A comparative study of generalized linear mixed modelling and artificial neural network approach for the joint modelling of survival and incidence of Dengue patients in Sri Lanka

IOP Conf. Series: Journal of Physics: Conf. Series **890** (2017) 012135 doi:10.1088/1742-6596/890/1/012135

A comparative study of generalized linear mixed modelling and artificial neural network approach for the joint modelling of survival and incidence of Dengue patients in Sri Lanka

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

[1]Department of Management Studies, The Open University of Sri Lanka, Nawala, Sri Lanka
[2] Department of Statistics, Faculty of Science, University of Colombo, Colombo 03, Sri Lanka

E-mail: jayanihp@gmail.com

Abstract. Survival time of patients with a disease and the incidence of that particular disease (count) is frequently observed in medical studies with the data of a clustered nature. In many cases, though, the survival times and the count can be correlated in a way that, diseases that occur rarely could have shorter survival times or vice versa. Due to this fact, joint modelling of these two variables will provide interesting and certainly improved results than modelling these separately. Authors have previously proposed a methodology using Generalized Linear Mixed Models (GLMM) by joining the Discrete Time Hazard model with the Poisson Regression model to jointly model survival and count model. As Aritificial Neural Network (ANN) has become a most powerful computational tool to model complex non-linear systems, it was proposed to develop a new joint model of survival and count of Dengue patients of Sri Lanka by using that approach. Thus, the objective of this study is to develop a model using ANN approach and compare the results with the previously developed GLMM model. As the response variables are continuous in nature, Generalized Regression Neural Network (GRNN) approach was adopted to model the data. To compare the model fit, measures such as root mean square error (RMSE), absolute mean error (AME) and correlation coefficient (R) were used. The measures indicate the GRNN model fits the data better than the GLMM model.

1. Introduction

Though, data on the survival time of a patient and the count (incidence of the disease) are a frequently encountered phenomenon in medical studies, it is not common in the literature to see these two variables considered together as a bivariate response in a joint model. In many cases, though, the survival times and the count can be correlated in a way that, diseases that occur rarely can have shorter survival times or vice versa [1].This is often further complicated by the data being of a hierarchical nature (in the form of clusters). An example of such a study is in infectious disease epidemiology of a life threatening disease where there is a geographical variation of the intensity of the disease. When the survival time of a patient and the frequency of a particular disease is correlated and the data has a clustered structure, it is better to model these two variables together.

In the process of developing a joint model, two main methods could be identified when searching literature: 1) using mixed modellfing approach and 2) using artificial neural network approach.

Authors have previously proposed a methodology using Generalized Linear Mixed Models (GLMM) by joining the Discrete Time Hazard model with the Poisson Regression model to jointly model survival and count model [2].

On the other hand, Artificial Neural Networks (ANNs) technique is widely used as an alternative approach for conventional statistical techniques recently, 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 [3].

Thus, the objective of this study is to develop a joint model of survival and incidence of dengue patients in Sri Lanka using ANN approach and compare its performance with the previously developed model based on GLMM approach.

2. Methodology

ANN is a computer program or hardwired machine that is designed to learn in a manner similar to the human brain. It's a data analysing technique that was introduced as a result of the information technology advancements and was first designed by McCulloch and Pitts [4] and further developed by many other inventors all over the world. It is an information processing paradigm that is established based on the workings of the biological nervous systems. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in union to solve specific problems. ANNs are also known as the most powerful computational tools to model complex non-linear systems [5].

However, ANN approach has advantages and disadvantages compared to the conventional statistical models such as GLMM. The advantages of ANN include, the requirement of less formal statistical training to develop, complex nonlinear relationships between independent and dependent variables can be implicitly detected, the ability to detect all possible interactions between predictor variables and development can be done using multiple different training algorithms. Disadvantages include, a "black box" nature, limited ability to explicitly identify possible causal relationships, requirement of greater computational resources, prone to overfitting, development is empirical, and many methodological issues remain to be resolved [6].

Among the different types of ANN approaches present, it is vital to identify which technique/approach is the most appropriate to model the data under consideration. As the response variables are continuous and discrete, the ANN technique that can address continuous variables should be chosen.

Feed forward neural network (FFNN) and generalized regression neural network (GRNN) which belong to a class of neural networks widely used for mapping continuous functions [7]. Researchers have compared the performances of FFNN and GRNN to model various scenarios and identified that the GRNN performs well in the case of continuous variables [8, 9]. Thus, the GRNN approach was adopted to use in this study as a better approach to model a continuous function which has been first introduced by Specht in 1990 [10].

GRNN was proposed by Donald F. Specht in 1990 and falls into the category of probabilistic neural networks [10]. This type of neural networks needs only a fraction of training samples a back propogational neural network would need in order to train the network [11].Therefore, the use of a probabilistic neural network is especially advantageous due to its ability to converge to the underlying function of the data with only few training samples available. The additional knowledge needed to get the fit in a satisfying way is relatively small and could be done without additional input by the user. This makes GRNN a very useful tool to perform predictions and comparisons of system performance in practice.

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 the each training sample can represent the position of prediction, X. (See equations (1) and (2)).

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

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

The smoothness parameter (σ) is the only network parameter of this procedure (i.e. the parameter of the Gaussian curves). The search for the value of the smoothness parameter has to take several aspects into account depending on the application the predicted output is used for. It was suggested to use the holdout method to select a good value of σ [10]. In the holdout method, one sample of the entire set is removed and for a fixed σ, GRNN is used again to predict this sample with the reduced set of training samples. The squared difference between the predicted value of the removed training sample and the training sample itself is then calculated and stored. The removing of samples and prediction of them again for this chosen σ is repeated for each sample-vector. After ending this process the mean of the squared differences is calculated for each run. Then the process of reducing the set of training samples and predicting the value for these samples is repeated for different values of σ. The σ for which the sum of the mean squared difference is a minimum is the σ that should be used for the predictions by using this set of training samples.

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 [12]. Hence, the same method has been used in this study.

Measuring performance of the developed models is a crucial factor to conclude the best model and conclusions. There are several statistical measures which are used to measure the model performance. This study considered the measures such as root mean square error (RMSE), absolute mean error (AME) and correlation coefficient (R) to compare the performances of two or more models, when at least one model under consideration is based on neural network approach [9, 13]. The correlation coefficient depicts the linear relationship between the actual output and the predicted output. Equations (3) to (5) are used to calculate the performance in each model, when there are *n* number of observations. To calculate the correlation coefficient, *x* and *y* are considered as the two variables, while $\bar{x}$ and $\bar{y}$ are their mean values respectively.

$$\text{Root mean square error (RMSE)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(I_i(observed) - I_i(predicted))^2} \quad (3)$$

$$\text{Mean absolute error (MAE)} = \frac{1}{n}\sum_{i=1}^{n}|I_i(observed) - I_i(predicted)| \quad (4)$$

$$\text{Correlation coefficient (R)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (5)$$

3. Analysis, results and discussion

The data for this study collated from the monthly records of Dengue patients in Sri Lanka from year 2006 to 2008, which have been collected by the Epidemiological Unit of Sri Lanka. The response variables in this study are, survival time of patients and the number of patients, while the explanatory variables are, climate factors (rainfall, first lag of rainfall, second lag of rainfall, humidity, first lag of humidity, second lag of humidity, temperature, first lag of temperature, second lag of temperature), year, month and district of the country.

All the explanatory variables do not have an equal impact on the response variable. Hence, all of them should not be considered equally significant when designing the best neural network. Thus, the

best combination of input variables should be used in order to design the best fitted neural network for the data.

Four input combinations were used to compare their performances and are provided in table 1. Here, the fourth input combination was based on the GLMM model developed by the same authors for the same data set [3].

The best model could be obtained when the best σ value is considered, in which the sum of square error is minimised. Here, 90% of the dataset was split to two categories, as training set (90% of data) and test set (10% of data). The ten-fold stratified cross validamethod was used to identify the best division of data and this has been carried out by using different columns of the input matrix to include in the training set and test set. As an example, the first combination would be 1 to 9 columns of input matrix is the training set while 10^{th} column is the test set and the second combination would be 1 to 8 columns and 10^{th} column of input matrix is the training set while 9^{th} column is the test set. Meanwhile, 10 different models have been considered for each input combination in order to choose the best division of the data.

Table 1. Input Combinations.

Input Combinations	Variables Included											
	R	**R1**	**R2**	**H**	**H1**	**H2**	**T**	**T1**	**T2**	**Year**	**Month**	**District**
1	X			X			X			X	X	X
2	X	X		X	X		X	X		X	X	X
3	X	X	X	X	X	X	X	X	X	X	X	X
4	X	X		X			X			X	X	X

*Note: Rainfall – R: Rainfall lag1 – R1: Rainfall lag2 – R2: Humidity – H: Humidity lag1 – H1: Humidity lag2 – H2: Temperature - T: Temperature lag1 – T1: Temperature lag2 – T2

Thus, for different sets of training data and test data divisions used in this study, the plots of σ values with their corresponding sum of square error (SSE) values were obtained separately for different input combinations. The best model is selected by considering the best input combination for the best σ values identified, which generates the smallest sum of squares value.

The summary of the recorded SSE values is presented in table 2 for best spread value and best input combination in each option and the minimum SSE value was identified accordingly. As presented in table 2, the minimum SSE value was recorded in 2^{nd} input combination for the 8^{th} division of training and test set combination.

Table 2. The SSE values for the four input combinations.

Input Combination	**1**	**2**	**3**	**4**
Best Spread	**1**	**0.8**	**1**	**1**
1	291.8360	41.0000	542.0000	67.3185
2	291.7045	37.9854	498.9998	44.0000
3	276.8312	47.9984	892.0000	33.0000
4	400.4834	35.8997	926.4045	28.8696
5	276.4105	28.4092	584.0000	48.4466
6	185.4227	33.9826	753.0000	41.8352
7	222.5113	36.0000	559.9363	36.9996
8	318.7454	17.0000	685.1784	48.9902
9	518.5261	43.0000	925.0000	54.5362
10	413.6808	36.4304	541.0000	40.0000
Minimum Value	185.4227	17.0000	498.9998	28.8696

The best GRNN model is given in figure 1.The performance measures such as root mean square error (RMSE), mean absolute error (MAE) and correlation coefficient (R) were obtained for both models and the measures are presented in table 3.

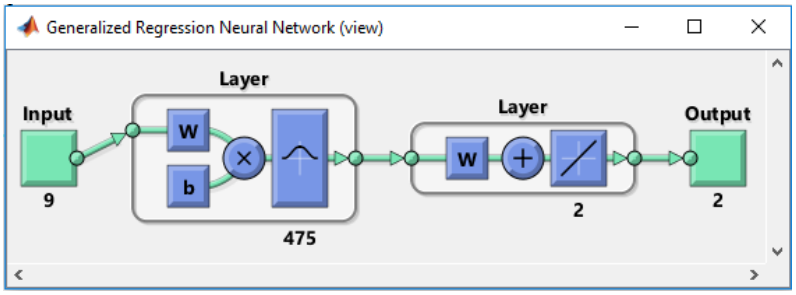

Figure 1. Developed GRNN.

Table 3. The performance measures of GLMM model and GRNN model.

	GLMM	**GRNN**
Root Mean Square Error (RMSE)	0.5770	0.1794
Mean Absolute Error (MAE)	0.5000	0.0284
Correlation Coefficient (R)	0.9324	0.9926

According to the values identified in Table 3, RMSE and MAE values are significantly lower in GRNN model than those of the GLMM model. That indicates the errors generating from GRNN model are very low when compared to those of the GLMM model. The R value indicates the correlation of the actual response value and the fitted response value. The correlation in both models are high, but GRNN model over performs GEE model with a slightly higher correlation coefficient.

4. Conclusion and future work

In this study, a GRNN model was developed as a joint model of survival and count variables of Dengue patients in Sri Lanka. The performance of the new model was compared to the previously developed GLMM model by the same authors. The comparison reveals that the GRNN model has a better fit. As this study is based on only one dataset, the conclusion cannot be generalized to all the scenarios. For the generalization purpose, a simulation study should be undertaken as a further research.

References

[1] Sunethra A A and Sooriyarachchi M R 2015 Joint Modeling of a Survival and a Count Response *Proc. of Annual Research Symposium 2015* (University of Colombo)
[2] Hapugoda J C and Sooriyarachchi M R 2016 Joint Modeling of Survival and Counts: Joining the Discrete Time Hazard Model with Poisson Regression Model *Proc. of the conf. of Int. Society of Clinical Biostatistics* (United Kingdom)
[3] Sehrawat R, Gupta P and Yadav R 2015 *Basic of Artificial Neural Network Journal of Computer Science and Engineering* **1**(5) 26-30
[4] McCulloch W and Pitts W 1943 A logical calculus of the ideas immanent in nervous activity *Bulletin of Mathematical Biophysics* **5** 115–33
[5] Koutsoyiannis D 2007 *Hydrological Sciences Journal* **52**(4) 832-39
[6] Tu J V 1996 *Journal of clinical epidemiology* **49**(11) 1225-31
[7] Timonin V and Savelieva E 2005 Spatial Prediction Of Radioactivity Using General Regression Neural Network *Applied GIS* **5** 1-19
[8] Comrie A C 1997 *Journal of the Air & Waste Management Association* **47**(6) 653-63

[9] Düzgün R 2010 *Int. Research Journal of Finance and Economics* **51** 59-70
[10] Specht D F 1990 Probabilistic neural networks *Neural Networks* **3** 109-18
[11] Specht D F 1991 *IEEE Transaction on Neural Networks* **2(6)** 568-76
[12] Kohavi R 1995 *Int. Joint Conference on Artificial Intelligence* **2** 1137-43
[13] Yay M and Akıncı E 2009 *Cypriot Journal of Educational Sciences* **4** 58-69