Science Journal of Applied Mathematics and Statistics
2018; 6(2): 49-57
http://www.sciencepublishinggroup.com/j/sjams
doi: 10.11648/j.sjams.20180602.11
ISSN: 2376-9491 (Print); ISSN: 2376-9513 (Online)

Correlated Spatiotemporal Data Modeling Using Generalized Additive Mixed Model and Bivariate Smoothing Techniques

Sabyasachi Mukherjee[1, *], Tapan Kumar Garai[2]

[1]Department of Mathematics, NSHM Knowledge Campus, Durgapur, India

[2]Department of Agriculture, Government of West Bengal, Medinipur, India

Email address:

sabyasachi99@gmail.com (S. Mukherjee), tapangarai@gmail.com (T. K. Garai)

[*]Corresponding author

To cite this article:

Sabyasachi Mukherjee, Tapan Kumar Garai. Correlated Spatiotemporal Data Modeling Using Generalized Additive Mixed Model and Bivariate Smoothing Techniques. *Science Journal of Applied Mathematics and Statistics*. Vol. 6, No. 2, 2018, pp. 49-57.
doi: 10.11648/j.sjams.20180602.11

Received: April 4, 2018; **Accepted**: April 28, 2018; **Published**: May 22, 2018

Abstract: *Background:* The present article tries to analyze a correlated spatiotemporal data using an advance regression modeling techniques. Spatiotemporal data contains the information of both space and time simultaneously. Naturally, it is very much complicated and not easy to model. This article focuses on some modeling techniques to analyze a correlated spatiotemporal agricultural dataset. This dataset contains information of soil parameters for five years across the twenty six different locations with their geographical status in term of longitude and latitude. Soil pH and fertility index are the two major limiting factors in agriculture. These two parameters are governed by many other factors viz. fertilizer use, cropping intensity, soil type, geographical location, soil health management etc. *Objective:* The present study has been set up to explore whether there is any spatial gradient in the average pH levels across the geographical locations while fertility index and cropping intensity are acting as possible confounder. *Methods:* Soil pH is the response variable which varies with respect to time and space generally has a correlated structure. Besides this, some random effects component with fixed effects having a nonlinear association with the response is observed here. Generalized additive mixed model (GAMM) regression and Bivariate Smoothing techniques have been exercised to arrive at a meaningful conclusion. *Conclusions:* It is found that the pH value varies with change in latitude. Besides this, year, fertility index of available potassium and phosphate are also significant cofactors of this study. Final model has been selected through minimum AIC value (204.9) and model checking plots.

Keywords: Spatiotemporal Data, Spatial Gradient, PH, Fertility Index, Cropping Intensity, GAMM, Bivariate Smoothing

1. Introduction

Soil is a dynamic natural body developed as a result of pedogenic processes during and after weathering of rocks, consisting of mineral and organic constituents, possessing definite chemical, physical, mineralogical and biological properties, having a variable depth over the surface and providing a medium for plant growth [5]. The Soil on the earth surface developed primarily from the weathering of rocks and minerals by the incessant action of rainfall, temperature, wind flow, earthquake etc. So, physicochemical characteristics of a soil in an area firstly depend upon the type of parent material from which this soil is developed and secondly upon the climatic feature of this area. In another process, soil is developed from the material brought by the river water from one place to another.

Besides this natural factor for the development of soil, now, several men made factor such as ploughing, application of chemical, high cropping intensity etc. largely influence the physicochemical characteristics of this natural gift either synergistically or antagonistically with the natural factor. Soil is developed by the action of the natural factor by thousand or more years. But manmade factor can change this soil within a very short period of time.

Soil pH is the negative logarithm of hydrogen ion concentration in the soil solution. It represents acidity/

alkalinity of soil and its theoretical value ranges from 0 – 14. pH value 7.0 is the neutral value, below and above of this value termed as acidity and alkalinity respectively. Soil pH value 6.5-7.5 is normal and suitable for all most all the agricultural crop. Causes of soil acidity generally are (i) Acidic parent material (ii) High rainfall (iii) Application of acid-producing fertilizer like- Urea, DAP, MAP etc (iv) High cropping intensity which means a high use of chemical fertilizer. Factor i) and ii) is natural and no role of human. Average rainfall may vary over the geographical areas. Cropping intensity and the use of chemical fertilizer may vary area to area. There is a definite relation between cropping intensity and fertilizer use. High cropping intensity means the high application of acid-producing fertilizer. High cropping intensity also removes basic cation (Ca, Mg etc.) from the soil in high rate. This basic cation can counteract soil acidity. In this point, the negative interaction of cropping intensity with soil pH is quite natural. [3]

The main objective of this study is to find out; using different statistical tools; whether there is any spatial gradient in the mean pH level across the geographical location while fertility index and crop intensity of the blocks are acting as possible confounders. Using the given data set, finding out pH gradient over time is also one of the objectives of the present study.

Most of the cases, in regression analysis, the coefficients are considered as fixed. However, there are cases in which it makes sense to assume some random coefficients. These cases generally occur in two situations, firstly when the main interest is to make inference on the entire population, from where some levels are randomly sampled and secondly when the observations are correlated. Like biological and medical studies, in agricultural studies often collected observations from the same units (e.g. individuals) over time. It may be reasonable to assume that correlations exist among the observations from the same individual. A model with both fixed and random effects is called mixed effects model. Fixed effects are parameters associated with an entire population or with certain repeatable levels of experimental factors, while Random effects are associated with unrepeatable individual experimental units, drawn at random from a population.

Before introducing Generalized additive mixed model (GAMM), it has been relevant that to introduced Generalized linear mixed models (GLMMs) first, because it is helpful to understand the structural ground of GAMM more easily. GLMMs [6] provide a unified likelihood framework for parametric regression of a variety of over-dispersed and correlated outcomes. Data of this type arise in many fields of research, such as longitudinal studies, survey sampling, clinical trials and disease mapping. A key feature of GLMMs is that they use a parametric mean function to model covariate effects while accommodating over dispersion and correlation by adding random effects to the linear predictor. However, this parametric mean assumption under GLMM may not always be desirable, since appropriate functional forms of the covariates may not be known in advance and the outcome variable may depend on the covariates in a complicated manner. It is hence of substantial interest to develop a nonparametric regression model for correlated data by incorporating a nonparametric mean function in GLMMs. This will allow more flexible functional dependence of the outcome variable on the covariates.

There are very many references on nonparametric regression with independent data using kernel and spline methods [12, 13]. The generalized additive models of Hastie and Tibshirani, [14, 22] are widely used and well understood. Regression analysis under various correlated structure has been studies by many author [9, 10]. However, a very limited work has been done on nonparametric regression when the data are correlated. Most researchers have restricted their attention to longitudinal data with normally distributed outcomes and a single nonparametric function [15, 25]. Several researchers have incorporated a nonparametric time function in linear mixed models [24, 29, 36, 37]. In 1999 Lin and Zhang proposed generalized additive mixed models (GAMMs), which are an additive extension of GLMMs in the spirit of Hastie and Tibshirani [14]. This new class of models uses additive nonparametric functions to model covariate effects while accounting for over dispersion and correlation by adding random effects to the additive predictor. GAMMs encompass nested and crossed designs and are applicable to clustered, hierarchical and spatial data.

Nonparametric functions by using smoothing splines, jointly estimate the smoothing parameters and the variance components by using marginal quasi-likelihood are estimated. This marginal quasi-likelihood approach is an extension of the restricted maximum likelihood (REML) approach used by Wahba (1985) [18] and Kohn et al., (1991) [30], in the classical nonparametric regression model and by Zhang et al., (1998) [37], Brumback and Rice [7] and Wang [31] in Gaussian nonparametric mixed models, where they treated the smoothing parameter as an extra variance component. Because numerical integration is often required by maximizing the objective functions, double penalized quasi-likelihood (DPQL) is proposed to make the approximate inference. Frequentist and Bayesian inferences are compared. A key feature of the method proposed is that it allows us to make systematic inference on all model components of GAMMs within a unified parametric mixed model framework. Specifically, our estimation of the nonparametric functions, the smoothing parameters and the variance components in GAMMs can proceed by fitting working GLMM using existing statistical software (R statistical software), which iteratively fits a linear mixed model to a modified dependent variable.

Finally, GAMM [11, 19] represents the model with higher flexibility and complexity, where mixed effects, smooth terms and a non-normal response are admitted. When the data are sparse (e.g. binary), the DPQL estimators of the variance components are found to be subject to considerable bias. A bias correction procedure is hence proposed to improve its performance. A detail discussion about GAMM has been reported in the next section [11, 19].

GAMMs applied here to study the relationship between 24

different blocks with their geographical locations (in terms of longitude and latitude) along with the information of various soil parameters nearly 5 years (2000-2005) and the Soil pH in Burdwan district, India. A mixed model (fixed and random effects) structure is important here because Soil pH may change with time and not only on time but also on the geographical location, fertility index of soil and crop intensity. The aim of this present study is to find this complex relationship between response and cofactors using semi-parametric regression technique [26] under GAMMs.

2. Materials and Methods

2.1. Materials

Soil analysis data from year 2000 to 2005 of Burdwan district has been collected from Soil Testing Laboratory, Burdwan. Burdwan district has 24 major agricultural blocks. Geographically they are situated between 23°53' and 23°56' N latitude to 88°25' and 86°48' E longitude. Total area of this Burdwan district is 7024 sq. km. and out of this approximately 60 percent area is agricultural land [23]. On an average 70 numbers of soil samples has been come from each block per year. Each sample observation measures four soil parameters namely- pH of soil, fertility index of nitrogen, phosphate and potassium. For soil pH 6.5-7.5 range is the normal condition, the below and above value of this range are designated as acidic and alkaline respectively. [2]. Cropping intensity is a soil parameter used in this study which indicates number of times a particular agricultural land cultivated in a year. Information of cropping Intensity for each block has been collected from the Agricultural Annual Report of Burdwan District [4] (Annual Report, 2005). Geographical location of each block in terms of longitude and latitude are also been incorporated in this dataset.

Table 1 shows the variable name, description and the nature of the variables namely - blocks, geographical location of each blocks in terms of longitude, latitude, fertility index of nitrogen, fertility index of phosphate, fertility index of potassium and cropping intensity. All of these six parameters are continuous in nature. Here we have considered Average soil pH (Mean pH) as the dependent variable (response variable) and the remaining others are treated as the independent or explanatory variables (cofactors). Taking all of this information along with the blocks and years a longitudinal dataset is presented here for possible analysis. Aim of this present study is to explore the relationship between Mean pH with the other cofactors and also to find the spatial gradient of soil pH.

Table 1. *Variable descriptions with their nature.*

Variable	Description	Variable nature
Block	Twenty four Blocks of Burdwan District	Discrete Variable
Year	Time period, from year 2000 to 2005	Discrete Variable
Mean pH	Average soil pH of each block per year	Continuous variable
Lat	Latitude of each block in minute.	Continuous variable
Lon	Longitude of each block in minute	Continuous variable
N	Fertility index of Nitrogen of each block per year	Continuous variable
P	Fertility index of Phosphate of each block per year.	Continuous variable
K	Fertility index of Potassium of each block per year.	Continuous variable
C.I	Cropping Intensity of each block in percentage.	Continuous variable

2.2. Methods

2.2.1. Generalized Additive Mixed Model (GAMM)

Consider n pairs of observations $(x_{i,}\, y_i)$, where y_i is an observation of random variable, Y_i with expectation, $\mu_i \equiv E(Y_i)$. Y is called response variable or dependent variable, while x is the predictor or independent variable. In the case of a fixed design, the simplest model which describes the relationship between x and y is:

$$y_i = \mu_i + \epsilon_i \tag{1}$$

where, $\mu_i = x_i\alpha$ and α is an unknown parameter while, $\epsilon_i's$, called random errors, are mutually independent random variables, supposed to be $\epsilon_i \in N(0, \sigma^2)$.

If there are more than one predictor, x_j, where $j = 1,2, \ldots, p$ is the number of different predictors, the equation (1), using matrix notation, becomes

$$y = X\alpha + \epsilon \tag{2}$$

where y is the n X 1 vector of the response, X is a n X p matrix of predictor variables, usually called the design matrix of the model, α is a p X 1 vector of unknown parameters and ϵ is n X 1 vector of random errors, with $\epsilon \in N(0, I\sigma^2)$. The vector 0 denotes a vector with n zero's and I is the identity matrix of order n X n.

The linear model in (2) is based on many limiting assumptions, which are:

Linearity: the dependence between variables can be described only by a straight line and it implies the estimation of parameters (the intercept and the slope parameters for each one of the independent variables);

Homoscedasticity: the error variance is the same whatever is the value of the explanatory variable, $Var(\epsilon|X = x) = \sigma^2 \; \forall \; x$;

Normality: the error is normally distributed, $\epsilon \in N(0, I\sigma^2)$;

Independence: the errors are uncorrelated.

All those assumptions are useful simplifications to carry out inference procedures, but in real cases if data do not comply with them, the model loses validity.

The incorporation of random effects generalizes in some

way the model (2). Let consider q vectors of predictor variables, z of length n. A Linear Mixed Model (LMM) can be easily built as an extension of the Linear Model (LM) and has the form

$$y = \mathrm{X}\alpha + \mathrm{Z}b + \epsilon \quad (3)$$

where b is a q X 1 vector containing random effects, $b \in N(0, G_\theta)$ while the vector of random errors has order n X 1 and $\epsilon \in N(0, \mathrm{R})$. Both b and ϵ are unobservable. The matrix Z is the design matrix for the random effects and has order n X q. The covariance matrix G_θ is positive definite and depends on unknown parameters θ, usually called variance components. Finally R is a positive definite matrix, sometimes used to model residual correlation. Usually it is equal to $I\sigma^2$ matrix. The basic assumptions for (3) are that the random effects and errors have mean zero and finite variances. Typically, the covariance matrices $G_\theta = cov(b)$ and $R = cov(\epsilon)$ involve some unknown dispersion parameters, or variance components. It is also assumed that b and ϵ are uncorrelated.

These models have the ability to model the mean structure (fixed effects) and the covariance structure (random effects and random errors) simultaneously.

LM and LMM permit only Gaussian responses. Generalized Linear Model (GLM) [21] (with only fixed effects) and Generalized Linear Mixed Model (GLMM) [6] (with both fixed and random effects) allow response to follow also some other distribution.

Thus, a GLM has the form

$$G(y) = \mathrm{X}\alpha + \epsilon \quad (4)$$

and a GLMM is represented as

$$\mathrm{G(y)} = \mathrm{X}\alpha + \mathrm{Z}b + \epsilon \quad (5)$$

where $\mathrm{G}(\cdot)$ is a monotonic link function. If $\mu^b \equiv E(\mathrm{y} \mid \mathrm{b})$, is the conditional mean of the response, model (5) can also be written

$$G(\mu^b) = \eta = \mathrm{X}\alpha + \mathrm{Z}b \quad (6)$$

and η is usually called linear predictor of the model, while, in the case of more than one covariate, $\eta_j = \mathbf{X}_j\alpha_j$, represents the partial effect of covariate x_j. The assumptions of this generalized model is firstly, response belongs to one exponential family distribution (that includes Gaussian and categorical responses), secondly, the mean of the observation is associated with a linear function of some covariates through a link function $G(\cdot)$ and finally the variance of the response is a function of the mean.

A GAMM is just a GLMM in which part of the linear predictor is specified in terms of smooth functions of covariates [19]. A GAMM [11, 19], represents the model with higher flexibility and complexity, where mixed effects, smooth terms and a not normal response are admitted. A GAMM has the following structure

$$G(\mathrm{y}) = \mathrm{X}^*\alpha + \sum_{j=1}^{p} f_j(x_j) + \mathrm{Z}b + \epsilon \quad (7)$$

where $\mathrm{G}(\cdot)$ is a monotonic differentiable link function, α is the vector of fixed parameters; X^* is the fixed effects model matrix, the f_j is the smooth function of covariate x_j (and it is centered), Z is the random effects model matrix, $b \in N(0, G_\theta)$ is the vector of random effects coefficients with unknown positive definite covariance matrix G_θ, $\epsilon \in N(0, \mathrm{R})$ is the residual error vector with covariance positive definite matrix R. The structure of those models allows element of the response vector, y, to be no longer independent [35].

In analogy to (5), the conditional mean of the response, μ^b, is linked to the linear predictor, η, and model (7) can be written

$$G(\mu^b) = \eta = \mathrm{X}^*\alpha + \sum_{j=1}^{p} f_j(x_j) + \mathrm{Z}b + \epsilon \quad (8)$$

And $\eta_j = f_j(x_j)$ is the partial effect of covariate x_j.

Model (8) encompasses various study designs, such as clustered, hierarchical and spatial designs. This is because it is possible to specify a flexible covariance structure of the random effects b. Note that in the generalized additive model framework, linear and polynomial models are specific cases of the more general additive model, when smooth effects reduce to linear.

The relationship between Mean.pH and all predictors was initially checked through all the frequentist and classical approaches using R statistical software (packages:\SemiPar, \mgcv", \nlme" and \MASS" as appropriate). REML algorithm was set as the estimation procedure of all models and the P-spline as the basis for smooth functions. Initially independent models were constructed for all covariates to check their effect on the Mean.PH and, if it resulted significant, its nature (linear or nonparametric). Finally GAMM has been applied to this longitudinal dataset for better output using R software with the help of the library packages \amer" and others supported packages. For this present study the GAMM model structure takes the form given below:

$$G(Mean.pH_{ij}) = Block_i + Year_j + Lat_i + f(Lon_i) + f(C.I_i) + K_{ij} + f(N_{ij}) + f(P_{ij}) + \epsilon_{ij} \quad (9)$$

$$\forall\, i = 1,2,\ldots,24;\ \forall\, j = 1,2,\ldots,5$$

After comparison between equation (8) and (9), it is quite understandable that the variable $Mean.pH_{ij}$ is a response which indicates the Average soil pH of *i-th* block for the *j-th* year. Similarly, other cofactors can be defined according to their usual meaning. $G(\cdot)$ is the monotonic link function and $f(\cdot)$ is the smooth function. This model formulation can be extended to include multiple smooth terms, other random effects and a linear predictor in the classical sense of linear regression: Just concatenate the unpenalized parts of the smooth terms to the design matrix of the fixed effects and the penalized parts of the smooth effects to the design matrix of the random effects. For smooth functions truncated power basis functions are applied here. The variable Block entered as a random effect part in this model whereas Year, Latitude,

Potassium (K) entered as a fixed effect and finally Longitude, Nitrogen, Phosphate and Cropping intensity entered as a smoothing covariates. For more detail about R software applications in this domain see [27, 28, 33].

2.2.2. Bivariate Smoothing

One of the more attractive features for Geostatistical data is to handling of the bivariate smoothing problem. As we shown in this section allows for bivariate smoothers to be incorporated into additive models.

In this section we consider the situation where data are available on a response $\mathcal{Y}$ (Mean pH) and bivariate predictors $\mathcal{X} \in \mathbb{R}^2$ (Longitude and Latitude) and we want to fit

$$y_i = f(x_i) + \varepsilon_i \quad (10)$$

Where 'f' is a smooth bivariate function. In many applications $\mathcal{X}_i$'s represents a geographical location, but may also represent two continuous predictors for which additivity is not reasonably assumed.

Estimation of 'f' in the above equation is done by using radial basis functions approximation; with the family of basis functions corresponding to the thin plate spline family. In the notation given there the case m = 1 corresponds to

$$f(x) = \beta_0 + \beta_1^T x + \sum_{k=1}^{K} \left(u_k \left|\left| \mathrm{x} - \kappa_k \right|\right|^2 \log \left\| \mathrm{x} - \kappa_k \right\| \right) \quad (11)$$

Here $\mathcal{K}_1$, ……., $\mathcal{K}_k \in \mathbb{R}^2$ are a set of knots that "cover" the space of the $\mathcal{X}_i$. For more details see [26, 27, 33].

This present article introduced bivariate smoothing using Mean.pH as a response variable and Longitude and Latitude are the possible cofactors. One of the major aim of this study is to find the spatial gradient in soil pH of district Burdwan which reported to the next section.

3. Results

Random effects and fixed effects of cofactors under GAMM are presented Table 2 and 3. Correlation of the fixed effects is presented in Table 4. Akaike information criterion (AIC) value for this selected model is 204.9, which is minimum with compare to others. It is well known that AIC selects a model which minimizes the predicted additive errors and squared error loss. It is not necessary that all the selected effects are significant [16].

3.1. Random Effects Part of Cofactors under GAMM

The Null hypothesis H_o: Random variance $\sigma^2 = 0$ Vs Alternative hypotheses $H_1 : \sigma^2 > 0$. Here all random variance viz. cropping intensity, P, N, longitude of the cofactors are greater than zero, so they have significant effects in terms of smoothing [1, 23, 32]. Blocks in this study are treated as the random intercept. The smoothing part of the cofactors is treated as random effects in GAMM model. It is observed from Table 2 that choice of these cofactors as a smoothing function is absolutely appropriate. Figure 2 shows plotting of the generated grid values plus the fitted values and confidence intervals of the smoothing variables against its original variable values. Simply, the smoothing terms of this model is plotted against their estimated function values with point wise confidence intervals. It is very clear from the Figure 2 that the smoother part of these cofactors has significant effects in the model fitting. Through this random effects the correlation between cofactors also been controlled by GAMM and which can be checked from Figure 3 (mainly the histogram and normal probability plot of residual values shows perfect Gaussian distribution).

3.2. Fixed Effects Part of Cofactors under GAMM

In Table 3, fixed effects of GAMM suggest that the positively significant effects of the linear terms are found with cofactors namely Latitude and Fertility Index of K, whereas Year is negatively significant. In linear part of the smoothing terms, Fertility Index of P is positively significant. The estimated values of the cofactors show the degree of effects to the corresponding response. Similarly the standard errors are very small for these cofactors which ensure the stability of the model [8, 34]. Finally, significant cofactors for this model are selected through p-value [8, 26, 35]. If the p-value of the corresponding cofactor is less than 0.05 then the cofactor can be treated as a significant factor for the response variable under 5% level of significance. Table 3 shows that some cofactors are significant for Mean pH. Correlation coefficient between fixed effects terms are shown in Table 4, which indicates that the fixed part of the cofactors alone with other non -random cofactors shows a correlated structure between them.

Figure 1 shows model summary plot which gives the plot between Cofactors and Response variable. In GAMM, it is described earlier that some cofactors like latitude, year and fertility index of K entered in the model parametrically or linearly and others cofactors like longitude, fertility index of N, P and C. I. entered non parametrically or non-linearly.

Table 2. *Results of the random effects and smoothing cofactors for soil testing data analysis from GAMM fit.*

Covariate	Variance	Standard deviation
Block (Intercept)	0.0624101	0.249820
f.C.I.	0.0524603	0.229042
f.P	0.0075691	0.087000
f.N	0.0025706	0.050701
f.Lon	0.0075997	0.087176
Residual	0.1477342	0.384362

Table 3. *Results of the fixed effects for soil testing data analysis from GAMM fit.*

Covariate	Estimate	Standard Error	t- value	P Value
Lat	0.003600	0.000398	9.043	0.000019***
Year	-0.134356	0.027636	-4.862	0.000067***
K	0.405114	0.141577	2.861	0.006665**
Lon.fx1	0.016340	0.233120	0.070	0.944554
N.fx1	-0.064563	0 .096694	-0.668	0.507967
P.fx1	0.31 4645	0.127157	2.474	0.017777*
C.I.fx1	-0.487333	0.266433	-1.492	0.135454

AIC value	BIC	logLik	Deviance	REMLdev
204.9	241.1	-89.43	147.8	178.9

Table 4*. Value of Correlation coefficient between fixed effects for soil testing data analysis.*

Covariate	Lat	YEAR	K	Ln.fx1	N .fx1	P.fx1
Year	-0.059					
K	-0.490	-0.236				
Ln.fx1	0.440	-0.038	-0.036			
N .fx1	0.210	-0.151	-0.019	-0.006		
P.fx1	0.261	-0.196	-0.113	0.099	0.066	
C.I.fx1	0.543	0.033	0.042	-0.139	0.059	-0.224

3.3. Model Checking Plot and Bivariate Smoothing Plot

Regression analysis method generally consists two parts, one is the numerical results and another is model checking plot. Analysis is certified after verifying both of these two. Figure 3 shows the model checking plots of GAMM. The first and second plot of Figure 3 shows the residual plot and the fitted value vs residuals plot respectively. Residuals and fitted value plot is very important in regression analysis if it shows any pattern or trends except randomness it will treated a bad fitting of the model. In both of these two plots have no such pattern to be identify and it is perfectly random which ensure that the describe model fits all the data well enough. Other two plots in Figure 3 show histogram and normal probability plot of residuals. The original distribution of the response variable i.e. mean pH follows Gaussian distribution. Here, the histogram of standard residuals is also distributed normally with mean zero, which indicates that the model is fitted well. Normal probability plot in figure 3 shows that the population quintile and sample quintile are exactly matches with each. It also shows that all the data are fitted in the model. So considering the entire model checking plots it is found that the GAMM fits the soil testing data very accurately.

Figure 4 which are generated by Bivariate Smoothing techniques shows, a significant change in Mean pH from south to north direction (Latitude wise) of the district Burdwan. From Table 3 and Figure 4 it has been established that there is a spatial gradient in the soil pH of Burdwan district.

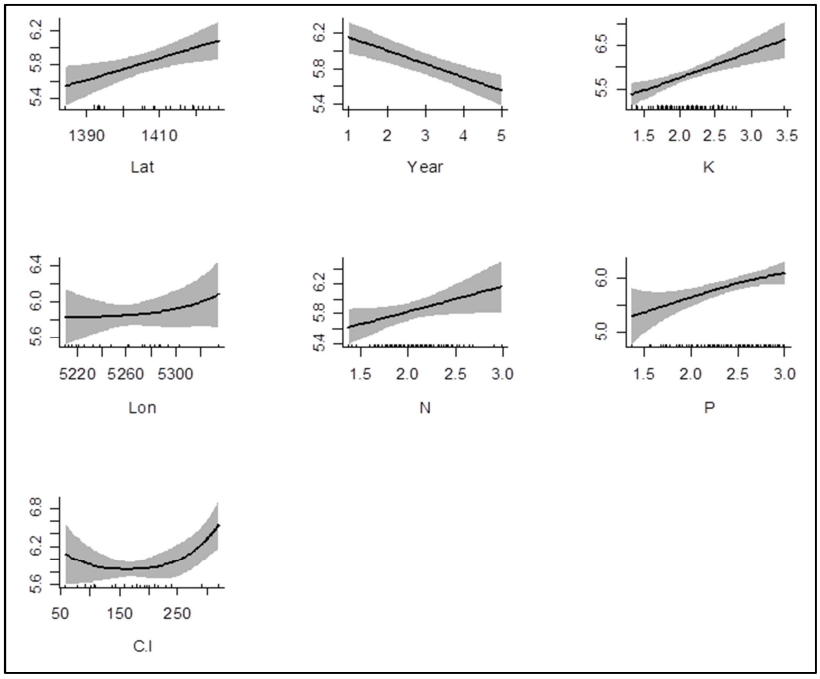

Figure 1. *Plot of the associations of different cofactors with Mean pH under GAMM fit.*

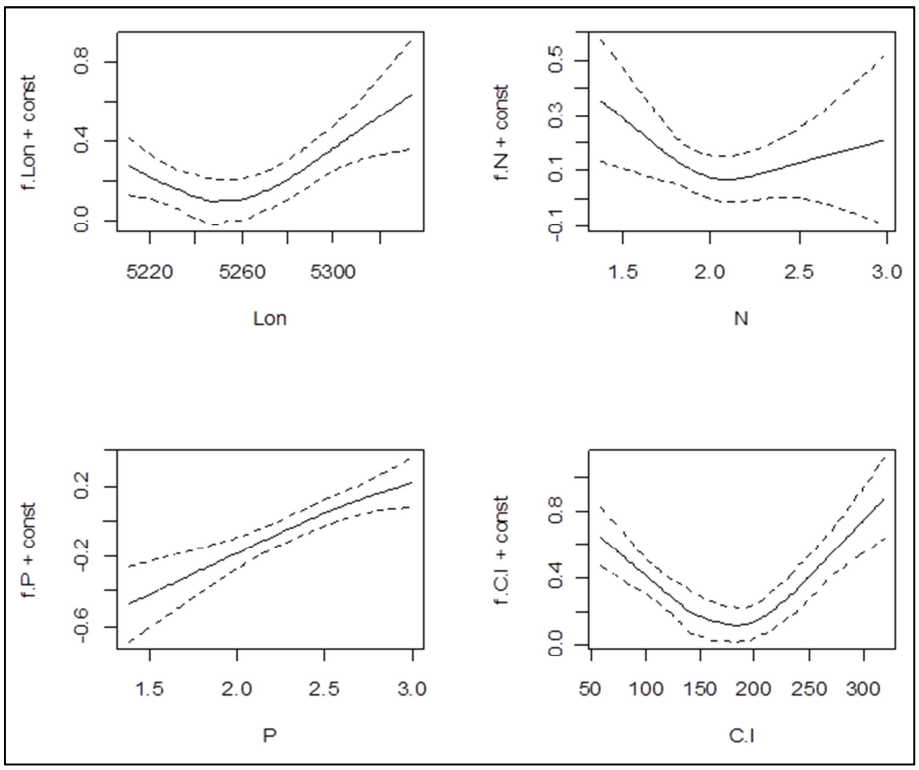

Figure 2. *Plot of smoothing functions and their estimated functional value under GAMM fit.*

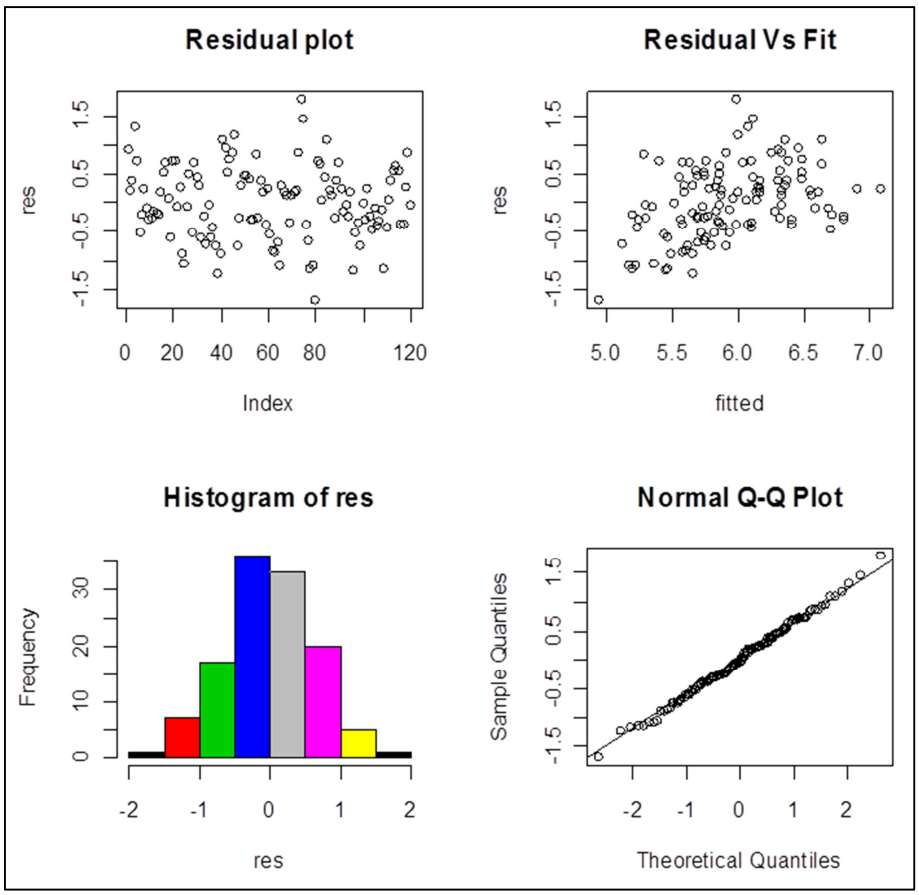

Figure 3. *Model Checking Plots under GAMM (a) Residuals plot (b) Residuals vs Fitted value plot (c) Histogram of residuals (d) Normal Probability plot of residuals.*

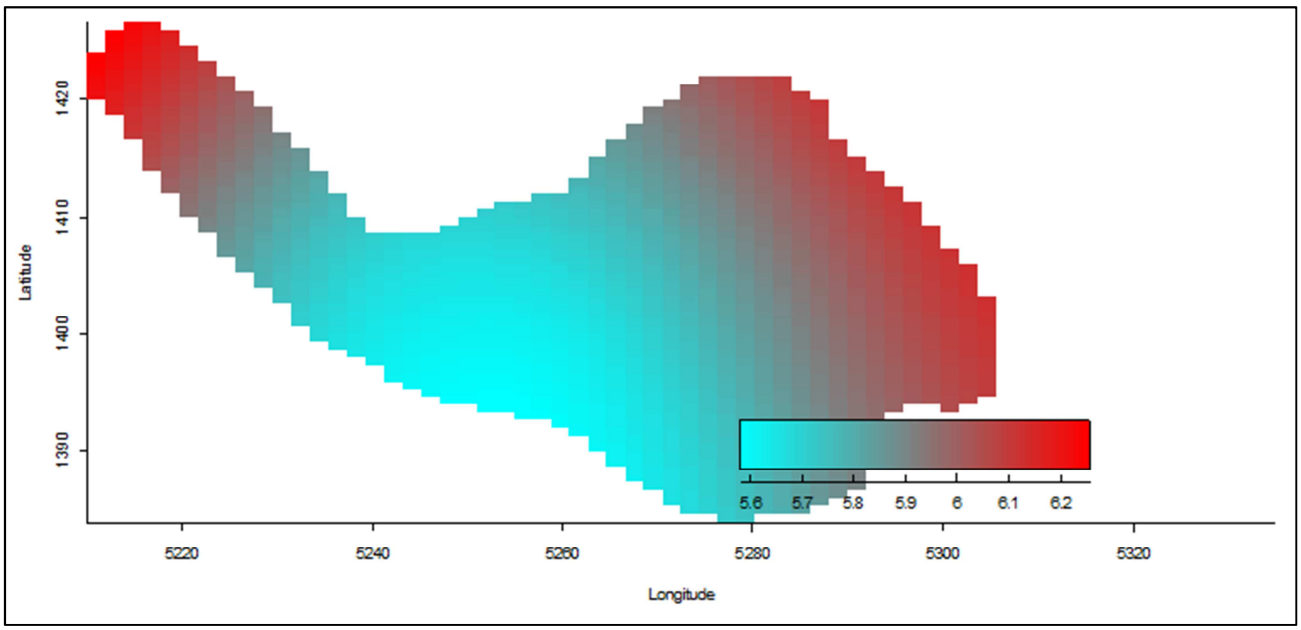

Figure 4. *Bivariate smoothing plot of Mean. pH with geographical locations of Burdwan district.*

4. Discussion

Cofactors in smoothing function have two parts in GAMM, random effect part and fixed effect part. In random effects part random variance and standard deviation of cofactors are produced and all the cofactors under smoothing is significant according to the hypothesis. Another part of the analysis shows the fixed effect of cofactors. Fixed effects suggest that the linear terms like Latitude, Fertility Index of K and year are significant and linear part of the smoothing terms like Fertility Index of P is significant. In regression analysis, interpretations of the estimated value of the parameters are well established [8, 34]. Also among all cofactors which are significant in terms of testing using p-value discussed in various regression books [8, 21, 34].

One unit change in fertility Index of P affects 0.31 amount increment in Mean pH. From the fixed effects estimation, it is found that one unit increment of fertility index of K increase pH by 0.4051 units. Time periods change in Year significantly decreases Mean pH by 0.1343 units. Fertility index of N decreases Mean pH as well as Crop Intensity also decreases Mean pH at a certain level though they are not significant factors. The effect of Latitude is significant for predicting Mean pH.

The most important finding in this study is that there is a significant change in Mean pH across the whole Burdwan district form south to north direction (latitude wise). Increase in one minute Latitude (moving from south to north) increases mean pH by 0.0036 amount.

Correlation coefficient lies between -1 to 1 but in case of positive correlation if this value is more than 0.7 and in case of negative correlation the value is less than -0.7 known as significant. Here correlation coefficient of fixed effects terms are out of this significant range that is why need not to consider this study under correlated structure.

5. Conclusion

Now a day's application of Generalized additive model (GAM) and Generalized additive mixed model (GAMM) is rapidly growing up in various research disciplines like- medical sciences, social sciences, engineering sciences, industrial fields and also in agricultural area [17, 20, 22]. One of the major reasons for this growing application may be the accessibility, accurate prediction power and complex relationship finding capabilities of GAMM. Generally, longitudinal dataset with time covariate (sometimes known as spatiotemporal data) has a correlated error structures which cannot be analyzed properly using ordinary regression techniques like multiple linear regression or Generalized linear model (GLM) except introducing mixed models [8, 21]. This present agricultural dataset is a very complicated spatiotemporal dataset, because the soil pH varies according to the different blocks as well as with the time periods presented through the variable years. The main beauty of this study is that, it is not limited to find only the relationship between soil parameters but also tries to find the soil pH gradient in the entire Burdwan district, which is a very new and modern idea in this agricultural field. Once we able to find this, then it is very helpful in cultivation because before starting to cultivate a land we already have an prior knowledge about the soil pH and the other soil parameters of that land through this model fitting. This present study gives an insight into the relationship between soil parameter with geographical location. Few relationship are well established like negative influences of year, fertility index of nitrogen and cropping intensity on soil pH. But the positive influences of fertility index of P and K on soil pH are not well established which need further research. Contribution of different parameters on soil pH can also be predicted from this model. Possible value of soil pH with a new set of cofactors can be found out using this model. Altering

values of different cofactors, targeted value of soil pH can be achieved from this study.

References

[1] Aerts, M., Claeskens, G., Wand, M. P., 2002. Some theory for penalized spline generalized additive models. Journal of Statistical Planning and Inference. 103, 455–470.

[2] Ali, S. J, 2005. Fertilizer Recommendation for Principal Crops and Cropping Sequences of West Bengal. Booklet No.1 Department of Agriculture, Covernment of West Bengal, Kolkata – 700001

[3] Antoaneta, A., 1995. Effects of Fertilizers Application And Soil pH on the Acidic and Sorptio Properties of Maize Trees. Bulg. J. Plant Physiol. 21(1), 52–57.

[4] Annual Report, 2005. Office of the Deputy Director of Agriculture, Burdwan District, West Bengal, India.

[5] Biswas, T. D., Mukherjee, S. K., 2006. Text Book of Soil Science. Tata-McGraw Hill Publishing Company Limited, New Delhi.

[6] Breslow, N. E., Clayton, D. G., 1993. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association. 88, 9–25.

[7] Brumback, B., Rice, J. A., 1998. Smoothing spline models for the analysis of nested and crossed samples of curves. J. Am. Statist. Ass. 93, 961-1006.

[8] Chatterjee, S., Hadi, A. S., 2006. Regression Analysis by Example, fifth ed. John Wiley & Sons, New Jersey.

[9] Das, R. N., Kim, J., Mukherjee, S., 2017. Correlated log-normal composite error models for different scientific domains. Model Assisted Statistics and Applications. 12, 39–53

[10] Das, R. N., Mukhopadhyay, A. C., 2017. Correlated random effects regression analysis for a log-normally distributed variable, Journal of Applied Statistics. 44:5, 897-915

[11] Fahrmeir, L., Lang, S., 2001. Bayesian inference for generalized additive mixed models based on Markov random field priors. J. R. Statist. Soc. 50, 201 - 220.

[12] Green, P. J., Silverman, B. W., 1994. Nonparametric Regression and Generalized Linear Models: A roughness penalty approach, first ed. Chapman and Hall, London.

[13] HaÈrdle, W., 1990. Applied Non-parametric Regression Analysis, first ed. Cambridge University Press, Cambridge.

[14] Hastie, T., Tibshirani, R., 1990. Generalized Additive Models, first ed. Chapman and Hall, London.

[15] Hart, J. D., 1991. Kernel regression estimation with time series errors. J. R. Statist. Soc. B, 53, 173-187.

[16] Hastie, T., Tibshirani, R., Friedman, J., 2001. The elements of statistical learning, first ed. Springer, USA

[17] Hastie, T., Tibshirani, R., 1995. Generalized additive models for medical research. Statistical Methods in Medical Research. 4, 187-196

[18] Kohn, R., Ansley, C. F., Tharm, D., 1991. The performance of cross-validation and maximum likelihood estimators of spline smoothing parameters. J. Am. Statist. Ass. 86, 1042-1050.

[19] Lin, X., D. Zhang, 1999. Inference in generalized additive mixed models using smoothing splines. Journal of the Royal Statistical Society. Series B, 61, 381–400.

[20] Mamouridis, V., 2011. Additive Mixed Models applied to the study of red shrimp landings: comparison between frequentist and Bayesian perspectives.

[21] McCullagh, P., Nelder, J. A., 1989. Generalized Linear Models, 2nd ed., Chapman and Hall, London.

[22] Mukherjee S., Kapoor S., Banerjee P., 2017. Diagnosis and Identification of Risk Factors for Heart Disease Patients Using Generalized Additive Model and Data Mining Techniques. J Cardiovasc Disease Res.; 8(4):137-44.

[23] Official website of Burdwan District. West-Bengal, India, bardhaman.nic.in/home.html

[24] Padoan, S. A., Wand, M. P., 2008. Mixed Model- based Additive Models for Sample Extremes. Statistics & Probability Letters. vol. 78, issue 17, 2850-2858.

[25] Rice, J. A., Silverman, B. W., 1991. Estimating the mean and covariance structure non-parametrically when the data are curves. J. R. Statist. Soc. B, 53, 233-243.

[26] Ruppert, D., Wand, M. P., Carroll, R. J., 2003. Semi parametric Regression, first ed. Cambridge University Press, New York.

[27] Scheipl, F., 2010. amer version 0.6.5.: Using lme4 to fit Generalized Additive Mixed Model. R package. http://cran.r-project.org.

[28] Scheipl, F., Munchen, L., 2010. amer: Using lme4 to fit Generalized Additive Mixed Models.

[29] Verbyla, A. P., 1995. A mixed model formulation of smoothing splines and test linearity in generalized linear model. Technical Report 95/5, Department of Statistics, University of Adelaide, Adelaide.

[30] Wahba, G., 1985. A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. Ann. Statist. 13, 1378-1402.

[31] Wang, Y., 1998. Mixed effects smoothing spline analysis of variance. J. R. Statist. Soc. B, 60, 159-174.

[32] Wand, M. P., 2003. Smoothing and mixed models. Computational Statistics. 18, 223–249.

[33] Wand, M. P., Coull, B. A., French, J. L., Ganguli, B., Kammann, E. E., Staudenmayer, J, Zanobetti, A., 2005. SemiPar 1.0 Functions for semiparametric regression. R package. http://cran.r-project.org.

[34] Wu. H., Zhang. J., 2006. Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed Effects Modeling Approaches, John Wiley & Sons, New Jersey

[35] Wood S. N., 2006. Generalized Additive Models: An Introduction with R, Chapman & Hall/CRC Press.

[36] Zeger, S. L., Diggle, P. J., 1994. Semi-parametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. Biometrics, 50, 689-699.

[37] Zhang, D., Lin, X., Raz, J., Sowers, M., 1998. Semi-parametric stochastic mixed models for longitudinal data. J. Am. Statist. Ass. 93, 710-719.