Spatio-Temporal Single Index Models for Correlated Data

Hamdy F. F. Mahmoud [1,2] *   Inyoung Kim[1]

[1] Department of Statistics, Virginia Tech University, VA 24061, USA
[2] Department of Statistics, Mathematics and Insurance, Assiut University, Assiut 71515, Egypt

Abstract

Modeling spatially-temporally correlated data using parametric models is common, however semiparametric models are very limited in this area. This article introduces a semiparametric spatial-temporal effects model, in which spatial effects are integrated into the single index function and temporal effects are additive to the single index function. We refer to this model as "semiparametric non additive spatio-temporal single index model" (NST-SIM). For estimation, Monte Carlo Expectation Maximization based algorithm is used. NST-SIM has many advantages demonstrated by simulation studies and real data application. It has smaller mean square error and higher accurate prediction compared to non-integrated spatial temporal single index model. It is applied to South Korean mortality data of six major cities and interesting results are found.

Key Words: MCEM algorithm, mixed model, single index model, spatio-temporal data

1. Introduction

Epidemiology has a long history of studying factors that affect the variability of mortality. Among the factors that affect mortality is the geographical (or spatial) variable that plays a crucial role in evaluating health care distribution. Spatio-temporal analysis has the additional benefits over spatial analysis because it allows the investigator to simultaneously study the existence of patterns over time. Because of the novel computational methods that allow for analyzing the large spatio-temporal databases, the spatio-temporal data analysis is an emerging research area.

Among many early authors who worked with spatio-temporal data and obtained statistical models are Goodall and Mardia (1994) and Cressie (1994). There are many others articles studied this type of data in the last two decades because of the increase of the availability of these data (see, Landagan and Barrios, 2007; Arcuti et al., 2013; Lekdee and Ingsrisawang, 2013; Sherman, 2011; Hayn et al., 2009; Nelson et al., 2009; Li et al., 2007; Genton et al., 2006; Kanevski and Maignan, 2004). Many parametric statistical models have been introduced by these articles to model spatio-temporal data, such as generalized linear mixed model, generalized linear additive model, and spatio-temporal auto-regressive model.

Mixed effects models and spatio-temporal models are different in terms of the form of the variance-covariance matrix structure that defines the type of correlation between random effects of time and spatial dependency. In spatio-temporal data, the spatial effects covariance matrix structure depends on the distance between any two locations or cities and it is assumed to follow some parametric function. There are many articles studied the estimation and prediction of the covariance functions (see, Hayn et al., 2009; Cressie and Wikle, 2011; Arcuti et al., 2013; Lekdee and Ingsrisawang, 2013). The common approach

*Corresponding author email: ehamdy@vt.edu

of modeling spatio-temporal data is using parametric models, however these parametric models have strong assumptions on the model. These assumptions in many cases are not satisfied to real data sets.

Nonparametric and semiparametric models relax these assumptions and as a result they are more appropriate to analyzing spatio-temporal data, however the semiparametric models are very limited in this area. This article introduces two semiparametric models that are more appropriate in modeling spatio-temporal data. The two models are extensions to the single index model (SIM) to incorporate the spatial and time effects into the model.

Ichimura (1993) introduced the single index model and many articles extensively studied estimation and inference of the SIM (Hridtache et al., 2001; Wang et al., 2010; Chang et al., 2010; Mahmoud et al., 2016; Mahmoud et al., 2016; Mahmoud and Kim, 2019) and applied in many different areas, such as economics, medicine, biostatistics, and epidemiology.

The SIM has many advantages over the parametric models that is because (1) it assumes that the function that describes the relationship between the response variable and the explanatory variables is an unknown function which make it avoids misleading results of misspecifying the link function (Horowitz and Hardle, 1996), (2) it doesn't assume specific type of error distribution, and (3) it avoid the curse of dimensionality problem by using the single index linear combination of the explanatory variables that reduces the $p-$dimension to only one dimension. In SIM estimation, we need to have a restriction on parameters to fix the identifiablility problem. One solution is to use the norm of the parameters equals to 1 (Xia et al., 2004; Lin and Kulasekera, 2007), and another solution is to set the first coefficient to be equal to 1 (Ichimura, 1993; Sheman, 1994). The second approach is used in this article.

Incorporating the random effects as additive effects into the single index function was introduced by Pang and Xue (2012), however random effects are assumed to be independent in a longitudinal data analysis. This article introduces two models: the first one has the correlated spatial and temporal effects are additive to the single index function and in the second model, the spatial effect is incorporated into the single index function and the spatial effects are additive to the unknown function. The first model is named "Additive spatio-temporal single index model (AST-SIM)" and the other model named "Non additive spatio-temporal single index model (NST-SIM)". These two models are applied to a real data set from Korea cover the time from 2000-2007. The data has many variables recorded daily: mortality, temperature, humidity, pressure, and time for six major cities in Korea (Busan, Seoul, Daejeon, Incheon, Gwangju, and Daegu). Figure 1 shows the location of these six cities.

This article mainly focus on the second model. To the best of our knowledge, this such model is not in the statistical literature. In Section 2, the additive and nonadditive models are displayed. Section 3 is dedicated for the algorithm that is proposed for the two models estimation. Section 3 describes the real data application and apply the two models on the motivating data. Conclusions are provided in Section 5.

2. Proposed Models

Two models are introduced to model the spatio-temporal data. In the first model, the spatial effects are additive to the single index function, we call it "Semiparametric additive spatio-temporal single index model", and in the second model, the spatial effects are non additive to the single index function and refer it to "Semiparametric additive spatio-temporal single index model". Let $Y_{s,t}$ is a discrete response variable at location s and time point t, u_s, s=1,$\dots$, S, be a spatial random process, w_t, $t = 1, \dots, n$, be a time random process,

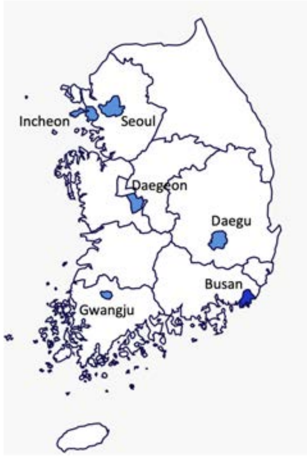

Figure 1: A map shows the location of the six major cities in South Korea

$\mathbf{x}_{1(s,t)}, \mathbf{x}_{2(s,t)}, \ldots, \mathbf{x}_{p_{(s,t)}}$ are p the explanatory variables at location s and time point t. The first model takes the form:

$$Y_{s,t}|\mu_{s,t} \sim \text{Pois}(\mu_{s,t}|u_s, \nu_t),\ \mu_{s,t}|u_s, \nu_t = g(X_{(s,t)}\beta) + u_s + w_t,$$

The second model takes the form:

$$Y_{s,t}|\mu_{s,t} \sim \text{Pois}(\mu_{s,t}|u_s, \nu_t),\ \mu_{s,t}|u_s, \nu_t = g(X_{(s,t)}\beta + u_s) + w_t,$$

where $g(\cdot)$ is an unknown function, β is a vector of coefficient parameters, and μ is the mean function. One of the differences of these models is that the first one needs a restriction on the coefficient parameters.

Gaussian kernel is used to construct the variance covariance matrix of the spatial correlated random effects that has mean $E(u_s) = \mathbf{0}$ and $cov(u_{s+a}, u_s) = C(a)$, where a is the distance between any to locations. For temporal effects, random walk with first order will be used to construct the variance covariance matrix of time effect.

3. Estimation Method

Models estimation is based on a modified Monte Carlo Expectation Maximization (MCEM) algorithm. It is modified for our semiparametric regression models. Initializing $\boldsymbol{\beta}$, $g(\cdot)$, u_s, $(s = 1, \ldots, S)$, w_t, $(t = 1, \ldots, T)$, σ_w^2, σ_u^2, and estimates of ρ_u and ρ_w ($\hat{\rho}_u$ and $\hat{\rho}_w$) is required to run the MCEM. The introduced algorithm, in this paper, to estimate AST-SIM and NST-SIM parameters and spatial and time effects, takes the following steps:

- I-step: parameters initialization:
  - $\sigma_u^{2(0)}$, $u_s^{(0)}$, $\sigma_w^{2(0)}$, and $w_t^{(0)}$ are initialized;
  - $Y_{s,t}^* = Y_{s,t} - u_s^{(0)} - w_t^{(0)}$;
  - Estimate $\boldsymbol{\beta}$, $\boldsymbol{\beta}^{(0)}$, and $\hat{g}(\cdot)^{(0)}$ by any semiparametric method, Ichimura's method is used.

- E-step: Generate a random sample from each location effect $(u_{s1}, u_{s2}, \ldots, u_{sN})$ given all the previous initials, and each time effect $w_{t1}, w_{t2}, \ldots, w_{tN}$ from $\log f[\mathbf{Y}, \mathbf{u}, \mathbf{w}|\boldsymbol{\mu} = f[X_{(s,t)}\boldsymbol{\beta}] + Z\mathbf{u} + Ww, \sigma_u^2\Sigma_{u,\hat{\rho}_u}, \sigma_w^2\Sigma_{w,\hat{\rho}_w}]$ via the Metropolis-Hastings algorithm.

- M-step Maximize $\frac{1}{N}\sum_{k=N_0+1}^{N}\log f\{\mathbf{u}_k|\sigma_u^2\Sigma_{u,\hat{\rho}_u}\}$ and $\frac{1}{N}\sum_{k=N_0+1}^{N}\log f\{w_k|\sigma_w^2\Sigma_{w,\hat{\rho}_w}\}$:
  - Get $\sigma_u^{2(1)}$ and $\sigma_w^{2(1)}$;
  - Calculate ${u_s}^{(1)} = \frac{1}{N}\sum_{k=1}^{N} u_{sk}$, ${w_t}^{(1)} = \frac{1}{N}\sum_{k=1}^{N} w_{tk}$ and $Y_{s,t}^{*} = Y_{s,t} - {u_s}^{(1)} - {w_t}^{(1)}$;
  - Estimate $\boldsymbol{\beta}$, $\boldsymbol{\beta}^{(1)}$, and $\hat{g}(\cdot)^{(1)}$.
- E-step and M-step are iterated until convergence achieved.

4. Mortality Data

The two spatio-temporal models that are introduced in this paper are applied to the South Korea data to find which model is more appropriate to describe the data based on some model selection criteria. For each model, we need to estimate the unknown function, the model parameters, and the dependence range. In the two models, the response variable is the mortality Y and the explanatory variables are temperature (x_1), pressure (x_2), and mean humidity mean temperature (x_3). The data is collected from six cities (Busan, Seoul, Daejeon, Incheon, Gwangju, and Daegu). The monthly data is used rather than daily data to avoid the problem of big "N" (Banerjee et al., 2004) that will make the computing time very big because of the variance covariance matrix rank, and the form of the mortality functions during the year can be studied. The time and location represent the temporal and spatial effects. The population sizes of the six cities are different, so we calculated the number of deaths per ten thousands of people.

Figure 2(a-c) show that the mortality-temperature relationship is negative, the mortality-humidity is positive, and the mortality-pressure relationship is positive. Figure 2(d) reveals that mortality is high at the beginning and at the end of the year and the highest mortality is of Busan city and the lowest is of Gwangju.

In addition to applying the two models on mortality data, we need to find the best model that is appropriate for this data based on fitting and prediction criteria. The following are the two models for our motivating data:

- AST-SIM $Y_{s,t}|\mu_{s,t} \sim \mathrm{Pois}[\mu_{\mathrm{s,t}}|\mathrm{u_s}, \mathrm{w_t}]$,
  $\mu_{\mathrm{s,t}}|\mathrm{u_s}, \mathrm{w_t} = \mathrm{g}[\mathrm{x}_{1\mathrm{s,t}}\beta_1 + \mathrm{x}_{2\mathrm{s,t}}\beta_2 + \mathrm{x}_{3\mathrm{s,t}}\beta_3] + \mathrm{u_s} + \mathrm{w_t}$,
- NST-SIM $Y_{s,t}|\mu_{s,t} \sim \mathrm{Pois}[\mu_{\mathrm{s,t}}|\mathrm{u_s}, \mathrm{w_t}]$,
  $\mu_{\mathrm{s,t}}|\mathrm{u_s}, \mathrm{w_t} = \mathrm{g}[\mathrm{x}_{1\mathrm{s,t}}\beta_1 + \mathrm{x}_{2\mathrm{s,t}}\beta_2 + \mathrm{x}_{3\mathrm{s,t}}\beta_3 + \mathrm{u_s}] + \mathrm{w_t}$,

The proposed Algorithm is used for AST-SIM and NST-SIM parameters estimation, and spatial and temporal effects estimation. For the initial values, generalized linear mixed effects model is used to find the initial values for the introduced algorithm. Metropolis-Hastings algorithm is run 5000 times to sample spatial and time effects after discarding 2% of the MCMC. To estimate the model parameters, MCEM algorithm is iterated until convergence achieved. Using variogram, the dependence range is estimated and found it is equal to 1, $\rho = 1.5$, however NST-SIM does not have this problem so we can estimate all the parameters.

Table 1 shows that the standard error of the parameter estimates of NST-SIM is smaller than the standard error of AST-SIM parameter estimates. In terms of log likelihood value, NST-SIM is much better than AST-SIM, and NST-SIM has higher R^2 value (0.94) compared to AST-SIM (0.83). The coefficient parameter of x_2 in AST-SIM is set to be equal to 1 to fix the identifiabiltiy problem.

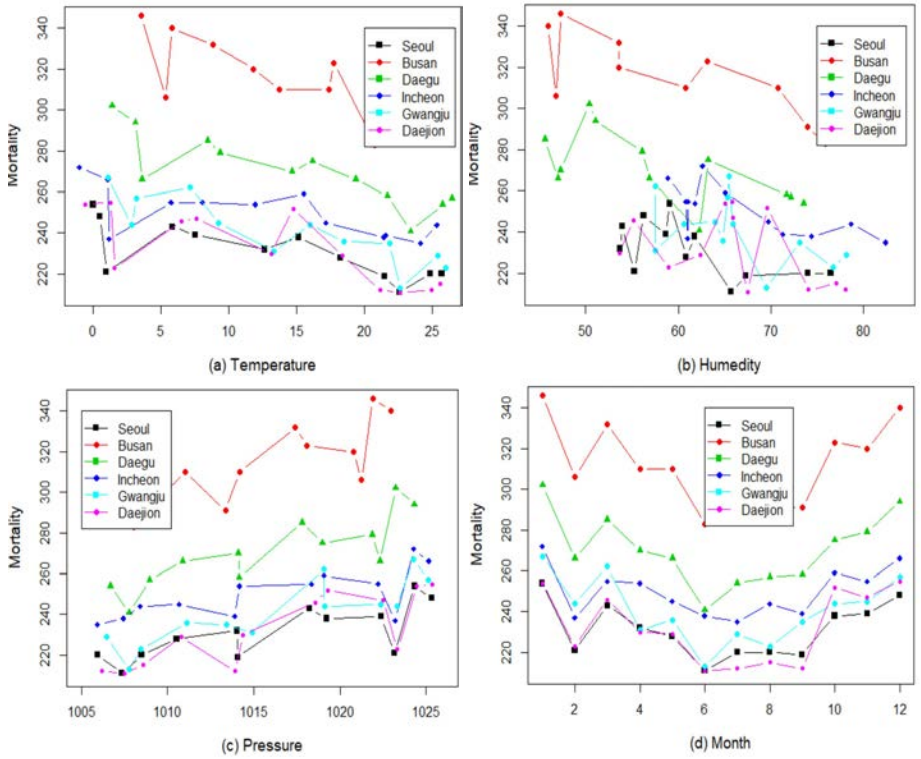

Figure 2: The relationship between temperature and mortality (a), the relationship between humidity and mortality (b), the relationship between pressure and mortality (c), and the relationship between month and mortality (d) for the cities.

Table 1: The AST-SIM and NST-SIM parameter estimates along with their 95% confidence intervals, R^2 values, and the log loglikelihood values

	AST-SIM		NST-SIM	
	Estimate	95% CI	Estimate	95% CI
x_1	-0.005	(-0.005, -0.005)	-2.173	(-4.23, -0.11)
x_2	0.005	(0.004, 0.0046)	1	0
x_3	0.003	(0.003, 0.0026)	3.740	(3.44, 4.03)
log Likelihood	-236.4		-313.8	
R^2	0.94		0.83	

Table 2 shows that both models confirmed that Busan city has the highest mortality and the lowest mortality is for Seoul city.

Figure 3 shows that the mortality functions of the six cities have the same form using the AST-SIM and the mortality functions of the six cities do not have the same form using the NST-SIM. This is one of the advantages of NST-SIM over AST-SIM, it is more flexible so it enables mortality functions to be different over locations.

To evaluate the performance of the models that this paper introduces, AST-SIM and NST-SIM, predicted mean square error (PMSE) criterion is used to evaluate their prediction performance, and mean square error (MSE) and R^2 criteria are used to evaluate their

Table 2: Spatial random effects estimates from the two introduced models (AST-SIM and NST-SIM)

	AST-SIM	NST-SIM
Busan	26.4381	0.197
Seoul	-17.1974	-0.071
Incheon	-13.1910	-0.031
Daegu	10.7913	0.103
Daejeon	-18.9257	-0.089
Gwangju	-12.4628	-0.064
σ_u^2	231.4672	0.001

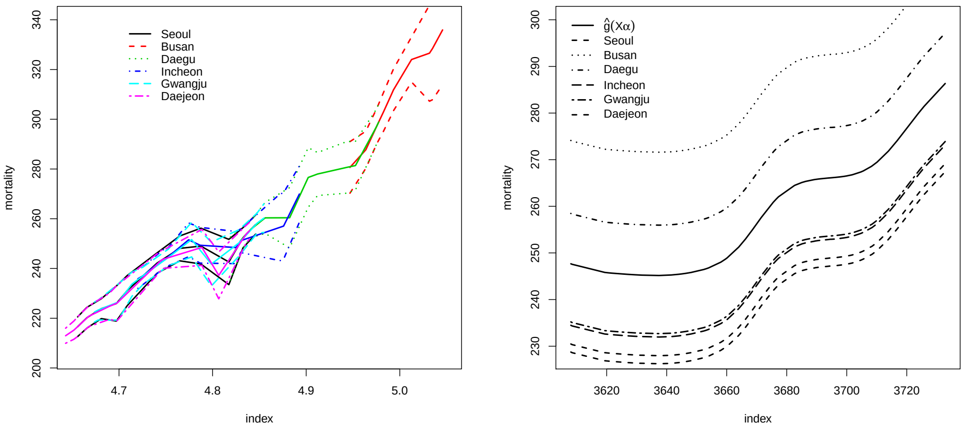

Figure 3: Smoothed mortality functions (NST-SIM at the left and AST-SIM at the right) of the cities and their confidence intervals at 95%.

appropriateness for describing the data. In terms of fitting the data, Table 3 shows that NST-SIM is better than AST-SIM, it has higher R^2 (0.945 vs 0.879), lower MSE (57.97 vs 180.19), and higher log likelihood (-236.54 vs -309.23). In terms of prediction, again NST-SIM is better than AST-SIM, it has higher PMSE median. The sample sizes of the data that is used in the evaluation are 2 and 6. In sum, NST-SIM performance is better than AST-SIM in terms of describing the mortality data and prediction of the South Korea mortality data.

5. Conclusion

This paper introduces two semiparametric models for modeling spatio-temporal correlated data. In one of them, the spatial effect is an additive effect to the single index model (AST-SIM) and in the other model, the spatial effect is a nonadditive effect (NST-SIM). An algorithm is proposed that is based on the Expectation Maximization algorithm to estimate the the two model parameters. It is found that the nonadditive model outperforms the additive model. The advantages of the NST-SIM are: (1) it does not need restriction on the coefficient parameters such as AST-SIM, (2) it does not enforce the mortality functions to have the same form over locations, such as AST-SIM. For the real data, it is found that NST-SIM is more appropriate to describe the mortality data of South Korea in terms of fitting and prediction, and the two models showed that Busan city has the highest mortality among the six cities.

Table 3: Median of the predicted mean square error (PMSE), mean of the mean square Error (MSE), the mean of $\log$ likelihood, and the mean of R^2 values of 250 estimates of fitting and prediction criteria of AST-SIM and NST-SIM under two cases: all data, and at different sizes of evaluating data sets ($n = 2 and 6$).

		Median PMSE	Mean $\log$ likelihood	Mean MSE	Mean R^2
All data	AST-SIM	—	-309.23	180.19	0.876
	NST-SIM	—	-236.54	57.97	0.945
n=2	AST-SIM	227.70	-257.2	169.20	0.88
	NST-SIM	184.60	-195.7	101.0	0.904
n=6	AST-SIM	284.2	-158.7	130.9	0.90
	NST-SIM	829.3	-104.7	131.2	0.87

REFERENCES

Cressie, N. (1994) Comment on “An approach to statistical spatial-temporal modeling of meteorological fields” by M. S. Handcock and J. R. Wallis. *Journal of the American Statistical Association*, **89**, 379-382

Goodall, C. and Mardia, K. V. (1994) Challenges in multivariate spatio-temporal modeling. *In Proceedings of the XVIIth International Biometric Conference*, Hamilton, Ontario, Canada, 8-12 August 1994, pp 1-17

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. 1st edition. Wiley, New Jersey.

Landagan, E. B., and Barrios, O. Z. (2007). An estimation procedure for a spatial-temporal model, *Statistics and Probability Letters*, **77**, 401-406.

Arcuti, S., Calculli, C., Pollice, A., D’Onghia, G., Maiorano, P. and Tursi, A. (2013). Spatio-temporal modelling of zero-inflated deep-sea shrimp data by Tweedie generalized additive. *Statistica*, **73**, 103-122.

Lekdee, K. and Ingsrisawang, L. (2013). Generalized linear mixed models with spatial random effects for spatio-temporal data: an application to dengue fever mapping. *Journal of Mathematics and Statistics*, **9**, 137-143.

Sherman, M. (2011). *Spatial Statistics and Spatio-Temporal Data*. New York: Wiley.

Hayn, M., Beirle, S., Hamprecht, F. A., Platt, U., Menze, B. H. and Wagner, T. (2009). Analysing spatio-temporal patterns of the global NO2-distribution retrieved from GOME satellite observations using a generalized additive model. *Atmospheric Chemistry and Physics*, **9**, 6459-6477.

Nelson, T. A., Duffus, D., Robertson, C., Laberfee, K. and Feyrer, L. J. (2009). Spatial-temporal analysis of marine wildlife. *Journal of Coastal Research*, **56**, 1537-1541.

Li, B., Genton, M. G. and Sherman, M. (2007). A nonparametric assessment of properties of space-time covariance functions. *Journal of the American Statistical Association*, **102**, 736-744.

Genton, M. G., Butry, D. T., Gumpertz, M. L. and Prestemon, J. P. (2006). Spatio-temporal analysis of wildfire ignitions in the St Johns River Water Management District, Florida. *International Journal of Wildland Fire*, **15**, 87-97.

Kanevski, M. and Maignan, M. (2004). *Analysis and modeling of spatial environmental data.*. Switzerland: EPFL Press.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, **58**, 71-120.

Hridtache, M., Juditski, A., and Spokoiny, V. (2001). Direct estimation of the single coefficients in a single-index model. *Annals of Statistics*, **29**, 595-623.

Wang, J. L., Xue, L. G., Zhu, L. X., and Chong, Y. S. (2010). Estimation for a partial-linear single-index model. *Annals of Statistics* , **38**, 246-274.

Chang, Z. Q., Xue, L. G., and Zhu, L. X. (2010). On asymptotically more efficient estimation of the single-index model. *Journal of Multivariate Analysis*, **101**, 1898-1901.

Horowitz, J. L. and Hardle, W. (1996). Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association*, **91**, 1623-9.

Lin, W. and Kulasekera, K. B. (2007). Identifiability of single index models and additive index models. *Biometrika*, **94**, 496-501.

Mahmoud, H.F.F., Kim, H., and Kim, I. (2016). Semiparametric single index multi change points model with an application of environmental health study on mortality and temperature. *Environmetrics*, **27**(8), 496-501.

Mahmoud, H.F.F. and Kim, I. (2019). Semiparametric spatial mixed effects single index models. *Computational Statistics & Data Analysis*, **136**, 108-122.

Xia, Y., Li, W. K., Tong, H., and Zhang, D. (2004). A goodness-of-fit for single index models. *Statistica Sinica*, **14**, 1-39.

Sherman, R.P. (1994). U-process in analysis of a generalized semi-parametric regression estimator. *Economic Theory*, **10**, 372-395.

Pang, Z. and Xue, L. (2012). Estimation of the single-index models with random effects. *Computational Statistics & Data Analysis*, **56**, 1837-1853.

Banerjee, S., Carlin, C. P. and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial*. London: Chapman and Hall.