SCHOOL OF STATISTICS
UNIVERSITY OF THE PHILIPPINES DILIMAN

WORKING PAPER SERIES

Robust Estimation of a Multilevel Model with Structural Change

by

Mary Jane Esmenda
and
Erniel Barrios

UPSS Working Paper No. 2016-02
February 2016

School of Statistics
Ramon Magsaysay Avenue
U.P. Diliman, Quezon City
Telefax: 928-08-81
Email: updstat@yahoo.com

Abstract: A spatiotemporal multilevel model is postulated and estimated using the forward search algorithm and maximum likelihood estimation imbedded into the backfitting algorithm. Forward search algorithm ensures robustness of the estimates by filtering the effect of 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. Simulation studies show that estimates are robust 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 settings.

Keywords: *multilevel model, spatiotemporal model, temporary structural change, forward search algorithm*

1. Introduction

Consider epidemics such as the spread of A(H1N1) which infects clusters of individuals. Outbreaks can lead to structural change in the behavior of the model since this creates severe fluctuations in the prevalence of the disease in affected areas. Infectious diseases are influenced by complex interactions among disease agents, socio-economic conditions, environmental and ecological factors, wildlife and humans. As an illustration, prevalence of a disease in the presence of outbreaks 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 spatial externalities indexed by geographic, demographic, health and social conditions. Neighboring 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 risk associated with the disease.

Estimation of prevalence rates of highly contagious diseases can be affected by factors based on physical and geographical 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 prevalence of a disease as it is generally dictated by conditions summarized through covariates. Also, group-level effect should be included since features of groups are often driven by the individuals they compose of, meaning that these individuals are influenced, in turn, by the 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 geographic conditions 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 (to the presence of temporary structural change) and is computationally viable. The estimation procedure is iterative and combines the forward search algorithm and estimation of a mixed model in the backfitting framework. The backfitting algorithm simplifies the estimation procedure that facilitates 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.

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. Multilevel Spatiotemporal Model

The prevalence of a disease in the presence of outbreaks is characterized by spatiotemporal clustering of infection among the susceptible population. Epidemic cases may take place in adjacent locations or areas that are close to each other. Prevalence rates in neighboring areas are expected to be in near approximations as they are similar in geographical distribution of population at risk and other factors that characterizes dynamics in infection. Presence of

diseases in the same area may be due to their common geographic, demographic, health, and social conditions. It is therefore logical to infer that these areas are homogeneous relative to 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 the 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.

A space-time interaction is necessary in understanding and characterizing the prevalence of a disease as it is generally dictated by conditions indexed by covariates. Also, group-level effect should be included since features of groups are often driven by the individuals they comprise, 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.

Given observations for *N* units at *T* time points, prevalence rate (Y_{ijt}) is postulated as a function of dependencies in space, time, and 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:

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

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

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

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

$$\varepsilon_{ijt} = \rho \varepsilon_{ijt-1} + a_t, \quad a_t \sim N(0, \sigma_a^2)$$

$\lambda_{0j^*} = 0$, if j[th] unit (higher level) does exhibit any outbreak episode

The parameters β, γ and ϕ are the original parameters while β^*, γ^* and ϕ^* are the temporary values due to the occurrence of an epidemic (structural change). 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. We assume that 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 on which clusters have been affected by the outbreak. Membership of N^d to the clusters is known, and that progression of epidemics in each cluster is homogeneous within but possibly heterogeneous across clusters.

Estimation Procedure

We propose a modified, iterative estimation procedure for spatiotemporal multilevel models by infusing the forward search algorithm and a mixed model (maximum likelihood estimation) into the backfitting framework. We also evaluate robustness of the method to presence of temporary structural change through a simulation study.

The general idea of the estimation procedure is to alternately estimate the parameters corresponding to covariates β for the individual-level, the parameters corresponding to the spatial parameter γ and the parameters corresponding to the covariates ϕ for the group-level through the imbedded forward search algorithm and mixed model estimation into the backfitting algorithm. The method can mitigate contamination that the ordinary least squares may possibly encounter during outbreaks, see for example Atkinson (2009) for further details on robustness of the forward search algorithm. The temporary structural change (outbreak) effect λ_0 and λ_1 are estimated using the maximum likelihood on the residuals after the effect of X_{ijt}, W_{ijt}, Z_{jt} and u_{jt} are removed from Y_{ijt}. Parameter ρ is then estimated by recomputing the residuals after the effect of the outbreak dynamics is removed from the previous residuals.

Our goal is to construct robust estimates of model parameters in the presence of contamination due to the temporary structural change caused by the outbreaks (interventions). Suppose that the time of the occurrence of an intervention like an outbreak is known a priori.

Vanishing structural change (temporary) characterized through outbreaks is 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. Incorporation of epidemics may result to alterations on the epidemic-free values of β, γ and ϕ, as reflected in Model (1).

An estimation procedure consisting of forward search and mixed model estimation imbedded in backfitting algorithm is described below:

Step 1: The parameters are estimated through forward search algorithm and mixed model estimation embedded into the backfitting algorithm.

Step 1a: ***Mixed Model Estimation***

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, e.g., a tolerance level ε.

Step 1b: ***Forward Search Algorithm***

Given the final estimates of the parameters of the mixed model, 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 maximum likelihood estimation since there is a closed-form structure of the disease dynamics. This is 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 $e_{ijt} = Y_{ijt} - \hat{Y}_{ijt}$ 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 neigborhood. These estimates are also averaged using 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 is performed on the residuals with its lagged values to estimate temporal parameter ρ. For each ij, estimate ρ using conditional least squares (CLS) in an *AR(1)* model of e_{ijt} , i.e., $e_{ijt} = \rho e_{ijt-1} + a_{ijt}$, say $\hat{\rho}_{ij}$. 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.

3. Simulation Study

The proposed model with the estimation procedure is evaluated using simulated data from the balanced (N = T) and unbalanced (T < N) scenarios 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.

The simulation study aims to exhibit an abstraction of the epidemic behavior and disease dynamics. Thus, investigation of robustness of parameter estimates is done on data sets that are nested on the following features: data with two clusters 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, 2 or 5, represents population divided into smaller number of susceptible

groups or otherwise. Consider a fixed number of *N* units, dividing the population into 2 and 5 clusters will look at setting where each neighborhood is comprised of large and small number of spatial units, respectively. The scope of the contamination over neighborhoods will be manifested by case where only a single cluster is affected and the case where all neighbors are affected by the epidemic. The case where a single cluster is infected may be viewed as the endemic case, where the infection is maintained in the population. The scenario where all clusters are suffering from the outbreak is parallel to those national or international concerns due to its high-risk transmissions. Short and long contamination periods were considered, some epidemics die down into the susceptible class faster than other epidemics. 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 the temporary structural change affects the covariate and spatial parameter, 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 larger the difference of β, γ and ϕ 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 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, we explored the behavior of the estimates for small and large sample sizes. The scenarios considered for balanced and unbalanced data sets are shown in Table 1 and Table 2, respectively.

Table 1. Simulated Data Scenarios on Balanced Data Sets

Two Balanced Data Sets (N = T = 20) or Small ; (N = T = 50) or Large																
Two Clusters								Five Clusters								
One-Cluster Contamination				All-Cluster Contamination				One-Cluster Contamination				All-Cluster Contamination				
Short Time Interval		Long Time Interval		Short Time Interval		Long Time Interval		Short Time Interval		Long Time Interval		Short Time Interval		Long Time Interval		
NC	WC	NC	WC	NC	WC	NC	WC	NC	WC	NC	WC	NC	WC	NC	WC	

where NC = no change in the original parameters
WC = with change in parameters

For the common data set where T < N, cases with 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 sizes.

Table 2. Simulated Data Scenarios on Unbalanced Data Sets

Six Common Data Sets (T = 10, N = 25/26) ; (T = 10, N = 30) ; (T = 10, N = 50) (T = 20, N = 25/26) ; (T = 20, N = 30) ; (T = 20, N = 50)																
Two Clusters								Five Clusters								
One-Cluster Contamination				All-Cluster Contamination				One-Cluster Contamination				All-Cluster Contamination				
Short Time Interval		Long Time Interval		Short Time Interval		Long Time Interval		Short Time Interval		Long Time Interval		Short Time Interval		Long Time Interval		
NC	WC	NC	WC	NC	WC	NC	WC	NC	WC	NC	WC	NC	WC	NC	WC	

where NC = no change in the original parameters
WC = with change in parameters

The response variable *Y* was computed from Equation (1). **Z** was sampled from Normal population (μ=5,000 and σ^2=100) while **X** was sampled from Normal population (μ=10,000 and σ^2=1000). Furthermore, the spatial units were divided into clusters/neighborhoods and spatial dependencies are introduced. Samples are generated in the neighborhood system variable **W** from Poisson distribution where each neighborhood would have mean $\mu_k = k$ *

100, k = 1,2 for the 2-cluster case and μ_k= k * 100, k = 1,2,…5 for the 5-cluster case. On the other hand, the error term was simulated from the AR(1), $\varepsilon_{ijt} = \rho\varepsilon_{ijt-1} + a_t, a_t \sim N(0,1)$ with $\rho = 0.5$ and the random component u_{jt} was simulated from the standard normal population. The values of the parameters were set at $\beta = 0.52$, $\gamma = 14.6$, $\phi = 0.61, \lambda_0 = 4{,}800{,}000$, $\lambda_1 = 2.5$. These values were chosen so that each component in the model would have significant contribution in the value of each response variable. The temporary structural change was manifested through the change in parameters of β, γ and ϕ to $\beta^* = 0.572$, $\gamma^* = 16.06$, $\phi^* = 0.671$, represent a 10% difference in the model parameter values. Higher disease severity rates were also considered, resulting to larger differences in the original and temporary values of the covariate and spatial parameters. Specifically, 20%, 30% and 40% difference were considered in the temporary values of β, γ and ϕ to $\beta^* = 0.624$, $\gamma^* = 17.52$, $\phi^* = 0.732$ and $\beta^* = 0.676$, $\gamma^* = 18.98$, $\phi^* = 0.793$ and $\beta^* = 0.728$, $\gamma^* = 20.44, \phi^* = 0.854$, respectively. Spatial variables and covariates were generated from the exponential function that was used to define outbreak dynamics (dies off over time). Response variable values (prevalence rate) dramatically increased at the beginning of the outbreak and returned to “normal” values as the outbreak dissipates.

4. Results and Discussion

Data generated from various simulation scenarios presented in the previous section are used in evaluating robustness of the estimated spatiotemporal multilevel model to presence of structural change. The performance of the hybrid algorithm was assessed by computing the absolute percent difference between estimates and simulated values of the parameters. The hybrid algorithm is benchmarked on the MLE estimates using the same simulated data.

Details of other scenarios for unbalanced data are no longer presented since the hybrid method yield similar results for various cases of unbalancedness.

4.1 No Structural Change

We also investigate the performance of the hybrid method in the absence of structural change (outbreak-free). In Table 3, balanced and unbalanced data sets were generated for small and large data sets, each divided into 2 or 5 clusters. Actual values of the parameters are specified similar to that in Section 3.

In both balanced and unbalanced data, the hybrid method provides desirable estimates for the parameters of group-level and individual-level covariates as well as for the spatial parameters. Parameter estimates are very close to the true values of the parameters, except for the temporal parameters that are estimated last in the backfitting algorithm. As Santos and Barrios (2012) noted, backfitting estimates poorly those parameters that are estimated towards the end in the iteration. See Table 3 for further details.

Table 3. No Structural Change

Balanced Data Set (T = N)								
	% Difference Between Estimates and True Parameters							
		β		γ		ϕ		ρ
Scenarios		Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid
Small Data Set (T=20, N=20)	2 clusters	0.111	0.000	0.009	0.005	0.197	0.000	81.208
	5 clusters	0.191	0.019	0.014	0.000	0.328	0.033	71.674
Large Data Set (T=50, N=50)	2 clusters	0.425	0.000	0.171	0.030	0.820	0.016	73.425
	5 clusters	0.222	0.000	0.001	0.008	0.377	0.000	94.926
Unbalanced Data Set (T < N)								
	% Difference Between Estimates and True Parameters							
		β		γ		ϕ		ρ
Scenarios		Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid
(T= 10, N = 25/26)	2 clusters	0.378	0.019	0.103	0.003	0.705	0.016	75.786
	5 clusters	0.317	0.019	0.010	0.000	0.557	0.066	88.211
(T= 10,	2 clusters	0.127	0.019	0.056	0.024	0.262	0.033	94.746

N = 30)	5 clusters	0.421	0.019	0.017	0.005	0.738	0.016	96.403
(T= 10,	2 clusters	0.179	0.000	0.089	0.050	0.361	0.033	87.195
N = 50)	5 clusters	0.236	0.000	0.007	0.010	0.410	0.016	86.966
(T= 20,	2 clusters	0.290	0.000	0.118	0.019	0.574	0.016	74.081
N = 25/26)	5 clusters	0.400	0.000	0.028	0.003	0.721	0.016	93.907
(T= 20,	2 clusters	0.216	0.000	0.041	0.015	0.410	0.033	85.349
N = 30)	5 clusters	0.083	0.000	0.011	0.010	0.164	0.016	92.924
(T= 20,	2 clusters	0.449	0.000	0.132	0.025	0.836	0.000	76.116
N = 50)	5 clusters	0.236	0.000	0.003	0.006	0.410	0.000	98.657

Absolute percent difference between estimates and their true parameter values are comparable for the hybrid method and the MLE. Regardless of the sample size, whether or not the panel is balanced, the hybrid method and MLE are comparable, both methods yield estimates that are near the true parameter values.

4.2 Presence of Structural Change

Structural change is simulated with the inclusion of outbreak parameters (λ_0 and λ_1) and included in hybrid estimation. This represents the temporary structural change that causes atypical observations in the dataset. The dynamics of epidemics have been simulated so that it represents cases where the outbreak poses a threat over a long period of time and those where outbreaks are easily controlled and the vulnerable population quickly recovers from the threat. Moreover, the outbreak can affect only a contained locale while it can also infest the entire population. Two case considered inducing the outbreak in only one cluster or in all clusters. With structural change, spatial parameters, group-level and individual-level covariates are affected. Several contamination levels were considered, namely 10%, 20%, 30% and 40% to illustrate the severity of the effect of the epidemic in the model. As the epidemic becomes quite severe, changes in the parameters becomes more remarkable. Efficiency of the procedure in generating robust estimators was also assessed relative to the clustering of the population into small or large number of neighborhoods.

4.2.1 Contamination in One Cluster

Contamination in one cluster represents the scenario that outbreak is endemic, i.e., contamination is confined in a specific location. Hybrid estimation method was used on both balanced and unbalanced data sets where onset of the outbreaks was infused in only one cluster. Generally, hybrid method provides robust estimates for both balanced and unbalanced data sets in one-cluster contamination, particularly when no structural change is present or short contamination periods are involved. Robust estimates are also achieved whenever the population is divided into large number of clusters. Moreover, in balanced cases, estimates close to the actual values are obtained even with minimal number of spatial units. For unbalanced data, increasing sample size produces comparable results, supporting further the efficiency of the method in small samples. This is especially useful in epidemiology where public health costs are ideally minimized with fewer individuals monitored and for shorter follow-up periods to avoid higher attrition rates. The system of taking subsets of parameters for simultaneous estimation minimizes the burden in convergence of most estimation methods. This is particularly true for the maximum likelihood estimation, divergence is often realized whenever large number of parameters are estimated. A comparison is made between the estimates of the hybrid estimation method and the maximum likelihood estimation. See Tables 4 and 5 for details of cases with one contaminated cluster in balanced data.

Table 4. Balanced, Small Data Set (T = 20, N = 20), Contamination in One Cluster

Two Clusters										
	% Difference Between Estimates and True Parameters									
		β		γ		$\emptyset$		λ_0	λ_1	ρ
		Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid	Hybrid	Hybrid
Case 1: contamination in 1 cluster, short period, no change in parameters										
		0.265	290.6	0.107	1682.3	0.508	281.2	0.002	0.001	20.7

Case 2: contamination in 1 cluster, short period, with change in parameters										
	10 %	0.286	288.9	0.177	1706.7	0.541	295.8	3.147	1.392	19.2
	20 %	0.221	288.3	0.093	1731.0	0.443	308.5	6.361	2.855	65.9
	30 %	0.176	287.2	0.110	1755.2	0.361	321.9	9.247	4.221	66.4
	40 %	0.221	286.1	0.094	1779.3	0.443	335.4	11.9	5.550	65.6
Case 3: contamination in 1 cluster, long period, no change in parameters										
		0.264	290.6	0.112	1682.3	0.525	281.2	0.002	0.001	20.8
Case 4: contamination in 1 cluster, long period, with change in parameters										
	10 %	1.724	270.9	19.891	1627.1	12.21	289.8	2.530	1.107	42.9
	20 %	2.007	12.6	35.403	3.821	19.8	850.3	5.079	2.256	42.3
	30 %	2.849	267.1	53.072	1714.2	29.34	337.29	7.430	3.344	42.09
	40 %	3.781	265.03	70.707	1757.74	39.08	361.18	9.673	4.410	41.97
Five Clusters										
	% Difference Between Estimates and True Parameters									
		β		γ		$\emptyset$		λ_0	λ_1	ρ
		Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid	Hybrid	Hybrid
Case 1: contamination in 1 cluster, short period, no change in parameters										
		0.280	79.80	0.062	159.903	0.295	251.88	0.017	0.007	9.224
Case 2: contamination in 1 cluster, short period, with change in parameters										
	10%	0.280	84.71	0.062	161.988	0.295	248.57	3.417	1.505	26.67
	20%	0.280	89.69	0.062	164.058	0.295	245.13	6.625	2.969	26.58
	30%	0.193	94.71	0.014	166.123	0.328	241.59	9.638	4.394	30.39
	40%	0.171	75.46	0.011	22.668	0.295	77.738	12.45	5.774	30.70
Case 3: contamination in 1 cluster, long period, no change in parameters										
		0.248	79.79	0.066	159.904	0.230	251.90	0.017	0.007	12.15
Case 4: contamination in 1 cluster, long period, with change in parameters										
	10%	4.304	75.71	2.187	155.467	12.557	247.77	3.226	1.416	9.412
	20%	8.939	11.44	4.306	4.387	25.492	792.13	6.268	2.798	9.464
	30%	13.61	76.59	6.348	164.327	38.311	267.13	9.137	4.145	9.437
	40%	18.13	52.86	8.464	31.828	51.033	18.164	11.82	5.450	9.428

Table 5. Balanced, Large Data Set (T = 50, N = 50), Contamination in One Cluster

Two Clusters										
	% Difference Between Estimates and True Parameters									
		β		γ		$\emptyset$		λ_0	λ_1	ρ
		Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid	Hybrid	Hybrid
Case 1: contamination in 1 cluster, short period, no change in parameters										

		0.205	12.33	0.164	767.60	0.443	384.62	0.002	0.001	6.000
Case 2: contamination in 1 cluster, short period, with change in parameters										
	10 %	0.205	12.8	0.258	752.8	0.492	377.9	3.298	1.452	46.45
	20 %	0.209	26.6	0.188	778.6	0.459	414.8	6.394	2.864	46.43
	30 %	0.584	34.2	0.312	805.4	1.197	441.1	9.298	4.234	50.37
	40 %	0.207	16.8	0.260	822.1	0.492	417.3	12.03	5.566	46.18
Case 3: contamination in 1 cluster, long period, no change in parameters										
		0.226	12.33	0.163	767.60	0.475	384.62	0.001	0.001	5.772
Case 4: contamination in 1 cluster, long period, with change in parameters										
	10 %	1.743	13.9	17.658	738.0	11.25	372.8	2.598	1.138	47.39
	20 %	3.277	28.9	34.99	782.8	21.95	420.4	5.072	2.253	47.51
	30 %	5.195	37.9	52.52	828.6	33.44	458.1	7.424	3.342	47.38
	40 %	6.373	23.2	69.87	864.7	43.51	447.9	9.659	4.405	47.38

Five Clusters

	% Difference Between Estimates and True Parameters									
		β		γ		$\emptyset$		λ_0	λ_1	ρ
		Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid	Hybrid	Hybrid
Case 1: contamination in 1 cluster, short period, no change in parameters										
		0.250	18.1	0.026	63.6	0.361	126.4	0.003	0.001	5.749
Case 2: contamination in 1 cluster, short period, with change in parameters										
	10%	0.250	16.8	0.026	62.9	0.361	126.8	3.343	1.473	13.69
	20%	0.249	16.2	0.002	65.1	0.426	133.0	6.477	2.904	14.04
	30%	0.250	15.7	0.026	67.3	0.361	139.3	9.408	4.289	13.56
	40%	0.246	15.1	0.003	69.6	0.426	145.8	12.17	5.639	13.90
Case 3: contamination in 1 cluster, long period, no change in parameters										
		0.279	17.7	0.024	62.1	0.410	123.4	0.002	0.001	5.901
Case 4: contamination in 1 cluster, long period, with change in parameters										
	10%	0.401	15.9	1.830	60.7	3.754	122.9	3.106	1.367	17.92
	20%	0.538	15.6	3.607	64.7	7.836	133.1	6.030	2.698	19.29
	30%	0.673	15.3	5.442	68.7	12.066	143.4	8.778	3.989	19.74
	40%	0.817	14.9	7.220	72.8	16.131	153.8	11.38	5.250	19.92

Effect of Number of Clusters (Balanced Data)

For balanced data, estimates are very close to true values of β, γ and ϕ except when contamination is prolonged with changes in the parameter values. In small balanced data,

when contamination is prolonged with associated changes in parameter values, spatial parameter γ is the most affected in two cluster case while in the five-cluster case, the group-level covariate ϕ is the most affected. More prominent structural change drives parameters estimates away from their true values, e.g., spatial parameter γ in two-cluster case, and group-level covariate parameter ϕ in the five cluster cases. Similar is true even in large balanced dataset for the two-cluster case. However, in five cluster case, the hybrid procedure provides robust estimates for all parameters.

Outbreak parameters are robustly estimated regardless of length of time series, sample size, and severity of structural change. The temporal parameter ρ, remains to be poorly estimated in two-cluster case, this is especially true whenever no change in β, γ and ϕ are considered in the data generating process.

Effect of Number of Clusters (Unbalanced Data)

In unbalanced data with two clusters, there is difficulty in estimating the spatial and the group-level covariate parameters whenever structural changes occur over a long period. As the severity of structural change increases, these parameter deviate away from their true values. Outbreak parameters are very well estimated even in unbalanced data. Temporal parameter is still poorly estimated when there are prolonged contaminations with or without change in parameters. Hybrid method is unable to properly estimate the temporal parameter as the disease become persistent. For five-cluster with unbalanced data, hybrid method provide robust estimates for the spatial, individual- and group-level covariates and temporal parameters. This illustrates the advantage of hybrid method in cases when larger numbers of

clusters are involved. Outbreak mitigation programs are made more efficient when population is divided into several geographical clusters, resulting into more efficient identification and prevention of diseases. See Table 6 for further details.

Estimates in the five-cluster scenario are comparable with those in two-cluster scenarios. In unbalanced data, number of clusters have minimal effect on the characteristics of the hybrid estimates.

Table 6. Unbalanced Data Set (T = 10, N = 50), Contamination in One Cluster

Two Clusters										
	% Difference Between Estimates and True Parameters									
		β		γ		$\emptyset$		λ_0	λ_1	ρ
		Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid	Hybrid	Hybrid
Case 1: contamination in 1 cluster, short period, no change in parameters										
		0.264	53.5	0.373	3326.2	0.656	1666.6	0.009	0.004	35.02
Case 2: contamination in 1 cluster, short period, with change in parameters										
	10 %	0.219	52.3	0.185	3347.9	0.475	1673.4	3.224	1.422	0.931
	20 %	0.264	109.3	0.373	3385.9	0.656	1793.8	6.235	2.796	1.041
	30 %	0.274	138.1	0.263	3428.0	0.851	1866.4	9.061	4.130	1.617
	40 %	0.265	53.0	0.446	3424.5	0.957	1710.6	11.71	5.424	1.807
Case 3: contamination in 1 cluster, long period, no change in parameters										
		0.204	60.2	1.936	3742.1	1.279	1875.0	0.077	0.033	60.41
Case 4: contamination in 1 cluster, long period, with change in parameters										
	10 %	0.907	52.9	16.49	3366.0	9.21	1682.8	2.635	1.155	55.03
	20 %	1.479	110.6	31.37	3420.3	17.11	1811.9	5.194	2.310	55.51
	30 %	2.121	139.9	46.09	3478.8	25.05	1893.1	7.629	3.439	55.54
	40 %	5.270	55.5	68.68	3491.6	41.16	1746.0	9.700	4.425	53.86
Five Clusters										
	% Difference Between Estimates and True Parameters									
		β		γ		$\emptyset$		λ_0	λ_1	ρ
		Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid	Hybrid	Hybrid
Case 1: contamination in 1 cluster, short period, no change in parameters										
		0.251	93.17	0.008	325.75	0.426	646.69	0.001	0.000	3.909
Case 2: contamination in 1 cluster, short period, with change in parameters										
	10%	0.251	92.71	0.008	328.45	0.426	654.05	3.271	1.443	1.197

	20%	0.250	92.15	0.011	331.16	0.528	661.59	6.324	2.839	1.966
	30%	0.251	91.63	0.024	333.86	0.658	669.08	9.182	4.190	0.989
	40%	0.252	90.50	0.007	336.57	0.485	677.57	11.86	5.500	20.36
Case 3: contamination in 1 cluster, long period, no change in parameters										
		0.270	82.37	0.134	289.69	0.115	575.93	0.020	0.008	10.89
Case 4: contamination in 1 cluster, long period, with change in parameters										
	10%	0.545	77.81	1.502	293.89	2.803	593.69	3.142	1.383	25.11
	20%	0.818	73.15	2.869	298.10	5.721	611.59	6.110	2.734	26.86
	30%	1.092	68.54	4.237	302.31	8.623	629.46	8.902	4.048	27.32
	40%	1.366	63.38	5.604	306.52	11.54	648.20	11.53	5.326	27.44

Effect of Sample Size

Increasing the number of spatial units *N* and time points *T* in the unbalanced case ensures robust performance of the hybrid estimation method. In two-cluster case, increasing the number of spatial units for a fixed time point produces comparable results. Regardless of the use of sample size (25/26) or large (30, 50) number of spatial time points, robust estimates for the parameters $\beta, \gamma, \phi, \lambda_0$ and λ_1 are obtained in cases with short contamination period or no structural changes in the parameters. Relatively large absolute differences between estimated spatial parameter γ and their true values for cases when severe structural changes are present as caused by longer epidemic episodes. Fixing the number of spatial units and increasing the number of time points results to better "forward searched" estimates.

In the five-cluster case, further improvement in estimates are realized with increasing sample size, consistent with usual asymptotic optimality. However, it should be noted that hybrid estimation method can produce generally robust estimated models even with small number of observations.

MLE vs. Hybrid Estimation

MLE is generally affected by structural change, specially, group-level and individual-level covariates and spatial parameters for both balanced and unbalanced data sets. In the presence of structural change, MLE is seriously influenced by atypical observations, causing distortion of the estimates. Estimates from the hybrid method on the other hand, exhibit robustness in the presence of temporary structural change.

4.2.2 Contamination in All Clusters

Hybrid estimation procedure is able to generate robust estimates for balanced and unbalanced data sets. The forward search algorithm is able to abate effects on the estimates of parameters β, γ and ϕ affected by structural change. Amidst variation caused by temporary structural change, proposed method is able recover true group-level and individual-level covariate and spatial parameters β, γ and ϕ, respectively.

In the two-cluster case, comparable results (relative to MLE) are achieved for some parameters. Minimal discrepancies in absolute percent differences are observed for both small and large sample sizes. For five-cluster comparison of small and large samples, better estimates are achieved for the small sample in cases with short contamination periods. Minimal absolute percent differences are noted for cases with longer contamination periods. See Tables 7 and 8 for further details.

Effect of Number of Clusters (Balanced Data)

The proposed method provides robust estimates for the group-level covariate and the individual-level covariate, spatial and outbreak parameters of the balanced data set. When β,

γ and ϕ are contaminated by 10%, 20%, 30% and 40% of its actual values, ρ is poorly estimated for small and large balanced data set. See Tables 7 and 8 for details.

Table 7. Balanced, Small Data Set (T = 20, N = 20), Contamination in all Clusters

Two Clusters										
	% Difference Between Estimates and True Parameters									
		β		γ		$\emptyset$		λ_0	λ_1	ρ
		Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid	Hybrid	Hybrid
Case 1: contamination in all clusters, short period, no change in parameters										
		0.099	0.923	0.044	8.992	0.311	950.3	0.043	0.019	99.96
Case 2: contamination in all clusters, short period, with change in parameters										
	10 %	1.276	3.135	1.133	7.436	1.492	1019.8	2.73	1.198	98.19
	20 %	2.453	7.115	2.310	2.184	2.656	897.2	5.35	2.385	98.09
	30 %	3.631	5.192	3.486	3.064	3.836	91.90	7.83	3.542	97.89
	40 %	4.808	3.962	4.662	4.923	5.016	97.26	10.19	4.670	97.65
Case 3: contamination in all clusters, long period, no change in parameters										
		0.150	0.923	0.027	8.991	0.230	950.29	0.045	0.019	99.28
Case 4: contamination in all clusters, long period, with change in parameters										
	10 %	4.24	5.981	4.079	4.578	4.39	1022.6	1.821	0.796	97.19
	20 %	8.37	12.577	8.197	3.821	8.49	850.28	3.618	1.597	96.91
	30 %	12.49	14.423	12.316	6.166	12.61	101.15	5.351	2.385	96.72
	40 %	16.61	19.058	16.435	10.83	16.71	105.15	7.022	3.160	96.55
Five Clusters										
	% Difference Between Estimates and True Parameters									
		β		γ		$\emptyset$		λ_0	λ_1	ρ
		Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid	Hybrid	Hybrid
Case 1: contamination in all clusters, short period, no change in parameters										
		0.249	0.365	0.025	7.569	0.000	878.21	0.047	0.020	82.49
Case 2: contamination in all clusters, short period, with change in parameters										
	10%	1.425	2.865	1.201	5.069	1.180	880.70	3.273	1.442	96.21
	20%	2.602	6.231	2.378	1.238	2.361	832.84	6.377	2.860	96.41
	30%	3.745	9.173	3.552	1.703	3.607	835.77	9.285	4.236	96.49
	40%	4.895	10.36	4.725	2.430	4.819	888.21	12.02	5.573	96.51
Case 3: contamination in all clusters, long period, no change in parameters										
		0.214	0.365	0.022	7.569	0.066	878.21	0.047	0.020	85.92
Case 4: contamination in all clusters, long period, with change in parameters										
	10%	4.342	5.635	4.140	1.830	4.180	831.85	2.192	0.960	93.17
	20%	8.469	11.44	8.258	4.387	8.279	792.13	4.332	1.920	93.19

	30%	12.59	16.23	12.38	8.762	12.38	842.83	6.378	2.861	93.22
	40%	16.69	21.52	16.49	14.052	16.51	847.72	8.339	3.783	93.23

Table 8. Balanced, Large Data Set (T = 50, N = 50), Contamination in All Clusters

Two Clusters										
	% Difference Between Estimates and True Parameters									
		β		γ		$\emptyset$		λ_0	λ_1	ρ
		Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid	Hybrid	Hybrid
Case 1: contamination in all clusters, short period, no change in parameters										
		0.215	0.038	0.068	1.914	0.246	30.721	0.016	0.007	17.31
Case 2: contamination in all clusters, short period, with change in parameters										
	10%	2.348	2.481	2.196	0.490	1.869	32.426	2.460	1.078	95.00
	20%	4.494	4.038	4.326	1.945	3.967	35.705	4.818	2.139	94.97
	30%	7.025	7.942	6.585	5.999	5.295	37.098	7.066	3.177	94.83
	40%	8.748	9.769	8.582	7.717	8.213	42.082	9.209	4.192	94.68
Case 3: contamination in all clusters, long period, no change in parameters										
		0.237	0.048	0.067	1.814	0.295	31.825	0.016	0.007	16.14
Case 4: contamination in all clusters, long period, with change in parameters										
	10%	4.929	5.038	4.750	3.005	4.377	35.836	1.671	0.729	89.96
	20%	9.620	9.135	9.433	6.912	9.033	42.705	3.302	1.454	90.07
	30%	14.71	16.192	14.25	14.309	12.92	44.607	4.881	2.168	90.03
	40%	19.01	20.577	18.80	18.531	18.34	52.902	6.408	2.871	89.94
Five Clusters										
	% Difference Between Estimates and True Parameters									
		β		γ		$\emptyset$		λ_0	λ_1	ρ
		Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid	Hybrid	Hybrid
Case 1: contamination in all clusters, short period, no change in parameters										
		0.252	1.096	0.004	3.616	0.279	376.52	0.013	0.006	8.488
Case 2: contamination in all clusters, short period, with change in parameters										
	10%	2.384	1.673	2.133	0.668	1.836	351.84	2.941	1.294	96.90
	20%	4.517	4.731	4.261	2.442	3.967	346.56	5.723	2.559	97.02
	30%	6.649	5.154	6.389	2.673	6.082	380.87	8.348	3.790	97.09
	40%	8.782	8.962	8.517	6.564	8.197	367.87	10.83	4.989	97.11
Case 3: contamination in all clusters, long period, no change in parameters										
		0.283	1.038	0.425	3.436	0.328	357.70	0.013	0.006	7.498
Case 4: contamination in all clusters, long period, with change in parameters										
	10%	4.978	3.712	4.689	1.315	4.328	362.62	2.002	0.877	91.72
	20%	9.673	9.269	9.370	7.037	8.984	343.15	3.937	1.742	91.76
	30%	14.37	13.15	14.05	10.790	13.64	370.07	5.796	2.592	91.77
	40%	19.06	17.96	18.73	15.564	18.30	376.87	7.584	3.427	91.77

Effect of Number of Clusters (Unbalanced)

Increasing sample size for a fixed set of 10 time points in a two-cluster case does yield improvements in parameter estimates. Extending the number of clusters to five, still no improvement on the quality of parameter estimates for an increase in the number of observations. With an increase from 30 to 50 spatial units, cases with short contamination period, comparable estimates are achieved for the individual-level covariate β, but the group-level covariate and the spatial parameters are not estimated well. Furthermore, for prolonged structural changes, better estimates are achieved for the individual-level covariate β and the group-level covariate ϕ. However, the spatial parameter γ yield more optimal estimates in the case of 30 spatial units

In five-cluster scenario, increasing sample size with short contamination period is involved, better estimates are obtained for all parameters except for the group-level covariate ϕ. While for the cases with prolonged contamination period, notable differences are detected for all the parameter but still comparable except for the outbreak parameters which does not vary over different sample sizes. Similar performance for both two-cluster and five-cluster are observed. Number of clusters does not affect the robustness of the estimates computed for all parameters.

Table 9. Unbalanced (T = 10, N = 50), Contamination in All Cluster

Two Clusters										
	% Difference Between Estimates and True Parameters									
		β		γ		$\emptyset$		λ_0	λ_1	ρ
		Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid	Hybrid	Hybrid
Case 1: contamination in 1 cluster, short period, no change in parameters										
		0.239	0.19	0.082	9.42	0.459	148.84	0.001	0.001	24.27
Case 2: contamination in 1 cluster, short period, with change in parameters										
	10 %	0.239	1.65	0.082	8.96	0.459	158.98	3.052	1.345	13.75
	20 %	0.359	3.08	0.187	7.53	0.367	160.41	5.917	2.649	14.03
	30 %	0.478	4.50	0.291	6.11	0.421	161.84	8.608	3.914	14.33

	40 %	0.456	10.19	0.072	0.58	0.367	158.84	11.14	5.143	14.64
Case 3: contamination in 1 cluster, long period, no change in parameters										
		0.245	0.19	0.066	9.44	0.656	149.2	0.107	0.046	85.73
Case 4: contamination in 1 cluster, long period, with change in parameters										
	10 %	3.276	4.50	2.962	6.12	3.197	162.97	2.148	0.940	91.19
	20 %	5.982	6.92	5.784	5.47	6.328	191.69	4.305	1.906	91.16
	30 %	8.851	10.25	8.644	2.13	9.164	195.02	6.369	2.853	90.99
	40 %	11.72	17.37	11.503	6.37	12.00	188.28	8.347	3.781	90.79

Five Clusters

	% Difference Between Estimates and True Parameters									
		β		γ		$\emptyset$		λ_0	λ_1	ρ
		Hybrid	MLE	Hybrid	MLE	Hybrid	MLE	Hybrid	Hybrid	Hybrid
Case 1: contamination in 1 cluster, short period, no change in parameters										
		0.252	5.23	0.016	17.16	0.326	1788.5	0.011	0.001	1.464
Case 2: contamination in 1 cluster, short period, with change in parameters										
	10%	0.262	1.48	0.045	13.41	0.426	1792.3	3.637	1.610	9.671
	20%	0.250	2.27	0.334	9.66	0.546	1796.0	6.999	3.159	9.605
	30%	0.251	2.21	0.206	11.18	0.485	2040.9	10.12	4.654	9.598
	40%	0.251	9.77	0.112	2.16	0.436	1803.5	13.02	6.096	9.645
Case 3: contamination in 1 cluster, long period, no change in parameters										
		0.261	5.98	0.010	19.61	0.639	2044	0.101	0.044	84.15
Case 4: contamination in 1 cluster, long period, with change in parameters										
	10%	3.121	0.94	2.867	9.69	3.492	1596.1	2.591	1.138	93.14
	20%	5.980	6.50	5.724	4.13	6.344	1601.7	5.139	2.290	93.37
	30%	8.841	6.54	8.581	6.89	9.197	2046.1	7.556	3.414	93.43
	40%	11.70	14.79	11.44	2.85	12.05	1809.3	9.851	4.512	93.42

Effect of Sample Size

For unbalanced data with ten time points, forward searched estimates of the group-level and individual-level covariates and spatial parameters are close to the true parameter values. The use of MLE in the estimation of the outbreak parameters is also beneficial as it generates optimal results, no large absolute percent differences are detected in three values of *N*, namely 25/26, 30 and 50. Moreover, the temporal component ρ has been well-estimated in the backfitting procedure in cases where short contamination periods are involved.

In unbalanced case, forward searched estimates for $\beta, \gamma, \emptyset, \lambda_0$ and λ_1 are comparable over different number of time points. It can be noted that the estimates from the unbalanced data sets with ten time points provides slightly better estimates than with twenty-time points. This is especially true for the temporal component. For the unbalanced case with twenty-time points, the temporal component ρ has been poorly estimated even in cases where short contamination periods are involved.

MLE vs. Hybrid Method

For the balanced data sets, hybrid method is more desirable over MLE specially for the group-level covariate parameter ϕ. This is also true for the unbalanced data. This can be attributed to the fact that the pure MLE procedure is affected by atypical observations that is distorting the results of the estimates in the presence of structural change.

5. Conclusions

With motivation from epidemiology, a generalized multilevel model is postulated, this is capable of summarizing spatial and temporal dependencies associated with the responses like prevalence rate. We proposed an estimation procedure based on the backfitting algorithm embedded with forward search algorithm and MLE of a mixed model to estimate the group-level covariate effect, individual-level covariate effect and the spatial parameters. A temporary structural change (e.g., those caused by disease outbreaks) is considered and robustness of the estimates are evaluated through a simulation study.

Simulation studies shows that the hybrid method and the MLE produced comparable estimates under scenarios of no structural change. Advantages are observed in favor of the hybrid estimation method in cases when there is a structural change. This advantage is highlighted 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 produce robust estimates in the hybrid method during episodes of temporary structural change. 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. (2009). Econometric applications of the forward search in regression: Robustness, diagnostics, and graphics. *Econometric Review* 28:21-29.

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: 453-555.

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

Santos, E., Barrios, E. (2012). Nonparametric Decomposition of Time Series Data with Inputs. *Communication in Statistics – Simulation and Computation* 41:9,1693-1710.