Research Article
Volume 10:9, 2021

Journal of
Applied & Computational Mathematics

ISSN: 2168-9679 Open Access

A Spatial-Nonparametric Approach for Prediction of Claim Frequency in Motor Car Insurance

Kipngetich Gideon*

Department of Pure and Applied Sciences, Pan African University, Nairobi, Kenya

Abstract

Spatial modeling has largely been applied in epidemiology and disease modeling. Different methods such as generalized linear models (GLMs), Poisson regression models, and Bayesian Models have been made available to predict the claim frequency for forthcoming years. However, due to the heterogeneous nature of policies, these methods do not produce precise and reliable prediction of future claim frequencies; these traditional statistical methods rely heavily on limiting assumptions including linearity, normality, predictor variable independence, and an established functional structure connecting the criterion and predictive variables. This study investigated how to construct a spatial nonparametric regression model estimator tor for prediction of claim frequency of insurance claims data. The study adopted a nonparametric function based on smoothing Spline in constructing the model. The asymptotic properties of the estimators; normality and consistency were derived and the inferences on the smooth function were derived. The simulation study showed that the estimator that incorporated spatial effects in predicting claims frequency is more efficient than the traditional Simultaneous Autoregressive model and Nonparametric model with Simultaneous Autoregressive error. The model estimator was applied to claims data from Cooperative Insurance Company insurance in Kenya with n = 6500 observations and the findings showed that the proposed model estimator is more efficient compared to the Local Linear fitted method, which does not account for spatial correlation. Therefore, the proposed method (Nonparametric spatial estimator) based on the findings has significant statistical improvement of the existing methods that are used for the prediction of claims. The study had a number of limitations, where the data used in the study is Lattice data (without a coordinate system); therefore, there was difficulty in classifying the claims to a specific area in the region (County).

Keywords: Nonparametric • SAR • Smoothing Spline • Claims • CIC • Spatial

Abbreviations

CIC: Cooperative Insurance Company

DP: Dirichlet Process

GB2: Generalized Beta of the Second Kind

PLS: Penalized Least Square

SAR: Spatial Autoregressive

LL: Local Linear

Introduction

Insurance has a fundamental role in providing financial protection and offering a transfer risk in exchange for an insurance premium. Therefore, estimation of the right premium for the policyholders is the noblest task in the insurance business. Insurance companies provide insurance to policyholders, and in turn, the policyholders have to pay insurance premiums through the agreed time (periodically). Due to competition in the market, charging a fair premium according to the policyholder's expected loss is profitable for the insurer. The company will attract more customers and boost customer relationships with the insurer. The amount of premiums paid by the policyholders is determined from estimates of their expected claim frequencies and the claims' severity. Setting precise and reliable estimates of claims frequencies has extreme importance.

***Address for Correspondence:** Kipngetich Gideon, Department of Pure and Applied Sciences, Pan African University, Nairobi, Kenya, Tel: 0708366735; E-mail: gideonlangat90@gmail.com*

Received 01 September 2021; **Accepted** 15 September 2021; **Published** 24 September 2021

Recent studies on spatial modeling have been rapidly applied in many fields; epidemiology, public health, and the Insurance sector. Different models commonly employed to fit current claims data and predict claim frequencies are Poisson regression models, generalized linear models, Credibility models, Bayesian Models. However, from the available literature, these models appear to be relatively inflexible. Essentially all models are wrong, but some are useful, which is true when the process being modeled is either not well understood, or the necessary data are not available [1]; the same problem of choosing the model is experienced in the modeling of the claims, Although the generalized linear models provide accurate and fast analysis of insurance data, they fall short because they are defined based on the assumptions, and an incorrect model assumption can cause model misspecification leading to erroneous results. Nonparametric models are deemed to minimize the shortcoming of these standard parametric models since fewer assumptions are made for the model, therefore, suitable for modeling insurance data which are nonlinear, nonparametric models perform better than generalized linear models (GLMs) [2], the only observable problems with modeling using Nonparametric models are the interpretation of some of their curves. When modeling claims and risks, we need to determine their behavior and spatial dependence, and spatial heterogeneity of the data so that the insurer can determine which areas are associated with a higher riskier when determining premiums amount to be paid.

[3] propose a Bayesian nonparametric approach for prediction of claims, here they found out that the model performs better compared to nonparametric GLMs in that it can capture the nonlinear random effects present in the data, [4] also propose a flexible nonparametric loss model for prediction of the claims, they found out that having flexible multivariate model may allow actuaries to estimate the dependence between different risk classes and different lines of business and this topic need to be explored further, and introduce the idea of using nonparametric data mining approach to modeling of the claims and prediction of risk, here the approach classify risk and predict claim size based on data. This study's research idea was to build based on the study by [5,6] where they introduce a nonparametric spatial regression model for prediction. This study's primary objective is to construct a spatial nonparametric estimator for the prediction of insurance claims. Therefore, this study's main contribution was to investigate the estimator's performance in the situation of additional covariates in the model and incorporate the

aspect of spatial dependence in constructing a nonparametric estimator for the prediction of insurance claim frequencies.

The main difference between this research and [6, 7] are as follows: (1) estimate a nonparametric spatial model where estimation of the unknown trend g (·) is based on smoothing spline (2) Rather than assuming that spatial correlation takes a particular form (as in SAR), spatial correlation and spatial heterogeneity are considered simultaneously with second order stationarity (3) the estimators' asymptotic properties under mild conditions.

Methods and Materials

In this section, we propose methods of estimating a claim frequency model for prediction. The basic claim frequency model is introduced, stressing the need to introduce a new method of estimating a more flexible claim model where the basic model restriction is relaxed. The spatial effects are incorporated to complete the proposed model.

Claim Model

Claim model is used a generalized linear model in modelling aggregate claims in non-life insurance [8]. This aggregate claim model can be improved by adding a more attractive feature on the way the fit is. Defined the following terms

1. Claim severity is the total claims divided by the number of claims (average size per claim)
2. Claim frequency is the number of claims divided by the duration (average number of claims per unit time)

Most traditional claim models have assumed that the claim follows a Poisson process with a rate ϡ ; on the other hand, claim severities follow a continuous model, and gamma distribution has been extensively used. The aggregate claim model is given by

$$S_t = X_1 + X_2 + \ldots + X_{Y_i} \tag{1}$$

Where S_t is the aggregate claim amount of a given trading yearn *t*, Y_i is random number that denotes the number of claims in a year *t*, X_i with *i = 1,2,...*Y_i is the amount of i^{th} claim realized in the year *t*. Some of the important assumptions as defined in the model equation (1) include

$Y_i \sim Poisson(\lambda)$ for some0 < ϡ < ∞ are iid $\forall i$

Y_i and X_i are independent $\forall i$

X_i are iid $\forall i$

Traditionally in insurance practice, the claim frequency has been modelled following Poisson distribution and X_i has the form of the loss distribution, i.e., separate gamma distribution for the claim severity [9]. Modeling the frequency of claims, let Y_i denotes the number of claims involving *i* policies at time *t*, then the total expected number of claims *Y* is expressed as

$$Y = \sum_{i=1}^{m} Y_i$$

m is the total number of observed policies, Y_i is mostly assumed to follow discrete distribution of which Poisson distribution has been commonly used in many models. Y_i depends on covariates such as the region where the policy was taken, age, sex, type of vehicle, number of claims per policy, years of policy ownership, insured cases number for a user and average claim size, then

$$Y_i \sim Poisson\left(\lambda_i^Y\right)$$

Where mean is given by

$$\lambda_i^Y = t_i exp\left(X_i \beta_i\right)$$

X_i is covariate vector for i^{th} observation and is assumed to have linear relationship with Y_i, t_i denotes the exposure time of policyholder *i* and β_i denotes a vector of unknown regression parameters.

The assumptions on X_i and Y_i are mispecifications, and if the data appear to be nonlinear, it will create a substantial modeling bias. Therefore, a nonparametric method is proposed; the main aim of this nonparametric technique is to relax highly restrictive regression function [10].

Model Estimation

The study proposes a nonparametric regression model to predict the number of claims Y_i, *i = 1,....n* observed in region *J* in order to relax restrictive assumption on the distribution of number of claims and X_i covariates vector for the i^{th} claim. Since claims in each region, *J* has nonlinear relation with the covariates $X_i^{'}s$. The nonparametric form of the model is given by the general form [11, 12].

$$y_i = g\left(x_i\right) + Z_i^T b + \dot{o}_i$$

G (.) is unknown nonparametric function used to model fixed effects, $Z_i^T b$ and ϵ_i for random unobserved effects. Since the form of $Z_i^T b = R_i$ for R_i is unknown, the study aims to estimate R_i that account for the spatial effects.

Let $I = (1,...,1)^T$ and $n = \left(n_1,\ldots,n_N\right)^T$ be two *N* dimensional vectors. We make assumption about the spatial model as

$$Y_i = g\left(X_i\right) + R_i\, var\left(R\right) = \sigma^2 \Sigma i \in \Lambda_n = \left\{1,\ldots,n_1\right\} \times \ldots \times \left\{1,\ldots,n_N\right\} \tag{2}$$

Where, $i = \left(i_1,\ldots,i_N\right)$ in Λ_n will be referred to as site, R_i cater for the spatial effects (Random effects) and the cardinality of Λ_n is $|\Lambda_n| = n$ [13].

Modeling spatial data as a finite realization of a vector stochastic process indexed by $R = \left(R_1,\ldots,R_n\right)^T$ follow joint Gaussian distribution here *E(R) = 0*, $\sigma^2 = var\left(R_i\right)$ and $\forall i \in \Lambda_n$, $\Sigma = \left[\rho\left(R_i, R_j\right)\right]$ correlation coefficient matrix that need to be estimated. For the vector $\left(X_{i1},\ldots,X_{id}\right) \in \Re^d$, $Y_i \in \Re$ and *g* (·), the unknown function *g* (·), need to be estimated for $x = \left(x_1,\ldots,x_d\right) \in \Re^d$, the response variable Y_i is claim frequency and X_i is six dimensional vector consisting of the following explanatory variables: gender, claim amount, age of the policyholder, gender, vehicle age, model of the vehicle and age category of the policyholder.

Estimating *g(x)* at some point $x \in \Re^d$, for X_i in the neighborhood of *x*, *g(X_i)* can be estimated by using smoothing spline [14]. The estimator $g\left(\cdot\right)$ of g $\left(\cdot\right)$ is the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - g\left(X_i;\beta\right)\right\}^2 + \lambda\beta^T\Omega\beta \tag{3}$$

Which can be represented as

$$\| Y_i - G\beta \|^2 + \lambda\beta^T\Omega\beta$$

where $G \in \mathbb{R}^{n\times n}$ is basis matrix defined as

$$G_{in} = \psi_n\left(x_i\right),\ i = 1,\ldots n$$

Where $\psi_1,\ldots\psi_n$ are the truncated power basis functions with knots at $x_1,\ldots$ x_n which is evaluated at the data values

$$\psi_n\left(x\right) = \begin{cases} x_i^n\left(0 \le n \le p\right) \\ \left(x_i - N_{n-p}\right)_+^p, \left(p+1\right) \le n \le N \end{cases}$$

$\left(x - N_{n-p}\right)_+^p = max\left(0, x_i - N_n\right)^p\, n \in \phi$ where *p* is compact interval. p is the degree of the spline and $n_i < \ldots < n_{N-p}$ are fixed points or knots in ϕ .

$\Omega \in \mathbb{R}^{n\times n}$ is the penalty matrix defined as

$$\Omega_{in} = \int g_i^{*}\left(x\right)\psi_n^{*}\left(x\right)dx,\ i = 1,\ldots n$$

Given the optimal coefficients $\hat{\beta}$ minimizing (3), the smoothing spline estimate at *x* is therefore defined as

$$\hat{g}(X_i)=\sum_{n=1}^{n}\hat{\beta}_n\psi_n(x) \quad (4)$$

The term affects shrinking components of estimation $\hat{\beta}$ towards zero. The parameter $\lambda \geq 0$ is the smoothing parameter.

Each computed coefficient $\hat{\beta}_n$ corresponds to a particular basis function ψ_n· the term $\beta^T\Omega\beta$ in (3) imparts more shrinkage on the coefficients $\hat{\beta}_n$ that correspond to wigglier functions $\psi_n(x)$. Hence, as we increase λ, we are shrinking away the wiggler basis functions. Similar, to least squares regression, the coefficients $\hat{\beta}$ minimizing (3) is

$$\hat{\beta}=\left(G^TG+\lambda\Omega\right)^{-1}G^TY=\left(X^TX+n\lambda D\right)^{-1}X^TY$$

The smoothing spline can be seen as linear smoother, therefore we represent as

$$\hat{g}(X_i)=g(x)^T\hat{\beta}=g(x)^T\left(X^TX+n\lambda D\right)^{-1}X^TY \quad (5)$$

which is linear combination of the points Y_i, $i = 1, \ldots n$ and X is a design matrix with entries x_i for $i = 1, \ldots, n$, Y, Y is a vector of the response variables, D is a diagonal matrix with $p+1$ zeros on the diagonal followed by N ones and $n\lambda D$ is a penalty term λ is estimated as

$$CV(\lambda)=\frac{1}{n}\sum_{i=1}^{n}w_i\left(Y_i-g_\lambda^{-i}(X_i)\right)^2 \quad (6)$$

Where g_λ^{-i} is the smoothing spline estimator fitted from the data less the i^{th} data Since $\sum$ in model equation (2) is unknown, we assume that R_i, $i = 1,2,\ldots, n$ is 2^{nd} - order stationary and isotopic process. Then let $C(h)$ and $\gamma(h)$ be covariogram and variogram respectively, where the distance between two locations is given by h [15,13] Then

$$C(h)=\sigma^2-\gamma(h)$$

$\sum=\left[\rho(R_i,R_j)\right]=\left[C(z_i-z_j)/\sigma^2\right]$, while z_i and z_j are the spatial locations associated with error values R_i and R_j, so it is appropriate to estimate $\gamma(h)$ for Σ [16,17]

$$2\hat{\gamma}(h)=\sum_{S(h)}\left[z_i-z_j\right]^2/N(h) \quad (7)$$

$S(h)=\left\{(x_i,x_j):\mid x_i-x_j\mid=h\right\}$, $N(h)$ is a number of distinct pairs in S (h). The characteristics of the estimator (7) can be estimated using Integrated square [18], given by

$$ISE(2\gamma)=\int_{h_1}^{h_k}\{2\hat{\sigma}(h)-2\gamma(h)\}^2dh$$

Where h_1 and h_k are the lags [16]. Therefore, since $\hat{g}(\cdot)$ has been estimated then

$$\hat{R}_i=Y_i-\hat{g}(X_i)$$

Since $R(z_i)$, the error at location z_i, is unobserved, this quantity must be estimated as well. We use the iterative procedure [13]:

Obtain $\hat{g}(\mathbf{x})$, $\hat{g}(\mathbf{X_i}),\mathbf{i}\in\Lambda_n$ by means of (4). Then put $\hat{R}_i^{(1)}=Y_i-\hat{g}(X_i)$

Obtain $\hat{\gamma}^{(1)}(h)$ for $h\in\left\{\mid z_i-z_j\mid,i,j\in\Lambda_n\right\}$ and $\hat{\sigma}^{2(1)}$.

Using (5), obtain $\hat{g}(X_i)=g(x)^T\left(X^TX+m\lambda D\right)^{-1}X^TY$ and $\tilde{g}^{(1)}(\mathbf{X}_i),\mathbf{i}\in\Lambda_n$.

Set $\hat{R}_i^{(2)}=Y_i-\tilde{g}^{(1)}(X_i)$ and go to 2. Step 3 then produces $\tilde{m}^{(2)}(\mathbf{x})$ and $\tilde{m}^{(2)}(\mathbf{X_i}),\mathbf{i}\in\Lambda_\mathbf{n}$.

Repeat this process to obtain $\tilde{g}^{(r)}(\mathbf{x})$ and $\tilde{g}^{(r)}(\mathbf{X}_i),\mathbf{i}\in\Lambda_n$.

The study selects $\epsilon > 0$, i.e., $\epsilon = 0.001$ and the procedure ends when $\mid\tilde{g}^{(r)}(\mathbf{x})-\tilde{g}^{(r-1)}(x)\mid<\epsilon$.

As proposed by [5, 6] R^2 is used to assess the performance of predictors, given by

$$R^2=1-\frac{\sum_{i=1}^{n}\left[g(x_i)-\hat{g}(x_i)\right]^2}{\sum_{i=1}^{n}\left[g(x_i)-\bar{g}\right]^2}$$

$\bar{g}$

Where $\bar{g}$ is the mean of $g(x_i)$, i=1,...,n.

Theoretical properties of Estimator

To obtain asymptotic results, we impose the following assumptions on model (2)

A1. The random field $\{X_i, i\in\Lambda_n\}$ is strictly stationary.

A2. The function g (·) is twice differentiable and its matrix of second derivatives at x denoted by g"(·) is continuous at all $x\in$.

A3. The process R_i, $i = 1,2,\ldots, n$ is 2^{nd} order stationary and isotopic, further $\exists$ a $\in > 0$ such that $E\left(\mid Y_i\mid^{2+\delta}\right)<\infty$ for $i = 1,2, \ldots, n$

Theorem 1: Asymptotic Normality

In addition to A1-A3, suppose that $\{\epsilon_i\}_{i=1}^n$ are iid with mean 0 and variance $\sigma^2 I_n$ then, for any $x_i\in$ and $k_o\geq C_n^{1/2m+1}$ for some constant C > 0 then

$$\frac{\hat{g}(\cdot)-\left(g(\cdot)+b(x_i)\right)}{\sqrt{var\left(\hat{g}(\cdot)\right)}}\xrightarrow{d}N(0,1)\quad as\quad n\to\infty \quad (8)$$

Where $b(x_i)$ is asymptotic bias [19], given by

$$b(x_i)=E(\hat{g}(x_i)-g(x_i)=b(x_i)$$

Proof of Theorem 1:

For m > 1, $S(m,t)$ is a set of spline functions with knots $t=\{0=t_0<t_1<\ldots<t_{k+1}=1\}$ of step functions with jumps at the knots and for m ≥ 2

$$S(m,t)=s\in C^{m-2}[0,1]:s(x)$$

Expressing function in $S(m,t)$ in terms of B-splines for fixed m and t, let

$$R_{i,m}(x)=(t_i-t_{i-m})[t_{i-m},\ldots,t_i](t-x)_+^{m-1},i=1,\ldots,k=k_0+M$$

Where $[t_{i-m}, \ldots, t_i]f$ denote the m^{th} – $order$ divide difference of the function f and $t_{i=}t_{min(max(i,0),ko+1)}$ for any $i = 1$- $M, \ldots, k$ we assume that

$$max_{1\leq i\leq k_0}\left|h_{i+1}-h_i\right|=o\left(k_0^{-1}\right)$$

And $h/min_{1\leq i\leq k_0}h_i\leq M$, where $h_i=t_i-t_{i-1},h=max_{1\leq i\leq k_0}h_i$ and M > 0 (predetermined constant) this assumption ensure that $M^{-1}<k_0h<M$, which is necessary for numerical assumptions

Let $D_n(x)$ be an empirical distribution function of $\left(x_i^n\right)_{i=1}$ with a positive continuous density $d(x)$ this implies $G(d)=\int_0^1R(x)R'(x)d(x)dx$ Then

$$\frac{E\left(\hat{g}(x)\right)}{\sqrt{var\left(\hat{g}(x)\right)}}-\frac{g(x)+b(x)}{\sqrt{var\left(\hat{g}(x)\right)}}=\frac{o\left(k_0^{-m}\right)}{\sqrt{k_0/n}}=o\left(n^{1/2}k_0^{-(m+1/2)}\right)=o(1)$$

Thus equation (8) follows if

$$\frac{\hat{g}(x)-E\left(\hat{g}(x)\right)}{\sqrt{\left(\hat{g}(x)\right)}}\xrightarrow[d]{}N(0,1)$$

we have

$$\hat{g}(x)-E\left(\hat{g}(x)\right)=R'(x)G_{k,n}^{-1}x_{\epsilon}=\sum_{i=1}^{n}a_i\epsilon_i$$

Where $a_i=R'(x)G_{k,n}^{-1}R(x_i)/n$, the required Lindeberg-Feller conditions, it suffices to verify that

$$max_{1\le i\le n}\left(a_i^2\right)=o\left(\sum_{i=1}^{n}a_i^2\right)=o\left(var\left(\hat{g}(x)\right)\right)$$

we also have

$$a_i^2n^2=R'(x)G_{k,n}^{-1}R(x_i)R'(x_i)G_{k,n}^{-1}R(x)$$

Finally, equation (8) follows from the assumption that $k_0/n\longrightarrow\infty$ hence the prove.

Theorem 2

Consistency: *From theorem 1, we can establish the asymptotic consistency of* $\hat{g}(\cdot)$ *where for* $\epsilon > 0$

$$\lim_{n\to\infty}P\left\{|\hat{g}(x)-g(x)|>\epsilon\right\}\le\frac{MSE\left(\hat{g}(x)\right)}{\epsilon^2}=0 \quad (9)$$

but $MSE\left(\hat{g}(x)\right)=var\left(\hat{g}(x)\right)+\left[bias\left(\hat{g}(x)\right)\right]^2$ and $\sup_{x\in[0,1]}|D_n(x)-D(x)|=o\left(k_o^{-1}\right)$

$$var\left(\hat{g}(x)\right)=\frac{\sigma^2}{n}R'(x)G^{-1}(d)R(x)+o\left((nh)^{-1}\right),$$

$$E\left(\hat{g}(x)\right)-g(x)=b(x)+o\left(h^m\right)\ b(x)=-\frac{f^{(m)}(x)h_i^m}{m!}B_m\left(\frac{x-t_i}{h_i}\right)$$

Where $B_0(x)=1$,

$B_i(x)=\int_0^x iB_{i-1}(z)dz+b_i$ and $b_i=i\int_0^1\int_0^x B_{i-1}(z)dzdx$ is the i^{th} Bernoulli number [20]. From equation (9)

$$\lim_{n\to\infty}P\left\{|\hat{g}(x)-g(x)|>\epsilon\right\}\le\lim_{n\to\infty}\frac{MSE\left(\hat{g}(x)\right)}{\epsilon^2}$$

$$\le\lim_{n\to\infty}\frac{\frac{\sigma^2}{n}R'(x)G^{-1}(d)R(x)+o\left((nh)^{-1}\right)+\left[-\frac{f^{(m)}(x)h_i^m}{m!}B_m\left(\frac{x-t_i}{h_i}\right)+o\left(h^m\right)\right]^2}{\epsilon^2} \quad (10)$$

As $n\longrightarrow\infty$ the numerator terms in RHS of equation (10) collapse to zero therefore

$$\le\lim_{n\to\infty}\frac{\frac{\sigma^2}{n}R'(x)G^{-1}(q)R(x)+o\left((nh)^{-1}\right)+\left[-\frac{f^{(m)}(x)h_i^m}{m!}B_m\left(\frac{x-t_i}{h_i}\right)+o\left(h^m\right)\right]^2}{\epsilon^2}=\frac{0}{\epsilon^2}=0$$

Hence the prove of equation (9), therefore we conclude that the estimator is consistent to the true function *g(x)* of equation model (2).

Data Description

The study used data from Cooperative Insurance Company's motor third-party liability insurance for 2018 and 2019. 6500 policies are included in the data. The following policy information was used: the area where the policy was purchased, policyholder age, gender, vehicle type, number of claims, years of policy ownership, claim amount, and insured cases number. Age is categorized into three categorize; Old (50 > years), young (<25 years) and Middle age (25-50) (Table 1).

Table 1 shows that there is a very large number of observations has no claims in claims dataset where the maximum number of claims made in a region were 4 in an observation.

Table 1. Claim frequency table for the data.

No. of Claims	Frequency of Observations	Percentage of Observations
0	4015	61.8
1	1967	30.3
2	431	6.6
3	71	1.1
4	16	0.2

Table 2. R^2 for model 1 over 100 simulations under different sample sizes.

Data	Sample sizes	
Method	10 × 10	20 × 20
SAR	0.7483	0.9587
NonPar(Wang 2017)	0.7585	0.9691
Proposed Method	0.8217	0.9962

Table 3. R^2 larger towards 1 the better) for claims data.

Method	LL	Proposed Method	1st iteration	2nd iteration	3rd iteration	4th iteration
R^2	0.7125	0.9484	0.9531	0.9543	0.9544	0.9544

The histogram in (Figure 1) of the observed claims shows that the distribution of claims is skewed to the right, therefore there is element of over-dispersion in the data due to large number of zeros.

Results and Discussion

A spatial estimation has been applied in many fields to assess the location effects on the prediction of observations. Many models have been proposed to measure the spatial autocorrelation [21]. We establish the conditions for convergence through Monte Carlo Simulation and verify the asymptotic properties and show how the estimator perform when applied to real insurance data set and interpret the results.

Simulation Results

This section describes the simulation process and the results of the method proposed in this research. This study proposed a model proposed by [6,13] given by

$$Y_{ij}=sin\left(X_{ij}\right)+R_{ij}\ i=1,\ldots,n_1,\ j=1,\ldots,n_2$$

Where Y_{ij} are the observations and X_{ij} is a saptial process which represents the explanatory variables and R_{ij} is the term for spatial effects? The semi-variogram is estimated as

$$\gamma(h)=\frac{3|h|}{4}-\frac{|h|^3}{16}\forall 0<h\le 2 \text{ otherwise } \gamma(h)=1\forall|h|>2$$

Where *h* is the distance between 2 locations z_i and z_j in 2-dimensional space X_{ij} is a spatial process and follows a mean 0 and second order stationary gaussian process, for this reason we use spectral method to simulate X_{ij} given by

$$X_{ij}=\left(\frac{2}{M}\right)^{\frac{1}{2}}\sum_{k=1}^{M}cos\left(w(1,k)*i+w(2,k)*j+r(k)\right)$$

$w(i,k)\ k=1,\ldots,M$ are iid normally distributed and are independent of *r(k)* iid uniform random variables on $[-\pi,\pi]$ as $n\longrightarrow\infty$, X_{ij} converges to a gaussian ergodic process [17]. The matrix of random errors

$R = \left(R_{11},\ldots,R_{n_1 n_2}\right)^T = B\epsilon$ where $BB^T = \sum = \left(1-\gamma(i-j)/C(0)\right)$.

$\epsilon = \left(\epsilon_{11},\ldots,\epsilon_{n_1 n_2}\right)^T \epsilon_{ij}$ $i = 1,\ldots,n_1$ $j = 1,\ldots,n_2$ are iid normally distributed.

We set the basis function as

$$\psi_n(x) = \begin{cases} x_i^n (0 \le n \le p) \\ \left(x_i - N_{n-p}\right)_+^p, (p+1) \le n \le N \end{cases}$$

$\left(x - N_{n-p}\right)_+^p = max\left(0, x_i - N_n\right)^p$ $n \in \phi$ where Φ is compact interval. p is the degree of the spline and $n_i < \ldots < n_{N-p}$ are fixed knots in Φ

To measure performance of estimators, we use R^2 given by

$$R^2 = 1 - \frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\left[g\left(X_{ij}\right) - \hat{g}\left(X_{ij}\right)\right]^2}{\sum_{i=1}^{n}\sum_{j=1}^{n_2}\left[g\left(X_{ij}\right) - \bar{g}\right]^2}$$

where $\bar{g}$ is the sample mean of $g(X_i)$, $i = 1, .., n$. The closer R^2 to 1 the better the performance of the estimators. Comparing the performance of the proposed method with the other immediate existing methods such as nonparametric spatial models with general correlation structures a method proposed by [13] where estimation was based on kernel estimation and the nonparametric spatial autoregressive (SAR) method under which $R = \rho WR + \epsilon$ proposed by [6].

Table 2 shows the simulation results of R^2 values from 100 simulations. We can see that R^2 for the proposed method is the largest and closer to 1; this demonstrates the superior performance of the proposed method.

Analysis for claims data

The study considers the claims data from CIC insurance observed in different parts of 42 counties of Kenya to illustrate the performance of the proposed method procedures, which exclude the 5 counties from the northern part of the country since the information regarding the 5 counties was not available. Since we are mainly interested in predicting the claims frequency, the study considers a set of 6500 observations data observed from 42 counties. Let Y_i denote the claim frequency, and $Xi = (X_1, ..., X_6)^{\intercal}$ be a vector which consists of the following explanatory variables: gender, claim amount, age of the policyholder, gender, vehicle age, model of the vehicle and age category of the policyholder. By using the following model where, we can predict claim frequencies.

$$Y\left(\mathbf{z}_i\right) = g\left(\mathbf{X}\left(\mathbf{z}_i\right)\right) + \mathbf{R}\left(\mathbf{z}_i\right), \quad Var\left(\mathbf{R}\right) = \sigma^2\Sigma, \qquad i = 1,\ldots,n$$

Where $Y(z_i): i = 1, \ldots, n$ is the observations (claims) in region z_i associated with independent variable $X(z_i)$ in region z_i, $R(z_i)$ is the unobserved error in region z_i and $g(.)$ is the estimated function. The observations are from a random process observed over a countable collection of spatial regions (county). The data at a particular location typically represent the entire region. Claims data resides on an irregular lattice, with each site representing an entire county [22]. Using R^2 to measure the performance of the predictor, R^2 is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left[Y\left(z_i\right) - \hat{Y}\left(z_i\right)\right]^2}{\sum_{i=1}^{n}\left[Y\left(z_i\right) - \bar{Y}\right]^2}$$

From the analysis of the data, the results were presented in the following table (3).

From table 3 the R^2 of LL (local linear method) without unobserved random correlation errors is 0.7125 the proposed method outperforms LL model with R^2 value of 0.9484. In presence of spatial effects, as the iterations increase the performance of the proposed method is significantly better than that of the LL. It is observed that R^2 increases very little in columns 5 and 6.

Discussion

The construction of the estimator aims at obtaining a robust method for the prediction of claims frequencies. The proposed method adopts a nonparametric approach where smooth function estimated using smoothing spline accounts for the fixed effects.

A random effect (spatial error) was added to the model to improve the prediction. From the simulation study, two methods; Nonparametric model with SAR error [6] and Nonparametric model with the correlated error, which uses kernel in the estimation [13] were compared to the proposed method. The results showed that the proposed method has a significantly better prediction performance of 0.9962 (99.62 efficiencies) compared to the two methods, which have R^2 values 0.9587(95.87% efficiency) and 0.9691(96.91% efficiency), respectively. The proposed method was applied to the Cooperative Insurance Company (CIC) claims data-set. The findings showed that the proposed model is more efficient with R^2 of 0.9544, which is closer to 1 (100% efficiency) in predicting the claims frequencies compared to the Local Linear fitted model with no account for spatial effects with R^2 of 0.7125(71.25% efficiency). The proposed method could now be applied to other classes of insurance claims due to its efficiency and capability to capture the nonlinear effects and spatial effects characterized by claims data.

Conclusion

The idea of deriving an appropriate estimator in predicting frequency claims in the insurance industry has gained more interest in finance and statistical research. Many researchers heavily rely on parametric estimators; however, the insurance datasets have some aspect of non-linearity. Hence, researchers in statistics are developing nonparametric estimators with spatial error to improve the prediction based on existing parametric models. The study constructed a spatial nonparametric estimator for predicting claim frequencies in motor insurance, and the theoretical properties of the estimators were derived. The study established the theoretical properties of the proposed estimators under mild conditions. The study found that the proposed nonparametric spatial method is more efficient from the simulation results and application of the method to claims data with 6500 observations than the Local Linear fitting models and nonparametric model with Spatial Autoregressive error in the prediction of claims frequencies. Therefore, based on the results, the proposed method can be applied in predicting claims frequencies due to its efficiency and capability to capture the nonlinear effects characterized by claims data.

Limitations & Recommendations for Further Studies

Some additional exogenous variables may influence claim frequency. For example, the environmental factors among other institutional factors. It is important to investigate how each of the factors affects the output variable. Thus, a more robust spatial estimator for studying the relationship between an output variable and input variables may be constructed in subsequent studies using the proposed method. The data set used in the study involves policies for cars taken by the CIC insurance company. However, in the insurance, there are also policies for life insurance. Hence, further studies should consider the use of other claims data sets (i.e., for life and property claims data sets) in demonstrating the applications of the constructed estimator and also consider developing a more efficient nonparametric package to map Lattice observations under this proposed method.

Limitations of the Study

The data used in the study is Lattice data (without a coordinate system);

therefore, there was difficulty in classifying the claims to a specific area in the region (County).

Acknowledgments

I would like to thank the Pan African University for their support and enabling environment also the CIC company for accepting to share their claims datasets which aid in accomplishing the objective of the study.

Conflicts of Interest

The authors declare no conflict of interest.

Appendix

The CIC insurance company did not provide URL link for accessing the data instead they provide an excel sheet containing the data.

References

1. Dursun A. "A comparison of the nonparametric regression models using smoothing spline and kernel regression." *WASET* 36 (2007): 253-257.
2. Barrow, David L, Chui C, Smith P, and Ward J. "Unicity of best mean approximation by second order splines with variable knots." *Math Comput* 32, no. 144 (1978): 1131-1143.
3. Box, George EP, and Draper N. Empirical model-building and response surfaces. *New York: Wiley* 424 (1987).
4. Cressie, Noel. Statistics for spatial data. *John Wiley & Sons* (2015).
5. Doupe, Patrick, Faghmous J, and Basu S. "Machine learning for health services researchers." *Value Health.* 22 (2019): 808-815.
6. Dudley, Claire. "Bayesian analysis of an aggregate claim model using various loss distributions." *Eidenburg: Disertasi Heriot-Watt University* (2006).
7. Fellingham, Gilbert W, and Kottas A. "Parametric and nonparametric bayesian methods to model health insurance claims costs." *University of California at Santa Cruz, Department of Applied Math and Statistics Technical Reports* (2007).
8. Fellingham, Gilbert W, Kottas A, and Hartman B. "Bayesian nonparametric predictive modeling of group health claims." *Insur Math Econ* 60 (2015): 1-10.
9. Green, Peter J, and Bernard W. Silverman. “Nonparametric regression and generalized linear models: a roughness penalty approach”. *Crc Press*, 1993.
10. Hall, Peter, and Opsomer J. "Theory for penalised spline regression." *Biometrika* 92 (2005): 105-118.
11. Hall, Peter, Kay JW, and Titterington DM. "Asymptotically optimal difference-based estimation of variance in nonparametric regression." *Biometrika* 77 (1990): 521-528.
12. Hong, Liang, and Ryan Martin. "A flexible Bayesian nonparametric model for predicting future insurance claims." *N Am Actuar J* 21 (2017): 228-241.
13. Huang, Chunfeng, Hsing T, and Cressie N. "Nonparametric estimation of the variogram and its spectrum." *Biometrika* 98 (2011): 775-789.
14. Karcher, Peter, and Wang Y. "Generalized nonparametric mixed effects models." *J Comput Graph Stat* 10 (2001): 641-655.
15. Kašćelan, Vladimir, Kašćelan L, and Buric M. "A nonparametric data mining approach for risk prediction in car insurance: a case study from the Montenegrin market." *Econ Res-Ekon Istraz* 29 (2016): 545-558.
16. Laslett, Geoffrey M. "Kriging and splines: an empirical comparison of their predictive performance in some applications." *JASA* 89 (1994): 391-400.
17. Li, Yinghua, Qin Y, and Li Y. "Empirical likelihood for nonparametric regression models with spatial autoregressive errors." *J Korean Stat Soc* (2020): 1-32.
18. Overmars, Koning D, and Veldkamp A. "Spatial autocorrelation in multi-scale land use models." *Ecol Modell* 164 (2003): 257-270.
19. Rice, John A, and Colin O. Wu. "Nonparametric mixed effects models for unequally sampled noisy curves." *Biometrics* 57 (2001): 253-259.
20. Shen, Xiaotong, Wolfe DA, and Zhou S. "Local asymptotics for regression splines and confidence regions." *Ann Stat* 26 (1998): 1760-1782.
21. Wahba, Grace. “Spline models for observational data”. *SIAM* 1990.
22. Wang, Hongxia, Wang J, and Huang B. "Prediction for spatio-temporal models with autoregression in errors." *J Nonparametr Stat* 24 (2012): 217-244.

How to cite this article: Kipngetich Gideon. "A Spatial-Nonparametric Approach for Prediction of Claim Frequency in Motor Car Insurance." *J Appl Computat Math* 9 (2021): 484.