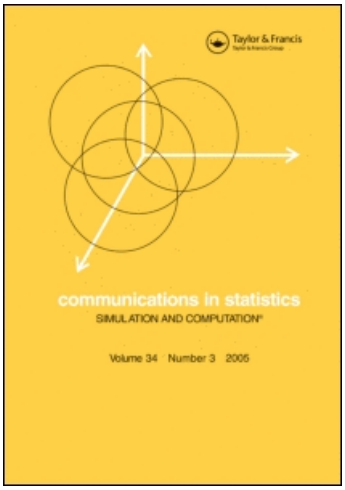
This article was downloaded by: *[Barrios, Erniel B.]*
On: *6 February 2011*
Access details: *Access Details: [subscription number 933193199]*
Publisher *Taylor & Francis*
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Communications in Statistics - Simulation and Computation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597237

Robust Estimation of a Spatiotemporal Model with Structural Change

Rowena F. Bastero[a]; Erniel B. Barrios[b]
[a] Department of Physical Sciences and Mathematics, University of the Philippines, Manila, Philippines
[b] School of Statistics, University of the Philippines, Quezon City, Philippines

Online publication date: 06 February 2011

To cite this Article Bastero, Rowena F. and Barrios, Erniel B.(2011) 'Robust Estimation of a Spatiotemporal Model with Structural Change', Communications in Statistics - Simulation and Computation, 40: 3, 448 — 468
To link to this Article: DOI: 10.1080/03610918.2010.543298
URL: http://dx.doi.org/10.1080/03610918.2010.543298

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.informaworld.com/terms-and-conditions-of-access.pdf

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Communications in Statistics—Simulation and Computation®, 40: 448–468, 2011
Copyright © Taylor & Francis Group, LLC
ISSN: 0361-0918 print/1532-4141 online
DOI: 10.1080/03610918.2010.543298

Downloaded By: [Barrios, Erniel B.] At: 23:39 6 February 2011

Robust Estimation of a Spatiotemporal Model with Structural Change

ROWENA F. BASTERO[1] AND ERNIEL B. BARRIOS[2]

[1]Department of Physical Sciences and Mathematics, University of the Philippines, Manila, Philippines
[2]School of Statistics, University of the Philippines, Quezon City, Philippines

A spatiotemporal model is postulated and estimated using a procedure that infuses the forward search algorithm and maximum likelihood estimation into the backfitting framework. The forward search algorithm filters the effect of temporary structural change in the estimation of covariate and spatial parameters. Simulation studies illustrate capability of the method in producing robust estimates of the parameters even in the presence of structural change. The method provides good model fit even for small sample sizes in short time series data and good predictions for a wide range of lengths of contamination periods and levels of severity of contamination.

Keywords Backfitting; Forward search; Robust estimates; Spatiotemporal model; Structural change.

Mathematics Subject Classification 62F35; 62J02.

1. Introduction

Modeling of disease prevalence and epidemics of infectious disease has focused on the dynamics over time and it was only recently that the spatial aspect was also considered. This is explained by the difficulty of obtaining data on the realization of an epidemic in the context of space and time dynamics. There are also computational difficulties hindering the parameterization of models including the interaction of spatial and temporal dependencies. With the availability of geographically indexed health and population data and advances in computing and statistical methodologies, a more realistic investigation of spatial variation in disease risk over time and space has become possible. There is an increasing interest in modeling of the spatiotemporal dynamics of prevalence data of infectious diseases in stochastic, spatially interacting populations. With space and time dependence incorporated in the model, statistical inference is more challenging, although this is necessary since it provides a more theoretically sound framework for modeling

Received November 16, 2010; Accepted November 23, 2010
Address correspondence to Erniel B. Barrios, School of Statistics, University of the Philippines, Quezon City 1101, Philippines; E-mail: ernielb@yahoo.com

that captures realistic process features and behavior (Gelfand, 2007) of the process involved.

An epidemic is the progress of a disease in time and space (Van Maanen and Xu, 2003). Occurrence of these epidemics or outbreaks in the population creates severe fluctuations in the prevalence of the disease in the general susceptible setting, resulting to possible structural change in the behavior of the model.

We postulate a model that takes into account the temporal and spatial dependencies like those exhibited by disease prevalence rates that are jointly determined by physical and geophysical conditions (covariates). This article develops an epidemic model that is flexible for both infected and non infected cases. We also propose an estimation procedure that is robust and computationally viable. The estimation procedure is iterative and combines the forward search algorithm and maximum likelihood estimation into the backfitting algorithm. The backfitting algorithm mitigates the convergence problem often encountered by the classical maximum likelihood estimation when there are numerous parameters in a nonlinear model. Atkinson and Riani (2007) emphasized the robustness of the forward search in a wide variety of statistical models. Buka et al. (1989) proved consistency and convergence of the backfitting algorithm in a relatively general class of smoothers in an additive model.

Insights in the dynamics of infectious diseases have gained much recognition as a key component in epidemiology and spatiotemporal modeling of infectious disease. The enormous public health concern inflicted by infectious diseases and outbreaks motivates the use of statistical modeling in increasing public awareness into its spread and transmission dynamics that can aid in mitigation. Modeling of the dynamics of disease prevalence enables the understanding of risk factors and consequently aids in the development of viable mitigation schemes, especially for future outbreaks or outbreaks in another location. Spatiotemporal modeling in epidemiology aims to understand the important determinants of epidemic development in order to develop sustainable schemes for strategic and tactical management of diseases. Developing countries usually experience some challenges in public health administration that requires space and time specific mitigation strategies, e.g., dengue and leptospirosis that becomes prevalent in depressed areas during heavy rainfalls.

Downloaded By: [Barrios, Erniel B.] At: 23:39 6 February 2011

2. The Backfitting Algorithm

The backfitting algorithm has been used in fitting an additive model. The algorithm cycles through the predictors and replaces each current function estimate by a curve based from smoothing a partial residual on each predictor (Hastie and Tibshirani, 1990). A smoother is used in summarizing the trend of a response measurement $y$as a function of one or more explanatory variables: $x_1, x_2, \ldots, x_p$. The smoother produces estimates that are less variable than y itself and could be non-parametric, allowing for a more "relaxed" estimation procedure since it does not assume a strict form for the dependence of the response variable $y$on the predictor variables $x_1, x_2, \ldots, x_p$.

In the additive model, the jth covariate has an associated component m_j, from the combinations of which the regression model is constructed. The m_j's are defined as arbitrary univariate functions, one for each predictor estimated through the

iterative scheme described as follows (Hastie and Tibshirani, 1990):

(i) Initialize: $\mu = ave(y_i)$, $m_j = m_j^0$, $j = 1, 2, \ldots, p$.
(ii) Cycle: $j = 1, 2, \ldots, p$.

$$\hat{m}_j = S_j \left[\frac{y - \sum_{k \neq j} m_k}{x_j} \right].$$

(iii) Continue the iterative process until the functions achieve convergence, that is, the functions no longer change. In this iteration, S_j is the smoothing matrix of the response variable against the different explanatory variables involved.

Buka et al. (1989) proved convergence and consistency of the backfitting in an additive model with a general smoother. Hastie and Tibshirani (1990) provided conditions under which the convergence of the backfitting algorithm is guaranteed. The backfitting procedure has been shown to work in the time-series context if the dependence structure is not quite strong (Chen and Tsay, 1993). Research on the backfitting algorithm has also dealt with its asymptotic properties and the convergence properties; see, for example, Opsomer (2000).

Downloaded By: [Barrios, Erniel B.] At: 23:39 6 February 2011

3. Forward Search Algorithm

The forward search algorithm is a powerful procedure for detecting multiple masked outliers, for discovering their underlying effects on models fitted to the data and for assessing the adequacy of the model (Atkinson and Riani, 2007). The method starts by fitting a small, robustly chosen subset of n observations from a total of N ($n < N$). The method moves "forward" to a larger subset by ordering the N residuals or other measure of closeness from the fitted model of n observations and using the $n + 1$ observations with the smallest closeness measure as the new larger subset. Usually, one observation is added to the subset at each step, but there are instances when two or more are added as at least one leave, an indicator of the introduction of some members of a cluster of outliers in the data set (Atkinson and Riani, 2002).

During the search, a series of parameter estimates are obtained which are robust from the beginning of the method to the usual least squares at the end. For outlier-free set of observations, it is expected that the parameters and plots of all N residuals continue to be stable as the number of unit in the subsets n increases. As a consequence of the search, observations that deviate from the fitted model are included towards the end of the search. These indicate the presence of potential outliers, influential subsets, or a systematic failure of the model presented.

4. Spatiotemporal Epidemic Model

The prevalence of a disease in the presence of outbreaks is characterized by spatiotemporal clustering of infection among the susceptible population. Certain epidemic cases may take place in adjacent locations or areas that are close to each other. The prevalence rates in close areas are expected to be in near approximations as they are similar in geographical distribution of population at risk and other scales defining the infection challenge. The occurrence of diseases on the same area may be due to their commonalities in terms of geographic, demographic, health, and social conditions. It is therefore logical to infer that these areas are homogeneous

Downloaded By: [Barrios, Erniel B.] At: 23:39 6 February 2011

in terms of environmental risks, quality of sanitation, population density and other socioeconomic factors. As a result of the dynamic nature of the outbreaks where the population at risk is constantly changing and the control treatments vary, it is imperative for these changes in spatial and temporal components of infection risk that occur over time to be included in the analysis. Hence, spatiotemporal models addressing the interactions between disease and the environment that is continuously evolving over time could be a useful tool in understanding and predicting the risk and spread of the disease.

The prevalence of highly contagious diseases can be affected by factors based on physical and geophysical conditions (covariates), information on the spread mechanism within the area with homogeneous conditions (spatial parameter) and a temporal measure that captures the temporary structural changes, as in the case of an epidemic outbreak at a specific time. A space-time interaction is necessary in understanding and characterizing the prevalence of a disease as it is generally dictated by conditions like covariates. Furthermore, the inclusion of structural change is necessary as there realistically exist in the dynamics of disease spread, that temporarily inflicts the population thereby affecting the disease rates at the susceptible setting.

4.1. *Postulated Model*

Given observations for N units and T time points, prevalence rate (Y_{it}) is postulated as a function of space, time and space-time interactions, represented by functions of X_{it}, W_{it}, and ε_{it}. We adopt the model of Landagan and Barrios (2007) to describe the condition of the general setting (epidemic-free) given by:

$$Y_{it} = \beta X_{it} + \gamma W_{it} + \varepsilon_{it} \quad i = 1, 2, \ldots, N, \;\; t = 1, 2, \ldots, T, \tag{1}$$

where,

Y_{it} = response variable from location i at time t
X_{it} = set of covariates from location i at time t
W_{it} = set of space and time interaction variables in the neighborhood system of location i at time t
ε_{it} = errors term for location i at time t.

The component βX_{it} assumes that spatial characteristics of the covariates will not vary significantly over time. As an example, if the covariate X_{it} is population density, this means that a highly densed area at $t = 0$ will remain highly dense at $t = 1, 2, \ldots, T$, hence, constant β over time. This assumption makes it possible to extend the time series using the same X_{it}. On one hand, W_{it} are variables that define the neighborhood system. Neighborhoods may be defined in terms of contiguous communities, e.g., mean regional expenditure on health facilities, number of community hospitals, amount of rainfall, mean regional population density, etc., can be considered as neighborhood variables with similar values among neighbors.

When outbreaks occur, it greatly influences volatility of the prevalence rate. Thus, Model (1) is modified to account for structural change as:

$$Y_{it} = \beta X_{it} + \gamma W_{it} + g_{i^*}(t^*, \lambda) + \varepsilon_{it} \quad i = 1, 2, \ldots, N, \;\; t = 1, 2, \ldots, T, \tag{2}$$

where $i^* \in N^d$, the set of spatial units experiencing the structural change (outbreak) and Y_{it} , $X_{it}W_{it}$, and ε_{it} are similarly defined as in Eq. (1) and $g(t^*, \lambda)$ is a function describing the dynamics of change (epidemics) in t^* contiguous time points between $t = 1, 2, \ldots, T$ where the change occurred, i.e., $t^* \subset t$. Further, $\delta_{it} = g_{i^*}(t^*, \lambda) + \varepsilon_{it}$ accounts for the temporal structure that incorporates the change dynamics into the model. Also, g_{i^*} is degenerate at zero for all units not affected by the intervention like an outbreak in an epidemics scenario.

Model (2) accounts for the oscillation in the time-space indexed response variable. The temporal component, defined as $\delta_{it} = g_{i^*}(t^*, \lambda) + \varepsilon_{it}$, is an additive term and a closed form of the progression in change caused by the intervention (e.g., growth dynamics) at time t^*. The error term accounts for disturbances in the model creating random shocks into the response variable (e.g., prevalence rate of the disease). Bacaer and Abdurahman (2008) reported that classical disease dynamics may be modelled as an exponential distribution during the latency and infectious rates. Thus, the function is defined as $g_{i^*}(t^*; \lambda) = \lambda_o \exp\{-\lambda_1 t^*\}$ where λ_0 is the baseline infectious rate and λ_1 is the infection rate from the susceptible to the infectious state $t^* = \{0, 1, 2, \ldots, \infty\}$ and assumes a zero value at the onset of the outbreak. It is be noted that as $t^* \to \infty$, the stochastic process $g_{i^*}(t^*; \lambda) \to 0$. This emphasizes that the structural change induced by the infectious period is temporary for the specific unit i, and the disease is non lethal. This distribution defines the jump in the realizations of the response variable that will eventually vanish over time.

In the presence of an outbreak, it is also logical to note that the progress of the disease will affect the community demographics and spatial features of the population. The model is further generalized as:

$$\begin{aligned}
Y_{it} &= \beta_d X_{it} + \gamma_d W_{it} + \lambda_{oi^*} e^{-\lambda_1 t^*} + \varepsilon_{it}, \quad \text{where} \\
\beta_d &= \beta I(i)_{\{i \in N^d\}} + \beta^* I(i)_{\{i \in N^d\}} \\
\gamma_d &= \gamma I(i)_{\{i \in N^d\}} + \gamma^* I(i)_{\{i \in N^d\}} \\
\varepsilon_d &= \rho \varepsilon_{it-1} + a_t \\
\lambda_{oi^*} &= 0, \quad \text{if the } i\text{th unit does exhibit any outbreak episode}
\end{aligned} \tag{3}$$

For $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$, model (3) combines models (1) and (2). The parameters β and γ are the original parameter values while β^* and γ^* are the temporary values due to the occurrence of an epidemic. This change in values of the parameters signifies the effect of the disease on the covariates and spatial dependencies of the model, respectively. The error component is investigated for temporal dependence or autoregression. Without loss of generality, assume that the error is an autoregressive process of order 1, given by

$$\varepsilon_{it} = \rho \varepsilon_{it-1} + a_{it}, \quad |\rho| < 1, \ a_{it} \sim IID\left(0, \sigma^2 a\right). \tag{4}$$

Moreover, it is assumed that clusters in N^d are identified a priori and that prior knowledge is available as to which clusters have been infected by the outbreak. The membership of N^d to the clusters is known, and that the progression of epidemics in each cluster is homogeneous within but possibly heterogeneous across clusters.

Downloaded By: [Barrios, Erniel B.] At: 23:39 6 February 2011

Downloaded By: [Barrios, Erniel B.] At: 23:39 6 February 2011

4.2. *Estimation in the Absence of Structural Change (Epidemic-Free Case)*

A modified iterative estimation procedure in estimating spatiotemporal models is proposed by infusing the forward search algorithm and maximum likelihood estimation into the backfitting algorithm. The performance of this procedure is evaluated on the postulated model with simulated data.

The general idea of the estimation procedure is to alternately estimate the parameters corresponding to the covariates β and the parameters corresponding to the spatial parameter γ through the forward search algorithm. The method can mitigate contamination that the ordinary least squares may possibly encounter during interventions; see Atkinson (2009). The temporary intervention (outbreak) effect λ_o, λ_1 are estimated using maximum likelihood on the residuals after the effect of X_{it} and W_{it} are set aside from Y_{it}. The parameter ρ is then estimated by recomputing the residuals after the effect of the outbreak dynamics is removed.

During a non infectious, epidemic-free time period, the prevalence rate may be modeled as a function of space and time with autocorrelated error terms, represented by model (1). The hybrid estimation procedure of backfitting and forward search algorithm is given below.

Step 1. The parameters β and γ are simultaneously estimated through the forward search algorithm. The forward search algorithm will generate robust estimates for β and γ. The steps are as follows.

i. From the N observations, choose a subset of size $n, n < N$, such that the sample is outlier-free and ideal to represent N locations. This is done by fitting the full data set on the model $Y_{it} = \beta X_{it} + \gamma W_{it}$. The choice of n observations corresponds to the n smallest residuals.
ii. Fit the model $Y_{it} = \beta X_{it} + \gamma W_{it}$ to the selected n observations and generate the parameter estimates $\hat{\beta}$ and $\hat{\gamma}$.
iii. Compute for the fitted Y_{it}, $\hat{Y}_{it}$ for all $i = 1, 2, \ldots, N - n$ and obtain the residuals $e_{it} = Y_{it} - \hat{\beta} X_{it} - \hat{\gamma} W_{it}$.
iv. From the $N - ne_{it}$'s computed, select one observation that corresponds to the smallest residual value.
v. Again, fit the model $Y_{it} = \beta X_{it} + \gamma W_{it}$ to the $n + 1$ observations. This procedure is repeated iteratively adding one observation at a time until all N locations have been included in the model or until the estimates are behaving differently based on some criteria, e.g., Cook's distance.

The forward search is used to obtain robust estimates of the spatial and covariates parameters. The forward search algorithm filters the influence of some observations in an iterative manner, upon convergence, the estimates that are not yet affected by a single or multiple influential observations are used. During periods of interventions like an outbreak, structural change could drive certain observations to exert influence on the model being estimated.

The residuals contain information on the temporal component that is initially ignored in this step. Optimality is achieved in this backfitting method since the covariate and the spatial dependencies are simultaneously estimated. One estimate of β and γ are computed for each time point and the T estimates are averaged to generate a single estimate $\hat{\beta}$ and $\hat{\gamma}$.

Step 2. Compute new residuals: $e_{it} = Y_{it} - \hat{Y}_{it}$, $\hat{Y}_{it} = \hat{\beta} X_{it} + \hat{\gamma} W_{it}$. Note that the resulting residuals will contain information on the true error and the temporal

parameter. Perform autoregression on these residuals to estimate the temporal parameter ρ.

One advantage of this estimation procedure is that it is able to optimize the parameters β and γ simultaneously. Furthermore, the convergence and uniqueness of the estimators for additive models are well-documented in the literature. In fact, it provides an exact solution to the projection equation, made suitable for any smoother matrix that is re-centered in nature (Opsomer, 2000).

4.3. *Estimation in the Presence of Structural Change (Epidemic Case)*

We aim to come up with robust estimates of model parameters in the presence of contamination due to the temporary structural change caused by the outbreaks (interventions). An outbreak is said to occur whenever disease levels exceed what is expected to be naturally occurring in a given community (e.g., neighbourhood, city, country, or region). The time of the occurrence of an intervention like an outbreak is assumed to be known. In epidemics, this could be proclaimed by the disease-monitoring committee.

This vanishing structural change characterized through outbreaks may be represented by an exponential infectious time $g(t*; \lambda) = \lambda_o \exp\{-\lambda_1 t*\}$. The mean value of the distribution is assumed to be equal to the removal rate of the disease in the epidemic model. Given the closed-form nature of the epidemic dynamic and its known likelihood function, the maximum likelihood method is optimal in estimating this model. Logically, incorporation of epidemics may result to alterations on the epidemic-free values of β and γ, as reflected by the generalized model (3). To investigate their behavior, an estimation procedure consisting of implementing a forward search and maximum likelihood procedures into the backfitting framework as described:

Step 1. The forward search algorithm is used to obtain robust estimate the parameters of β and γ simultaneously. The steps of the algorithm are as follows.

i. Choose a subset of size n, $n < N$ from N observations that is ideal and outlier-free for all the given locations. Fit the full data set on the model $Y_{it} = \beta X_{it} + \gamma W_{it}$. The choice of n observations corresponds to the n smallest residuals.
ii. Fit the model $Y_{it} = \beta X_{it} + \gamma W_{it}$ to the selected n observations and generate the parameter estimates $\hat{\beta}$ and $\hat{\gamma}$.
iii. Compute for the fitted Y_{it}, $\widehat{Y_{it}}$ for all $i = 1, 2, \ldots, N - n$ and obtain the residuals $e_{it} = Y_{it} - \hat{\beta} X_{it} - \hat{\gamma} W_{it}$.
iv. From the $N - ne_{it}$'s computed, select 1 observation corresponding to the smallest residual, without throwing away the information generated on the n observations initially considered.
v. Again, fit the model $Y_{it} = \beta X_{it} + \gamma W_{it}$ to the $n + 1$ selected observations. This procedure is repeated iteratively adding one observation at a time until all N locations have been included in the model or until the model is behaving wildly based on some diagnostic measure. In this case, the Cook's D is observed as the search progresses. The Cook's D is said to be influential it its value exceeds $\frac{4}{n}$ where n is the number of observations. The algorithm then stops if the Cook's D is no longer influential to the model based on this threshold.

The residuals still contain information on the temporal component and temporary structural change that is initially ignored in this step. On the assumption

Downloaded By: [Barrios, Erniel B.] At: 23:39 6 February 2011

of model additivity, optimality is expected in this backfitting method since the simultaneity of the covariate and the spatial dependencies are aptly accounted for. Estimates of β and γ are computed for each time point and the T estimates are averaged to generate a single estimate $\hat{\beta}$ and $\hat{\gamma}$.

Step 2. The parameters of the temporary structural change is then estimated through the maximum likelihood estimation using residuals from the previous step as the dependent variable. This will be implemented only on neighborhoods that are infected by the disease. It is therefore imperative that prior knowledge of the infected areas is available. A new set of residuals is computed as $e_{it} = Y_{it} - \widehat{Y}_{it}$ where $\widehat{Y}_{it} = \hat{\beta}X_{it} + \hat{\gamma}W_{it}$, and $\hat{\beta}$ and $\hat{\gamma}$ are the averaged estimates across all time points. For infected areas, we note that these residuals e_{it} will contain information on the temporary structural change and temporal component initially ignored in the Forward Search Algorithm in Step 1. The maximum likelihood estimates of λ_0 and λ_1 is generated only on infected neighborhoods. These estimates are also averaged through the computation of the harmonic mean of the raw estimates. The final residuals may then be computed as $e_{it} = Y_{it} - \widehat{Y}_{it}$ where $\widehat{Y}_{it} = \hat{\beta}X_{it} + \hat{\gamma}W_{it} + \hat{\lambda}_0 \exp\left\{-\hat{\lambda}_1 t\right\}$ for areas with outbreaks. Otherwise, the final residuals are defined by $e_{it} = Y_{it} - \widehat{Y}_{it}$, where $\widehat{Y}_{it} = \hat{\beta}X_{it} + \hat{\gamma}W_{it}$.

The MLE is used in this step due to its optimality given that the function is in closed-form and thus, have a known likelihood function. As a consequence, numerical maximization can be obtained easily. The estimates of λ_1 and λ_2 are also robust since the exponential function postulated to explain disease dynamics is quite flexible.

Step 3. Autoregression parameters will be estimated based on the residuals from Step 2.

These steps are implemented iteratively until parameters do not vary significantly. Also, the estimates are said to be robust if the estimates do not vary significantly from the true parameters even in the presence of temporary structural change.

In model (3), temporary structural changes are introduced, illustrated by the presence of an epidemic. In this procedure, the algorithm from the non epidemic case is further infused with the maximum likelihood to estimate the outbreak parameters. The forward search algorithm assures that the observations used in the estimation of covariate and spatial parameters, β and γ, respectively, are only those that exclude the temporary perturbations caused by the outbreak. This procedure is beneficial for this model since atypical observations are expected during the occurrence of an epidemic. The forward search algorithm guarantees robust estimation of the covariates and spatial parameters since outliers caused by the outbreak are eliminated. Similar to the epidemic-free algorithm, backfitting is computationally efficient since it minimizes the estimation load by only considering subsets of model parameters. The alternate removal of covariate, spatial and outbreak effects in the model also provides robust estimation of the temporal component that has been initially ignored in the process of backfitting estimation.

5. Simulation Study

The model along with the estimation procedure will be evaluated using simulated data in the balanced ($N = T$) and unbalanced ($T < N$) scenarios. Typical panels

Downloaded By: [Barrios, Erniel B.] At: 23:39 6 February 2011

involve a short span of time points for several individuals, i.e., unbalanced case. This means that asymptotic arguments are heavily reliant on the number of individuals approaching infinity (Hsiao, 1986). Also, in reality, it is difficult to compile long time-series since chance of attrition is heightened.

The simulation study aims to recreate the reality of the epidemic behavior and disease dynamics. An investigation of the robustness of parameter estimates is conducted on data sets that are nested on the following features: data with two vs. five clusters, all clusters are contaminated vs. only one cluster is contaminated, infection over short vs. long periods of time, changes in parameters of the covariates vs. no apparent change in parameters. The number of clusters, two or five, depicts the performance of the procedure whenever the population is divided into smaller number of susceptible groups or otherwise. Considering a fixed number of N units, dividing the population into two and five clusters will look at setting where each neighborhood is comprised of large and small number of spatial units, respectively. Meanwhile, the scope of the contamination over the neighborhoods are depicted by making a single cluster infectious or infected while likewise considering the case where all neighbors are affected by the epidemic. The instance where a single cluster is infected may be viewed as the endemic case, where the growth of disease occurs only within a confined locale. The scenario where all clusters are suffering from the outbreak is parallel to those disease shoot-ups that have been treated as national or international concerns due to its high-risk transmissions. In terms of the length of time, short and long contamination periods were considered. This presents the reality that some epidemics die down into the susceptible class faster than other epidemics. In this study, long contamination periods are defined by 50% of the time points affected while short contaminations are defined whenever the disease persist only during 25% of the time points. The introduction of a temporary structural change affects the covariate and spatial parameter. This is manifested by the change of value in the original parameter which may in fact serve as the indicator for disease severity. It is expected that the longer the difference of β and γ is to the actual value, the more severe the disease is, i.e., causing more deviant effects on these parameters. The simulation study will also look at the possibility that the epidemic will not affect any of the covariate and spatial features of the population. As a consequence, the case wherein no change is made to the parameters will also be included.

Furthermore, the behaviors of the estimates are considered for small and large sample sizes. For the common data set where $T < N$, cases on $T = 10, 20$ and $N = 25/26, 30, 50$ will be investigated. These six combinations generated from the values of T and N for the common data set feature the small and large sample cases.

Simulated data sets have eight different settings of time points and spatial units which depict the balanced and unbalanced, as well as, the small and large size, common features of most data sets. For each of these cases, eight unique set-ups, shall be investigated for both circumstances when the population is divided into two and five clusters. The simulation scenarios represents the following cases: (1) contamination in 1 cluster, short period, no change in parameters; (2) contamination in 1 cluster, short period, with change in parameters; (3) contamination in 1 cluster, long period, no change in parameters; (4) contamination in 1 cluster, long period, with change in parameters; (5) contamination in all clusters, short period, no change in parameters; (6) contamination in all clusters, short period, with change in parameters;

Downloaded By: [Barrios, Erniel B.] At: 23:39 6 February 2011

(7) contamination in all clusters, long period, no change in parameters; (8) contamination in all clusters, long period, with change in parameters.

The response variable $\boldsymbol{Y}$ was computed using Eq. (3). $\boldsymbol{X}$ was sampled from the Normal population with mean 10,000 and variance 1,000. To introduce spatial dependencies, the spatial units were divided into clusters or neighborhoods. There are cases where the units are divided into two clusters and in some cases, into five clusters. This was done by generating samples for the neighborhood system variable $\boldsymbol{W}$ from the Poisson distribution where each neighborhood would have a mean of 100 and 200 for the 2-cluster case and 100, 200, 300, 400, and 500 for the 5-cluster case. On the other hand, the error term was simulated from the AR(1) process with $\varepsilon_{it} = \rho\varepsilon_{it-1} + a_t$, $a_t \sim N(0, 1)$ with $\rho = 0.5$. The values of the coefficients were set as $\beta = 0.52$, $\gamma = 14.6$, $\lambda_0 = 4{,}80{,}000$, and $\lambda_1 = 2.5$. These values were chosen so that each component in the model would have significant contributions in the value of the response variable. The temporary structural change was induced through the change in values of β and γ as $\beta^* = 0.572$, $\gamma^* = 16.06$, depicting a 10% difference in the model parameters. Higher disease severity rates were also considered which results to larger differences in the original and temporary values of the covariate and spatial parameters. Specifically, 20%, 30%, and 40% differences were considered transforming β and γ to $\beta^* = 0.624$, $\gamma^* = 17.52$, and $\beta^* = 0.676$, $\gamma^* = 18.98$, and $\beta^* = 0.728$, $\gamma^* = 20.44$, respectively.

6. Results and Discussion

The performance of the hybrid algorithm of forward search, backfitting, and MLE was assessed by computing the absolute percent difference between the estimates and actual values of the parameters of the simulated data. Meanwhile, another set of estimates was obtained from the same simulated data using the MLE which treats the epidemic model as non-linear regression model. The same success measure was calculated for these estimates. Also, the predictive abilities of the two algorithms were compared by the mean absolute prediction error (MAPE).

6.1. *Absence of Structural Change*

Considering Model (1) that depicts no structural change (absence of an outbreak), 16 datasets were simulated. These data represent the benchmark case and will be used to investigate the efficiency of the proposed method in the absence of structural change. Table 1 show the MAPE and absolute percent difference of the estimated parameters from the true values.

In both balanced and unbalanced data sets, the hybrid estimation method produces desirable estimates for the covariate and spatial parameters. The forward search method clearly provides optimal estimates for β and γ, as seen by the minimal absolute percent difference between the estimates and true parameter values. However, the hybrid procedure failed to generate robust estimates for the temporal parameter ρ, for balanced and unbalanced data sets. The estimation of the temporal parameter is performed poorly as evident in the large absolute percent differences for the true value and estimate of ρ, even leading to a 100% underestimation of its true value in some cases. However, the small values of the MAPE indicate good predictive ability of the model in both types of data.

Downloaded By: [Barrios, Erniel B.] At: 23:39 6 February 2011

Table 1
Estimates under the non-epidemic case

Balanced data set ($T = N$)									
		Percent difference between estimates and true parameters (%)							
		β		γ		ρ		MAPE	
Scenarios		Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid	MLE
Small data set	2 clusters	0.017	0.001	0.005	0.002	103.036	14.707	0.017	0.019
($T = 20, N = 20$)	5 clusters	0.003	0.001	0.002	0.000	94.946	4.243	0.014	0.001
Large data set	2 clusters	0.003	0.000	0.001	0.001	93.965	1.569	0.016	0.000
($T = 50, N = 50$)	5 clusters	0.001	0.000	0.001	0.000	89.055	2.642	0.012	0.001
$T = 10$,	2 clusters	0.006	0.001	0.002	0.001	102.590	3.630	0.017	0.001
$N = 26/25$	5 clusters	0.005	0.000	0.003	0.001	105.392	23.493	0.011	0.001
$T = 10, N = 30$	2 clusters	0.011	0.002	0.003	0.003	102.561	27.949	0.019	0.004
	5 clusters	0.001	0.003	0.000	0.002	97.906	17.947	0.015	0.001
$T = 10, N = 50$	2 clusters	0.005	0.002	0.000	0.006	90.226	7.098	0.017	0.005
	5 clusters	0.002	0.001	0.001	0.000	84.803	8.704	0.014	0.001
$T = 20$,	2 clusters	0.010	0.005	0.003	0.007	85.713	0.378	0.017	0.003
$N = 26/25$	5 clusters	0.003	0.001	0.003	0.000	111.583	2.126	0.012	0.001
$T = 20, N = 30$	2 clusters	0.011	0.000	0.007	0.000	100.080	6.687	0.016	0.001
	5 clusters	0.008	0.002	0.006	0.002	89.839	3.844	0.014	0.002
$T = 20, N = 50$	2 clusters	0.010	0.004	0.007	0.008	101.233	4.260	0.014	0.002
	5 clusters	0.003	0.001	0.000	0.001	88.375	2.912	0.012	0.001

Downloaded By: [Barrios, Erniel B.] At: 23:39 6 February 2011

In general, this establishes that the forward search offers optimal solutions to the estimation of β and γ.

Focusing on balanced data sets and the effect of sample size on the hybrid method, smaller yet comparable absolute percent differences are realized over large, balanced data sets than small, balanced data sets for all parameters involved in the estimation method. This displays the efficiency of the proposed method in estimation when larger number of observations and longer time periods are involved. This is consistent with large sample theory for panel data. In the epidemiological setting, large sample theory is difficult to achieve since it requires larger cohorts, longer follow-ups and better review programs of health status in several geographical areas. Nonetheless, robust estimates are generated, regardless of sample size, on the spatial and covariate parameters. Although the temporal component remains poorly estimated, the model fit, evaluated through the MAPE, indicates good predictive ability of the hybrid model in small and large data sets. This emphasizes the advantage of the proposed method as it provides an efficient estimation procedure even with small sample sizes which is easier to collect in the epidemiological setting.

In unbalanced data, robust estimates for the covariate parameter β and spatial parameter γ are obtained for all combinations of N spatial units and T time points considered in the simulation study. This signifies the capability of the forward search to estimate the parameter values given small number of time points and observations. An increase in N and T provide comparable estimates to the MLE.

The temporal component ρ remains poorly estimated. This exhibits the failure of the proposed method in properly estimating the temporal aspect of the epidemic model in an epidemic-free state. It is possible that temporal dependencies had been properly accounted by the epidemic component of the model, leaving the residuals almost a white noise when the parameter ρ was estimated.

In general, the MLE procedure demonstrates a slight advantage in estimating the parameters in the absence of structural change (epidemic-free model). It must be noted however, that the estimation of β and γ through the forward search yield comparable values to the MLE, which ascertains its capacity to provide decent estimates. The use of forward search holds more promise in the non epidemic case since it only utilizes observations that do not greatly affect the model fit. Comparable MAPE's are likewise computed which indicate no apparent advantage of the MLE in terms of model fit. Furthermore, while the MLE procedure may have a slight advantage over the hybrid method in the estimation of the temporal component, it can easily suffer as the model is filled with too many variables. The MLE algorithm suffers from convergence problems when several parameters are involved. Hence, the proposed hybrid method poses to be more beneficial especially in the extension of the epidemic-free model to more covariate and spatial indicators.

Downloaded By: [Barrios, Erniel B.] At: 23:39 6 February 2011

6.2. *With Structural Change (Epidemic Case)*

In the epidemic case, estimation of outbreak parameters λ_0 and λ_1 has been incorporated into the two algorithms. This component represents the temporary structural change that causes atypical values in the data. The dynamics of the epidemics have been recreated in such a way that it illustrates the instances where the outbreak poses a threat over a long period of time and those where the outbreaks are easily treated and the population quickly recovers from the threat. Also, there are cases when the outbreak becomes concentrated only on certain areas while there are those that infest the entire population. This is incorporated by allowing the simulation to induce the outbreak in only one or in all clusters (neighborhood). Another setting that happens in the presence of an epidemic is that the covariates and spatial parameters are affected. The outbreak could cause a change in the effect of household size (covariate) or the mean family expenditure of the neighborhood (spatial parameter) in explaining the prevalence rate.

6.2.1. Contamination in One Cluster. The case when only one cluster is contaminated looks at the reality that the outbreak is endemic within a certain neighborhood. This could be due to isolated incidents leading to a sudden increase in magnitude of cases and prevalence rates in certain communities. For instance, the outbreak in leptospirosis may be realized only in the community (cluster) where heavy floods occurred, sparing those areas that have not suffered from flooding.

For balanced data, the estimates are generally robust, i.e., small absolute percent differences of estimates and true values of β and γ for all cases except whenever the contamination occurs over a long period of time with changes in the parameter values and the population is divided into two clusters only. Moreover, the estimates become poorer as the epidemic affects the covariates and spatial parameter more seriously. This difficulty is not encountered whenever the population is divided into more clusters. Small percent differences are computed between the estimates and the true values of the outbreak parameters. With respect to the temporal parameter

in the two-cluster case, small differences are also computed whenever no change in β and γ are assumed in the data simulation. Remarkably large percent differences between the estimates and the true values are computed whenever a change in covariate and spatial parameters exist. Meanwhile, when five clusters are involved, robust estimates are generated. This implies that the hybrid procedure has a slight advantage when more clusters are predefined and in effect, fewer spatial units are infected by the outbreak. When the population is divided into more neighborhoods, the chance of contamination declines since units in neighborhoods will be isolated from the contamination contained in another neighborhood. Although the MAPE is within the acceptable range, it can be concluded that the predictive ability of the model relatively decreases whenever the epidemic occurs within a long period of time and the covariate and spatial parameters are affected; see Tables 2–3 for details.

Downloaded By: [Barrios, Erniel B.] At: 23:39 6 February 2011

Unbalanced data sets with two clusters encounter the same difficulty in estimating the spatial and covariate parameters whenever structural changes occur in its true parameter due to an outbreak that exists for a long time. As the contamination rates on the said parameter increases, wilder estimates are achieved as evident in the increasing absolute percent differences between the true values and estimates for β and γ. The outbreak parameters, on one hand are estimated with minimal percent difference from the true value. Still, the hybrid estimation provides poor estimates of the temporal parameter when changes in the parameters are involved or in some instances, when long contamination periods are realized. This means that as the disease become more persistent (long contamination periods) or high-risk (high contamination rates on the spatial and covariate) in the two-cluster segregation of the population, the hybrid method is unable to properly estimate the temporal parameter. However, the estimated models generated through the hybrid method are superior in terms of model fit regardless of disease persistence and risk.

Meanwhile, the unbalanced data sets with five clusters generally provide robust estimates for the spatial, covariate, outbreak and temporal parameters, especially in cases involving five clusters. This shows the advantage of the hybrid method in cases when larger numbers of clusters are involved. Such clustering scheme is more realistic in epidemiology. Outbreak programs are made more efficient when the population is divided into several geographical clusters, which aids in more efficient identification, declaration and prevention of disease schemes. Thus, the hybrid method presents beneficial results as it is proven to be computationally-efficient and robust even in the presence of structural changes during instances when more pre-defined clusters are involved. The MAPE also conveys the predictive gain in the use of the hybrid method. The small MAPE values show that predicted responses of the estimated models from hybrid procedure are close to the actual observations of the response variable. Hence, the proposed estimation method is indeed optimal; see Table 4 for details of results for unbalanced data.

Meanwhile, it is noted that the poor estimates of the temporal component may be attributed to the fact that backfitting produces relatively better estimates for parameter subsets estimated at the initial stage than those estimated during the later stages of the iterative process.

The stability of the estimates during the five-cluster scenario over the two-cluster scenario may be attributed to the fact that given the former, less number of spatial units is infected with the disease. Given a fixed sample size N, the five-cluster case would have $\frac{N}{5}$ number of infected units as opposed to the two-cluster case where $\frac{N}{2}$ spatial units are affected by the disease. Hence, greater number of

Downloaded By: [Barrios, Erniel B.] At: 23:39 6 February 2011

Table 2
Estimates in small balanced data ($T = 20$, $N = 20$)

Scenarios		Percent difference between estimates and parameters (%)											
		β		γ		λ_0		λ_1		ρ		MAPE	
		Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid	MLE
							Two clusters						
Case 1:		0.00	0.00	0.01	0.00	0.00	0.00	3.00	0.00	4.40	4.68	0.00	0.00
Case 2:	10%	0.01	0.00	0.02	0.01	1.98	1.98	0.87	0.87	46.08	45.30	0.59	0.59
	20%	0.00	0.00	0.01	0.00	3.90	3.90	1.73	1.73	44.94	46.78	1.08	1.09
	30%	0.00	0.00	0.00	0.01	5.78	5.78	2.58	2.58	33.52	33.17	1.51	1.51
	40%	0.00	0.00	0.00	0.01	7.55	7.55	3.41	3.41	37.99	39.14	1.87	1.87
Case 3:		0.01	0.01	0.04	0.01	0.00	0.00	0.00	0.00	19.59	19.39	0.00	0.00
Case 4:	10%	11.98	11.73	20.70	20.56	1.04	1.08	0.45	0.47	51.05	51.16	2.35	2.33
	20%	22.80	22.72	38.67	38.30	2.13	2.13	0.94	0.94	49.06	43.19	4.47	4.49
	30%	36.59	37.11	61.41	64.07	2.96	2.99	1.30	1.31	50.78	45.10	6.89	6.80
	40%	45.59	46.57	77.84	81.38	4.38	4.39	1.93	1.93	34.39	32.93	8.61	8.54
							Five clusters						
Case 1:		0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	9.94	10.16	0.00	0.00
Case 2:	10%	0.00	0.00	0.00	0.00	1.99	1.99	0.87	0.87	8.92	8.82	0.24	0.24
	20%	0.01	0.00	0.01	0.01	3.92	3.92	1.74	1.74	11.50	11.52	0.43	0.43
	30%	0.01	0.00	0.01	0.00	5.83	5.83	2.60	2.60	3.20	2.41	0.60	0.60
	40%	0.00	0.00	0.00	0.00	7.35	7.35	3.32	3.32	7.96	7.97	0.74	0.74
Case 3:		0.01	0.01	0.02	0.01	0.00	0.00	0.00	0.00	9.32	8.20	0.00	0.00
Case 4:	10%	2.15	4.93	1.90	4.36	1.75	1.43	0.77	0.62	3.25	6.19	0.99	1.37
	20%	4.16	9.59	3.69	8.41	3.58	3.00	1.58	1.32	13.01	14.03	1.84	2.59
	30%	6.32	14.75	5.52	12.83	5.09	4.15	2.27	1.84	30.62	31.34	2.66	3.87
	40%	8.18	18.82	7.22	16.57	6.77	5.66	3.03	2.51	14.69	17.36	3.32	4.85

Case 1: contamination in 1 cluster, short period, no change in parameters.
Case 2: contamination in 1 cluster, short period, with change in parameters.
Case 3: contamination in 1 cluster, long period, no change in parameters.
Case 4: contamination in 1 cluster, long period, with change in parameters.

Downloaded By: [Barrios, Erniel B.] At: 23:39 6 February 2011

Table 3
Estimates in large balanced data ($T = 50$, $N = 50$)

Scenarios		Percent difference between estimates and parameters (%)										MAPE	
		β		γ		λ_0		λ_1		ρ			
		Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid	MLE
						Two clusters							
Case 1:		0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	1.19	1.36	0.00	0.00
Case 2:	10%	0.00	0.00	0.01	0.01	2.02	2.02	0.88	0.88	46.72	46.78	0.97	0.97
	20%	0.00	0.00	0.01	0.00	3.90	3.90	1.72	1.72	52.76	52.08	1.77	1.77
	30%	0.00	0.00	0.01	0.00	5.77	5.77	2.58	2.58	48.69	48.49	2.45	2.45
	40%	0.01	0.00	0.01	0.01	7.57	7.57	3.41	3.41	47.94	48.30	3.04	3.04
Case 3:		0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.55	0.33	0.00	0.00
Case 4:	10%	23.21	11.70	39.49	20.19	0.09	1.04	0.03	0.45	30.63	37.56	3.07	2.50
	20%	22.98	23.52	39.21	40.10	2.22	2.18	0.96	0.95	36.90	21.68	4.82	4.84
	30%	34.52	34.86	58.89	60.31	3.30	3.31	1.44	1.44	33.09	34.86	6.98	6.93
	40%	47.27	46.34	80.83	80.38	4.12	4.44	1.82	1.97	34.58	29.84	9.13	8.97
						Five clusters							
Case 1:		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	6.21	6.22	0.00	0.00
Case 2:	10%	0.00	0.00	0.00	0.00	2.03	2.03	0.89	0.89	4.71	4.70	0.39	0.39
	20%	0.00	0.00	0.00	0.00	3.95	3.95	1.74	1.74	4.02	4.03	0.71	1.55
	30%	0.00	0.00	0.00	0.00	5.85	5.85	2.62	2.62	0.47	0.48	0.98	0.98
	40%	0.01	0.00	0.01	0.00	7.55	7.55	3.41	3.41	6.48	6.48	1.21	1.21
Case 3:		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.93	4.85	0.00	0.00
Case 4:	10%	0.01	4.85	0.01	4.23	2.02	1.45	0.88	0.63	3.31	2.01	0.82	1.44
	20%	0.01	9.77	0.01	8.62	3.89	2.78	1.73	1.23	9.96	13.23	1.51	2.79
	30%	0.00	14.67	0.01	12.84	5.80	4.25	2.59	1.87	5.17	5.84	2.09	4.09
	40%	0.01	18.74	0.01	16.28	7.47	5.39	3.38	2.41	4.44	0.51	2.59	5.17

Case 1: contamination in 1 cluster, short period, no change in parameters.
Case 2: contamination in 1 cluster, short period, with change in parameters.
Case 3: contamination in 1 cluster, long period, no change in parameters.
Case 4: contamination in 1 cluster, long period, with change in parameters.

Downloaded By: [Barrios, Erniel B.] At: 23:39 6 February 2011

Table 4
Estimates in unbalanced data ($T = 10, N = 26/25$)

Scenarios		Percent difference between estimates and parameters (%)											
		β		γ		λ_0		λ_1		ρ		MAPE	
		Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid	MLE
						Two clusters							
Case 1:		0.01	0.00	0.02	0.01	0.00	0.00	0.00	0.00	2.36	1.89	0.00	0.00
Case 2:	10%	0.02	0.00	0.03	0.01	1.96	1.96	0.86	0.86	18.26	18.08	0.29	0.29
	20%	0.01	0.01	0.04	0.01	3.84	3.84	1.70	1.70	0.98	5.59	0.54	0.53
	30%	0.00	0.00	0.01	0.00	5.62	5.62	2.51	2.51	0.75	1.47	0.75	0.75
	40%	0.01	0.00	0.01	0.01	7.46	7.46	3.37	3.37	16.25	16.29	0.96	0.96
Case 3:		0.35	0.36	0.64	0.66	0.03	0.03	0.01	0.01	52.10	54.72	0.06	0.06
Case 4:	10%	14.34	13.50	26.30	24.09	0.94	0.98	0.42	0.43	78.01	77.40	2.20	2.15
	20%	26.45	26.21	47.97	47.15	2.10	2.09	0.91	0.91	64.98	63.19	4.07	4.07
	30%	37.25	37.98	66.32	68.59	3.24	3.23	1.40	1.39	23.76	26.80	5.84	5.87
	40%	49.25	49.17	85.73	84.90	4.14	4.12	1.82	1.81	59.05	62.39	7.79	7.83
						Five clusters							
Case 1:		0.01	0.01	0.02	0.01	0.00	0.00	0.00	0.00	9.02	10.68	0.00	0.00
Case 2:	10%	0.01	0.01	0.01	0.01	1.91	1.91	0.84	0.84	14.59	14.59	0.12	0.12
	20%	0.01	0.00	0.01	0.00	3.87	3.87	1.72	1.72	17.10	16.74	0.22	0.22
	30%	0.00	0.01	0.00	0.01	5.55	5.55	2.49	2.49	31.86	31.52	0.31	0.31
	40%	0.01	0.00	0.01	0.00	7.44	7.44	3.37	3.37	6.34	6.16	0.38	0.38
Case 3:		0.09	0.14	0.08	0.13	0.01	0.02	0.01	0.01	19.40	16.50	0.02	0.03
Case 4:	10%	3.27	5.30	2.93	4.73	1.59	1.35	0.70	0.59	15.27	17.28	0.97	2.13
	20%	6.32	10.26	5.57	9.00	3.28	2.85	1.45	1.25	19.72	18.68	1.82	2.42
	30%	10.02	16.00	9.20	14.65	4.65	4.00	2.07	1.77	7.76	7.69	2.70	3.59
	40%	12.75	20.50	11.27	18.00	6.15	5.31	2.75	2.36	30.21	13.57	3.41	4.58

Case 1: contamination in 1 cluster, short period, no change in parameters.
Case 2: contamination in 1 cluster, short period, with change in parameters.
Case 3: contamination in 1 cluster, long period, no change in parameters.
Case 4: contamination in 1 cluster, long period, with change in parameters.

Downloaded By: [Barrios, Erniel B.] At: 23:39 6 February 2011

spatial units is affected whenever the population is divided into smaller number of neighborhoods. As an inherent consequence, a higher number of atypical observations is observed during such clustering scheme and as such, may lead to poorer estimates. This implies that the efficiency of the five-clustering scheme may be attributed to the minimal number of spatial units contaminated due to the outbreak which produces smaller fluctuations in the data set relative to that of the two-clustering scheme of the population.

Furthermore, looking at the results of the balanced data sets, comparable results between small and large data sets are detected for the two-cluster case. Therefore, same robustness levels are produced for the spatial, covariate, and outbreak parameter in both small and large data sets. However, the two data sets encounter similar estimation difficulty of the temporal parameter when the observation suffers from structural changes. Nonetheless, the small MAPE values assure good model fit of the estimated models acquired through the proposed estimation method. Meanwhile, in the five-cluster case, a notable gain in the estimation of β and γ for large data sets is produced by the hybrid method over the small data sets in the case that the outbreak exists for a long time and has posed structural changes on the said parameters. Hence, the forward search is able to perfectly capture the actual parameter values whenever large balanced data sets are collected with a large number of clusters is involved. The temporal estimation also is more beneficial given large data sets. This supports the efficiency of the proposed method for large sample sizes. In terms of the model fit, both small and large sample sizes are comparable given its closely similar MAPE values. These results indicate that while the hybrid method has minor advantages for large samples over small samples, the estimates are generally robust and efficient for both natures of the sample size.

The efficiency of the hybrid method in small sample sizes and short time periods is among the advantages of this method. This is especially useful in the field of epidemiology where public health costs are ideally minimized through the number of individuals studied and shorter follow-up periods are proposed to avoid higher attrition rates.

Comparison is made between the estimates of the proposed hybrid estimation method and the maximum like estimation method that treats the generalized epidemic model as a nonlinear regression model.

For balanced and unbalanced data sets, comparable estimates are computed for both estimation methods as seen in the negligible differences between absolute percent differences and MAPE of the two procedures. However, some simulation scenarios demonstrate better estimates generated by the hybrid method over the MLE. All of these scenarios assume that the epidemics occurred over a long period of time, infecting a single cluster where the spatial and covariate parameters have been affected through different contamination rates. This includes the case of an unbalanced data set with 10 time points and 25 spatial units divided into 5 clusters. The forward searched estimates illustrate at most 8% reduction in absolute percent differences over the ML estimates. This signifies the superior capability of the proposed method to produce estimates that are robust especially when there is a structural change. Another scenario that illustrates this point is depicted in the case of an unbalanced data set with 10 time points and 50 spatial units divided into 5 neighborhoods. Better spatial and covariate parameters are attributed to the proposed hybrid method, there is a 15% reduction in absolute percent differences when contamination rates are more severe across β and γ. The MAPE, on one hand,

has a 2% improvement in favor of the backfitting method. The last scenario with apparent superior yields for backfitting estimates over ML estimates occur in five cluster division of a population of size 50 observed through 20 time points. At most, 20% is reduced from the absolute percent difference of the ML estimate by the backfitting method for the spatial and covariate parameters. Similar improvement of 2% in a 40% contamination rate is observed in the backfitting method.

6.2.2. Contamination in All Clusters. Another consideration in this study is the contamination of all clusters. Such incidents pertain to epidemics that are easily transmitted, making it more widespread. As a consequence, outbreaks may occur in all clusters of the population. Such is the case for the AH1N1 outbreak where almost all Asian countries have been widely infected. This is the scenario being investigated by this simulation study where all clusters have been considered as infectious and infected.

Downloaded By: [Barrios, Erniel B.] At: 23:39 6 February 2011

The hybrid estimation method of the forward search algorithm and the maximum likelihood estimation into the backfitting procedure was applied on both balanced and unbalanced data sets where onset of the outbreaks was infused in all clusters. The length of time for the outbreak to die down and the structural changes it presents on the covariate and spatial parameters are among the conditions investigated alongside the wide scope of neighborhood contamination. The proposed method provides robust estimates for the covariate, spatial and outbreak parameters of the balanced data set. Minimal absolute percent differences are achieved indicating that the estimate values and actual parameter values do not differ much in magnitude. In terms of the temporal component, good estimates are achieved in cases when no temporary structural change is realized in the presence of an outbreak. However, when the parameters of β and γ are contaminated by 10%, 20%, 30%, and 40% of its actual values, the estimates of ρ are poorly estimated. In fact, the absolute percent differences computed for this estimate is at least 90% implying a major problem on the estimation of the temporal component. However, the MAPE values are within acceptable range and thus, the estimated model produces predicted responses that are nearly alike that of the actual observations. This supports further the benefit of the hybrid method in terms of providing robust estimates and good model fit in balanced data sets. The performance of the hybrid method and the MLE for cases where all clusters are contaminated; see Table 5.

Given unbalanced data sets with ten time points, the forward searched estimates of the covariate and spatial parameters produces close approximations of the actual parameter values since small percent differences are calculated. The use of the MLE on the estimation of the outbreak parameters is also beneficial as it generates optimal results since no large percent differences are detected in all three variations of N, namely 25/26, 30, and 50. The temporal component ρ has been well-estimated in the backfitting procedure in cases where short contamination periods are involved. However, when prolonged epidemic episodes are realized, regardless of the presence of structural change in the model, the hybrid method fails to capture the true temporal parameter values. These poor estimates are evident in the absolute percent differences that are at least 90% in value. In terms of predictive ability, the hybrid method is able to produce estimated models with good predictive capacity. The small MAPE values support this conclusion. It is also noted that the same performance behaviour is established for both two-cluster and five-cluster division of the population. In this instance, the number of clusters defined does not affect the robustness of the estimates computed for all parameters in the epidemic model.

Downloaded By: [Barrios, Erniel B.] At: 23:39 6 February 2011

Table 5
Estimates in small balanced data ($T = 20$, $N = 20$)

Scenarios		Percent difference between estimates and parameters (%)										MAPE	
		β		γ		λ_0		λ_1		ρ			
		Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid	MLE
							Two clusters						
Case 1:		0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	7.40	8.91	0.00	0.00
Case 2:	10%	0.01	0.00	0.02	0.01	2.22	2.22	0.97	0.97	100.32	100.34	1.19	1.19
	20%	0.01	0.01	0.00	0.01	4.34	4.34	1.93	1.93	96.93	97.28	2.18	2.18
	30%	0.00	0.01	0.03	0.00	6.46	6.46	2.90	2.90	97.13	97.30	3.03	3.04
	40%	0.01	0.01	0.01	0.01	8.36	8.36	3.79	3.79	98.41	97.45	3.78	3.77
Case 3:		0.01	0.01	0.00	0.01	0.00	0.00	0.00	0.00	5.64	4.68	0.01	0.00
Case 4:	10%	5.00	5.00	0.06	0.03	1.42	1.42	0.62	0.62	98.95	98.96	4.00	4.00
	20%	9.99	10.00	5.55	0.03	2.53	2.89	1.11	1.27	96.20	85.14	7.93	7.64
	30%	15.00	15.01	0.05	0.02	4.23	4.23	1.87	1.87	93.96	86.76	11.01	11.01
	40%	19.99	20.00	14.08	19.35	4.63	4.32	2.06	1.92	94.70	83.59	15.07	15.41
							Five clusters						
Case 1:		0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	13.67	13.36	0.00	0.00
Case 2:	10%	0.00	0.01	0.01	0.00	2.82	2.82	1.24	1.24	96.83	96.84	1.21	1.21
	20%	0.00	0.00	0.01	0.00	5.58	5.58	2.49	2.49	99.44	102.93	2.23	2.23
	30%	0.00	0.01	0.01	0.00	8.06	8.07	3.65	3.65	97.00	97.10	3.10	3.10
	40%	0.00	0.00	0.01	0.00	10.70	10.70	4.89	4.89	92.93	92.22	3.81	3.81
Case 3:		0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	13.44	4.98	0.01	0.00
Case 4:	10%	5.00	5.00	0.02	1.54	2.11	1.91	0.93	0.84	97.54	96.63	3.92	4.02
	20%	10.00	10.00	9.79	9.99	2.89	2.86	1.27	1.26	92.53	80.23	8.18	8.19
	30%	15.00	15.00	0.02	0.01	5.84	5.85	2.63	2.63	102.24	105.86	10.65	10.65
	40%	19.62	19.62	0.02	11.79	8.03	6.62	3.63	2.97	93.94	84.11	13.42	24.26

Case 1: contamination in 1 cluster, short period, no change in parameters.
Case 2: contamination in 1 cluster, short period, with change in parameters.
Case 3: contamination in 1 cluster, long period, no change in parameters.
Case 4: contamination in 1 cluster, long period, with change in parameters.

Downloaded By: [Barrios, Erniel B.] At: 23:39 6 February 2011

Hence, the proposed estimation procedure is able to generally provide robust estimates for balanced and unbalanced data sets. The forward search algorithm generates robust estimates for the parameters β and γ, which are affected by structural change. This suggests that amidst the fluctuations caused by the temporary outbreaks in the population, the proposed method is able to reveal the actual non epidemic value of the covariate and spatial parameters β and γ, respectively. The simultaneous estimation of these parameters also provides additional optimality in estimation. The small absolute percent differences in outbreaks parameters likewise show the efficiency of the proposed method in estimating this term. Although cases exist where poor estimates are derived for the temporal component, the hybrid method is still considered beneficial. These poor estimates of the temporal component may be attributed to the fact that it is the last parameter estimated in the backfitting procedure. Results such as these are often expected. The proposed method also offers additional gain as seen in the small MAPE values, indicating superior predictive ability of the estimated models obtained from the infusion of the three algorithms.

The pure MLE procedure, that treats the epidemic model as a nonlinear regression, yield estimates that are generally comparable to the proposed method. Considering the complexity of the MLE estimation in a nonlinear regression model, and the difficulties in convergence in cases where there are too many predictors, the proposed method is an ideal alternative. While the proposed method is computationally simpler, it produces estimates comparable to the more optimal MLE estimation of a nonlinear regression model.

7. Conclusions

A generalized model for epidemics capable of summarizing spatial and temporal dependencies of the population, was postulated. The model also incorporates a temporary structural change caused by disease outbreaks. We propose an estimation procedure based on the backfitting method that integrates the forward search method and maximum likelihood estimation. The simulation study shows that the hybrid method produces comparable results to the maximum likelihood estimate of the model treated as a nonlinear regression under the epidemic-free general and with structural change scenarios. Advantages are detected in favour of the backfitting method in cases where there is severe structural change (e.g., epidemic outbreak). This is exemplified whenever long contamination periods are realized and whenever the contamination results to temporary values in the covariates and spatial variables that are significantly different from the true parameter values. The forward search algorithm is able to induce robustness to the proposed estimation method during the period of structural change (epidemic episodes). Furthermore, backfitting is more computationally beneficial as it provides higher chances of convergence when several parameters are involved. The postulated model is a robust abstraction of the epidemic dynamics that can capture the general features not affected by erratic fluctuations during an outbreak.

References

Atkinson, A. (2009). Econometric applications of the forward search in regression: Robustness, diagnostics, and graphics. *Econometric Review* 28:21–39.

Atkinson, A., Riani, M. (2002). Forward search added-variable *t*-tests and the effect of masked outliers on model selection. *Biometrika* 89:939–946.

Atkinson, A., Riani, M. (2007). Building regression models with forward search. *Journal of Computing and Information Technology – CIT* 15:287–294.

Bacaer, N., Abdurahman, X. (2008). Resonance of the epidemic threshold in a periodic environment. *Mathematical Biology* 57:649–673.

Buja, A., Hastie, T., Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics* 17:453–555.

Chen, R., Tsay, R. (1993). Nonlinear additive arx-models. *Journal of the American Statistical Association* 88:955–967.

Gelfand, A. (2007). Guest editorial: spatial and spatio-temporal modeling in environmental and ecological statistics. *Environmental Ecological Statistics* 14:191–192.

Hastie, T., Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman and Hall.

Hsiao, C. (1986). *Analysis of Panel Data*. Cambridge, MA: Cambridge University Press.

Landagan, O., Barrios, E. (2007). An estimation procedure for spatiotemporal model. *Statistics and Probability Letters* 77:401–406.

Opsomer, J. (2000). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis* 73:166–179.

Van Maanen, A., Xu, X. (2003). Modeling plant disease epidemics. *European Journal of Plant Pathology* 109:669–682.

Downloaded By: [Barrios, Erniel B.] At: 23:39 6 February 2011