A Simulation Study for Performance Comparison between Generalized Linear Mixed Modeling (GLMM) and Generalized Regression Neural Networks (GRNN) in Joint Modeling of Mixed Responses

J.C. Hapugoda[1*] and M.R. Sooriyarachchi[2]

[1]*Department of Organizational Studies, Faculty of Management Studies, The Open University of Sri Lanka, Nawala, Nugegoda, Sri Lanka.*

[2]*Department of Statistics, Faculty of Science, University of Colombo, Sri Lanka.*

ABSTRACT

Joint modeling of mixed responses has become a popular research area due to its applicability in many disciplines. When there is an association between two responses, a joint model will provide improved results than modeling the responses separately. In this study, the joint modeling of survival and count variables was carried out using Generalized Linear Mixed Modeling (GLMM) and Generalized Regression Neural Network (GRNN) to compare their performances under the setting of clustered data. A joint model of survival and count variables that was developed by joining the Discrete Time Hazard Model (DTHM) and Poisson Regression model was used in this study as the GLMM model. A simulation study was carried out under three different sample sizes; *n* = 20, 100, and 500, and for three levels of correlations between two responses: low (r = 0.30), moderate low (r = 0.30), and high (r = 0.30). The root mean square error, absolute mean error and correlation coefficient between actual and predicted response data were calculated to compare the performances of GLMM and GRNN models. The results revealed that the GRNN has a better fit in general, but under large sample sizes and high correlations between response variables, GLMM outperformed the GRNN. Application of these methods to a data set from poultry industry further confirmed the suitability of Generalized Regression Neural Network fit for joint modeling of real-world data.

*Corresponding author : jchap@ou.ac.lk

INTRODUCTION

Joint modeling of data with mixed responses has become a popular research area in recent times. Many researchers have used joint modeling techniques in various disciplines (Rizopoulos, 2012; Gardiner, 2013; Elashoff *et al.*, 2016; Hapugoda and Sooriyarachchi, 2017a; 2017b; 2018). When there is an association between the two responses, a joint model will provide interesting and improved results than modeling the responses separately.

In the literature, many approaches of joint modeling could be identified (Rizopoulos, 2012; Gardiner, 2013). Two such methods include mixed modeling approach and artificial neural network approach (Hapugoda and Sooriyarachchi, 2017b). It will be beneficial to be aware of the performances of these approaches in joint modeling, while concluding the better fit of the models to the data.

Many researchers have been carried out on comparative studies to identify the performances of different models using different techniques (Comrie, 1997; Yay and Akinci, 2009; Hapugoda and Sooriyarachchi, 2017a). Survival time data (i.e. the duration of time between the initial point up to the occurrence of a particular event) and count data (i.e. the number of observations) are some of the commonly recorded data in many fields of studies. In many cases these two variables can be correlated (Hapugoda and Sooriyarachchi, 2017b). In addition, the hierarchical nature (i.e. in the form of clusters) of data could also be seen in many fields of studies. If a set of data consists of all the above-mentioned forms, joint modeling of these two variables can be carried out considering the clustering effect. In that case, it is also better to identify which approach can be utilized to obtain more accurate results.

Thus, the objective of this study was to compare the performances of the models obtained by using the mixed modeling approach and artificial neural network approach for such data. A simulation study was carried out for this purpose. This concept was then applied to some data from the poultry industry of Sri Lanka.

The overview of the paper is as follows: the methodologies of the two approaches used in this study are explained in the 'Methodology' section, the details of the simulation study carried out are explained in the 'Simulation study' section, and the application of the concept to the poultry industry is described with a discussion in the 'Results and Discussion' section.

MATERIALS AND METHODS

Generalized Linear Mixed Modeling Approach

A joint model of survival and count was developed by joining the Discrete Time Hazard Model (DTHM) and Poisson Regression model by the same authors of this article (Hapugoda and Sooriyarachchi, 2017a; 2017b; 2018). The DTHM was used to convert continuous survival times with the censoring information combined to it into a binary response variable with time indicators as explanatory variables. The underlying theory of the joint model is as follows:

The responses of analysis were: Y_{ij1} (binary – survived/ not survived status of the batch of eggs) and Y_{ij2} (count - number of eggs), where i stands for the batch of breeders. Here, the hatchability occurs in time interval j for the ith batch, given that the eggs from i^{th} batch survived in interval j-1 (hazard at time T_j), and x_{ij} are covariates at the batch level i where T_{ij} contains the unit-coded time indicators for the time periods assessed in the study (for an observation at time j, $T_{ij} = 1$ and for a time k where $k \neq j$, $T_k = 0$). Variables that impact $Y = (Y_1, Y_2)$ are the explanatory variables denoted by X_{ij} (i=1: N). To formulate a joint model, Generalized Linear Mixed Models (GLMM) can be used to form marginal models for each response by considering mean $E(Y_l/X_{ij})$ and variance $Var(Y_l/X_{ij})$ where l=1,2. The approach to link the responses is by structuring a covariance matrix $Var(Y_l/X_{ij})$ to include potential correlations.

In GLM, $l_k(E(Y_{ijk}/X_{ij})) = X_{ijk}'\beta_k$, k=1,2 and l_k is the link function. Here, $l_1(u)$ is the logit link and $l_2(u)$ is the log link. A structural formulation of the model is given as:

$$l_1(y'_{ij1}) = Logit(\pi_{ij}) = \beta_0 + u_{0j} + \Sigma\alpha_j T_{ij} + \beta X_{ij} \text{ where } u_{0j} \sim N(0, \sigma_u^2) \tag{1}$$

and

$$l_2(y'_{ij2}) = \log(\mu_{ij2}) = \log(E_{ij}) + \beta_{0j} + \beta X_{ij} \text{ where } \beta_{0j} = \beta_0 + u_{0j} \text{ and } u_{0j} \sim N(0, \sigma_u^2) \tag{2}$$

For simplicity, we assume that both random effects are the same (u_{0j}) and have the same variance (σ_u^2). The joint model variance matrix is as follows:

$$\begin{bmatrix} \sigma_u^2 & \rho_{12}\sigma_u^2 \\ \rho_{12}\sigma_u^2 & \sigma_u^2 \end{bmatrix} \tag{3}$$

GLIMMIX will structure the variance matrix of Y = (Y_1, Y_2) as follows:

$$Var(Y_i/X_i) = A_i^{1/2} R_i A_i^{1/2} \qquad (4)$$

where R_i is a user specified 2x2 covariance structure and A_i is the diagonal matrix of the variances of (Y_1, Y_2).

Neural Network Approach

Artificial Neural Networks (ANNs) technique is widely used as an alternative approach for conventional statistical techniques, due to its remarkable ability to derive meaning from complicated data by extracting patterns and detecting trends that are too complex to be noticed by either humans or other computer intensive techniques (Sehrawat *et al.*, 2015)

Among the different types of neural network techniques present, it is vital to identify the most appropriate technique to model the data under consideration. As the response variables are continuous (as the original continuous survival variable was considered in neural network) and discrete (count), the ANN technique that can address such variables should be chosen.

Generalized Regression Neural Network (GRNN) approach was examined to study whether it is a better approach to model survival and count variables together.

GRNN was proposed by Donald F. Specht in 1990 and falls into the category of probabilistic neural networks (Specht, 1990). The probability density function used in GRNN is the Normal Distribution. For each training sample, X_i is used as the mean of a Normal Distribution for n number of samples. The distance, D_i, between the training sample and the point of prediction, is used as a measure of how well each training sample can represent the position of prediction, *X*. (See equations (5) and (6)).

$$Y(X) = \frac{\sum_{i=1}^{n} X_i \exp(-D_i^2/2\sigma^2)}{\sum_{i=1}^{n} \exp(-D_i^2/2\sigma^2)} \qquad (5)$$

$$D_i^2 = (X - X_i)^T . (X - X_i) \qquad (6)$$

The smoothness parameter (σ) is the network parameter of this procedure (i.e. the parameter of the Gaussian curves). It was suggested to use the holdout method to select a good value of σ (Specht, 1990).

The method to obtain the best neural network is based on selecting the best parameters and best input combination. It was identified that the ten-fold stratified cross validation method is a better approach to select the best model in terms of the input combination (Kohavi, 1995; Cui *et al.*, 2008). Hence, the same method has been used in this study.

Performance Comparison

Measuring performance of the models is required to compare the two models and conclude the best model. There are several statistical measures which are used to measure the model performance. According to literature, the measures such as root mean square error (RMSE), absolute mean error (AME) and correlation coefficient between actual response data with predicted response data (R) to compare the performances of two or more models, when at least one model under consideration is based on neural network approach (Yay and Akinci, 2009; Hapugoda *et al.*, 2013; Hapugoda and Sooriyarachchi, 2017a). Equations (7) to (9) are used to calculate the performance in each model, when there are 'n' number of observations.

Root Mean Square Error (RMSE)

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(I_i(observed) - I_i(predicted))^2}$$
$$(7)$$

Mean Absolute Error (MAE)

$$\frac{1}{n}\sum_{i=1}^{n}|I_i(observed) - I_i(predicted)|$$
$$(8)$$

Correlation Coefficient (R)

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (9)$$

RESULTS AND DISCUSSION

Simulation Study

To assess the performance of the two models, a simulation study was carried out and the details are presented in this section.

A joint model should be developed when there exists any kind of association/ correlation between the response variables. A correlated survival time

and count variables were simulated using correlated Exponential and Poisson variables based on Multivariate Normal distribution (Smart, 2015). Later the continuous survival data was categorized in a biologically plausible way. Here, to represent the cluster effect in the simulated data, each dataset was divided into three equal sections to represent three clusters. The data generation for this simulation study was based on the previous work by the same authors (Hapugoda and Sooriyarachchi, 2018).

In different scenarios, the correlation among the two response variables can be low, medium, or high. However, to generalize the developed model for any situation, the simulation study was carried out under these three cases for three different sample sizes (n = 20, 100, and 500).

The correlations considered were, r = 0.3, 0.5 and 0.75 for low, moderate, and high correlations respectively. The significances of the correlation coefficients with respect to the sample size were tested at 5% significance level, which reveals significance in all cases. Under each case, 1000 simulated data sets were generated using R software. PROC GLIMMIX procedure in SAS software was used to model the data using mixed modeling approach while Matlab software was used to model the data using a neural network approach.

RMSE, MAE and correlation coefficient (R) between actual response data with predicted response data were recorded for both models in each dataset. Here, the model based on generalized linear mixed modelling approach was named as GLMM model and the model based on the neural network approach was named as GRNN model. The results of these are presented in Tables 1 and 2.

Table 1 illustrates the overall mean values for RMSE, MAE and R in both models in all simulation categories. According to these values, in all categories except with the largest sample size (500) together with high correlation, mean RMSE values of GRNN model were lower than those of GLMM model.

The mean R values of GRNN model were higher than those of GLMM model under all situations. In low correlations between response variables when the sample sizes were 100 and 500, the mean MAE values of GRNN model were higher than those of GLMM values. Also, in moderate correlations for sample sizes of 500, the difference in MAE were

negligible. In the other combinations MAE values of GRNN model were lower than those of GLMM values. In sample sizes of 100 and 500 the differences in mean values of RMSE, MAE and R were negligible, whereas in sample size of 20, the differences were notably higher.

Table 2 provides the details on the number of times the GRNN model has performed better than the GLMM model based on RMSE, MAE and R values. In moderate correlations between the responses for all the sample sizes, GRNN model recorded better performance in nearly all cases. In low correlations between the responses, for all sample sizes, GRNN had a better performance according to the RMSE and R values, but with respect to MAE values in sample sizes of 100 and 500, the number of times the GRNN fitted better was lower.

Given high correlations between the responses for all sample sizes, GRNN model recorded a better performance in significantly higher number of times. However, when the sample size is largest (500) and correlation is taken to be high, GRNN performed worse than the GLMM a few times. These few cases resulted in extremely large values of RMSE and MAE for GRNN as the mean was larger for these statistics. This shows that under this scenario the GRNN is not dependable.

It can be seen that the GRNN model has a better fit than the GLMM model for cases when sample size is small or moderate, especially when the correlation is high or moderate. For the two larger sample sizes, both models were comparable for a small correlation. However, when the sample size was large, and the correlation was high the GLMM model outperformed the GRNN model according to the RMSE and MAE. The difference in R under this scenario was negligible.

It can be understood that the GRNN model has a better fit than the GLMM model in most of the cases. Though, the differences of mean values of RMSE, MAE and R were notably diverse in the small sample size (20) (in Table 1), both models were comparable for the two larger sample sizes when correlation was small or moderate. But the strength of the GLMM model is indicated for large samples and high correlation. One may argue that this is the situation in which a joint model would really be more efficient than univariate models.

Table 1: Mean values obtained in simulation study

Sample Size	Correlation between responses	GRNN			GLMM			Difference of mean values between GRNN and GLMM		
		RMSE	MAE	R	RMSE	MAE	R	RMSE	MAE	R
500	High	3.89	3.73	0.80	3.07	2.85	0.71	0.82	0.88	0.09
	Moderate	1.61	1.85	0.91	2.10	1.87	0.54	-0.49	-0.02	0.37
	Low	2.31	2.70	0.84	2.78	2.42	0.32	-0.47	0.28	0.52
100	High	3.91	3.72	0.81	5.00	4.16	0.71	-1.09	-0.44	0.10
	Moderate	1.55	1.75	0.92	3.23	2.89	0.51	-1.68	-1.14	0.41
	Low	2.19	2.52	0.85	2.83	2.47	0.34	-0.64	0.05	0.51
20	High	3.41	3.15	0.85	16.04	13.20	0.72	-9.79	-10.05	0.13
	Moderate	1.35	1.40	0.94	11.82	9.72	0.54	-10.47	-8.32	0.40
	Low	1.86	1.90	0.89	7.50	6.19	0.43	-5.64	-4.29	0.46

Table 2. Number of times the GRNN Model performed better than GLMM model

Sample Size	Correlation between responses	RMSE	MAE	R
500	High	997	814	1000
	Moderate	1000	1000	1000
	Low	1000	4	1000
100	High	968	757	986
	Moderate	1000	1000	1000
	Low	995	332	1000
20	High	988	965	933
	Moderate	1000	997	1000
	Low	998	973	1000

Application to Poultry Data

To perform the comparison study in a real data set, the following analysis was carried out. This work is an extension to previous study by the same authors, based on the data in poultry industry (Hapugoda and Sooriyarachchi, 2017a). The data set consisted of 632 observations.

The data from a broiler grandparent farm from Sri Lanka were available for nine batches of breeders maintained over a two- and half-year period. The variables were breeder age, fertility rate, length of storage period for eggs, the number of hatching eggs and the number of non-hatching eggs for each batch at different time points. The description on the variables of the data are shown in table 3.

In the poultry industry, the accurate prediction of 'the number of chicks that can be obtained by hatching a known number of eggs' (number of hatching eggs) is the most important factor for the economic success of a breeder operation. This prediction is dependent on breeder factors such as breeder age and fertility rate (Reis *et al.*, 1997; Lapao *et al.*, 1999; Tona *et al.*, 2002; Petek and Dikmen, 2006; Hapugoda *et al.*, 2013).

Though the storage of hatching eggs is a necessary part of the commercial incubation, dramatic decline in the number of hatching eggs is observed in broiler breeder eggs stored for extended periods and there is an interaction between length of storage and breeder age (Reis *et al.*, 1997; Lapao *et al.*, 1999; Petek and Dikmen, 2006). In practice, the eggs are hatched based on the demand (selling of one-day-old chicks is the objective of breeder operation). Here, the eggs from older breeders with less fertility are given the priority, irrespective of the length of storage (in days). This is due to the practice of the farm from which the data were obtained, based on the assumption that the hatchability is dependent on breeder age.

Table 3: Description of data

Variable	Category	Code
Length of storage	<14 days (t1)	1
	14 – 20 days (t2)	2
	>20 days (t3)	3
Survival status	Hatched	1
	Not hatched	0
Breeder age	Continuous	
Fertility rate	Continuous	
Egg Count	The number of eggs from each batch in each day	

Fitting a GLMM Model

If a joint model could be developed to estimate the number of hatching eggs with their length of storage considering breeder factors such as age and fertility rate, it will be beneficial for the success of the operation of poultry industry. Such a model has been developed in a previous study using generalized linear mixed modeling approach by the same authors (Hapugoda and Sooriayarachchi, 2017a). The Spearman correlation coefficient between these 2 variables is 0.60 which indicates that a joint model is more plausible than 2 univariate models. The purpose of this analysis is to compare the performance of the GLMM model with the model developed using neural network.

Fitting a GRNN Model

To achieve this purpose, a GRNN model was fitted to the data in the following manner. The best model could be obtained when the best σ value was considered, in which the sum of square error was minimized. In this analysis, the best σ value was recorded as 0.5.

For the analysis, the dataset was split to two categories, as training set (90% of data) and test set (10% of data). The ten-fold stratified cross validation method was used to identify the best division of data and this was carried out by using different columns of the input matrix to include in the training set and test set. For example, the first combination was 1 to 9 columns of input matrix as the training set while 10th column was taken as the test set. Likewise, 10 different models were considered to choose the best division of the data. Thus, for different sets of training data and test data divisions used in this analysis, the plots of σ values with their corresponding sum of square error (SSE) values were obtained. The best model was selected by considering the best σ values identified, which generated the smallest sum of squares value. Here, the best division of data was identified as; taking the 1st column of input matrix as test set, while the rest of the data were considered as the training set. The recorded SSE value for this combination was 38.01.

The obtained neural network consisted of 4 input nodes, 2 output nodes and 2 hidden layers, where the two hidden layers consisted of 568 nodes and 2 nodes, respectively.

Performance Comparison

RMSE, MAE and correlation coefficient between actual response data with predicted response data were recorded for both models for the data. Here again, the model based on generalized linear mixed modelling approach was named as GLMM model and the model based on neural network approach was named as GRNN model. The results of these are presented in table 4.

Table 4: Performance Comparison

Measure	GLMM model	GRNN model
RMSE	5.92	0.27
MAE	5.82	0.14
R	0.55	0.92

In our simulation scenario this is the case of large sample size with moderate correlation between the response variables. The simulations indicated that all 3 statistics were better for the GRNN in this scenario. RMSE and MAE values were significantly lower in GRNN model than those of the GLMM model. This indicated the errors generated from GRNN model were low when compared to those of the GLMM model. The R value in GRNN model over performed GLMM model with a significantly higher correlation among the actual values and predicted values of the response variables.

Due to the better fit of the GLMM under ideal conditions and due to the black box nature of the neural networks, Statisticians often prefer to use GLMMs over GRNNs, though Computer Scientists and Data Scientists may have different views. This is also because, interpretations on the effects of variables cannot be carried out using a GRNN model, though it can be found out which variable is more important in defining the model output. The parameter estimates of the final GLMM model can be interpreted using the parameter estimates of the model obtained.

CONCLUSION

This study compared the performance between GLMM and GRNN in joint modeling of mixed responses. A simulation study was carried out to assess the performance using measures of RMSE, MAE and correlation coefficient between actual response data with predicted response data. This simulation study revealed better fit of GRNN compared to GLMM for small to moderate samples under all conditions and for large samples under low and moderate correlation. However, for the ideal case of joint modeling, that is when the samples are large, and the correlation is high the GLMM model outperformed the GRNN model. This could possibly be due to the method of estimation in GLMM which is Maximum Likelihood via Laplace approximation. Maximum Likelihood Estimates

are asymptotically unbiased and asymptotically Normally distributed. Regarding high correlation, the bivariate model would really gain maximum superiority over univariate models only when the correlation between responses are high.

These methods were applied to a practical data set having a large sample size and moderate correlation. As expected, based on the simulations, this exercise revealed better fit of GRNN model, under the tested scenario.

REFERENCES

Comrie, A.C. (1997). Comparing Neural Networks and Regression Models for Ozone Forecasting. Journal of the Air and Waste Management Association. 47(6), 653-663.

Cui, G., Wong, M.L., Zhang, G., and Li, L. (2008). Model selection for direct marketing: performance criteria and validation methods. Marketing intelligence and planning. 26(3), 275-292.

Elashoff, R., Li, G. and Li, N. (2016). Joint Modeling of Longitudinal and Time-to-Event Data. CRC Press.

Gardiner, J.C. (2013). Joint Modeling of Mixed Outcomes in Health Services Research. In SAS Global Forum, San Francisco, CA.

Hapugoda, J.C., Perera, D.B.U.S. De S. and Deshani, K.A.D. (2013). A Comparative Study of Generalized Estimating Equations (GEE) and General Regression Neural Networks (GRNN) to Assess the Production of Poultry Breeders, Proceedings of the Sri Lanka Association for The Advancement of Science (SLAAS) Annual Conference, Sri Lanka, p. 146.

Hapugoda, J.C. and Sooriyarachchi, M.R. (2017a). Joint Modeling of Mixed Responses: An Application to Poultry Data, Proceedings of 5th Annual International Conference on Operations Research and Statistics (ORS), Singapore, 182-185.

Hapugoda, J.C. and Sooriyarachchi, M.R. (2017b). A Comparative Study of Generalized Linear Mixed Modeling and Artificial Neural Network approach for the joint modeling of survival and incidence of Dengue patient. Journal of Physics Conference Series. 890(1), 012135 (DOI: 10.1088/1742-6596/890/1/012135).

Hapugoda, J.C. and Sooriyarachchi, M.R. (2018). Joint Modeling of Discrete Time Hazard Model with Poisson Regression Model: A Simulation Study. Proceedings of Jaffna University International Conference, Sri Lanka, pp. 551 - 554.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. International Joint Conference on Artificial Intelligence. 2, 1137-1143.

Lapao, C., Gama, L.T. and Soares, M.C. (1999). Effects of Broiler Breeder Age and Length of Egg Storage on Albumen Characteristics and Hatchability. Poultry Science. 78(5), 640–645.

Petek, M. and Dikmen, S. (2006). The effects of prestorage incubation and length of storage of broiler breeder eggs on hatchability and subsequent growth performance of progeny. Czech Journal of Animal Sciences. 51(2), 73–77.

Rizopoulos, D. (2012). Joint Models for Longitudinal and Time-to-Event Data, CRC Press.

Reis, L.H., Gama, L.T. and Soares, M.C. (1997). Effects of Short Storage Conditions and Broiler Breeder Age on Hatchability, Hatching Time, and Chick Weights. Poultry Science. 76(11) 1459–1466.

Sehrawat, R., Gupta, P. and Yadav, R. (2015). Basic of Artificial Neural Network. Journal of Computer Science and Engineering. 1 (5), 26-30.

Smart, F. (2015). Easily generate correlated variables from any distribution (Retrieved from http://www.econometricsbysimulation.com/2014/02/easily-generate-correlated-variables.html)

Specht, D.F. (1990). Probabilistic neural networks Neural Networks. Neural Networks. 3 (1), 109-118.

Tona, K., Bamelis, F., De, K., Bruggeman, V. and Decuypere, E. (2002). Effect of induced molting

on albumen quality, hatchability, and chick body weight from broiler breeders. Poultry Sciences. 81(3), 327-332.

Yay, M. and Akıncı, E. (2009). Application of Ordinal Logistic Regression and Artificial Neural Networks in a Study of Student Satisfaction. Cypriot Journal of Educational Sciences. 4, 58-69.