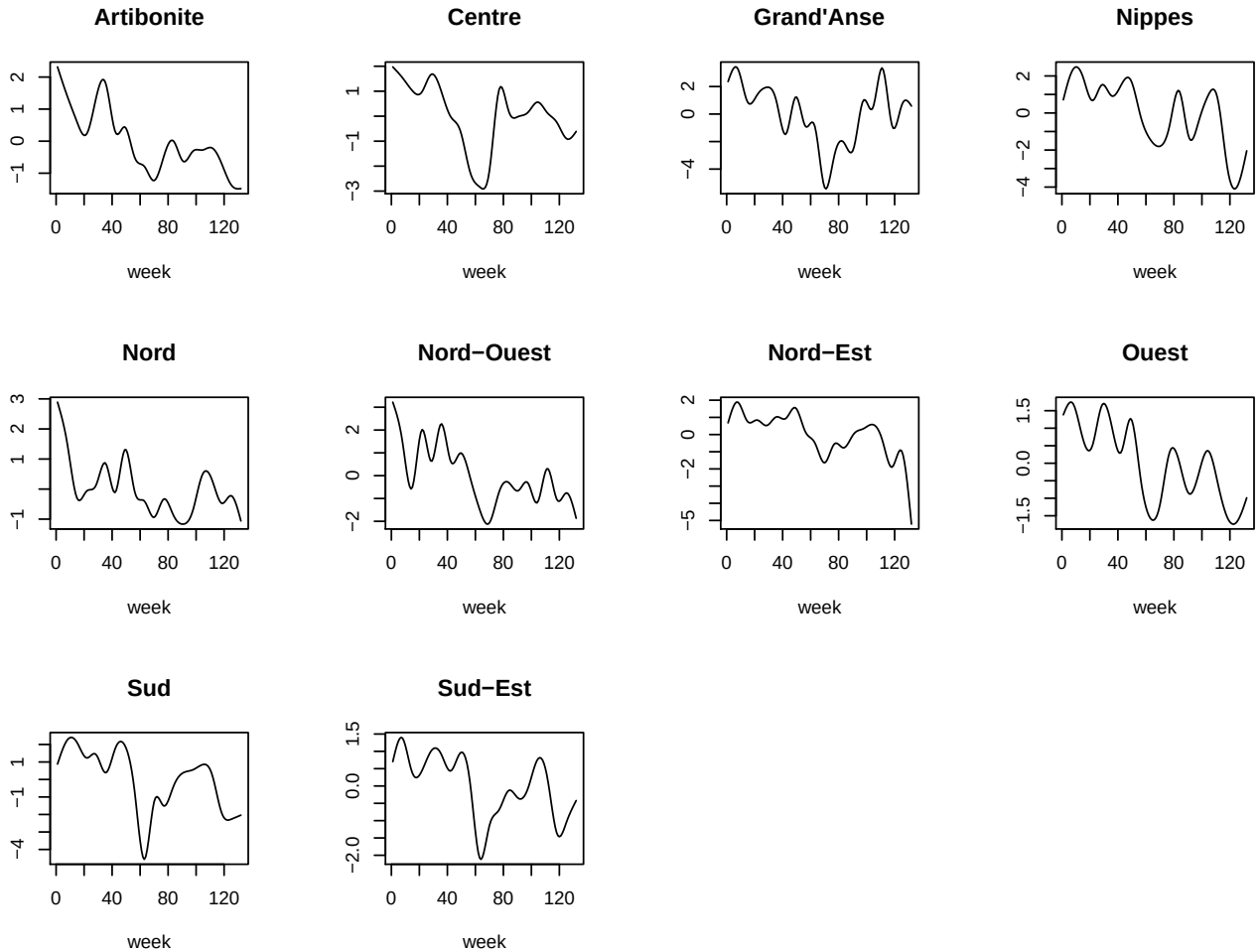

Figure 6: Fitted temporal trends of the zero-inflated Poisson regression model

Results

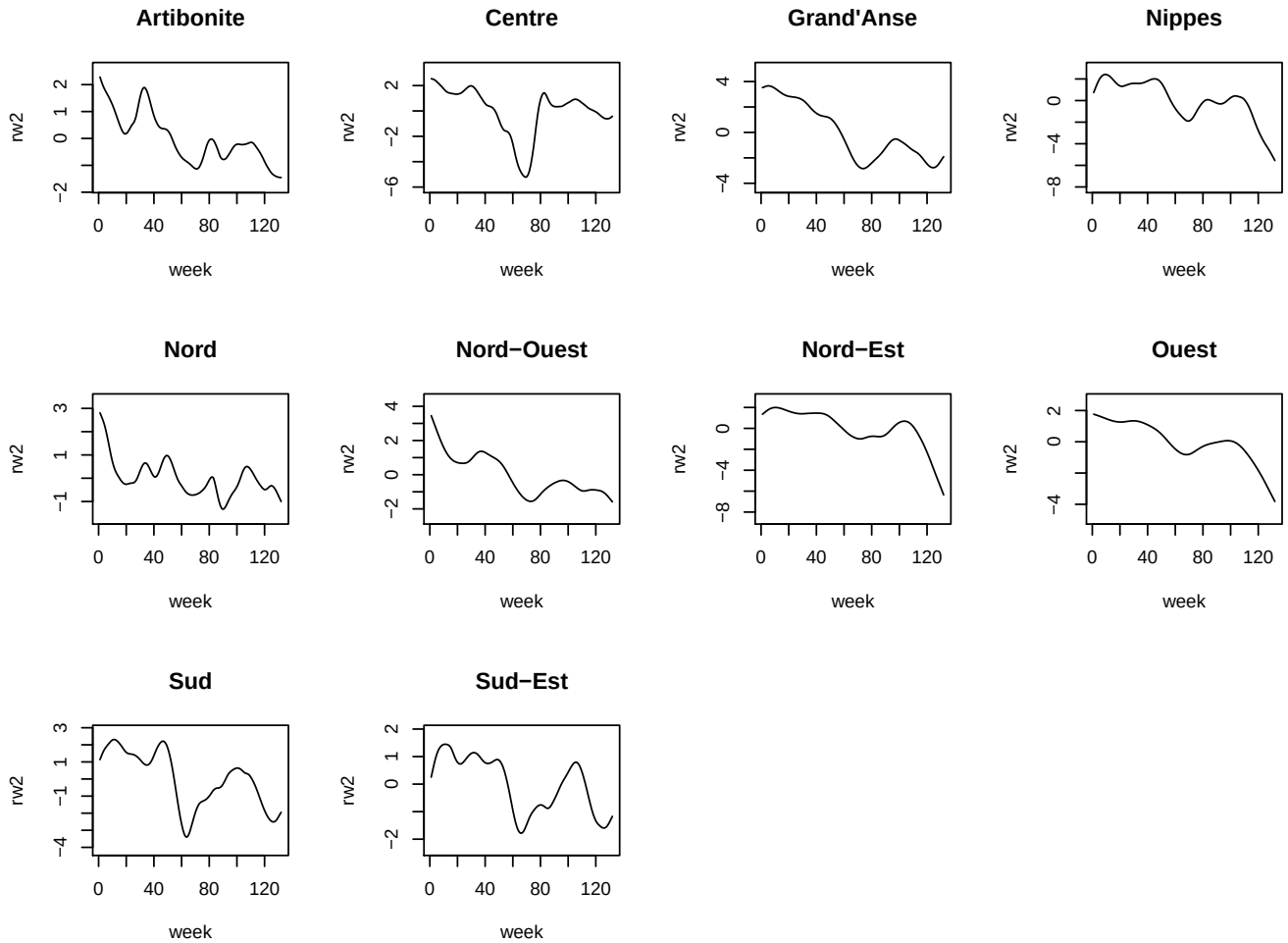

Figure 7: Fitted temporal trends of the latent Gaussian model

Results

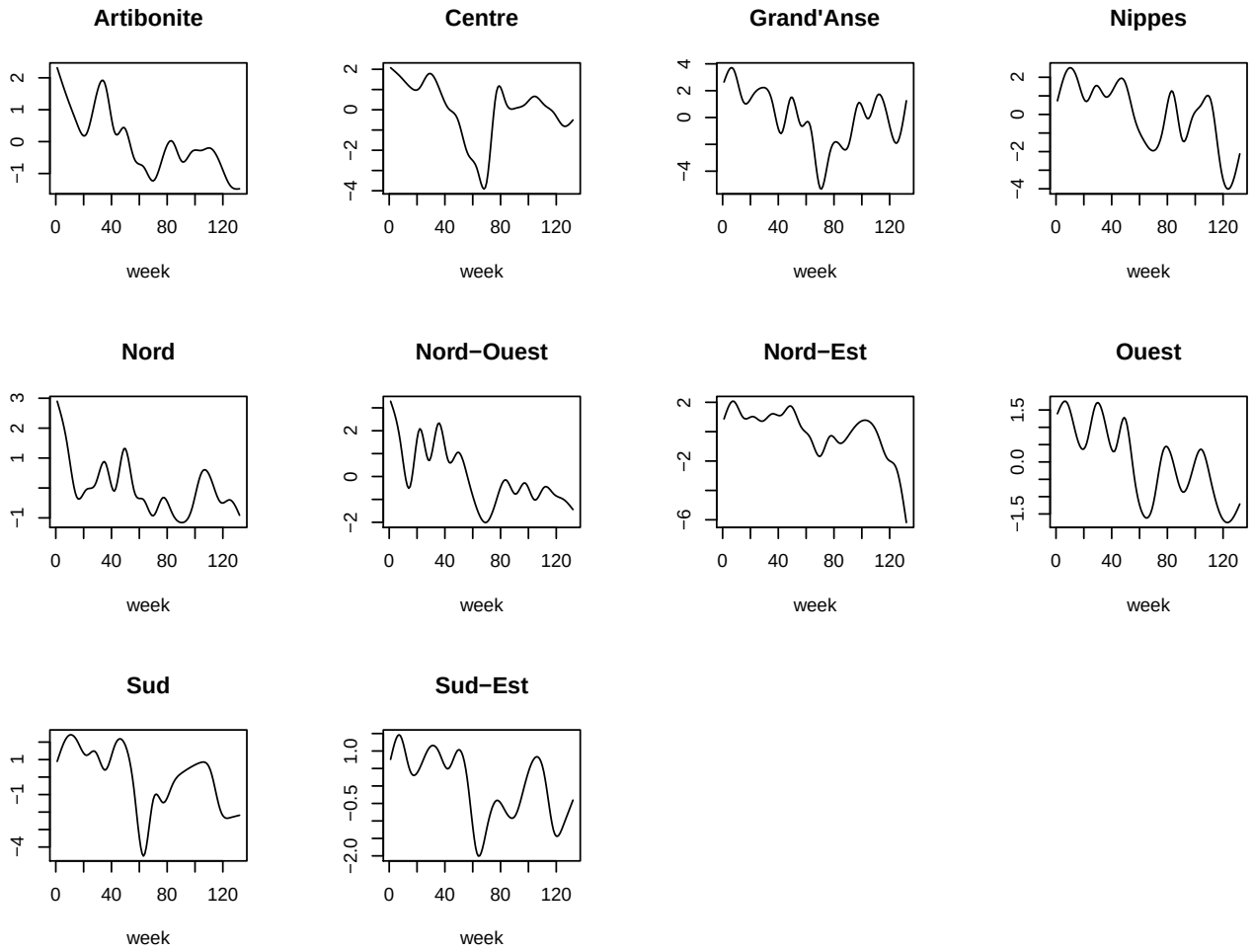

Figure 8: Fitted temporal trends of the first Bayesian linear mixed model

Conclusion

- For each department the fitted temporal trends obtained from all these models are quite similar. It suggests that these methods capture the real temporal trends consistently and they are implemented correctly.
- There is a dramatic decrease in cholera cases around week 70 for almost all departments.
- Some neighboring departments have very similar trends such as Sud and Sud-Est, Nord and Nord-Ouest. It implies that there exists some spatial correlation among departments.

Discussion

- Since there are only 10 departments in the cholera data, all the analysis are conducted for each department separately. If the number of areas we are dealing with is much larger, it is necessary to study the spatial relations and switch to models with spatial-temporal interactions.
- The curve of $\widehat{\text{MSPE}}$ is not smooth and it becomes quite unstable when K is large. Also, the values of $\widehat{\text{MSPE}}$ in the cross-validation are very large, which reflects that the prediction performance based on temporal trend only is not good. This might be caused by extrapolation. If we want to use these models to predict future cases, including more covariates might help.

Acknowledgments

- Dr. Nikolay Bliznyuk, Masters Advisor
- Dr. Larry Winner, Committee Member
- Xueying Tang and Hunter Merrill
- Thomas Weppelmann