SCHOOL OF STATISTICS
UNIVERSITY OF THE PHILIPPINES DILIMAN

WORKING PAPER SERIES

Semiparametric Poisson Regression Model for Clustered Data

by

Eiffel A. De Vera and **Erniel B. Barrios**

UPSS Working Paper No. 2011-03
January 2011

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

ABSTRACT

A semiparametric poisson regression is proposed to model a spatially clustered count data. A heterogeneous covariate effect across the clusters and a random clustering effect are integrated into the model. A nonparametric model is used to account for the heterogeneous covariate effect. Two estimation procedures are proposed: (1) the parametric and nonparametric parts are estimated simultaneously through a penalized least squares method; and (2) the parametric and nonparametric parts are estimated iteratively in a backfitting framework. The simulation study exhibited the advantages of these methods over ordinary poisson and linear models (with transformation) when the aggregate covariate effect is negligible, i.e., sensitivity to the covariate is minimal or the data-generating model is not linear. In general, the two methods are advantageous over the traditional approaches when the linear model fit is poor. In cases where there is a good linear fit, the proposed methods are at par with the traditional methods, but they can still be advantageous when there are several covariates involved since backfitting yield computational simplicity in the estimation process.

Keywords: *Poisson Regression; Clustered Data; Nonparametric Regression; Backfitting; Random Effects; Generalized Additive Models*

1. INTRODUCTION

Variable resulting from counts of the number of times an event happened at a given time or space is common in epidemiological studies, environmental studies, among others. Poisson regression analysis is an appropriate strategy in analyzing count data generated by such variables as it predicts the average value of the count variable conditional on one or more factors. When the sample size is large, the central limit theorem is often invoked so that classical linear regression is still used in modeling the heterogeneous mean. This method, however, is prone to bias and often inconsistent and inefficient because of the stringent assumptions of classical regression. When used for discrete variables, skewness, nonnegative values of the response, and heteroscedasticity often occur and violates the assumptions of classical regression analysis. With malaria incidence data, Ruru and Barrios (2003) demonstrated that poisson regression has edge over classical regression of count data in terms of parsimony since classical regression requires more variables in the equation to achieve as much fit as poisson regression does.

The simplest poisson regression model $Y_i \sim Po(\mu)$ or $\log E(Y_i) = \mu$ assumes a homogeneous mean. In reality, however, the mean of Y is affected by other factors as well. Thus, other explanatory variables (X_i's) are postulated to affect the mean leading to the model $Y_i \sim Po(\mu + X_i'\beta)$ or $\log E(Y_i) = \mu + X_i'\beta$. The expected mean of the response variable in this model is heterogeneous and it depends on the explanatory variables. However, in phenomenon that assumes spatial dependence like those in epidemiology, the cluster where the observation belongs can contribute homogeneous effect on the response variable.

In the same study on malaria incidence by Ruru and Barrios (2003), they had to cluster analyze first the data before applying poisson regression since it did not describe the data well when analyzed as a whole. Three distinct clusters were identified and are modeled independently by poisson regression. It was reported that each cluster produced poisson regression model with varying set of significant predictor variables.

Observations in clusters may be correlated because of demographic similarity and other spatial dependency-inducing phenomena. Even the conditional mean model with covariates may not suffice in this case. Clusters must be considered in the model because membership within a cluster can have a significant effect on Y_i. Aiming for parsimony, Demidenko(2007) proposed a poisson regression that can account for cluster effect, i.e., $Y_{ij} \sim$ Po ($\mu_i + X_{ij}'\beta$) where Y_{ij} refers to the j^{th} observation in the i^{th} cluster and μ_i is the cluster-specific intercept, a random component. Thus, the link $\log E(Y_i|\ \mu_i) = \mu_i + X_{ij}'\beta$ implies that the random cluster-specific intercepts and the covariates with fixed coefficients jointly explain the heterogeneous means. However, this did not take into account the possibility that the covariates for each cluster could differ as illustrated in Ruru and Barrios(2003).

It is hypothesized that because of clustering, the effects of X_{ij} vary across the clusters. Given this information, the model becomes $Y_i^k \sim Po\left(\mu_k + X_{ij}^k \beta_j\right)$. This model is highly vulnerable to overparametrization and we proposed to resolve this issue by transforming the model into an additive combination of parametric and nonparametric specifications, i.e. $\log\left(Y_i^k / \mu_k\right) = \mu_k + f\left(X_{ij}^k\right)$. The cluster-specific intercept μ_k is formulated parametrically through a random effects model while the effects of the covariates are specified in a nonparametric way. The semiparametric model is then estimated iteratively through the backfitting algorithm.

This study proposes an alternative modeling strategy for count data in clusters using poisson regression. As clustering of data may cause values of intercepts and coefficients of the covaiates of the outcome variable to vary, we provide a parsimonious semiparametric poisson regression model. Real life data are not easy to model not only because of its natural variability but also due to the heterogeneous effect of certain factors across groups of observations. The parametric part of the semiparametric model will take advantage of inherent homogeneity within clusters. While the nonparametric part will induce flexibility into the function to mitigate the overparameterization that can result when dynamic model is used instead.

Considering additivity of the postulated model, backfitting is proposed. This will be advantageous over simultaneous estimation of all the components when more than one predictor is involved. For two or more predictors, simultaneous estimation of the nonparametric functions requires thin plate smoothing splines whose convergence rate declines as the number of predictor increases further. This will not be a problem in backfitting since each term, including the nonparametric functions are estimated one at a time.

2. METHODOLOGY

We proposed a model that can keep track of the spatial distribution of epidemics. Consider the spread of the A(H1N1) virus where it infects individuals by clusters of vulnerable groups, e.g., school of young children. If schools are considered as clusters and the goal is to predict the prevalence rate of flu, then poisson regression would be appropriate to model this count data. It is possible that school 1 belongs to a cluster with higher incidence of flu and that the significant covariates within this school are age and gender. While in school 2, possibly an area with lower incidence and covariate could be socio-economic status. The factors that may significantly affect the number of flu incidences are the covariates X_{ij}^k's of the individuals and the school where he/she belongs.

Given count data Y_i^k and covariates X_{ij}^k in n clusters, i = 1,…, n_k , j = 1,…, p, k = 1,…, n, the goal is to explain variation in Y_i^k in terms of X_{ij}^k. If clustering exists among the observations such that Y_i^k is the i^{th} observation in the k^{th} cluster, k = 1,…, n, i = 1,…,n_k so that within the cluster, the effect of the covariates X_{ij}^k are homogeneous, but between two different clusters the impact of the covariates on the count data Y_i^k could vary, then poisson regression will not work. Clustered data can possibly be endowed with spatial autocorrelations within the clusters. While observations across the clusters can still be independent, within the clusters they can be spatially correlated, i.e, nearby observations exhibit higher correlations than those farther from each other. It is also possible that correlations are homogeneous among observations within a cluster, e.g., in epidemiological settings. This can be addressed when the parameters representing the effect of the covariates on Y_i^k will be allowed to vary across the clusters. As a consequence, this might lead to substantial overparametrization that will impose so many constraints on either ordinary least squares or on maximum likelihood estimation. In lieu of varying parameters across the clusters, the effect of X_{ij}^k on Y_i^k is postulated in a nonparametric way. By relaxing the functional form of the effect of X_{ij}^k on Y_i^k, this can address the varying effect across clusters.

Thus, given n clusters, each with n_k elements, k = 1, …, n, we postulate the model

$$\log\left(Y_i^k / \mu_k\right) = \mu_k + f\left(X_{ij}^k\right) \qquad (1)$$

where:

i = 1,…, n_k, j = 1, …, p

n_k is the cluster size (may be equal or unequal)

n is the total number of clusters

μ_k is a random variable with $\mu_k = u_k + \varepsilon_k$, $\varepsilon_k \sim N(0, \sigma_\varepsilon^2)$, u_k is the cluster-specific intercept

X_{ij}^k is the value of j^{th} explanatory variable of the i^{th} observation within the k^{th} cluster

$f\left(X_{ij}^k\right)$ is a smooth function of X_{ij}^k (nonparametric)

Y_i^k is the i^{th} value of the response variable within the k^{th} cluster

Assumptions:

1. Y_i^k is a count response variable such that $Y_i^k \sim Po\left(\mu_k + X_{ij}^k \beta_j\right)$.

2. X_{ij}^{k} could be qualitative or quantitative independent variable. The X_{ij}^{k}'s are the attributes of the observations that may be quantitative or qualitative in nature. The postulated model could be extended to more than one independent variable as $\log(Y_i^k/\mu_k) = \mu_k + f_1(X_{i1}^k) + ... + f_p(X_{ip}^k)$ where $f_1(X_{i1}^k)$, $f_2(X_{i2}^k)$, …, $f_p(X_{ip}^k)$ are the smooth functions of the explanatory variables. The smooth functions of the X_{ij}^{k}'s are used in lieu of varying coefficients of the predictors between clusters.
3. μ_k is a random cluster intercept. This component is used to address the effect of clustering of the observations. The degree of spatial autocorrelations among observations within the same cluster is assumed to be homogeneous. This is common in epidemiological settings where all individuals in a neighborhood are either vulnerable or not vulnerable to an epidemic threat. This is assumed to follow a random effect model, i.e., normal distribution with mean u_i (cluster-specific intercept) and variance σ^2_{ε}.
4. Clusters are independent. The clusters should be independent and mutually exclusive; each observation must belong to only one cluster. Cluster independence implies that only the elements within the cluster can exhibit spatial dependencies but elements between clusters are spatially independent.
5. Cluster sizes may vary. The postulated model is assumed to work for data with equal and unequal sizes of clusters.

Ruru and Barrios(2003) illustrates the premise of the postulated model. In studying malaria incidence, observations were cluster analyzed into three groups: low malaria incidence, moderate malaria incidence and high malaria incidence, then poisson regression analysis were performed in each cluster since analysis of the whole data yield poor model fit. In low malaria incidence group, there were 5 significant risk factors, 7 for the moderate malaria incidence group, while 18 predictors for the high incidence group. Clustering was apparent in this study and that the covariates vary from one cluster to another. Instead of the separate poisson regression model for each cluster, a semiparametric model is proposed. The term μ_k in the model will take into account the random cluster effect (the three groups of malaria incidence) and the term $f(X_{ij}^k)$ will take into account the varying risk factors for each cluster resulting to only one parsimonious Poisson regression model.

2.1 Estimation Procedure

Taking advantage of the additive model formulation, backfitting is invoked to estimate the parametric and nonparametric parts of the model

$$\log E(Y_i^k \mid \mu_k) = \mu_k + f(X_{ij}^k) \qquad (2)$$

The model is transformed into an additive form so that each parameter can be estimated separately, some can be estimated parametrically and some nonparametrically.

Two estimation procedures are proposed: Method 1, estimates the parametric and nonparametric parts are estimated simultaneously in a semiparametric context; Method 2 estimates the nonparametric part first, and then the parametric part is estimated from the residuals (backfitting).

Semiparametric Estimation (Method 1)

The generalized additive model (GAM) in equation (2) is capable of estimating simultaneously the parametric and nonparametric components of the model. It separates the linear trend from any general nonparametric trend during the estimation process. The parametric and nonparametric parts are estimated simultaneously in a semiparametric context. GAM fits the parametric linear model to account for the cluster effect and spline nonparametric function to estimate the nonparametric function $f\left(X_{ij}^{k}\right)$.

Smoothing splines are used to fit the nonparametric part of the model, i.e., consider a simple additive regression model

$$Y = f(X) + \varepsilon \tag{3}$$

where $E(\varepsilon) = 0$ and $var(\varepsilon) = \sigma^2$, $a<X_1<\ldots<X_n<b$ then $E(Y) = f(X)$. The goal of spline smoothing is to estimate the function f as a solution of the penalized least squares problem given by

$$\min_f S(f) = \sum_{i=1}^{n}[Y_i - f(X_i)]^2 + \int_a^b \lambda[f''(x)]^2 dx \tag{4}$$

Where the smoothing parameter $\lambda > 0$. The first term of (4) measures the goodness of fit and the second term served as the penalty for lack of smoothness in f. The smoothing parameter λ controls the tradeoff between smoothness and goodness of fit. Large λ emphasizes smoothness of f over model fit, while small λ put higher leverage on model fit rather than on smoothness of f. As λ approaches 0, the solution to f is an interpolation of the data points. To choose a value of smoothing parameter λ, generalized cross validation (GCV) is used. It chooses a λ value that minimizes the generalized cross validation mean squared error given by

$$GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i - f_\lambda(x_i)}{1 - tr(S_\lambda)/n}\right\}^2 \tag{5}$$

If there are several functions to be estimated in the model, a thin plate smoothing is used. It approximates smooth multivariate functions observed with noise. It allows greater flexibility in the possible form of the regression surface. It also uses the penalized least squares method to fit the data with a flexible model with the advantage that the multidimensional data could be used. GCV is also used as criteria for choosing the best smoothing parameter, see Hardle, et. al. (2004) for further details.

Backfitting of the Semiparametric Model (Method 2)

In Method 2, the parametric and nonparametric parts of the model (2) are estimated separately in the context of backfitting. First, $f(X^k_{ij})$ is estimated nonparametrically using spline smoothing. As discussed in the previous section, a spline smoothing is used since there is only one function to be estimated. Then the partial residual $\hat{e}_i$ is computed as $\left(Y_i^k - \exp\left[\hat{f}\left(X_{ij}^k\right)\right]\right)$. This partial residual contains information on cluster effect and thus, will be used to estimate the cluster-specific intercept u_k from a random effect formulation of μ_k. The predicted value of the response variable is the sum of the estimates of the parameters and the nonparametric function, i.e., $\hat{Y}_i^k = \hat{u}_k + \exp\left[\hat{f}\left(X_{ij}^k\right)\right]$.

This can be further generalized if there are two or more covariates involved following the same procedure. Each $f_j\left(X_{ij}^k\right)$ is smoothed separately, after each smoothing steps, partial residuals will be computed and smoothed on the next covariate until all the covariates are exhausted. The last set of residuals is used to estimate the cluster-specific intercepts in a random effect model. The advantage of Method 2 over Method 1 is that it is more computationally simpler since the components are estimated one at a time. In Method 1, thin plate smoothing splines should be used once there are two or more predictors involved whose convergence rate decreases as the number of predictors increases.

2.3 Simulation Studies

A simulation study is conducted to evaluate the performance of semiparametric poisson regression model for clustered data. Each data set was composed of n clusters of size n_k, k = 1, …,n. Y_i^k was generated following:

$$\log Y_i^k = \mu_k + f\left(X_{ij}^k\right) + w\varepsilon \qquad (6)$$

where μ_k is the cluster-specific random intercept, $\left(X_{i1}^k, X_{i2}^k, ..., X_{ip}^k\right)$ are explanatory variables and are subject-level covariates, f is a smooth function, Y_i^k is the response variable (rounded-off to a whole number). The constant w is used to induce misspecification error.

The simulated data are used to compare the proposed semiparametric model (and two estimation procedures) to existing methods like the ordinary poisson regression and classical linear regression through their mean absolute percentage error (MAPE) and root mean square error (RMSE). Table 1 summarizes the simulation settings.

Table 1. Boundaries of Simulation Study

1. distribution of μ_k	• $\mu_k \sim N(u_k, 2)$, $u_k = 5$, increases by 5 for the succeeding clusters • $\mu_k \sim Po(u_k)$, $u_k = 2$, increases by 2 for the succeeding clusters
2. number of clusters, n	• small size – 5 clusters • medium size – 10 clusters • large size – 20 clusters
3. cluster size, n_k	• equal number of cluster size: o small size– 5 o medium size– 10 o large size – 20 • unequal number of cluster size: o small variation– randomly select from 1 to 5 o medium variation– randomly select from 1 to 10 o large variation– randomly select from 1 to 20
4. distribution of X_{ij}^k	$X_{ij}^k \sim U(10,50)$
5. functional form of $f(X_{ij}^k)$	• $f(X_{ij}^k) = \beta X_{ij}^k$ • $f(X_{ij}^k) = \exp(\beta X_{ij}^k)$
6. value of β in $f(X_{ij}^k)$	• $\beta = 0.10$ • $\beta = 2$
7. distribution of ε	$\varepsilon \sim N(0, 1)$
8. misspecification w in the model	• w =1 • w =5

μ_k was generated so that the cluster effect is homogeneous for each cluster. Two distributions of μ_k were considered, normal and Poisson. We also considered the number of clusters and cluster size. An increase in cluster size could increase efficiency as noted by Arceneaux and Nickerson (2009). A simple linear function of X_{ij}^k was considered along with a more complicated exponential function. These were done to evaluate whether the method could capture nonlinearities on the dependence of Y_i^k on X_{ij}^k. A small value of $\beta = 0.10$ and large value of $\beta = 2$ were considered to simulate the minimal or the dominating effect of the covariate to the dependent variable. Estimates from the two semiparametric methods were compared to ordinary normal regression (estimated through ordinary least squares) and poisson regression model (without cluster effect) estimated through maximum likelihood method.

From the scenarios where Methods 1 and 2 yield lower MAPE, comparison of the performance of Methods 1 and 2 were done by considering further tow predictors. In the two predictor setting, the predictor with larger value of β was estimated first in the backfitting algorithm. The residuals left was then used to estimate u_k parametrically to account for the cluster effect.

Table 2 summarizes the features of β for the two predictors setting.

Table 2. Values of β for Two Predictors Setting

Set	β_1	β_2
Set 1	0.05	0.05
Set 2	0.10	0.10
Set 3	0.7	0.3

3. Results and Discussion

The predictive performance of the proposed semiparametric poisson regression model estimated using two procedures are compared with ordinary linear regression and ordinary poisson regression in simulated and actual data [malaria incidence data by Ruru and Barrios(2003)].

3.1 Cluster Size

The cluster sizes are categorized into: small (5 observations per cluster under equal cluster size and small variation of cluster sizes under unequal cluster size), medium (10 observations per cluster under equal cluster size and medium variation under unequal cluster size) and large (20 observations per cluster under equal cluster size and large variation of cluster sizes under unequal cluster size).

The average MAPE values under equal cluster size, as shown in Table 3 illustrates the advantage of semiparametric and backfitting methods over ordinary poisson and linear regression for all levels of cluster size. The two proposed methods are also fairly robust to cluster size. From 5 elements per cluster to four times the number of elements per cluster, MAPE increases only by around 2%. Furthermore, if there is large cluster size, backfitting method has a lower MAPE than semiparametric method. This is also true under unequal cluster size as shown in Table 4.

Table 3. MAPE for Equal Cluster Size

Cluster Size	Semiparametric Method	Backfitting Method	Ordinary Poisson Regression	Ordinary Linear Regression
Small (n_k = 5)	14.60	14.99	17.66	36.29
Medium (n_k = 10)	16.69	16.32	19.58	40.73
Large (n_k = 20)	16.40	15.09	19.24	39.99

Table 4. MAPE for Unequal Cluster Size

Cluster Size	Semiparametric Method	Backfitting Method	Ordinary Poisson Regression	Ordinary Linear Regression
Small (n_k ranges from 1 to 5)	13.70	16.01	16.70	22.13
Medium (n_k ranges from 1 to 10)	14.22	15.13	16.40	22.08
Large (n_k ranges from 1 to 20)	16.74	16.26	18.51	26.28

3.2 Number of Clusters

The number of clusters is categorized into three groups: small (5 clusters), medium (10 clusters) and large (20 clusters).

The MAPE values, as shown in Table 5, of semiparametric, backfitting and ordinary Poisson regression are comparable but lower than those from ordinary linear regression for all levels of number of clusters. As the number of clusters increases, MAPE of all methods decreases, confirming the observations of Arceneaux and Nickerson (2009) that adding clusters, and not increase of cluster size, leads to increase in efficiency in poisson regression. Also, backfitting method has lower MAPE than semiparametric method if the number of clusters is large.

Table 5. MAPE for Varying Number of Clusters

No. of Clusters	Semiparametric Method	Backfitting Method	Ordinary Poisson Regression	Ordinary Linear Regression
small (n = 5)	16.22	17.18	19.32	36.93
medium (n = 10)	15.59	15.77	17.97	31.25
large (n = 20)	13.99	13.23	16.04	23.85

3.3 Functional Form of $f\left(X_{ij}^{k}\right)$

Two functional forms of $f\left(X_{ij}^{k}\right)$ are considered, linear function $[\beta X_{ij}^{k}]$ and exponential function $[\exp(\beta X_{ij}^{k})]$. The MAPE for all methods with varying functional form of X_{ij}^{k} are shown in Table 8. For a linear function of X_{ij}^{k}, MAPE values for all methods are comparable. Ordinary linear regression is advantageous when the linear effect of the covariate is dominant. The proposed methods however are at par with the existing methods even in cases where these existing methods are known to perform optimally. However, in an exponential function of X_{ij}^{k}, semiparametric, backfitting and ordinary poisson regression are relatively advantageous over the ordinary linear regression. Furthermore, the semiparametric and

backfitting methods are more flexible in capturing nonlinear forms of the relationship between the covariates and Y_i^k .

Table 6. MAPE for Varying Functional Form

Functional Form		Semiparametric Method	Backfitting Method	Ordinary Poisson Regression	Ordinary Linear Regression
Linear	β =.1	20.85	22.58	22.43	22.32
	β =2	5.35	6.08	6.86	4.08
Exponential	β =.1	16.78	15.65	20.66	51.66

3.4 Degree of Importance of the Covariate

The relative importance of a covariate in the model is reflected in the magnitude of β, small values indicate minimal role of the covariate in explaining the dependent variable, while large values indicate the dominating role of the covariate. As long as the value of β is small, i.e., $\beta = 0.10$, semiparametric and backfitting methods are comparable to ordinary Poisson and linear regression (see Table 6). However, as β becomes large, i.e., the covariate prominently explains the variations of Y, ordinary linear regression is superior but still comparable with the proposed methods.

3.5 Distribution of μ_k

Two distributions of cluster effect μ_k are considered namely: normal distribution and poisson distribution. They are simulated in such a way that there will be apparent clustering and homogeneous cluster effect in each group. Table 9 shows that normally distributed μ_k produces lower MAPE than a Poisson distributed μ_k for all methods. Also, semiparametric and backfitting methods have lower MAPE than the ordinary poisson and linear regression for the two distributions. Moreover, the two proposed methods have much smaller MAPE than ordinary linear regression if the distribution of cluster effect is poisson. This is explained by the fact that linear regression assumes a normally distributed response variable.

Table 7. MAPE by Distribution of the Random Cluster Effect

Distribution of μ_k	Semiparametric Method	Backfitting Method	Ordinary Poisson Regression	Ordinary Linear Regression
Normal	12.88	13.18	15.50	26.62
Poisson	17.55	17.59	19.99	35.01

3.6 Model Misspecification

Misspecification was simulated using a constant multiplier to the error terms. Model misspecification usually leads to residuals that exhibit large variance. When a constant is multiplied into the simulated error terms, the bigger part of the variation in Y cannot be explained by the covariates. MAPE are summarized in Table 8. MAPE values without misspecification are lower than those with misspecification for all methods which is expected to happen. However, whether there is misspecification or not, the two proposed methods are always superior over the ordinary Poisson and ordinary linear regression models. It shows the robustness of the proposed methods to misspecification errors, as inherent advantage of nonparametric models.

Table 8. MAPE for Varying Misspecification Error

Level of Misspecification Error w	Semiparametric Method	Backfitting Method	Ordinary Poisson Regression	Ordinary Linear Regression
No misspecification (w=1)	10.85	11.25	13.44	26.84
With misspecification (w=5)	19.25	19.19	21.72	34.41

3.7 Comparison of the Two Semiparametric Poisson Regression Methods

The values of MAPE for two predictors setting are summarized in Table 9. It shows that semiparametric method is superior to other methods, whether model misspecification exists or not. For a linear function, backfitting method is at par to ordinary linear regression if misspecification exists while for exponential function when β is 0.10, backfitting method has a lower MAPE value than ordinary linear regression.

Table 9. MAPE for Models with Two Covariates

Functional Form	B	Misspecification	Semiparametric Method	Backfitting Method	Ordinary Poisson Regression	Ordinary Linear Regression
Linear	0.05	w = 1	13.25	21.17	17.27	17.45
		w = 5	27.84	34.87	31.67	33.72
	0.1	w = 1	10.5	16.11	13.78	13.75
		w = 5	21.54	26.87	24.54	26.08
	0.7	w = 1	4.45	6.33	5.68	5.05
		w = 5	8.3	9.7	9.52	9.93
Exponential	0.05	w = 1	8.29	10.78	9.53	10.35
		w = 5	16.66	19.38	18.86	19.04
	0.1	w = 1	15.4	26.29	21.45	33.48
		w = 5	16.87	27.34	23.23	38.24

3.8 Illustration

The two semiparametric methods are applied to the data on malaria incidence study by Ruru and Barrios (2003). The data consist of 125 observations divided into 3 clusters: cluster 1 (low malaria incidence group) consisting of 49 villages, cluster 2 (moderate malaria incidence group) consisting of 52 villages and cluster 3 (high malaria incidence group) consisting of 24 villages. There were 22 covariates used.

When the data was analyzed as a whole (clustering is ignored), poisson regression resulted to 5 important risk factors with MAPE value of 338.62% while classical regression resulted to 4 risk factors with MAPE value of 327.92%. When poisson regression was done for each cluster, cluster 1 yield 5 predictors with MAPE value of 48.666%, cluster 2 had 7 predictors with MAPE value of 8.9989%, and cluster 3 had 18 predictors with MAPE value of 1.9407%.

Applying the two proposed methods and the benchmark methods resulted to the MAPE values in Table 10. If all risk factors are included, ordinary regression has the lowest MAPE of 57.47% but closely followed by semiparametric method with MAPE value of 60.7%. The backfitting method resulted to the largest MAPE of 190.61%. But analyzing further the data, by looking into the residual of backfitting method, five outliers were identified. Removing these outliers causes the MAPE to drastically decrease for all methods resulting to semiparametric method to have the lowest MAPE. Reducing the risk factors into five, as what was done in Rurua and Barrios(2003), ordinary regression has still the lowest MAPE followed again by the semiparametric method though their values increased compared to the analysis where all risk factors are involved. Also, the MAPE of backfitting method has the largest value of 137.14% but it is the only method that decreases when there is reduction in the number of predictors and its MAPE is still much lower than 338.62%, the MAPE of the original poisson regression analysis. Removing the same set of outliers, semiparametric method has the lowest MAPE followed closely by ordinary linear and poisson regression.

Table 10. MAPE for Malaria Data with Three Clusters

	MAPE (22 risk factors)		MAPE (5 risk factors)	
Methods	**With Outliers**	**Without Outliers**	**With Outliers**	**Without Outliers**
Semiparametric Method	60.7	8.73	90.51	14.25
Backfitting Method	190.61	44.29	137.14	45.26
Ordinary Poisson Regression	78.92	16.56	96.65	17.37
Ordinary Linear Regression	57.47	19.13	75.48	17.31

Further analysis is conducted where each cluster was subdivided into two groups resulting now to six clusters. The results are shown in Table 11. Considering all risk factors, all MAPE values decrease except for the ordinary linear regression. Semiparametric method has still the lowest MAPE value of 49.45% while backfitting has the largest MAPE. However, backfitting method poses the largest decrease in MAPE from 190.61% to 157.19% when the number of clusters is increased. Furthermore, removing the five outliers in the dataset causes all the methods to have a large decrease in MAPE. Semiparametric method is still superior compared to other methods with the lowest MAPE value of 6.95%.

Reducing the risk factors into 5 significant ones causes the MAPE values to increase except for the backfitting method. This shows that as the number of clusters increases, backfitting method yield improvements in its predictive ability while the opposite happens for ordinary linear regression. The two proposed semiparametric methods are more stable when number of clusters is increased. This is further validated by comparing the MAPE values of 3 clusters (see Table 10) to the MAPE of 6 clusters (Table 11). Still, the semiparametric method has the lowest MAPE value of 11.4% followed by ordinary poisson regression while MAPE values of backfitting and ordinary linear regression are comparable.

Table 11. MAPE for Malaria Data with Six Clusters

	MAPE (22 risk factors))		MAPE (5 risk factors))	
Methods	**With Outliers**	**Without Outliers**	**With Outliers**	**Without Outliers**
Semiparametric Method	49.45	6.95	58.46	11.4
Backfitting Method	157.19	42.26	130.47	43.38
Ordinary Poisson Regression	64.77	12.46	66.23	14.61
Ordinary Linear Regression	58.31	11.98	123.3	41.32

4. Conclusions

The postulated semiparametric Poisson regression model for spatially clustered count data given n clusters with n_i observations, is given by $\log\left(Y_i^k / \mu_k\right) = \mu_k + f\left(X_{ij}^k\right)$. The parameter μ_k accounts for the cluster effect while $f(X^k_{ij})$ can account for the possibly varying coefficient of the covariates across the clusters. Two methods are proposed in the estimation of the postulated model. The semiparametric method estimates the nonparametric part $f(X^k_{ij})$ and parametric part μ_k simultaneously while the backfitting method takes advantage of the additivity of the model and uses the backfitting algorithm.

Both the simulated and actual data exhibited the advantages of the two proposed methods over ordinary poisson and linear regression models when the covariate effect is

negligible, i.e., sensitivity to the covariate is minimal or the data-generating model is not linear. The two methods are generally advantageous over the traditional approaches when the linear model is inadequate, i.e., in cases of misspecification or nonlinearity. As the number of clusters increases, the predictive ability of the two methods also increases. In cases where there is a good linear fit, the proposed methods are still at par with the traditional methods, but they can be advantageous when there are several covariates involved especially since backfitting can induce computational simplicity in the estimation process.

References

Arceneaux, K. and Nickerson, D. (2009). Modeling Certainty with Clustered Data: A Comparison of Methods. *Political Analysis*. **17**, 177-190.

Demidenko, E. (2007). Poisson Regression for Clustered Data. *International Statistical Review*, **75**, 96-113.

Hardle, W., Muller, M., Sperlich, S., Werwatz, A. (2004). Nonparametric and Semiparametric Models, Berlin: Springer.

Ruru, Y. and Barrios, E. (2003). Poisson Regression Models of Malaria Incidence in Jayapura, Indonesia. *The Philippine Statistician*, **52**, 27-38.