Robust Estimation of a Multilevel Model with Structural Change

Erniel Barrios *and* ***Mary Jane Esmenda***
University of the Philippines Diliman

Abstract: A generalized spatiotemporal multilevel model is postulated and estimated using the backfitting algorithm imbedded with forward search algorithm and maximum likelihood estimation. The forward search algorithm ensures robustness of the estimates, filtering the effect of the temporary structural changes in the estimation of the group-level covariate parameters, the individual-level covariate and the spatial parameters. Backfitting algorithm provides computational efficiency of the estimation procedure assuming an additive model. Using simulation studies, the estimated model is shown to be capable of producing robust estimates even in the presence of structural changes induced for example by temporary epidemic outbreak. The model also produced robust estimates even for small sample sizes and short time series common in epidemiological setting.

Keywords*: multilevel model, spatiotemporal model, robust method, forward search algorithm*

1. Introduction

Infectious diseases are influenced by complex interactions among disease agents, socio-economic, environmental and ecological factors, wildlife and humans. Consider for instance, prevalence of a disease in the presence of outbreaks that is characterized by spatiotemporal clustering of infection among the susceptible population. Prevalence rates in neighboring areas are expected to be correlated as they are similar in geographical distribution of population at risk and other scales defining the spread of the infection. The occurrence of the disease 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 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 multilevel 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 spread and the risk associated with the disease.

The estimation of 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. Also, group-level effect should be included since features of groups are often driven by the individuals they contain, which means that these individuals are influenced, in turn, by the “emerged” additive feature of the group to which they belong. Furthermore, the inclusion of structural change is necessary as there realistically exist in the dynamics of disease spread, that temporarily inflicts the population density affecting the disease rates at the susceptible setting.

We postulate a model that takes into account the group-level effect, individual-level effect, temporal and spatial dependencies in a multilevel analysis typically exhibited by disease prevalence rates that are jointly determined by physical and geophysical condition and group-level factors (covariates). This paper develops an epidemic multilevel model that is flexible for both infected and non-infected cases. We propose an estimation procedure that is robust and computationally viable. The estimation procedure is iterative and combines the forward search algorithm and a mixed model in the backfitting framework. The backfitting algorithm simplifies the estimation procedure and facilitating convergence. Atkinson and Riani (2007) emphasized robustness of the forward search algorithm in a wide variety of statistical models. Buja et. al. (1989) proved consistency and convergence of the backfitting algorithm in a relatively general class of smoothers in an additive model.

Modeling of the dynamics of disease prevalence enables the understanding on how certain diseases are transmitted and studying trends of diseases to identify sources of infections and make recommendations for abating their spread. The study of outbreaks can facilitate the development of viable mitigation schemes for subsequent improvement of public health. Spatiotemporal multilevel 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 rainfall.

2. Methodology

Given observations for *N* units and *T* time points, prevalence rate (Y_{ijt}) is postulated as a function of dependencies in space, time, space-time interactions. In the presence of an outbreak, we account for the group-level factors (community demographics and spatial features of the population) that contribute to disease outcomes:

$$Y_{ijt} = \beta_d X_{ijt} + \gamma_d W_{ijt} + \phi_d Z_{jt} + u_{jt} + \lambda_{0j*} e^{-\lambda_1 t^*} + \varepsilon_{ijt} \qquad (1)$$

where $\beta_d = \beta I(i)_{\{i \epsilon N^d\}} + \beta^* I(i)_{\{i \epsilon N^d\}}$

$\gamma_d = \gamma I(i)_{\{i \epsilon N^d\}} + \gamma^* I(i)_{\{i \epsilon N^d\}}$

$\phi_d = \emptyset I(i)_{\{i \epsilon N^d\}} + \emptyset^* I(i)_{\{i \epsilon N^d\}}$

$\varepsilon_{ijt} = \rho \varepsilon_{ijt-1} + a_t, a_t \sim N(0,1)$

$\lambda_{oj*} = 0$, if the j[th] unit does exhibit any outbreak episode

The parameters $\beta, \gamma \; and \; \emptyset$ are the original parameters' values while $\beta^*, \gamma^* \; and \; \emptyset^*$ are the temporary values due to the occurrence of an epidemic (structural change). 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 auto-regression. We assume that the error is an autoregressive process of order 1. 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.

Estimation Procedure

Our aim is to have robust estimates of model parameters in the presence of contamination due to the temporary structural change caused by the outbreaks (interventions). An outbreak is a sudden rise in the occurrence (the number of cases) of a disease in a given community (e.g. neighbourhood, city, country, or region). Also, 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 $g(t^*;\lambda) = \lambda_0 \exp\{-\lambda_1 t^*\}$. 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 $\beta, \gamma \; and \; \phi$, as reflected in Model (1). To investigate their behavior, an estimation procedure consisting of implementing an imbedded backfitting and forward search algorithm in a mixed model is described below:

Step 1: The parameters are simultaneously estimated through the imbedded backfitting and forward search algorithm in a mixed model.

Step 1a: ***Backfitting Phase***

i. Fit the model $Y_{ijt} = \beta X_{ijt} + \gamma W_{ijt} + \nu_{ijt}$ using all N observations. Compute the residuals e_{ijt}. The residuals contain information on other parameters.

ii. Estimate ϕ and the random components u_{jt} using the residuals in Step i in a multilevel model.

iii. Given the estimates of ϕ and the random components u_{jt} in Step ii, compute new residuals and iterate from Step i using these new set of residuals in place of Y_{ijt}.

The amount of bias is minimized as the iteration progresses. The iteration then stops when the succeeding estimate values are not very far from the preceding estimate values with 0.0001 or 0.01% differences.

Step 1b: ***Forward Search Phase***

Given the final estimates of the parameters in the Backfitting Phase, compute the residuals.

i. Choose n observations corresponding to the n smallest residuals.

ii. Fit the model $Y_{ijt} = \beta X_{ijt} + \gamma W_{ijt} + \nu_{ijt}$ using all n observations. Compute the residuals e_{ijt}.

iii. Estimate ϕ and the random components u_{jt} using the residuals in Step ii in a multilevel model.

iv. Given the estimates of ϕ and the random components u_{jt} in Step iii, compute new residuals and iterate from Step ii using these new set of residuals in place of Y_{ijt}.

Step 2: The parameters of the temporary structural change will be estimated through the maximum likelihood estimation due to the closed-form nature of the disease dynamics. This will be implemented only on neighbourhoods 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 $e_{ijt} = Y_{ijt} - \hat{Y}_{ijt}$ where $\hat{Y}_{ijt} = \hat{\beta} X_{ijt} + \hat{\gamma} W_{ijt} + \hat{\phi} Z_{jt}$, $\hat{\beta}$, $\hat{\gamma}$ and $\hat{\phi}$ are the averaged estimates across all time points. For infected areas, we note that these residuals e_{ijt} will contain information on the temporary structural change and temporal component initially ignored in the Forward Search Algorithm in Step 2. The Maximum Likelihood estimates of λ_0 and λ_1 are generated only on infected neigbourhoods. These estimates are also averaged through the computation of the harmonic mean of the raw estimates. The final residuals may then be computed as $e_{ijt} = Y_{ijt} - \hat{Y}_{ijt}$ where $\hat{Y}_{ijt} = \hat{\beta} X_{ijt} + \hat{\gamma} W_{ijt} + \hat{\phi} Z_{jt} + \hat{\lambda}_0 \exp\left\{-\hat{\lambda}_1 t\right\}$ for areas with outbreaks. Otherwise, the final residuals are defined by $e_{ijt} = Y_{ijt} - \hat{Y}_{ijt}$ where $\hat{Y}_{ijt} = \hat{\beta} X_{ijt} + \hat{\gamma} W_{ijt} + \hat{\phi} Z_{jt}$.

Step 3: Another regression will be performed on the residuals with its lagged values. This will estimate the temporal parameter ρ. For each ij, estimate ρ using Conditional Least Squares (CLS) in an *AR(1)* model of e_{ijt}, i.e. $e_{ijt} = e_{ijt-1} + \varphi_{ijt}$. Let the estimate be denoted by $\hat{\rho}_{ij}$. Compute the average of $\hat{\rho}_{ij}$ for all i, j,. i.e. $\hat{\rho} = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \hat{\rho}_{ij}$.

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.

Simulation Study

The proposed model in this study with the estimation procedure will be evaluated using simulated data from the balanced (N = T) and unbalanced (T < N) scenarios. The scenarios are adopted from (Bastero and Barrios, 2011). The balanced case is considered since in panel data analysis, this is the setting where most optimal characteristics of existing methods were observed. However, typical panels involve a short span of time 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 and the chance of attrition is heightened.

3. Results and Discussions

In general, the hybrid estimation procedure produced robust estimates for the group-level covariate ϕ, the individual-level covariate β and the spatial parameter γ in all scenarios where structural change is observed. Hence, the forward search is able to filter the bias induced by the temporary structural change observed during epidemic outbreaks. Moreover, the incorporation of the MLE in the backfitting algorithm also provided optimal estimates for the outbreak parameters λ_0 and λ_1. The outbreak dynamics is postulated to follow the exponential distribution. In terms of the temporal parameter ρ estimation, the backfitting is optimal for cases where structural change does not affect the group-level covariate, individual-level covariate and the spatial parameters. However, poor estimates are produced when the parameters β, γ and $\emptyset$ are contaminated in the presence of an outbreak. Also, poorer estimates are produced when the epidemic occurs in all neighbourhoods as compared to the event when the disease is endemic.

Another advantage of the proposed method is in relation to computational efficiency. This is due to the backfitting algorithm where the parameters are alternately estimated, which de-loads the computational burden imposed by the MLE where all the parameters are estimated all at once. Furthermore, when numerous parameters are involved, the MLE is prone to divergence or have lower convergence rates. This is sufficiently addressed by the backfitting method, which produces optimal solutions particularly in additive models, such as the generalized epidemic multilevel model postulated. The infusion of the forward search also assures computational gain as it minimizes the bias induced by the presence of outbreaks. It enables the procedure to utilize observations that are temporarily altered by the temporary disease growth in the population.

We provide an example of relative bias comparison for the balanced, small data case in Table 1.

Table 1. Balanced, Small Data Set (T = 20, N = 20)

Two Clusters										
	Percent difference between estimates and parameters (%) {(True Value - Estimated Value) / True Value} * 100									
		β		γ		ϕ		λ_0	λ_1	ρ
Scenarios		Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid	Hybrid	Hybrid
Case 1		0.265	290.615	0.107	1682.329	0.508	281.246	0.002	0.001	20.754
Case 2	10%	0.286	288.923	0.177	1706.712	0.541	295.852	3.147	1.392	19.244
	20%	0.221	288.327	0.093	1731.027	0.443	308.525	6.361	2.855	65.973
	30%	0.176	287.192	0.110	1755.205	0.361	321.984	9.247	4.221	66.384
	40%	0.221	286.096	0.094	1779.315	0.443	335.393	11.966	5.550	65.638
Case 3		0.264	290.615	0.112	1682.329	0.525	281.246	0.002	0.001	20.795
Case 4	10%	1.724	270.865	19.891	1627.123	12.213	289.836	2.530	1.107	42.864
	20%	2.007	12.577	35.403	3.821	19.770	850.279	5.079	2.256	42.319
	30%	2.849	267.077	53.072	1714.247	29.344	337.295	7.430	3.344	42.090
	40%	3.781	265.038	70.707	1757.740	39.082	361.180	9.673	4.410	41.970
Five Clusters										
	Percent difference between estimates and parameters (%) {(True Value - Estimated Value) / True Value} * 100									
		β		γ		$\emptyset$		λ_0	λ_1	ρ
Scenarios		Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid	Hybrid	Hybrid
Case 1		0.280	79.808	0.062	159.903	0.295	251.885	0.017	0.007	9.224
Case 2	10%	0.280	84.712	0.062	161.988	0.295	248.574	3.417	1.505	26.666

	20%	0.280	89.692	0.062	164.058	0.295	245.131	6.625	2.969	26.584
	30%	0.193	94.712	0.014	166.123	0.328	241.590	9.638	4.394	30.395
	40%	0.171	75.462	0.011	22.668	0.295	77.738	12.455	5.774	30.702
Case 3		0.248	79.788	0.066	159.904	0.230	251.902	0.017	0.007	12.148
Case 4	10%	4.304	75.712	2.187	155.467	12.557	247.770	3.226	1.416	9.412
	20%	8.939	11.442	4.306	4.387	25.492	792.131	6.268	2.798	9.464
	30%	13.614	76.596	6.348	164.327	38.311	267.131	9.137	4.145	9.437
	40%	18.130	52.865	8.464	31.828	51.033	18.164	11.827	5.450	9.428

**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.*

4. Conclusions

A generalized multilevel model for epidemics was postulated and shown to be capable of summarizing spatial and temporal dependencies of the population. This model also incorporates a temporary structural change caused by disease outbreaks in the population. We also propose an estimation procedure based on the forward search algorithm and a mixed model embedded into the backfitting algorithm to estimate the group-level covariate, individual-level covariate and the spatial parameters and the maximum likelihood to estimate the temporary outbreaks in the backfitting framework.

Simulation studies shows that the hybrid method and the MLE produced comparable estimates under the epidemic-free scenarios. Advantages are detected in favor of the hybrid estimation method in cases when there is an epidemic outbreak. This is exemplified whenever the contamination effect is temporary in the group-level covariate, individual-level covariate and spatial variables that are highly different from the true parameter values. The forward search algorithm is able to yield robust estimates through the hybrid method during 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 outbreak dynamics that can capture the general features not affected by erratic fluctuations during an outbreak.

REFERENCES:

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

Bastero, R., Barrios, E. (2011). Robust Estimation of a Spatiotemporal Model with Structural Change. *Communication in Statistics – Simulation and Computation* 40(3): 448-468.

Buja, A., Hastie, T., Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics* 17(2): 453-510.

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