SEMI-PARAMETRIC MODELING OF TEMPORAL TRENDS FOR CHOLERA OUTBREAK IN HAITI

By Shaojun Zhang

University of Florida

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.

In this report, under the framework of semi-parametric modeling, smoothing spline Poisson regression, zero-inflated Poisson regression, a latent Gaussian model and two Bayesian linear mixed models are applied to study the temporal trend of cholera cases for each department in Haiti. I find that the trends fitted by different models are similar to each other and departments that are geographically close tend to have similar temporal patterns.

1. Introduction. The ongoing Haiti cholera outbreak is the worst epidemic of cholera in recent history, according to the U.S. Centers for Disease Control and Prevention [1]. After the 2010 earthquake, in little over two years, as of August 2013, it had killed at least 8231 Haitians and hospitalized hundreds of thousands more while spreading to neighboring countries including the Dominican Republic and Cuba [2]. However, it's still not clear why Haiti suffered a cholera break after being apparently free of the disease for more than 100 years.

Some research have been done to model the cholera transmission dynamics. Gaudart et al. 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 [3]. Kirpich et al. applied an extended SIR model to study the transmission mechanism in Ouest department [4].

In this study, I use semi-parametric methods to quantify the temporal dynamics of cholera cases for each department in Haiti. Smoothing spline Poisson regression, zero-inflated Poisson regression, a latent Gaussian model and two Bayesian generalized linear mixed models are applied. The design matrix is formed by the basis functions of time and the dimension of the basis is determined by cross-validation with the smoothing spline Poisson regression model.

The outline of this report is as follows. In Section 2, the cholera data set is described. In Section 3, the statistical models are explained in detail. In

Section 4, the main results are given. In Section 5, some existing problems are discussed.

2. Data.

Haiti is divided into $N = 10$ departments as shown in Fig 1 and Table 1. 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) [5]. There are exactly $T = 132$ weeks in this time period.

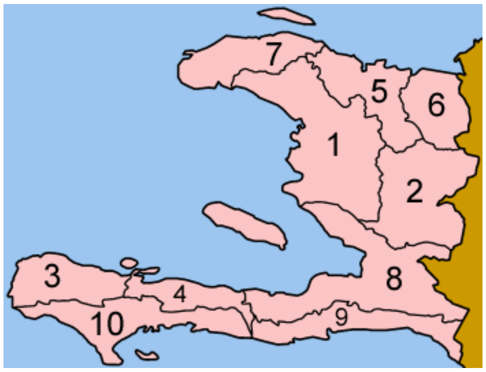

FIG 1. *Departments of Haiti*

https://en.wikipedia.org/wiki/Departments_of_Haiti#/media/File:Haiti_departments_numbered.png

TABLE 1
Names of departments

#	Department
1	Artibonite
2	Centre
3	Grand'Anse
4	Nippes
5	Nord
6	Nord-Est
7	Nord-Ouest
8	Ouest
9	Sud-Est
10	Sud

3. Methods.

3.1. *Smoothing spline Poisson regression.* Let Y_{it} denote the number of cholera cases in department i at week t for $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$.

Then the smoothing spline Poisson regression 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 [6].

Let $\{z_i(\cdot)\}_{i=1}^K$ denote the collection of nonconstant cubic spline basis functions where K is the dimension of the basis and let $\mathbf{Z}$ denote the design matrix consisting of the basis functions evaluated at each time point, that is

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

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 $\log \boldsymbol{\mu_i} = (\log \mu_{i1}, \log \mu_{i2}, \dots, \log \mu_{iT})^T$, $\mathbf{1} = (1, 1, \dots, 1)^T$ and $\boldsymbol{\beta_i} = (\beta_{i1}, \beta_{i2}, \dots, \beta_{iK})^T$. 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.

The parameters β_{i0} and $\boldsymbol{\beta_i}$ are estimated by penalized iteratively reweighted least squares (P-IRLS) and the smoothing parameter λ_i is chosen by the generalized cross validation (GCV) score [7].

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

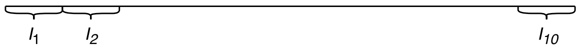

FIG 2. *Time intervals for cross-validation*

width, denoted by $\{I_j\}_{j=1}^{10}$ in chronological order as shown in Fig 2. The mean square prediction error (MSPE) is estimated as

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

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 $\widehat{\mathrm{MSPE}}$s for K from 10 to 40 are calculated and the K corresponding to the minimum $\widehat{\mathrm{MSPE}}$ is selected.

3.2. *Zero-inflated Poisson regression.* In the time period considered, there are a number of zero cases, especially for department Grand'Anse and Nord-Ouest as shown in Fig 3. Therefore, zero-inflated Poisson regression might be appropriate. Compared to Poisson regression, it puts an additional point mass of probability p_i at zero for department i. The model can be expressed as

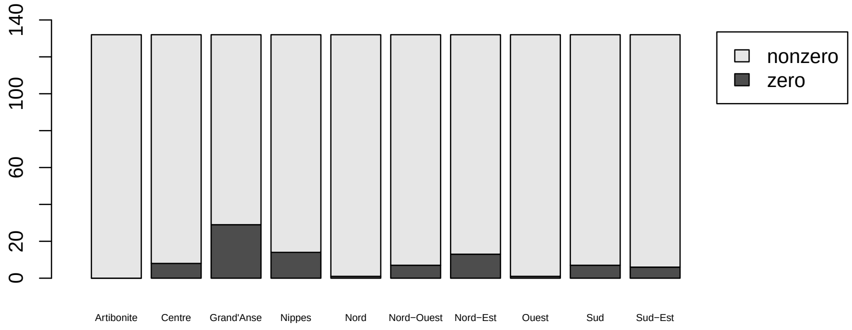

FIG 3. *Number of zero cases for each department*

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

[8] with log link function

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

3.3. *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 $g_i(\cdot)$.

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 [9]:

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

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 [10], which serves the same goal as the penalty term in the smoothing spline Poisson regression.

3.4. *Bayesian generalized linear mixed models.* Two Bayesian generalized linear mixed models are applied. Both models 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}

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

Their differences lie in the priors for $\boldsymbol{u_i}$. The first model assumes different covariance matrices for different departments,

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

with hyperprior

$$\sigma_{ui}^2 \sim \text{Gamma}(1, 1),$$

while the second model assumes common covariance matrix for all departments,

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

with hyperprior

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

Mathematically, the covariance matrix of $\boldsymbol{u_i}$ corresponds to the penalty term in the smoothing spline Poisson regression model. Therefore, from this perspective, their differences can be interpreted as that the first model allows different penalty terms for different departments and the second model uses the same penalty term for all departments.

4. Results. The number of basis functions K is determined by cross-validation for the smoothing spline Poisson regression model. The $\widehat{\text{MSPE}}$s for K from 10 to 40 are shown in Fig 4. It can be seen that the minimum is achieved when $K = 19$. The fitted temporal trends for all departments are shown in Fig 6.

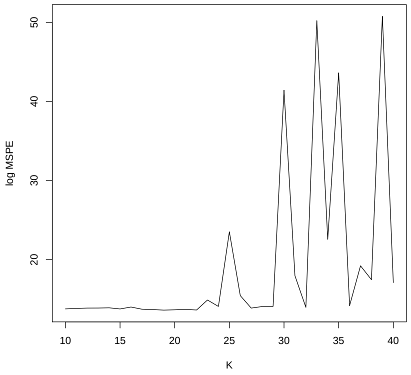

FIG 4. *Log $\widehat{MSPE}$ for different K*

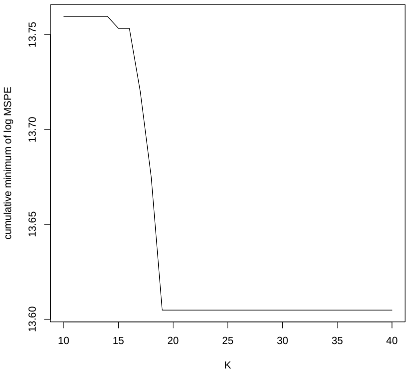

FIG 5. *Cumulative minimum of log $\widehat{MSPE}$*

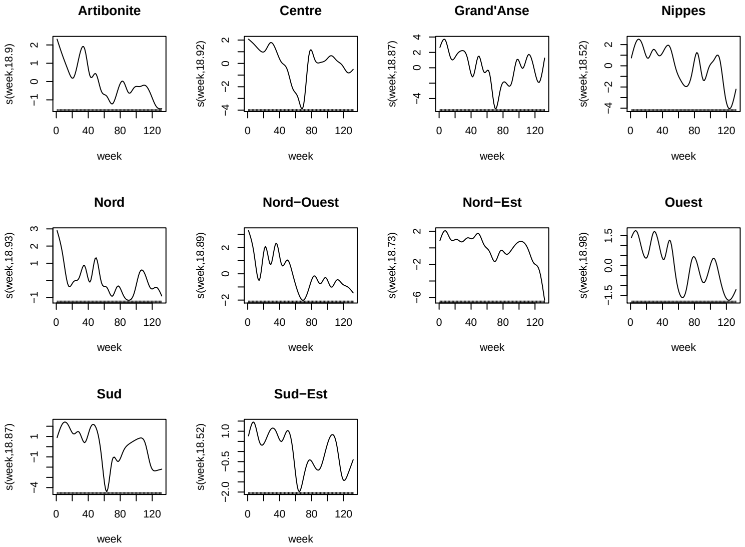

Fig 6. *Fitted temporal trends of the smoothing spline Poisson regression model*

The zero-inflated Poisson regression model is applied and Vuong's non-nested tests are conducted to check whether zero-inflated Poisson regression and standard Poisson regression are indistinguishable. The fitted point mass at zero and p-value for each department are shown in Table 2. Roughly speaking, p-value decreases as the fitted point mass at zero increases. Note that there are no zero cases for department Artibonite and thus Vuong's test is inapplicable and of course unnecessary. The fitted temporal trends for all departments are given in Fig 7.

Table 2
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

The latent Gaussian model is fitted by using integrated nested Laplace approximations (INLA), which enable us to perform computationally efficient approximate Bayesian inference [11]. The fitted temporal trend for each department is shown in Fig 8. As we can see, the trends are similar to those of the previous two models, but a little smoother.

Two Bayesian linear mixed models are implemented by MCMC with JAGS (Just Another Gibbs Sampler). The results of the two models are

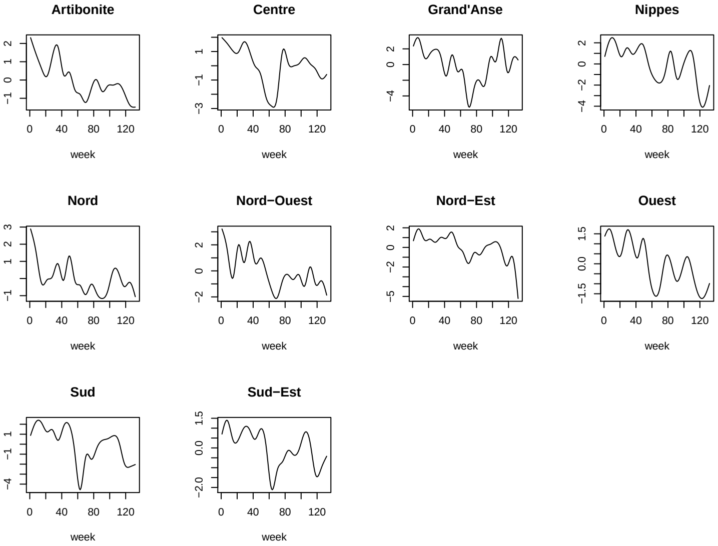

Fig 7. *Fitted temporal trends of the zero-inflated Poisson regression model*

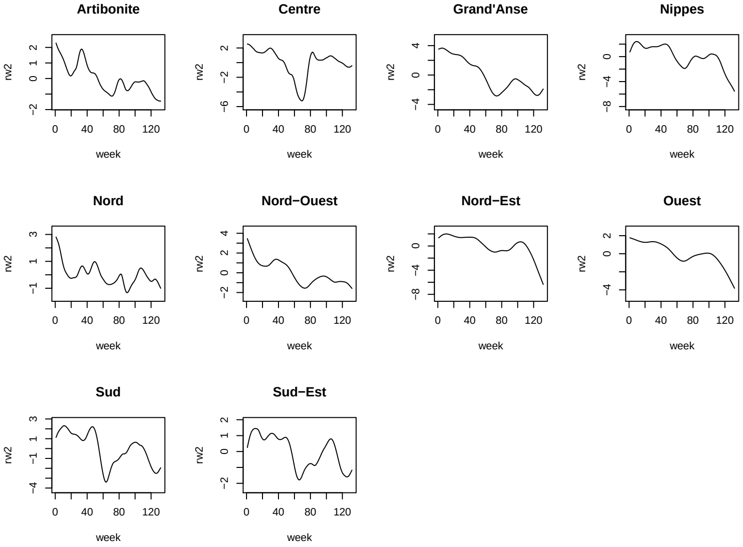

Fig 8. *Fitted temporal trends of the latent Gaussian model*

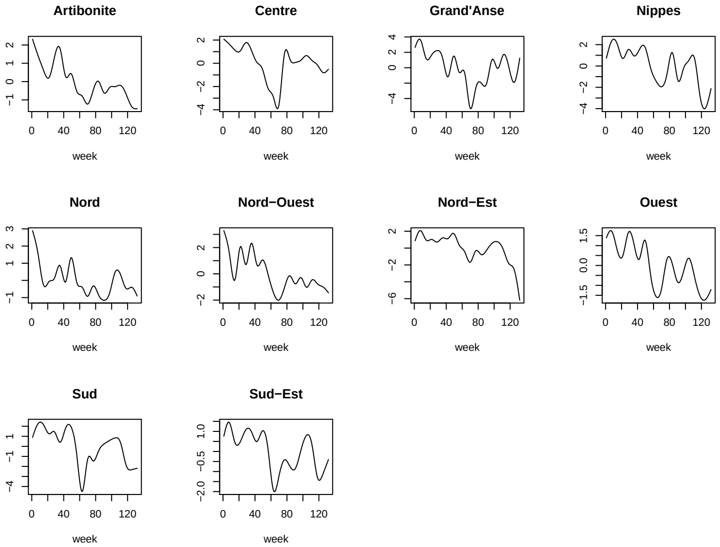

Fig 9. *Fitted temporal trends of the first Bayesian linear mixed model*

very close. The fitted temporal trends of the first model are shown in Fig 9.

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. Also, it can be seen that there is a dramatic decrease in cholera cases around week 70 for almost all departments.

Furthermore, 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.

5. Discussion. Since there are only 10 departments in the cholera data set, 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.

All computation for this report cost about 27 minutes on my laptop (2.5 GHz Intel Core i5, 4GB 1600 MHz DDR3). Most of the time is spent on

MCMC, where 3 parallel chains are constructed with 5000 iterations for adaptation and 5000 iterations for monitoring.

Acknowledgments. I would like to thank my Masters advisor, Nikolay Bliznyuk for his mentoring in statistical modeling. I would also like to thank Xueying Tang and Hunter Merrill for helping in Bayesian computation and Thomas Weppelmann for providing the cholera data set.

References.

[1] [Online]. Available: http://www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=28070+&Itemid=999999&lang=fr

[2] [Online]. Available: http://www.mspp.gouv.ht/site/downloads/Rapport%20%20Web%2012.08_Avec_Courbes_Departementales.pdf

[3] J. Gaudart, S. Rebaudet, R. Barrais, J. Boncy, B. Faucher, M. Piarroux, R. Magloire, G. Thimothe, and R. Piarroux, "Spatio-temporal dynamics of cholera during the first year of the epidemic in haiti," *PLoS neglected tropical diseases*, vol. 7, no. 4, p. e2145, 2013.

[4] A. Kirpich, T. a. Weppelmann, Y. Yang, A. Ali, J. G. Morris, and I. M. Longini, "Cholera Transmission in Ouest Department of Haiti: Dynamic Modeling and the Future of the Epidemic," *PLoS neglected tropical diseases*, vol. 9, no. 10, 2015.

[5] [Online]. Available: http://new.paho.org/hq/images/Atlas_IHR/CholeraHispaniola/atlas.html

[6] Analysis of mortality data using smoothing spline poisson regression. [Online]. Available: http://soa.org/library/research/actuarial-research-clearing-house/2006/january/arch06v40n1-ix.pdf

[7] S. N. Wood, *Generalized additive models : an introduction with R*. Chapman and Hall/CRC, 2006.

[8] D. Lambert, "Zero-Inflated Poisson Regression , with an Application to Defects in Manufacturing," *Technometrics*, vol. 34, no. 1, pp. 1–14, 1992.

[9] Random walk model of order 2 (rw2). [Online]. Available: http://www.math.ntnu.no/inla/r-inla.org/doc/latent/rw2.pdf

[10] F. Lindgren and H. Rue, "On the second-order random walk model for irregular locations," *Scandinavian Journal of Statistics*, vol. 35, no. 4, pp. 691–700, 2008. [Online]. Available: http://dx.doi.org/10.1111/j.1467-9469.2008.00610.x

[11] H. Rue, S. Martino, and N. Chopin, "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations," *Journal of the Royal Statistical Society: Series B*, vol. 71, no. 2, pp. 319–392, 2009.

[12] Tokyo rainfall data. [Online]. Available: https://www.math.ntnu.no/~hrue/r-inla.org/examples/tokyo/tokyo.pdf

[13] H. R. Merrill, "Bayesian Geostatistical Prediction of Soil Nitrogen in Florida, USA," Master's thesis, 2014.