

Comparing Time series, Generalized Linear Models and Artificial Neural Network Models for Transactional Data analysis

Joseph Twagilimana, University of Louisville, Louisville, KY

ABSTRACT

The aim of this paper is to compare the Autoreg Procedure for fitting Time Series Models, the Glimmix procedure for fitting Generalized Linear Models and the Artificial Neural Network for the analysis of medical data. This comparison will be illustrated by the Analysis of Length Of Stay (LOS) at a Hospital Emergency Department (ED). Almost all medical records contain a date and a time stamp to record events. Unfortunately the arrival of patient at a Hospital Emergency Department doesn't happen at regular interval of time which makes the variable Length of Stay (LOS) transactional than a Time Series. Using the SAS HPF procedure, transactional data can be transformed into Times series.

For further LOS analysis, Time Series Models, or Generalized Linear Models or Data Mining techniques such as Artificial Neural Network can be applied. What these techniques have in common is that they can handle autocorrelated variables. In this paper, we show how these methodologies can be applied and we compare their results.

Keywords: Generalized linear mixed models, Text mining, Decision trees, Neural network, Mining medical data, transactional time series.

INTRODUCTION

When analyzing data, there is no a priori best model. The aim of this paper is to show how several candidate models can be used before deciding which one provide better results. Transactional series and Time series have the particularity of having autocorrelated observations and the SAS AUTOREG procedure, the GLIMMIX procedure are designed to handle this type of data. Artificial Neural Network, are data mining techniques that do not make any assumptions about the data and can be applied to analysis of interval variables. In this paper we apply and compare these three methodologies for the analysis of the length of stay (LOS) at a hospital emergency department.

Preliminary studies have shown that the length of stay (LOS) at a Hospital Emergency Department (ED) is closely related to the time of triage, the process of determining which patients are the most critical and have to be treated first. Triage can happen at any time as the patients walk into the ED. These random arrivals correspond to random exits, making the variable LOS transactional. Ordinary time series analysis techniques cannot be applied to transactional data as they require time to be defined as fixed intervals. SAS has recently developed the procedure HPF (high-performance forecast), which allows the analysis of transactional data. Using the HPF procedure transactional data can be accumulated to a regular time interval to form time series data. By choosing an accumulation interval of one hour, one may be able to predict LOS for each of the 24 hours of the day. With an accumulation interval of 4 hours, or 6 hours, one may be able to predict LOS for the 4 hours, or 6 hour periods. A long accumulation interval tends to produce data that are more correlated than those produced by a short accumulation interval as this can be seen on the correlogram in Figure1.

A correlogram, is the plot of the set $\{\rho_0, \rho_1, \dots, \rho_k\}$ where $\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}$ and

$\hat{\gamma}_k = \frac{1}{N} \sum_{t=1}^{N-k} (x_{t+k} - \bar{x})(x_t - \bar{x})$ is the autocovariance coefficient at lag k. .

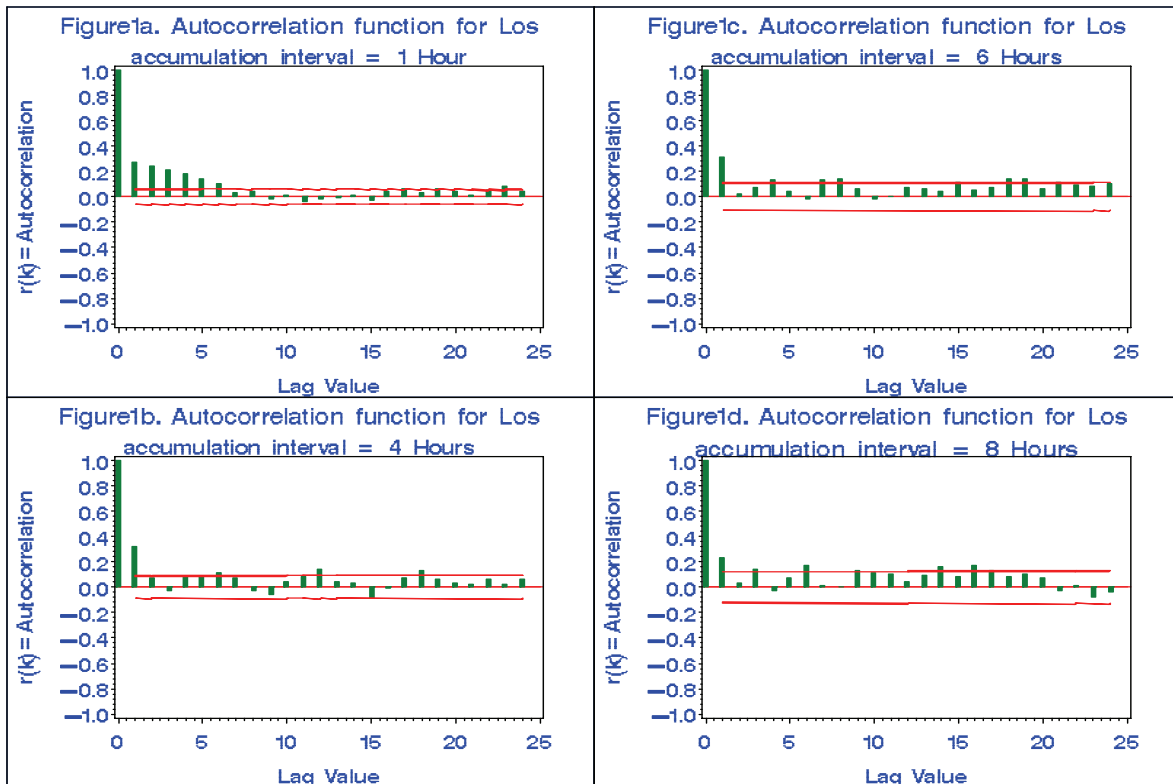


Figure 1 Correlogram of accumulated LOS for a 1 Hour, 4 Hours, 6 Hours and 8 Hours accumulation interval. A short accumulation interval tends to produce time series that are more autocorrelated.

ACCUMULATING TRANSACTIONAL DATA TO A TIME SERIES

Once the accumulation interval is decided, the SAS high performance forecast procedure (PROC HPF) can be used to transform the transactional data into a multivariate time series. The proc HPF is very important as an automated forecasting procedure, especially in the following situations:

- A large number of forecasts must be generated.
- Frequent forecast updates are required.
- Time-stamped data must be converted to time series data.
- The forecasting model is not *a priori* known for each time series.
- Future values of the independent variables are needed to predict the dependent variable.

The big challenge with the HPF procedure is that it doesn't handle nominal variables. But with medical data, the most important variables are nominal; for example, complaints, diagnoses, charges, and gender. Instead of leaving them out of the analysis, we recoded them using 0 and 1 dummy variables. As this may be a tedious task if there are several nominal variables with several classes, we recommend to the SAS software developer that they incorporate an automatic dummy recoding into the statistics and data mining components. For example, the variable Cluster1 is a numerical binary variable with value 1 if the observation belongs to Cluster 1 and 0 otherwise. Some other SAS procedures, such as proc GLM or Proc MIXED, perform automatically a nominal recording, but not PROC HPF.

When invoking the procedure HPF, for accumulation purposes, no forecasts are needed, and the option **lead** must be set to 0. The following code shows how the procedure can be used:

```

proc hpf data=Two out=Three lead=0 ;
  id Triage interval=Hour1. accumulate=Total;
  forecast LOS Age visits ChargesCount;
  forecast Cluster1 - Cluster8 MDCCode1 - MDCCode8
    RN_Code1 - RN_Code32 Disposition_Rec1 - Disposition_Rec4
    Time00 - Time23 Male Female Emergent Urgent NonUrgent
    / Model=idm ;/*idm= intermittent time series */
run;
quit;
data sasuser.HPF2IbexFinal_Clus;
  set Three ;
  LOS=round(LOS/visits,1);
  Age=round(Age/visits,1);
run;
Quit;

```

Accumulating the transactional variable LOS by one hour intervals leaves us with a time series with 25% missing values and many zeroes. Such time series are called intermittent time series. These time series are mainly constant values except for relatively few occasions. With Intermittent series, it is often easier to predict when the series departs from the constant value and by how much from the next value. The HPF procedure uses special methods in handling this kind of data. Intermittent models decompose the time series into two parts: the interval series and the size series. The interval series measure the number of time periods between departures. The size series measures the magnitude of the departures. This is specified in the procedure by using the option “*model=idm*” in the forecast statement.

Components of the Time Series LOS and Predictions.

Time series have one or more variation components: Trend, Cyclic variation, Seasonal, and Irregular variation. A trend shows a shift variation in the level of the mean. A trend can be linear, having a constant rate or increase or decrease; or it can present a periodic variation (Figure 2 (a)). The trend main effect is in the increase of the decrease of the mean. If a time series oscillates at regular intervals, we say that it has a cyclic component or a cyclic variation (Figure 2 (b)). *Seasonal variation* is a cyclic variation that is controlled by seasonal factors. Water consumption has a seasonal high in summer and a low in winter. It happens that it is sometimes possible to disassociate trend and cyclic components. An Irregular component is an irregular fluctuation about the mean. The components can be additive or multiplicative. Decomposition of a time series into its components can be done automatically using the SAS software. The figures below show the multiplicative components of the time series LOS: the trend-cyclic component (Figure 2 b), the seasonal component (Figure 2 c) and the irregular component (Figure 2 d).

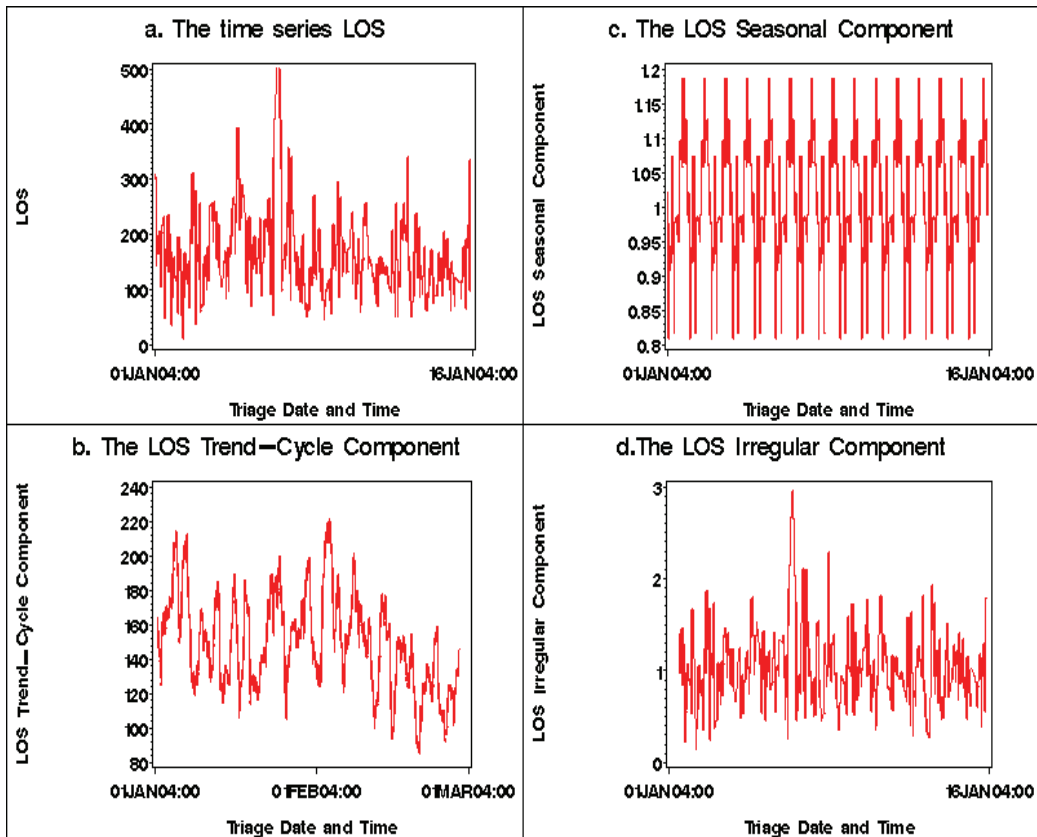


Figure 2 Decomposition of the time series LOS into its components: The Trend-cycle (b), the Seasonal (c) and the irregular (d). The general trend shows that the LOS tends to decrease from January to March.

Los Predictions with Proc AUTOREG

Among the time series components, only the irregular component is random. Using the SAS AUTOREG procedure, we predicted the irregular components and then recombined all the components to obtain the final predictions. A Plot of LOS versus its predictions is shown in figure 3.

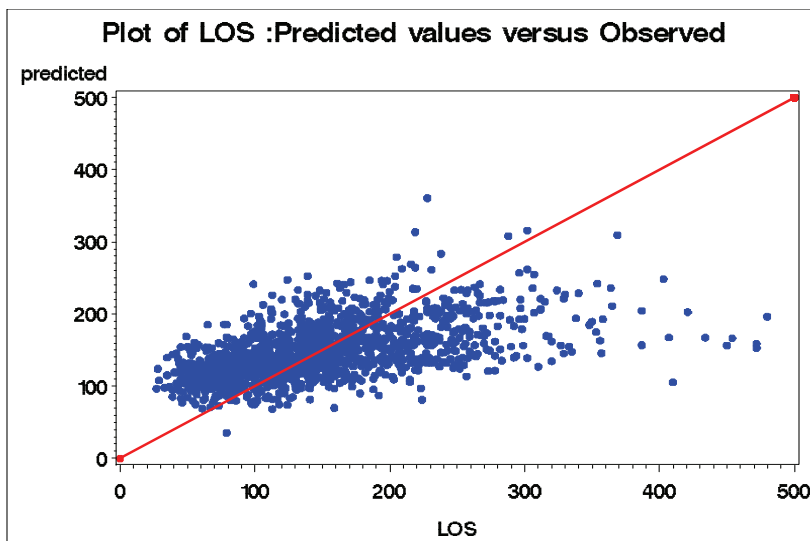


Figure 3. Plot of LOS versus its predictions. When the LOS becomes too long, it is hard to predict since the scatter points spread further from the 45 degree line (red).

Generalized Linear Mixed Models

Generalized Linear Models were fit using the SAS procedure, Proc Glimmix, which is still an experimental procedure. The GLIMMIX procedure doesn't require that the response be normally distributed. It doesn't require a constant variability, nor does it require observations to be independent. The only requirements are that the response has a distribution that belongs to the exponential family, and that the relationship is linear. The Glimmix procedure can fit models with only fixed effects as well as models with random effects or both. The code used is as follows:

```
proc glimmix data=[dataset];  
class [List of Nominal Variables];  
MODEL LOS = [Fixed effect inputs variables] / link=identity point ;  
random [random effects]  
nloptions technique=[Optimization techniques];  
Output Out=Glimmixout Pred=P Resid=Residual;  
run;
```

A plot of the observed versus the predicted values of LOS by the Glimmix procedure is shown below in Figure 4.

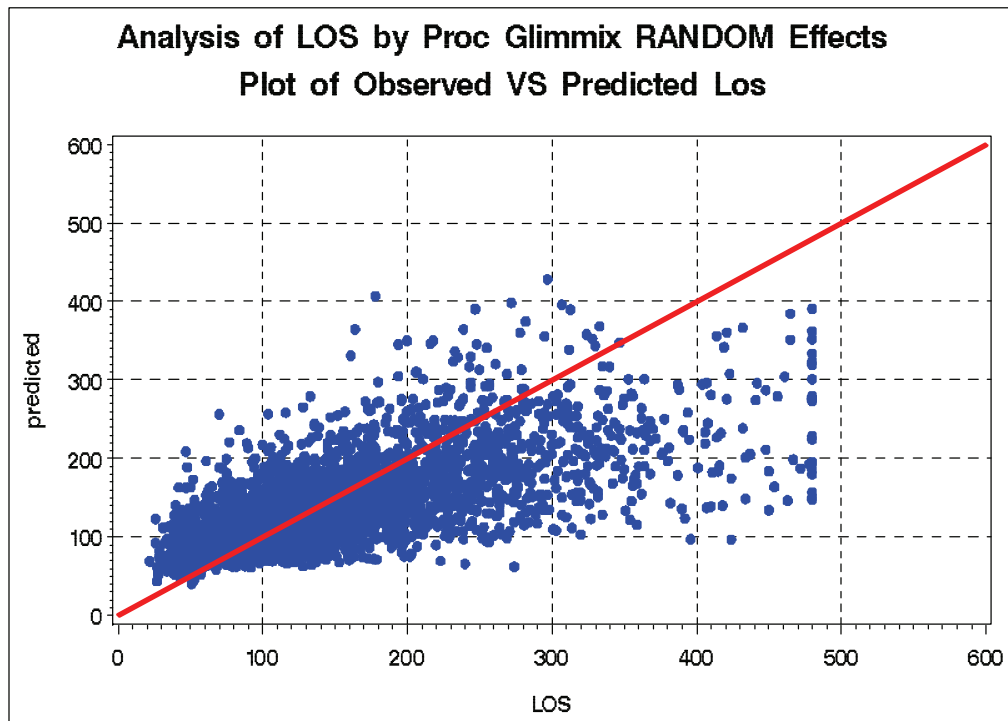


Figure.4. Plot of observed values versus the predicted values by Proc Glimmix.

SAS Enterprise Miner Artificial Neural Network

An Artificial Neural Network (ANN) is an information-processing system that has certain performance characteristics in common with biological neural networks. It is a computing process that mimics the neurophysiology of the human brain. Similar to the brain, in the ANN, information is processed in many processing units (neurons or nodes) interconnected by means of directional links, each with an associated

weight or strength W_{ij} , W_{kl} (Figure 5). The first index refers to the neuron, and the second to the input to which the weight refers.

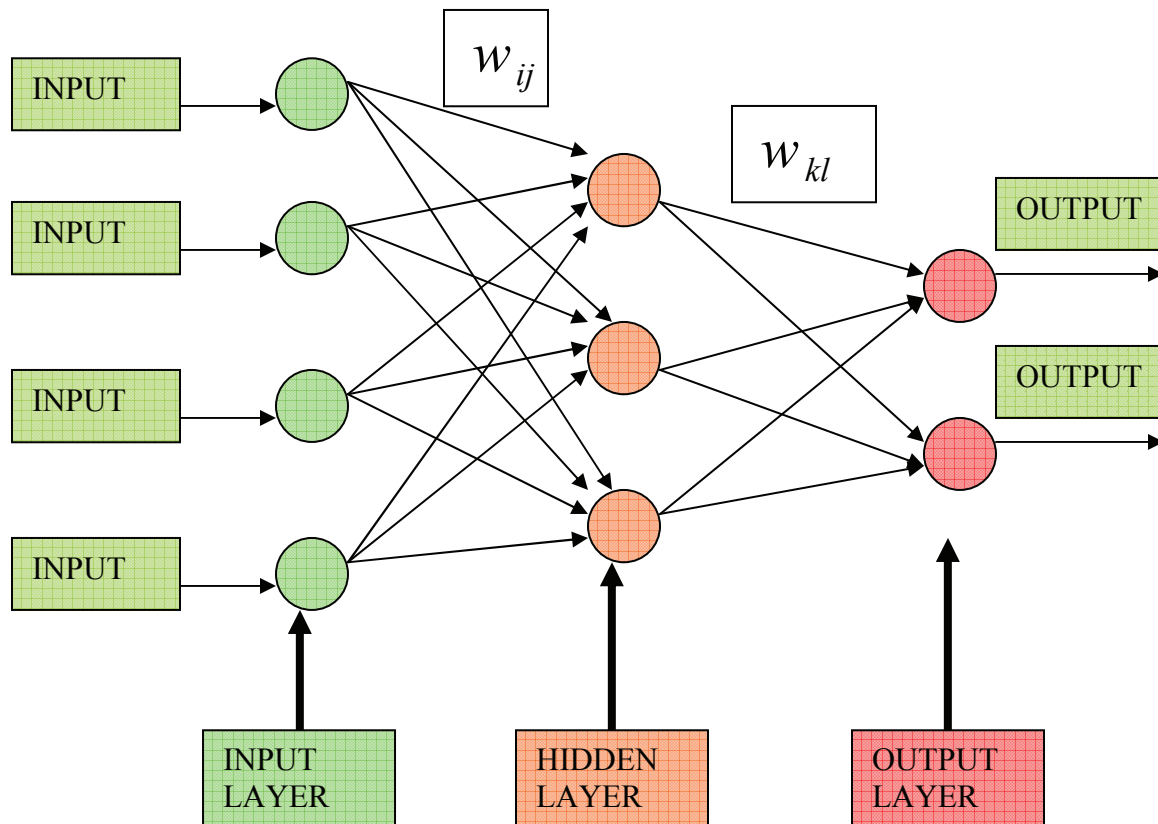


Figure 5. Architecture of an Artificial Neural Network.

An Artificial Neural Network is applied to predictions (classification and regression). For the regression model, we only have one output neuron. For a K-class classification, there are K output neurons. In the domain of Statistics, Artificial Neural Networks are non-linear statistical data modeling tools.

The Neural Network Learning Process

To start this process, the initial weights are chosen randomly. Then the training, or learning, begins. During the learning process, data cases (rows) are presented to the network one at a time. The network processes the records in the training data one at a time, using the weights and activation functions in the hidden layers, and then produces predicted values. The predicted values are compared to the target values. The differences between outputs and target values constitute the error function. Training techniques are aimed to minimize this error function by adjusting the initial weights. The process starts over until some stopping criteria are met. Most error functions are based on the maximum likelihood principle, although computationally, it is the negative log likelihood that is minimized. Using SAS Enterprise Miner, we applied the ANN to the predictions of LOS.

METHODS COMPARISONS

We compared the Glimmix procedure, the time series procedure Proc Autoreg that fits Time series models, and the Artificial Neural Network. From Figures 6 and 7 below, we conclude that the time series models applied to the accumulated data performed better than the Glimmix procedure when applied to the same data, and that both performed better than the Artificial Neural Network.

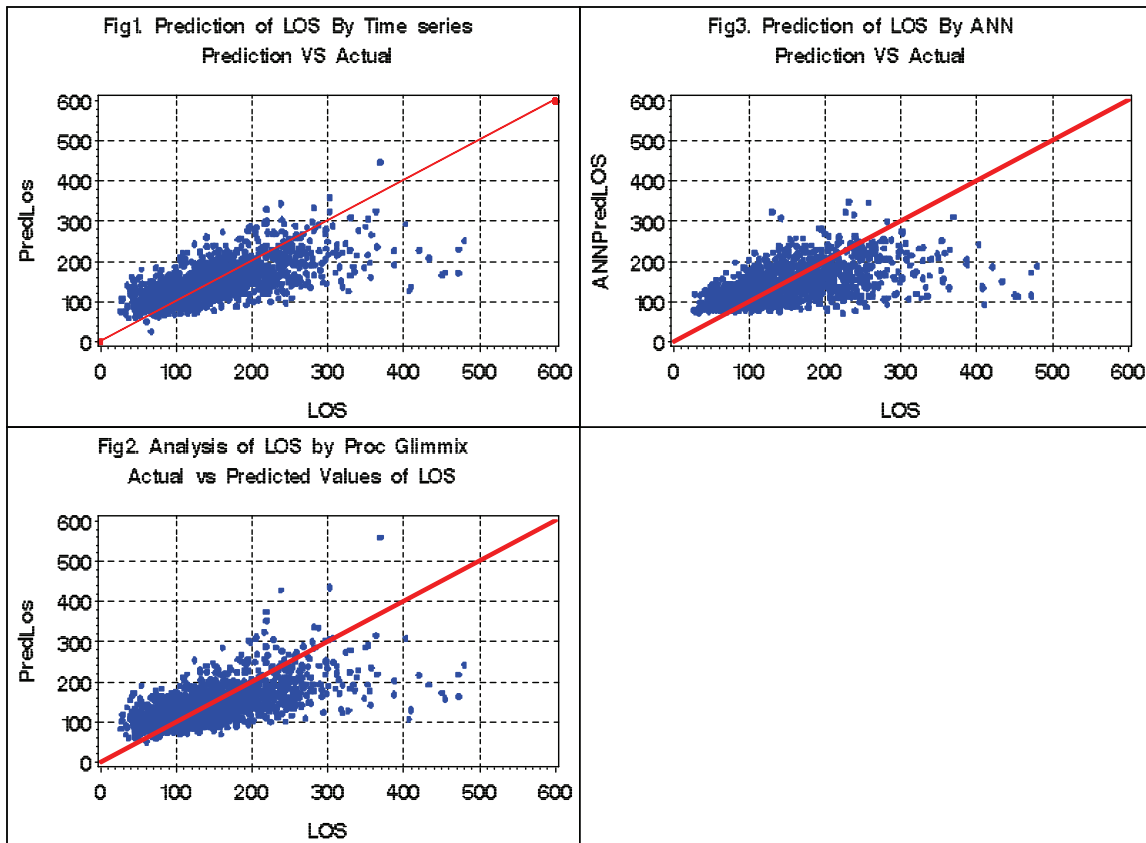


Figure 6. Comparison of Glimmix procedure, Time series models (Proc Autoreg) and Artificial Neural Network.

The graphs in the Figure 6 show the predicted values of LOS plotted against the observed ones. These graphs show that the predicted values by the Autoreg procedure are closer to the observed ones. In fact dots in the plot are closer to the red line which the 45 degree lines with the equation *predicted=Observed*.

The fact that the Autoreg procedure perform better than the other models is also confirmed in Figure seven showing the residuals of the three models. The mean of the Autoreg procedure is closer to zero than the mean of the other models, and we also have the lower variance in the case of the autoreg procedure.

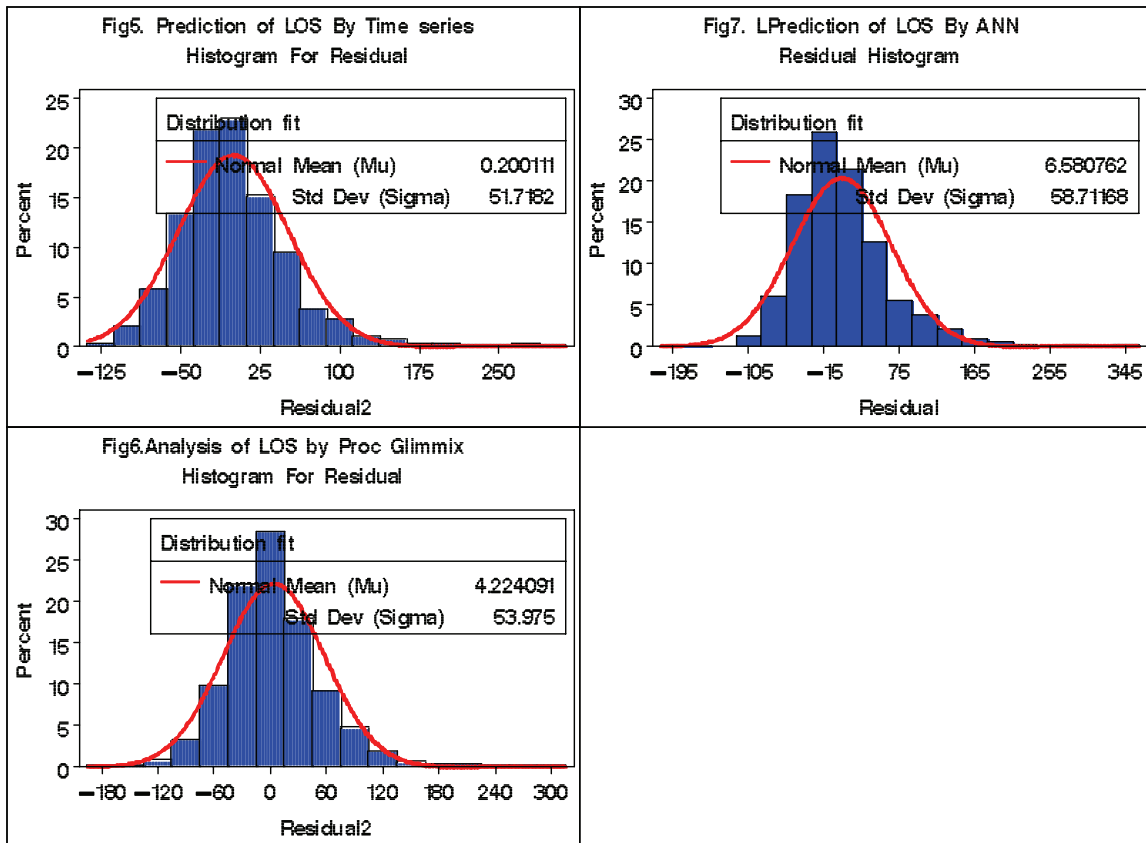
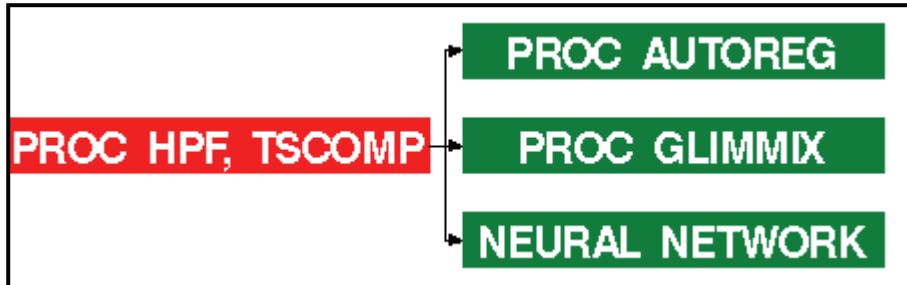


Figure 7 Comparison of Residual of Glimmix procedure, Time series models (Proc Autoreg) and Artificial Neural Network.

Conclusion

When analyzing time series that are nonstationary, nonnormally distributed and with nonconstant variance, Autoregression models, Generalized Linear Models and Artificial Neural Network models can be applied in order to make the right choice on the final model. In the case of transactional series the HPF procedure must be applied first in order to transform the transactional series into time series. The following diagram is a summary of the process.



When analyzing data, we recommend that all candidate models be explored and then the optimal be chosen. In some cases, methods may be combined.

REFERENCES

- [1] Michael J.A Berry, Gordon S. Linoff, *Data Mining Techniques*, second edition, Wiley Publishing, Inc, Indianapolis 2004.
- Relationship Management*. New York: John Wiley
- [2] Mohsen Pourahmadi (2001) "Foundation Of Time Series Analysis and Prediction Theory"
- [3] The Glimmix Procedure, Nov 2005 <http://support.sas.com/rnd/app/papers/glimmix.pdf>
- [4] SAS 9.1.3 High-Performance Forecasting, User's Guide, Third Edition http://support.sas.com/documentation/onlinedoc/91pdf/sasdoc_913/hp_ug_9209.pdf

CONTACT INFORMATION

Joseph Twagilimana
Department of Mathematics
University of Louisville
Louisville, KY 40292
502-852-6826
j0twag01@louisville.edu