8-2013

Forecasting Emergency Department Volumes Using Time Series and Other Techniques

Uchechukwu A. Nwoke
St. Cloud State University

FORECASTING EMERGENCY DEPARTMENT VOLUMES USING TIME SERIES AND OTHER TECHNIQUES

by

Uchechukwu A. Nwoke

B.Sc., Obafemi Awolowo University Ile-Ife, Osun State Nigeria, 2004

A Thesis

Submitted to the Graduate Faculty

of

St. Cloud State University

in Partial Fulfillment of the Requirements

for the Degree

Master of Science

St. Cloud, Minnesota

August, 2013

This thesis submitted by Uchechukwu A. Nwoke in partial fulfillment of the requirements for the Degree of Master of Science at St. Cloud State University is hereby approved by the final evaluation committee.

____
Chairperson

____

____

____
Dean
School of Graduate Studies

FORECASTING EMERGENCY DEPARTMENT VOLUMES USING TIME SERIES AND OTHER TECHNIQUES

Uchechukwu A. Nwoke

The aim of this research is to forecast patient volumes in the Emergency Department of a regional hospital in Minnesota, which eventually will aid in addressing the issue of registered nurse staffing fluctuation, more specifically, productivity and capacity planning in the ED. Several methods are applied to forecast arrival patient volume, and cumulative patient volume to evaluate each model's performance. The methods considered are linear regression, time series models and dynamic latent factor method. Long term forecast for as long as six months ahead is the goal here due union regulations that only allows for significant changes in registered nurse staffing schedule be put in place six months in advance. This long term forecast will enable administrators implement effective and timely changes to enhance productivity.

The patient arrival count, where each patient is counted once in the system, is analyzed to see how many patients the department encounters hourly. Also, cumulative patient count which gives us an idea of how many patients are in the department at any given time was also considered, here patients are counted for every hour they are in the emergency department (ED). Patient who come to the ED are categorized by their acuity level. Of all the patients that came to the ED, 52% need urgent care; this group is also analyzed to predict their arrival volume.

Lastly data was simulated with different patterns and the forecasting results from the different methods were compared and estimated. The forecast accuracy and performance for these models is then evaluated using out-of-sample forecasts for up to six months ahead. Mean square error (MSE), Root mean square error (RMSE) and mean absolute error (MAE) were utilized tosee which method is most reliable and also consistent.

____
Month Year

Approved by Research Committee:

____
Xu, Hui Chairperson

ACKNOWLEDGEMENT

I would like to express my profound gratitude to Dr. Xu for his guidance, patience and support during all these years. Thank you for being an incredible advisor and friend. Nothing of this would have happened without your encouragement and advice.

I am also very grateful for the help and advice from my thesis committee Dr. Zhang and Dr. S Warner. I could not have asked for a better committee. I also want to commend and appreciate Rob Hyndman and David Matteson for their work which this paper is anchored heavily on, and also their timely and friendly correspondence.

I want to thank the head nurse of Buffalo hospital emergency department Margo Binsfeld and Jill Mahoney for providing me with all the resources needed for this thesis and for all your kindness shown to me during the process.

To all my professors at St Cloud state university, thank you for all the knowledge and wisdom impacted in me, it is deeply appreciated.

I want to sincerely thank my family for believing in me even when I had stopped believing in myself, without your support and love this would not have been possible.

To all my friends who sat in the library with me and proofread my papers I am deeply thankful, you all are a blessing.

Last but not the least to Him who gave His life for me, My Lord Jesus Christ, the I AM THAT I AM, to you I owe it all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Chapter 1

INTRODUCTION

The health care system has experienced an increased interest in and recognized appreciation of the essential role nurses play in patient care.[4, 46] During a time in which health care resources are becoming limited, overwhelmed, and financially taxing, the key focus has become productivity and capacity planning. This problem is multi-dimensional, due to the fact that administrators must carefully consider their operations. Some of which include; adequately staffing registered nurses and allocating resources. The objective is to ensure quality patient care, while avoiding overstaffing and thus avoiding unnecessary expenditure.[10, 29, 32]

> Operational studies have been successfully implemented in several areas to improve patient experience: reduced wait time, more accurate patient record keeping, patient satisfaction "surveys," open and frequent communication, and forecasting.[25] HealthCare has seen a lot of improvement over the years but there is still room for more. Planning and staffing is of the utmost importance because of its direct impact on patient and employee safety.[2]

Understanding staffing fluctuation and patient volume could help improve the health care delivery system across every level but it appears to be more difficult for the Emergency Department (ED). In a clinic or surgery setting staffing is fairly predictable because patients make appointments and so the departments know what to expect and can plan ahead, but this is not so for the ED. Due to The Emergency Medical Treatment and Labor Act (EMTALA) [1], non-profit emergency departments must provide medical screening for every patient. Many people that do not have insurance utilize the ED as a place to receive primary care.[41] On the other hand, there are times when the ED experiences a low volume of patients, having more than required staff increases health care expenditure and cost, also reduces the overall efficiency of the department.[2]

When staffing and planning is effective and efficient, employees will have necessary resources to do their job well and productivity can be maximized. This in turn improves positive patient outcomes and experiences, patient and optimum throughput, employee satisfaction, and reduces unnecessary spending (see figure below).

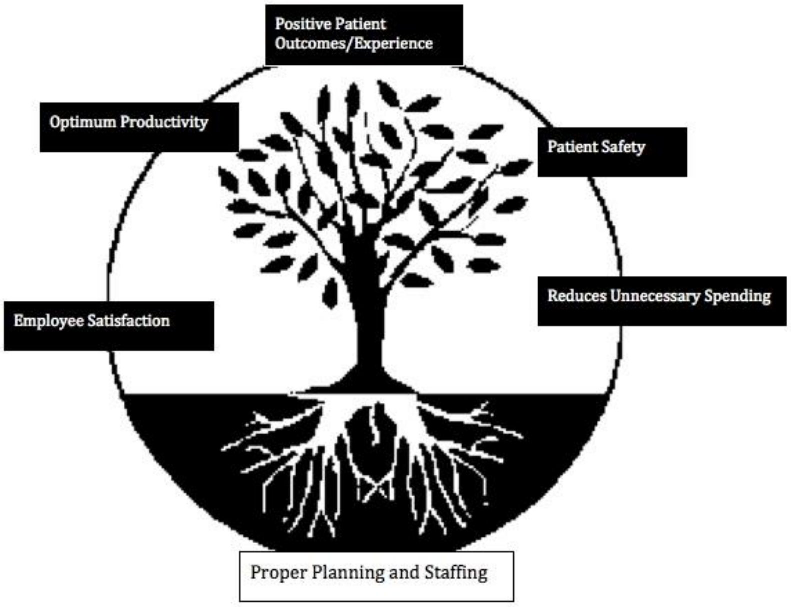
Positive Patient Outcomes/Experience
Optimum Productivity
Patient Safety
Reduces Unnecessary Spending
Employee Satisfaction
Proper Planning and Staffing

Figure 1: Capacity Planning Chart

First, Forecasting can be defined as, "the process of making statements about events whose actual outcomes (typically) has not yet been observed. A commonplace example might be estimation of some variable of interest at some specified future date."[3]

In other words it is trying to estimate a variable before it is observed, or to “foresee the future”. A very common example of forecasting is weather forecast. Forecasting is widely used in marketing, securities analysis and, it has evolved into a multidisciplinary science.[5, 20, 26] It is an essential instrument in most industries requiring scientific planning. There are several cases where forecasting can be applied; it might be whether to forecast when the sun will rise tomorrow or what a house bought today will be worth in five years, whatever the case may be, forecasting is a vital tool that facilitates proficient and effective planning and productivity.[26] the predictability of an event or a variable relies on various factors including.[26]

i. How much data is collected

ii. How accurately is the data collected

iii. If the contributing factors can be adequately explained or understood.

iv. Will the event or variable be affected by forecast values?

An example is , if a patient family medical history is known and the patient lifestyle is closely monitored the possibility of having a heart attack might be highly accurate compared to that patient being involved in an accident. In the latter case the data most likely isn’t collected and all the contributing factors are not understood. Sometimes the forecast can in itself affect the outcome, and this is one of the dangers of forecasting .[4] For example say there is a forecast for increase in the price of a commodity, this will in most cases drive consumers to increase their demand. When demand surpasses supply, this in turn will lead to price increase. One really has to keep in mind the limitations and choose rather to err on the side of caution, when applicable.

Next, forecasting method is "a procedure for computing forecasts from present and past values".[4] A good forecast is based on the assumption that the factors involved are changing and aims to capture the way the things are changing. Forecasts method can be simple like linear regression or complex like artificial neural networks. Various forecasting methods have been utilized in the quest for proper planning: linear regression, artificial neural network, time series, etc. In this paper even though other methods are considered, the main focus will be on time series methods.

Then, time series is defined by Chatfield as "a collection of observations made sequentially through time".[2] Examples are daily temperature of a city, number of babies born every hour in a hospital, etc. Time series forecast involves using data collected sequentially to make predictions. The aim of using time series methods is to predict future values based on data collected in the past and present.[22] Time series forecasting amongst other methods is a tool that has been be applied in predicting patient volumes and other variables (example length of stay) that are peculiar to the ED. Various studies have been carried out using both univariate and multivariate methods. Univariate methods depend solely on previous values of the series being forecasted while a multivariate series relies on additional explanatory variables.[4] Examples of these methods include; historical average, linear regression, time series models which includes; auto regressive integrated moving average (ARIMA) models and multivariable time series.[22, 24, 30, 33, 36-38] The ARIMA model has most widely being used in predicting patient volumes, length of stay, etc. One limitation of ARIMA models is that it does not accommodate series with multiple seasonal patterns as our data suggests. In this paper the aim to is to apply ARIMA models that have been modified to include multiple seasonality, some innovative exponential smoothing methods proposed by Taylor,[39, 40] Gould et al.[14] and De Livera,[9] and also a factor latent model based on Poisson process proposed by Matteson ·These methods are being are considered because the series for Patient volume in ED is characterized by multiple seasonal patterns. We will compare these to the previously used methods to see if there is increased accuracy.

Chapter 2

A REVIEW OF LITERATURE

In recent years, many research studies have been done in forecasting daily patient volumes in acute care hospitals. The legislation passed in California in 1999 has led to a series of questions and close monitoring of registered nurse staffing.[4, 10, 21] Emergency departments are one of the most used providers of acute care in the health sector; the study of which can play a vital role in the development of the subdivision and the entire industry in general.[16]

The number of emergency departments in the US declined by 425 departments in the years 1993 to 2003. Despite this decrease, the patient volume has increased by 26% in visits.[21] Also between 1997 and 2007 there was an increase in patient volume of 12.5% and a decrease of 189 departments. This development has made the planning and effective allocation of resources crucial.[32, 38, 46] One way to tackle this problem is the use of models to produce accurate forecasts to help ensure that supply meets demand. Several authors have used statistical techniques to build models to forecast different ED behaviors like patient volume, length of stay or patient acuity with or without covariates.[22, 24, 30, 33, 36-38]

There have been a lot of publications on Emergency departments in recent years, and we would be looking at some of them to answer these questions. This review attempts to answer some of the following questions:

a) What forecasting methods have been studied?

b) What factors were considered and why?

c) How effective were these models?

d) Are there other factors that should have been considered?

e) Is there any need for new forecasting methodology?

f) How suitable are these methods especially for long term forecasts?

Emergency Medicine Papers

Jones et al. (2007)[22] used multiple linear regressions as a benchmark model while comparing several other models; for instance, time series models such as SARIMA (Seasonal Autoregressive Integrated Moving Average), exponential smoothing, time series regression, and Artificial neural network to predict daily patient volumes in the ED. The variables considered in the linear regression model were calendar variables (weekday, month and holiday). This was done using dummy variables and a “near holiday” variable was also considered, Climatic variables were put in the model as well, but only the time series regression used these variables. Twenty-seven months’ worth of data was collected for the analysis. The goal was to predict 1- 30 days in advance and compare the benchmark model to the others to see if the any of the new models achieved better forecast accuracy.

The time series regression model showed some improvement from the linear regression model but offered only little improvement in post forecast accuracy. All others (SARIMA, exponential smoothing and artificial neural network) failed to provide consistently accurate forecasts for ED volumes. This study also confirmed the widely held belief that there are weekly and seasonal patterns found in patient volume but did not take this property into account when modeling the time series data. Jones et al. concluded that even though time series regression provided slightly more accurate forecasts of ED, they violated a major assumption in linear regression. The regression based model that incorporated calendar variables and accounted for site-specific, special day effects and also allow for residual auto-correlation, provided the most informative and consistently accurate predictions of daily ED volumes. In other words the regression model was preferred to the time series model but long term forecasts were not considered in this study.

Schweigler et al. (2009)[33] also applied statistical models to predict overcrowding of the ED. Historical averages were considered reliable for long term forecast, but short term forecasts were also desired. In developing a model, two main factors were considered: the ability for wide usage, and simple models yet accurate forecast in making predictions. Three different locations were used in the analysis data was collected hourly. Two methods, namely; a 24-hour SARIMA model and a sinusoidal model with an AR structured error term, were compared with the historical average method as the benchmark. The historical average (HA) method was basically the mean occupancy for each site each hour of the day. The two AR (seasonal and sinusoidal) models were chosen because they were the accounted most conservatively for the 24-hour cycle and had a strong correlation between the previous and the next hour's occupancy. The HA showed the best goodness of fit but using the AIC (Akaike's Information Criterion), which is basically a measure of relative goodness of fit, SARIMA performed best because the HA requires more parameters than the AR models. On the other hand, forecast accuracy measured using RMSE (root mean square error), which is calculated by summing the difference between the observed and predicted values, showed that the AR models performed better.

While AR models are an improvement from the historical average method, it does not account for other cycles such as seasonal cycles, weekly cycles etc. and other complex season's patterns that characterize the patient volume in an ED. In simpler terms: times series models provide a better statistical fit than other models such as linear regression or historical experience, but performance against future behavior has not typically been dealt with. Also time series methods have not yet been used to directly investigate overcrowding but have been used to model related behaviors such as patient arrival per minute.

Sun, Heng, and Seow, (2009)[36] carried out a study in Singapore intended to identify local factors associated with daily patient volume and develop coordinating prediction models. Patient acuity levels were taken into consideration. Variable selection was based on literature, local weather factors and availability of data.

ARIMA models were applied to the three categories of acuity and overall data. The three categories: P1 (resuscitation and those in imminent danger), P2 (major emergency, with severe symptoms) and P3 (minor emergency with moderate symptoms). Mean absolute percentage error (MAPE) and Ljung test was used to choose the best-fit model. The best-fit model for P1 was ARIMA (0, 1, 1) and it did not show any weekly or yearly periodicity and was only predicted by ambient air quality, while for P2 was ARIMA (1, 1, 1) (1, 0, 1) showed weekly cycles and was significantly correlated with public holidays. For P3 was ARIMA (0, 1, 1) (1,0,1) showed strong correlation with day of the week, month, public holiday and ambient air quality of PSI(pollution standard index)>50. The MAPE for P1, P2, P3 and total attendances were 16.9%, 6.7%, 8.6% and 4.8%, respectively.

The authors concluded that even though there was a high variability in the data, the predictions had a good accuracy; despite P1 having the highest MAPE, it still demonstrated acceptable forecasting abilities. It was observed that weather did not have a significant impact on the models unlike previous studies, and this might be due to the fact that Singapore is in the tropics. P3 factors predicted higher attendances. This model was effective for both short-term forecasts (weekly) and long term (three months).

The limitations of this study include, other lurking variables not identified and studied, and the use of average daily temperature, also other forms of explanatory variables were not studied(quadratic, log, etc.). It would have proven more beneficial to predict hourly rather than daily patient volume. Another limitation is that only one year of information was used of this study and so annual trends cannot be captured, also long term forecasts were not considered.

Kam, H. J. (2010)[24] investigated the possibility of building a model to predict the number of patient visits to a regional ED per day. Analyses were done using moving average; univariate and multivariate seasonal auto regressive integrated moving average (SARIMA) models. These results were later compared and evaluated. For the moving average method, past time series data was used to calculate the arithmetic mean; its main advantage is its capability to remove non-conforming changes or periodic factors.

The seasonal ARIMA is an extended ARIMA model that allows for seasonal factors. When utilizing this method, the trend and seasonality are removed to “stabilize” the series before forecasting. This was seen to be effective in short term forecasting while the multivariate SARIMA model incorporates explanatory variables. Weather and calendric information were used as explanatory variables in building the model. The results suggest that the moving average method was flat as it returned the value of the mean rendering it inadequate. The SARIMA models were more accurate than the MA. The multivariate ARIMA was most accurate in predicting the daily volume. The authors suggested incorporating weather information (temperature and rain) to predict daily volumes, and further recommended that local, geographical and cultural factors be considered, and long term forecasts was not the focus here.

Rathlev (2011)[30] focused on analyzing length of stay and using staffing as a covariate. The authors analyzed the relationship between several covariates and length of stay per 8- shift. The covariates include: ED nurses on duty, ED discharged (defined as patients who went home, were transferred or admitted), ED discharge on previous shift, resuscitation cases, admissions and ICU admissions. This study was carried out in 8-hour shifts, 7.00 am -3.00pm, 3.00pm-11.00 pm, and 11.00 pm-7.00am. Patients were assigned based on their time of departure rather than initial presentation. Hospital occupancy was measured based on a 24 hour period. Due to the correlation of length of stay (LOS) (since the outcomes are not independent) ARIMA model was used to analyze the data. AIC was use select the best model and other relevant diagnostics were carried out. A full model was used and later all insignificant terms were dropped but there was no significant difference in the results. ARIMA (2, 2) was the best fit for the model, however, most of the covariates were found to be insignificant except for the number of ED admissions which was significant for all three shifts. ICU admissions on shift 1 were also significant and this can be explained by the fact that these patients require more nurses. Fewer than three ICU admissions were also seen as insignificant.

Statistical Papers

There have been recent innovations in time series modeling that are groundbreaking and stimulating.

Taylor, J. W. (2003)[40] first proposed that double seasonality can be applied to a time series to capture both seasonalities. Here the data was seen to possess intraday and intraweek patterns. Multiplicative Seasonal ARIMA and the Holt-Winters exponential smoothing formulation were applied with the latter adjusted to accommodate both seasonalities. The multiplicative seasonal ARIMA had earlier been proposed by Box et al .[3] and can be easily extended to accommodate three or more seasonalities.

Prior to this time no literature had considered extending the Holt-Winters method which was quite suitable for one seasonal pattern to accommodate double seasonality. In Taylor's paper, empirical analysis were carried out to compare the newly proposed double seasonal Holt-Winters method with the standard Holt-winters and also to compare it with the double multiplicative double seasonal ARIMA model. It was observed that the new model outperformed the traditional method. It was also improved by the inclusion of an AR (1) model for residuals and this was optimal when the parameters were estimated in the same process as the exponential smoothing technique. It also outperformed the well-specified double seasonal ARIMA model and so the author concluded that this new formulation has great potential.

Gould et al. (2007)[14] in their paper focused on modeling time series with multiple seasonal patterns and different lengths. This study introduced a new method applying the innovation space models which forms the basis for all exponential smoothing methods. Holt Winters (HW) exponential smoothing method and ARIMA methods of Box et al.[2] are most frequently used but they do not have the capability to account or detect day to day patterns and also it treats all days as same and does not pick up the varying patterns of different days. The double seasonal exponential smoothing method (DS) proposed by Taylor is a major improvement as it allows us to nest a cycle within a cycle but its major drawback is it assumes the same intraday cycle for all days of the week.

Thus a major objective of this new model called multiple seasonal (MS) processes is to allow for the seasonal terms that represent a seasonal cycle to be restructured more than once within a cycle if the need arises. For example in an hourly data there are 24 potential sub cycles, however if all the hours from 1am to 7am have a similar structure, it might be simpler to use the same sub-cycle for these 7 hours and the models be updated more frequently to improve accuracy also different smoothing parameters may be applied to different sub-cycles. This also helps reduce the number of sub-cycles. This model was developed for both additive and multiplicative seasonal patterns and was applied to a utility dataset obtained from a company in Midwestern United states and also to traffic data (hourly vehicle counts) for the Monash freeway in Victoria, Australia; both of them were recorded hourly.

In general the MS models provided more accurate forecasts than the HW method and DS methods and were also better suited to capture the changes in seasonality in the data. Several of the MS models were used with different restrictions and varying parameters and a model selection criterion was applied to select the best one using a combination of the mean square forecast error (MSFE), number of parameters and seed values in each model. In conclusion the MS model is an improvement to from the HW and DS because of its flexibility. It also allows for reducing the number of parameters and seeds required by the full MS model and missing values were adequately handled in both cases.

Taylor, J. W. (2010)[39] proposed to extend three of the more successful models than accommodated double seasonality to include triple seasonality. The three models are double seasonal ARMA model, Holt- Winters exponential smoothing (HWT) and the multiple seasonal (MS) method earlier proposed by Gould et al. Three cycles were considered; intraday, intraweek and annual cycles, and was used to forecast short term electricity demand on a British and French load series which consists of half hourly data collected for five years. Artificial Neural Network Model was also included in this study as the benchmark model.

In the ARMA and Holt-Winters methods, a single model was first considered using the intraweek cycle and this was further expanded to include the intraday and another the annual cycles thus for the double seasonal ARMA and exponential smoothing two series are proposed; one is the intraday and intraweek cycle the other is the intraweek and annual cycle. Finally, the intraday-intraweek model was extended to include the annual cycle, forming the triple seasonal models.

The MS model renamed the “intra cycle exponential smoothing method” (IC) here due to its emphasis on the intraday cycle also, only models that include the intraday cycle are considered. A common model is proposed for days that exhibit comparable patterns.

When certain restrictions are made this model becomes very similar to the double seasonal HW method for intraday and intraweek cycle.

The worth of extending the various models was estimated and it was observed that there was evident improvement in forecast accuracy when using double instead of single and a further substantial improvement when using the triple seasonal model. This was also seen in the Holt-Winters method. In the ARMA approach there was little difference in the double seasonal models but in the HWT method was a significant difference, with the intraday-intraweek model having an increased accuracy over the intraweek-intrayear cycles. An autocorrelation adjustment term was also included in the HWT and IC methods; and compared to models without the adjustment. Results show that it leads to significant improvement in the IC method, and even though the results were similar for the HWT methods this adjustment is needed.

On comparing the various methods it was seen that the HWT and the IC methods show strong similarities and also the triple seasonal versions. Double seasonal ARMA model did better than the double seasonal HWT method for the intraweek and intrayear but for the intraday and intra week double seasonal HWT was a little more precise. Both triple seasonal methods performed alike. When compared with the benchmark method, all models were seen to outperform the benchmark model.

Although forecast accuracy is of great significance, it is not the only benchmark to use when selecting a forecasting method. In comparison, HWT is superior to the ARMA model because the latter requires extensive specification and a more demanding optimization due to far larger number of parameters. It is also the same problem with the IC method and also there is no clear way to decide upon the number of unique cycles to be used. In other words since the HWT method is as good and less complex; the triple seasonal HWT model carries the day.

De Livera et al. (2011)[9] introduced a state space modeling framework for modeling complex seasonal periods which incorporates Box-Cox transformations, Fourier representations and time varying coefficients and ARMA error correction. A major attribute of this framework is that it is expedient to a wide range of applications and this is shown in three empirical studies. This is important because most time series models are designed to accommodate simple seasonal patterns with a small integer-valued period but are sufficiently developed to deal with time series with multiple patterns and non-linear patterns. The new method proposed here is stipulated to be a more versatile approach than previous existing models; it allows for multiple nested and non-nested patterns, handles potential nonlinearities and is able to produce better forecasts than previously existing models. It is also more suitable to handle complex seasonal patterns like non-integer seasonality, calendar effects and non-nested seasonal patterns.

The models proposed are the BATS (Box-Cox transform, ARMA errors, Trend and Seasonal components) and TBATS (Trigonometric Box-Cox transform, ARMA errors, Trend and Seasonal components) models are acronyms for the key features of the model. BATS model includes a Box-Cox transform parameter, ARMA (auto regressive moving average) errors parameters and seasonal periods. It is the most obvious generalization of traditional seasonal innovations model to accommodate multiple seasonal periods, however, it cannot be adapted for non-integral seasonality amongst other drawbacks.

The TBATS model is obtained by replacing the seasonal component in the BATS model with a trigonometric seasonal function, because of this it can be used to model non-integer seasonal frequencies. There are some advantages to ascribe to using this model which includes it can accommodate typical non-linear features that are often encountered in real time series and it involves a much simpler yet efficient procedure.

The model selection is based on the following:

- AIC (Aiake Information Criterion) is used to choose between models and provide the best basis for automated model selection. Other methods can also be used.
- The forecast for the TBATS model depend on the number of harmonics used for the seasonal component. This is needed because it and it is impracticable to consider all the possible combinations possible. A method was proposed to select the best model and it was based on a regression model using an approach based on multiple linear regressions.
- Suitable values for the ARMA orders are selected using a two-step approach and subsequent study[40] indicated that this approach provided the best out of sample prediction for the ARMA models compared to several alternatives.

The proposed models were applied to three complex time series; weekly gasoline data which is an example of non-integer seasonal periods, 5-minute interval retail banking calls data; an example of multiple nested seasonal periods and daily electricity demand in turkey an example of multiple non-nested and non-integer seasonal periods. The results from these models were compared by out-of-sample performance using the root mean squared error (RMSE). In all three, the TBATS models had a lower RMSE and so it was concluded that it outperformed the BATS model.

The authors suggested that other explanatory variables may be applied to the BATS and TBATS models, thus allowing more information to be included in the models. This approach was also seen to be general and can be used for any innovations in the state space model. It was also seen that the adaptability of the TBATS model is an improvement from previously existing models.

Matteson, David S (2011),[27] used a method which involves combining integer-valued time series model with a dynamic factor structure. Here, an integer valued time series model is introduced with a dynamic latent factor structure with day of the week and week of the year effects, accounted for as simple constraints on factor loadings. This factor structure allows for a substantial reduction in the number of parameters in the model. This model is claimed to lead to better short term forecast accuracy because it models unambiguously the remaining serial dependence. This is done by introducing the covariates (Day of the week and week of the year effects) using simple constraints on the factor loadings. Smoothing splines are used to estimate the model by imposing smooth evolution the factor levels of loading. Factor levels account for the non-stationary pattern in the intraday call arrivals while the time series model depicts the remaining relationship in the process. The data used in this study is call arrival data received by Toronto EMS between January 1, 2007 and December 31, 2008 for which ambulances were dispatched. This analysis was carried out using 2007 data as training data and 2008 as validation data and vice versa.

To estimate the intraday arrival rate model, a thin plate regression splines with a ten dimensional basis, the Poisson family and the log-link functions are used through the GAM function. Thin plate regression splines are low rank isotropic smoothers possessing some beneficial properties like, not needing to decide on the placement of knots and can be applied efficiently for large datasets.[44.] The amount of smoothness for the factors and the loading function are allowed to be automatically estimated by generalized cross validation (GVC). The time series plot of the multiplicative residuals from this factor model, appear to be stationary but reveal some sequential dependence. Time series models for the latent conditional intensity inflation rate (CIIR) process to account for this dependence. A GAM[45] model is considered here with some restrictions and also an integer-GARCH (1, 1) model is applied. If this models sufficiently explains the dependence then and autocorrelation plot of the multiplicative residuals is expected to be statistically independent for all lags. Three nonlinear generalizations are also considered as they may better characterize the sequential dependence; namely; Exponential autoregressive model, piecewise linear threshold model and a model with regime switching at deterministic times.

Out-of-sample comparison was done carried out by fitting models to the 2007 training data and using 2008 as validation and vice versa. A series of models were considered; simple prediction, factor models (FM) without constraints with K= 1... 6, FM with constraints and FM with constraints and smoothing splines and the latter FM with k=4 and the inclusion of the CIIR process with the various time series models. The RMSE and other residual types were considered.

The FM models did slightly worse than the SP models, the FM with constraints was a substantial improvement, while the FM with constraints and smoothing splines also presented extra improvement. Also with the addition of the intGARCH model for the CIIR process to FM=4, the RSME improved slightly again. This model has the best performance for both sets.

In conclusion, it is observed that the factor model estimation with smoothing splines significantly increases forecast performance. This model was able to capture the nonstationary behavior exhibited in call arrivals. Also the introduction of the CIIR process allowed adaptive forecasts of deviations from this diurnal pattern. There are also some limitations to this model; there is no prediction interval for the predictions and also it assumes that the there is no change in pattern between the observed and predicted time frame.

Chapter 3

DATA DESCRIPTION

The data used in this study was provided by a non-profit regional medical center in Wright County, Minnesota that provides care to about 70,000 patients every year.[2] The data consists of daily observations from 2009 January 1st-December 31st 2012, inclusive. The data contains 84,329 patients but only 65,535 observations was be used for analysis and 18,794 observations will be used for validation. Our empirical analysis used the first three years of data to estimate forecasting methods parameters and 2012 data was used to evaluate post-sample forecast accuracy. We will deal with only the test data set for now and include the validation dataset post-analysis.

The variables in our data include:

- Arrival Datetime: time of patient's arrival
- ED Depart Datetime: time of patient's departure from ED
- We use the difference between ED departure and arrival times to compute length of stay in ED
- Hospital Discharge Datetime: time of patient's discharge from hospital (same as ED depart time if patient was not admitted.

Acuity Level: this can be defined as "The measurement of the intensity of care required for a patient accomplished by a registered nurse"[15].This plays a major role in determining how much nursing care a patient needs. The levels are:[13]

1. Resuscitation: This group of patients requires immediate lifesaving intervention or are in an unresponsive state.
2. Emergent: The patients in this category are in a high risk condition and might be confused, lethargic, disorientated in distress or in severe pain
3. Urgent: Patients in a high risk situation but with stable vitals. This group requires several resources like , I.V, lab tests , X-rays etc
4. Semi-Urgent: Patients in a stable condition requiring one or two resources
5. Non Urgent: Patients not requiring any resources.

First Assigned Nurse Start Datetime: this is the time when the nurse started attending to the patient (the difference with arrival time gives us the wait time) .

Age at Admit: Age of patient at time of admits.

Gender: Sex of patient.

Inpatient Admit Datetime: time the patient was admitted.

Ready for Discharge Datetime: time the patient was ready to be discharged.

Ready for Inpatient Admit Datetime: time the patient was ready to be admitted.

Roomed Datetime: time the patient was put in a room

Transfer Datetime: time of patient transfer to another facility.

Descriptive Statistics

On average, there were about 58 daily ED visits from January 2009 to December 2012.

In 2009 the mean was 64, 61 in 2010, 56 in 2011 and 52 in 2012; we observe that there is a decline in patient count, from 64 to 52 within four years.

Table 1: Average Daily ED Daily Attendances

YEAR	2009	2010	2011	2012	Overall Mean
January	62	57	60.4	49.3	57.2
Feburary	70.7	59.5	73.9	49.3	63.3
March	62.5	54.9	69.4	48.8	58.9
April	66.4	59.1	61.7	48.3	58.9
May	69.1	63.5	62.7	51.1	61.6
June	62.3	61.9	49.2	55.4	57.2
July	61.2	63.4	53.0	55.5	58.3
August	59.2	63.7	50.4	49.4	55.7
September	61.8	61.9	50.6	52.8	56.8
October	72.7	61.2	49.7	49.2	58.2
November	56.5	58.0	47.9	49.7	53.0
December	56.1	58.9	44.0	58.9	54.5

We graph the total count of patients for each month by year:

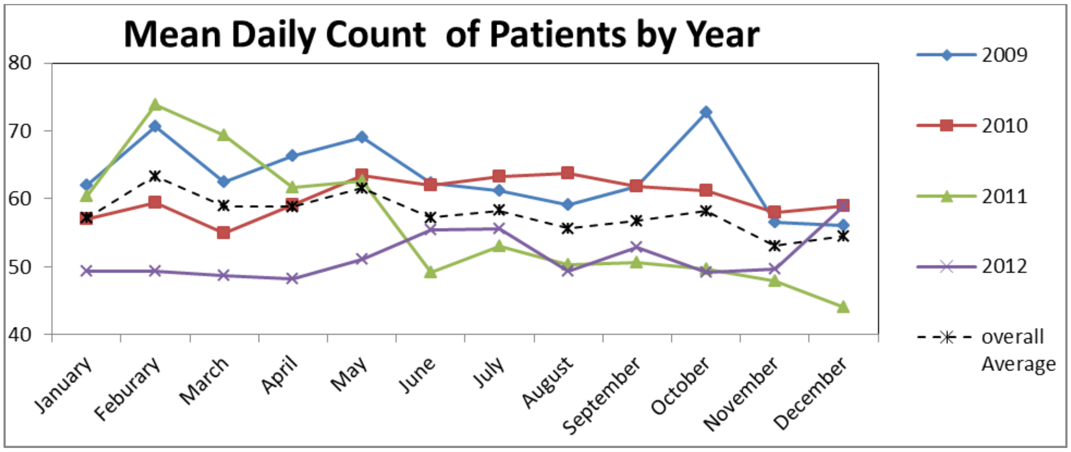

Figure 2: Graph of Mean Daily Count by Year

From the above graph we see that 2009 and 2010 track closely, 2011 tracks closely with the previous years until May but then we notice a decline and this decline continues till 2012. We also observe that the overall mean drops after May. This drop in patient count might be due to certain factors which are beyond the scope of this study. Also we see a similar pattern of behavior of the curves. We can say that our data shows a monthly or seasonal pattern. We also plot the data for each day of the week:

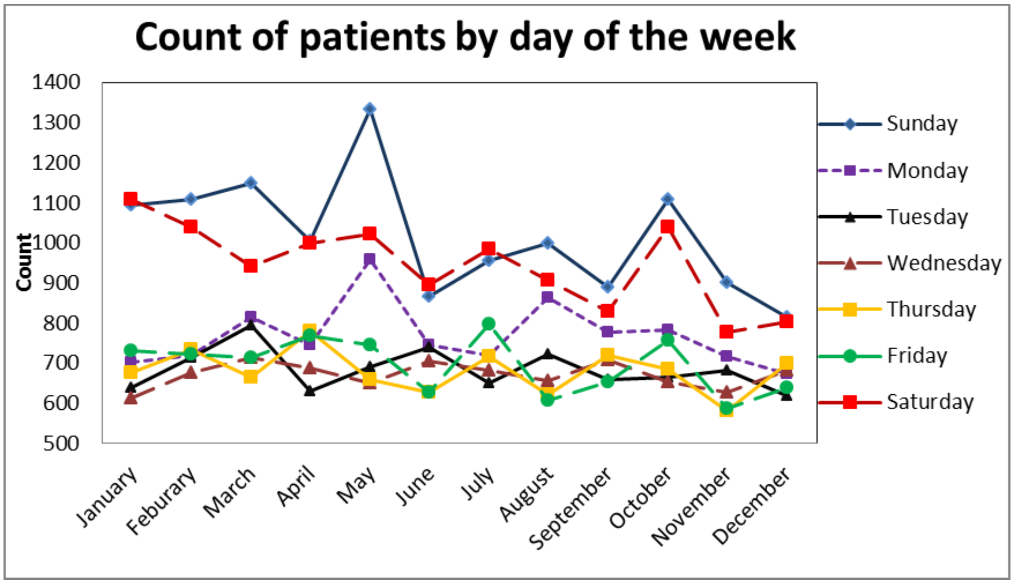

Figure 3: Graph of Patient Count by Day of the Week

Here also we see can identify patterns and trends;

Saturday and Sunday have the highest patient count significantly higher than the week days and this might be due to the fact the hospital is situated in a residential area and most people are home on the weekend as opposed to week days when most residents are away at work in the metro area. Also we see that Monday has a higher volume than the Tuesday, Wednesday, Thursday and Friday. This leads us to assume our data has a weekly pattern. We graph the hour of the day for each day of the week to see if there is any intraday patterns for our data:

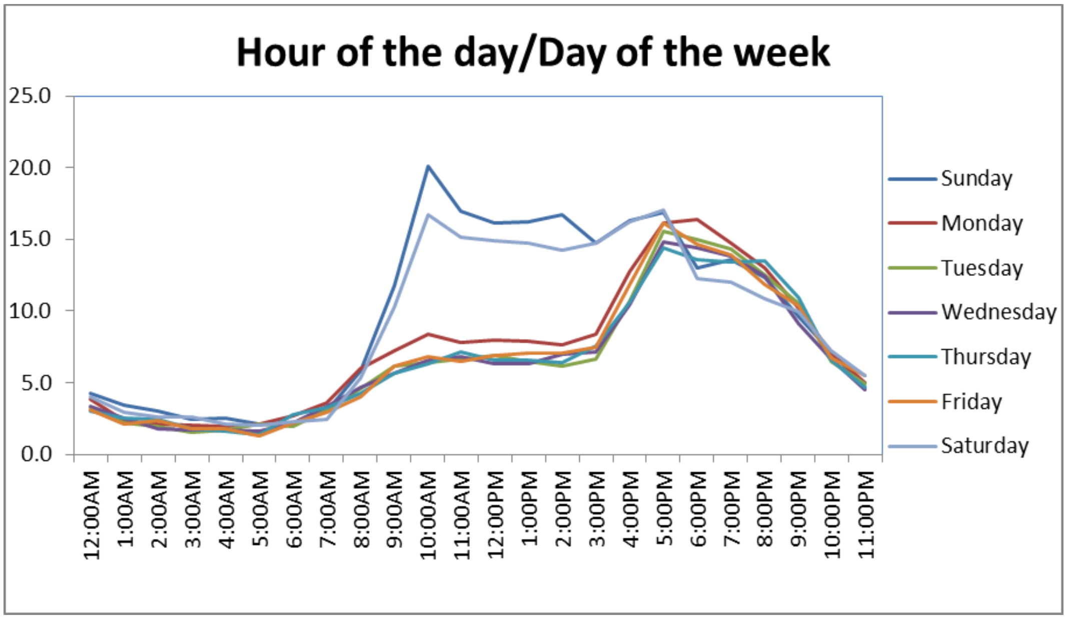

Figure 4: Graph of Mean Patient Count by Hour of Day and Day of the Week

From our graph we see that all days of the week behave similarly from 1 am till 7am, the average patient count within that time is about one. After 8 am on weekends (Saturday and Sunday) we see a spike in patient tally and the average patient count at this time is approximately five patients and the peaks occurs about 10 am and continues till about 5pm where we see notice a slight dip between the hours of 6-9pm mostly on Saturdays apart from this we see a sort of "merge" in pattern, further investigation reveals that the count decreased significantly between 3pm to 11pm in 2011 but the pattern remains the same.

For week days we begin to notice an increase in patient count at 8 am, but here there is an average increase of one patient as opposed to five on the weekend, then at 3pm we see another increase this time with an average of two patients increase. At about 6pm we observe that there is a merge with the weekend data.

We see from our graph that again there is a difference in the weekends and weekdays, also we can assume that our data has an intraday cycle.

Acuity

The proportion of patients based on acuity for 2009-2011 is given in the table below:

Table 2: Acuity Level Proportions

Acuity levels	Proportion
1 (Resuscitation)	0.16%
2 (Emergent)	9.89%
3 (Urgent)	51.93%
4 (Semi-Urgent)	33.87%
5 (Non Urgent)	3.21%
Blank	0.95%

From we table it is observed that 52% of the patients who come to emergency are of level 3 acuity (Urgent) while 34% are of the semi-urgent category , together both groups account for 86% of the patients arriving at the ED, while emergent accounts for 10% , Resuscitation is the least encountered category.

The proportion based on acuity is plotted by hour of the day to observe the distribution.

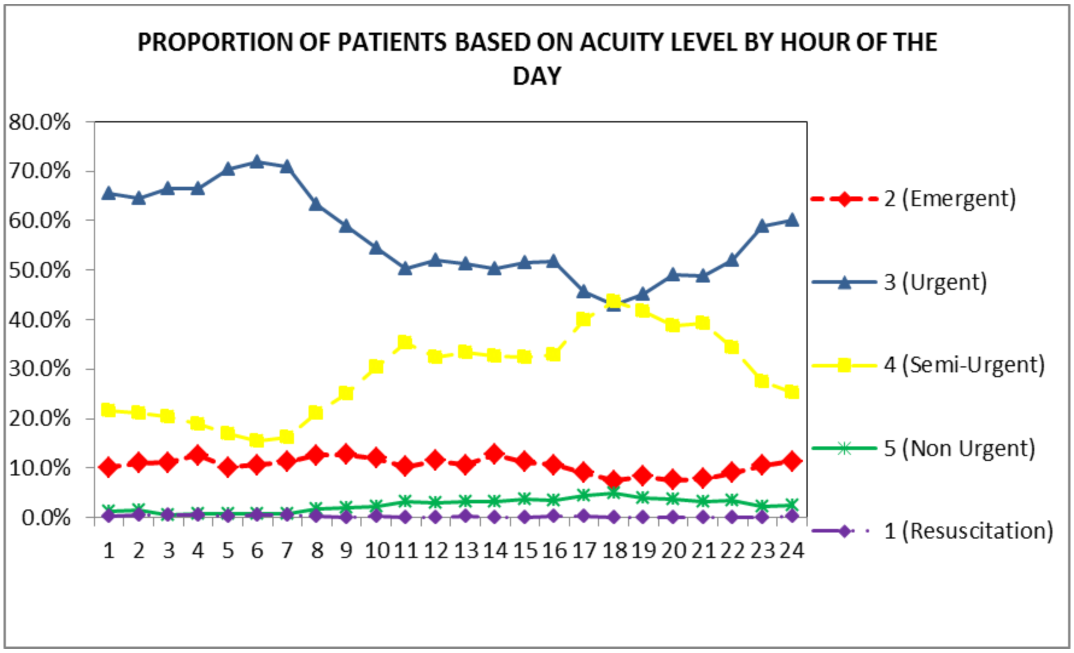

Figure 5: Proportion of Patients Based on Acuity Level by Hour of the Day

It is seen here again that the largest proportion of patients are urgent and semi urgent, with urgent being at 70% at midnight and reduce gradually to about 50% at 11 am, drops to 40% at 6pm and gradually rises again. The semi urgent patients on the other hand; at midnight the proportion for this group is about 20%, this drops a little at 6 am and gradually begins to rise to 40% at 11 am , is steady till 4 pm, peaks at 6pm the begins to decline again. This implies that patients with more severe illness come in at night while those whose symptoms are not as severe prefer to come in during the day. All the other acuity levels are steady throughout the day with Emergent at about 10%, non-urgent and resuscitation is about 5% and this is similar to what was obtained by Sun et al.[36]

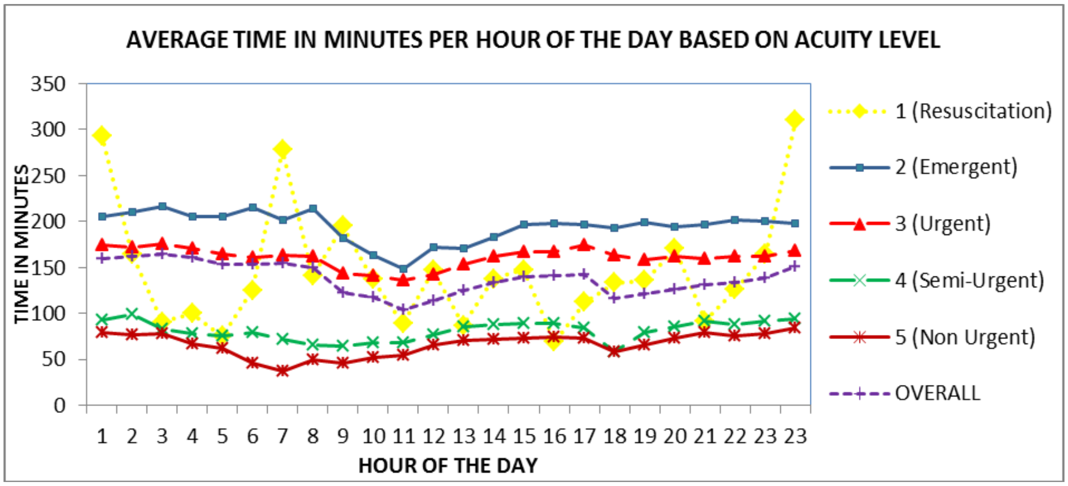

Figure 6: Graph of Mean LOS in Minutes Based on Acuity Level

The overall average length of stay (LOS) in the ED at any given time of the day is 135 minutes, with a standard deviation of 17 minutes. For table 2 we see that the mean LOS for emergent category is 33 minutes more than the LOS of urgent category. Semi-urgent spend 80 minutes less time than the Urgent category.

Table 3: Mean LOS by Acuity Level

Acuity levels	Mean LOS in Minutes
1 (Resuscitation)	152
2 (Emergent)	196
3 (Urgent)	163
4 (Semi-Urgent)	83
5 (Non Urgent)	71
OVERALL	135

Chapter 4

RESEARCH QUESTIONS

The main goal of this study is to attempt to help attain more efficient allocation of human resources in the ED to maximize productivity. This is to be done by forecasting how many nurses are needed to efficiently run the ED at a given time. This is to ensure that there are enough nurses in the department to effectively take care of patients needs and maximize productivity. This study will attempt to answer the following questions:

- Can patient arrival volume be predicted accurately?
- Using the same methods for predicting patient arrival, can cumulative patient volume also be accurately forecasted?
- How much data is required to make the most accurate predictions?
- How accurate will six months predictions be?
- Which method(s) is most suitable for our data?
- Can we predict urgent acuity patient arrival volume?
- What forecast methods can handle multi seasonality?
- If there is a trend (steady decline or increase) in the data which forecasts method will most successfully capture it?
- How easily can these methods be implemented in the ED?

The following forecasting methods will be used to build models to forecast ED arrival patient volume, cumulative patient volume, simulated data and urgent acuity patient arrival volume. Data collected for 2012 was used for validation and long term forecast of about 180 days is considered. Forecast accuracy will be estimated using the mean square error (MSE), root mean squared error (RMSE) and mean absolute error (MAE).

- Linear regression
- Seasonal auto regression integrated moving average (ARIMA)
- Exponential smoothing methods which include; Holt-Winters exponential smoothing method (HWT) , Box-Cox transform, ARMA errors, Trend and Seasonal components (BATS),proposed by De Livera[9] and TBATS (Trigonometric Box-Cox transform, ARMA errors, Trend and Seasonal components) methods also proposed by De Livera.[9]
- Factor latent structure model.

Chapter 5

HOW MUCH DATA IS NEEDED?

A major factor in determining the accuracy of our data is how much data is needed to build the model. In exponential smoothing more weight is put on the most recent observations but how much of this data is useful in the analysis.

Regression: Here three year data was also more appropriate that using just one year or two years and it also helps stabilize the variance in the data.

Time Series Models: We plot the out of sample root mean square error (RSME) for our three time series models using one month, three months, six months, nine months, twelve months, twenty four and thirty six months to forecast one month ahead (744 observations).

Factor Latent model: For this model we use the data from the average of the three years to build our latent factor model. This is to stabilize the data and reduce the effect of the decline experienced from June 2011. In other words, using only 2011 data had more average than using the three year hourly average for each day,

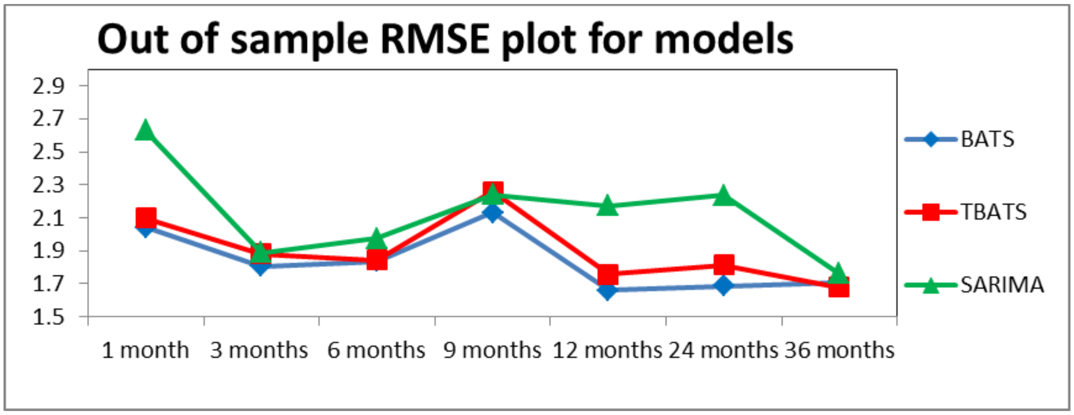

Figure 7: Out of Sample RMSE for Time Series Models for Different Time Periods

From the plot it is observed that using twelve months of data is as effective as using twenty-four months or thirty six for the BATS and TBATS models but for SARIMA three years of data is a better choice, it performs as good as the other models at this point. For our models three years of data was used, except for the dynamic latent factor model where and average of the three year data was utilized.

Chapter 6

METHODS

The different hourly data series that will be analyzed include:

- Patient arrival volume
- Patient cumulative volume
- Simulated data
- Urgent acuity arrival data volume

The methods previously outlined will be evaluated.

SECTION I: PATIENT ARRIVAL COUNT

Regression Model

A regression model tries to model or explain the relationship between a response or dependent variable and one or more predictor or explanatory variables.[12, 31] This relationship might be either associative or causative. The response must be a continuous variable but the predictors can be nominal or continuous. There are several reasons for regression modeling which includes:[31]

- Prediction of future observations (forecasting)
- Assessment of the relationship between explanatory and response variables
- General description of data structure

- Parameter estimation
- Variable selection

Here we are mainly concerned in using regression for forecasting.

The basic form of a regression equation is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + e$$

The parameters $\beta_0, \beta_1, \beta_2, ..., \beta_k$ are called regression coefficients with β_0 known as the intercept e accounts for the variation in y not explained by the $x's$. The error terms are assumed to be independent and identically distributed. The betas measure the effect of each of each covariate, after taking into account all other covariates in the model [26.] The best estimates of beta are the ones which minimizes the sum of the squared errors, this implies we find the values of betas that minimize; [31]

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1,i} - ... - \beta_k x_{k,i})^2.$$

Fitting the Regression Model

The covariates or explanatory variables used for fitting a regression model are categorical variables for, hour of the day, day of the week and month of the year. For day of the week variables, Wednesday is the reference category while for hour of the day 12.00 am is he reference Category and for Month of the year December is the reference category.

There are 40 explanatory variables in total, with 1 or 0 values and using 40 degrees of freedom.

The corresponding regression equation can be seen in appendix(Site reference here).

After fitting the regression model we plot the residuals to check if the conditions for regression are satisfied:

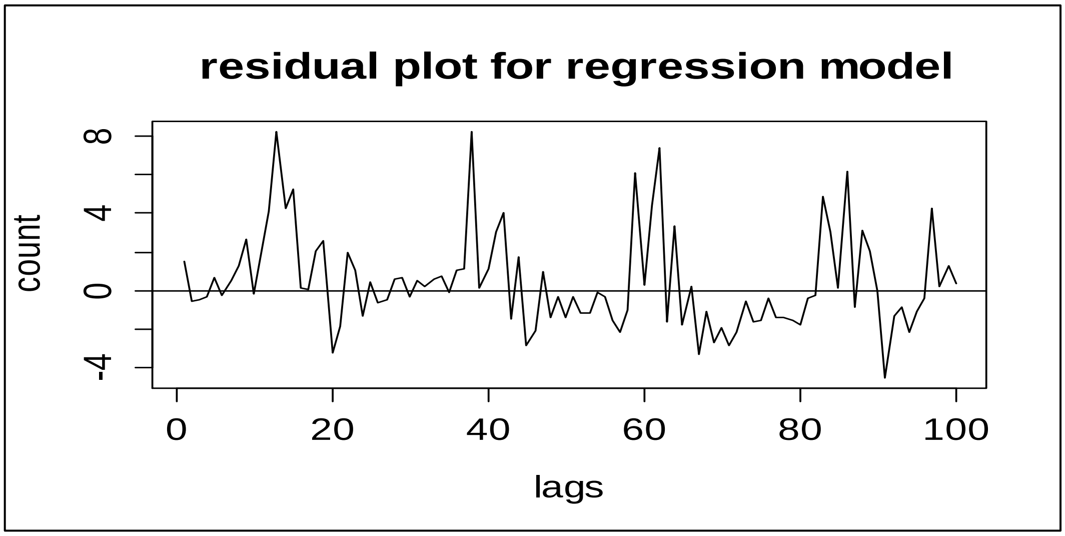

Figure 8a: Regression Residual Plot for Patient Arrival Count Data

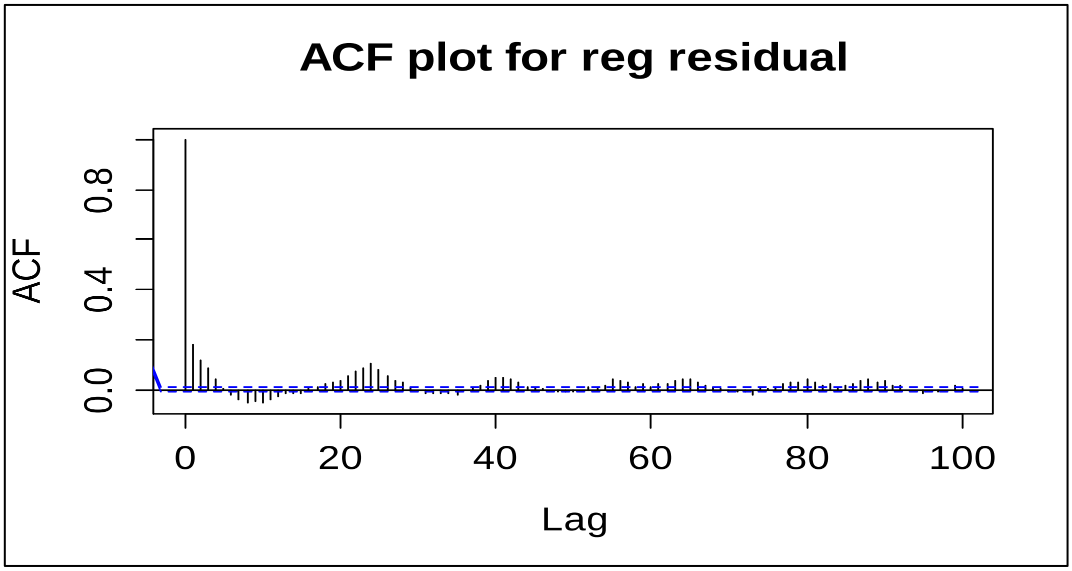

Figure 8b: Regression Residual ACF Plot for Patient Arrival Count Data

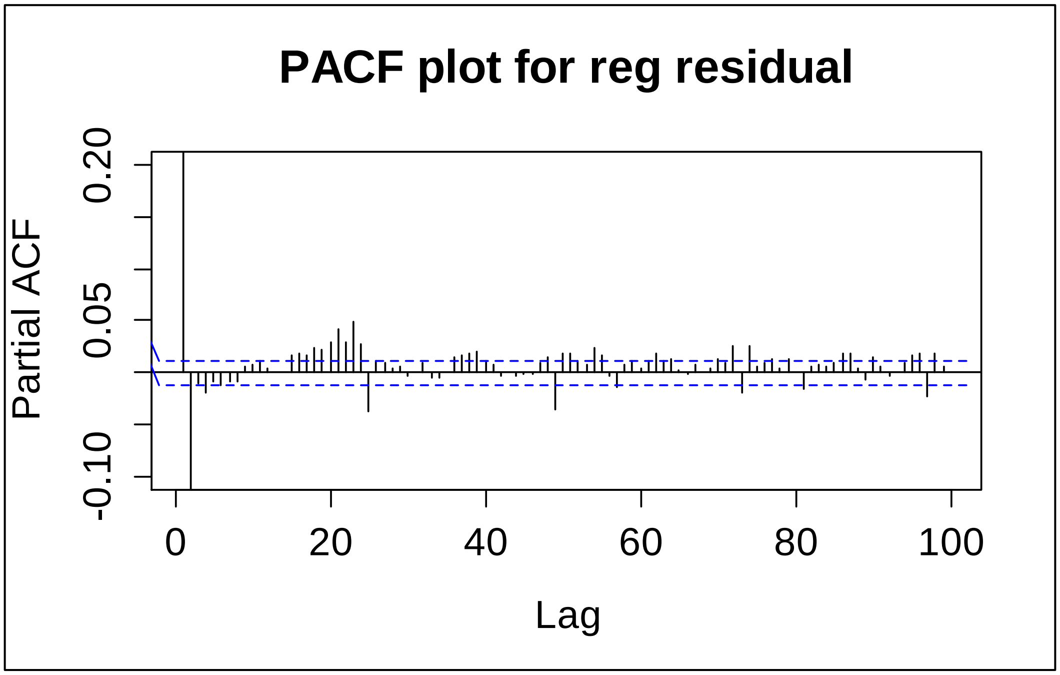

Figure 8c: Regression Residual PACF Plot for Patient Arrival Count Data

From the ACF and PACF plot of the residuals, it can be deduced that there is still remaining serial dependence after the regression has explained 72.3% variation between patient count and the covariates. a*uto.arima* function is applied to the residuals to model the remaining relationship, and then the residual is forecasted and added to the regression prediction.

The ARIMA model used to model the residuals is:

```
Series: arrival regression residuals
ARIMA(2,0,2) with non-zero mean

Coefficients:
ar1       ar2       ma1       ma2  intercept
1.3941  -0.5040  -0.5199  -0.0741     0.1112
s.e.  0.0622   0.0433   0.0626   0.0145     0.0517

sigma^2 estimated as 5.156:  log likelihood=-58843.05
AIC=117698.1   AICc=117698.1   BIC=117747.2
```

This means the non-seasonal ARMA model has the following coefficients; MA (2) and AR (2) with zero differencing and non-zero mean, the AR coefficients are 1.39 and -.50 and the MA coefficients are -0.52 and -0.74. This model is selected based on AIC.

Time Series Methods

We plot the first our data as a time series:

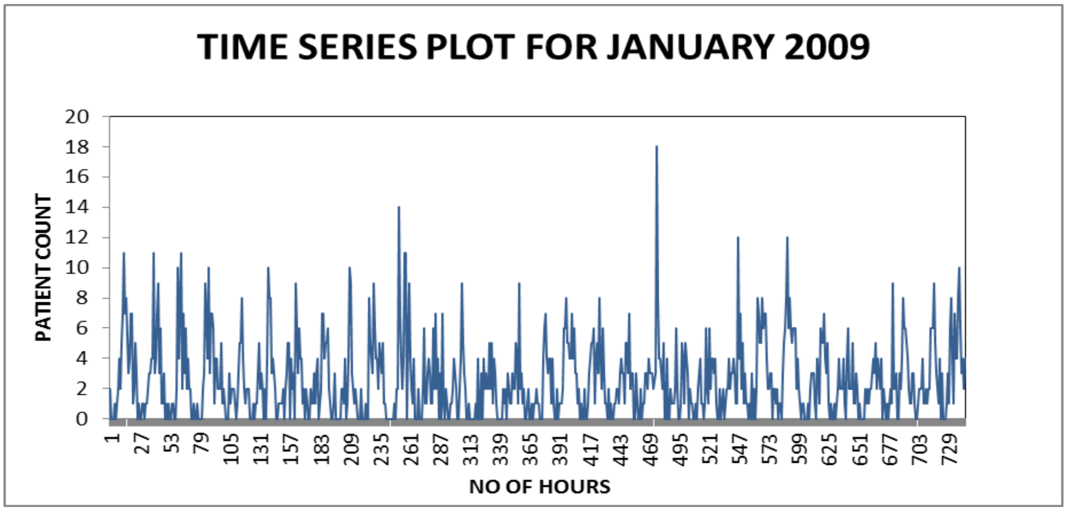

Figure 9a: Time Series Plot for Patient Arrival Data for January 2009

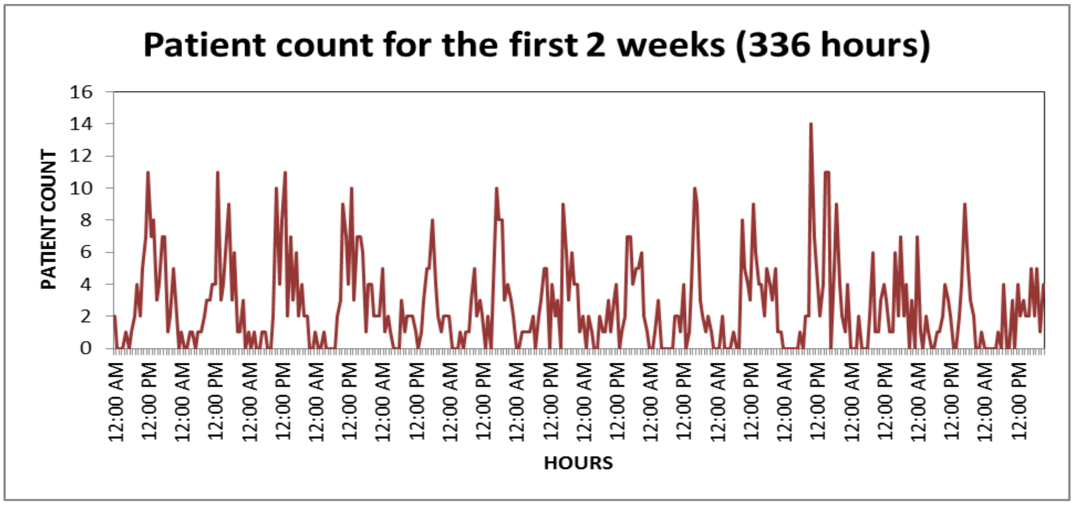

Figure 9b: Time Series Plot for Patient Arrival Data for First Two Weeks of 2009

From out time series plot of January (744 hours), it can be seen that our time series exhibits multiple seasonal patterns. These multiple seasonality is more visible in Figure 6b, plot for the first two weeks of January 2009. Intraday and weekly cycles are observed from the plots. These cycles are not uniform(Figure 3), Saturday and Sunday have a similar pattern, Monday tracks closely while the rest of the weekdays exhibit a similar pattern. The underlying levels of the daily patterns also vary from week to week but are highly correlated with the levels of the days immediately preceding. An effective model for this data must take into account this features without being too complicated *msts*.

The *msts* command in the forecast package[17] in R is used to plot our data so as to capture the multi-seasonality feature. This command develops from the popular *ts* class but it has an added feature which contains the vector of seasonal periods. All procedures that work on the *ts* class also work on this class.[17]

Also we plot the ACF and PACF graphs for our data:

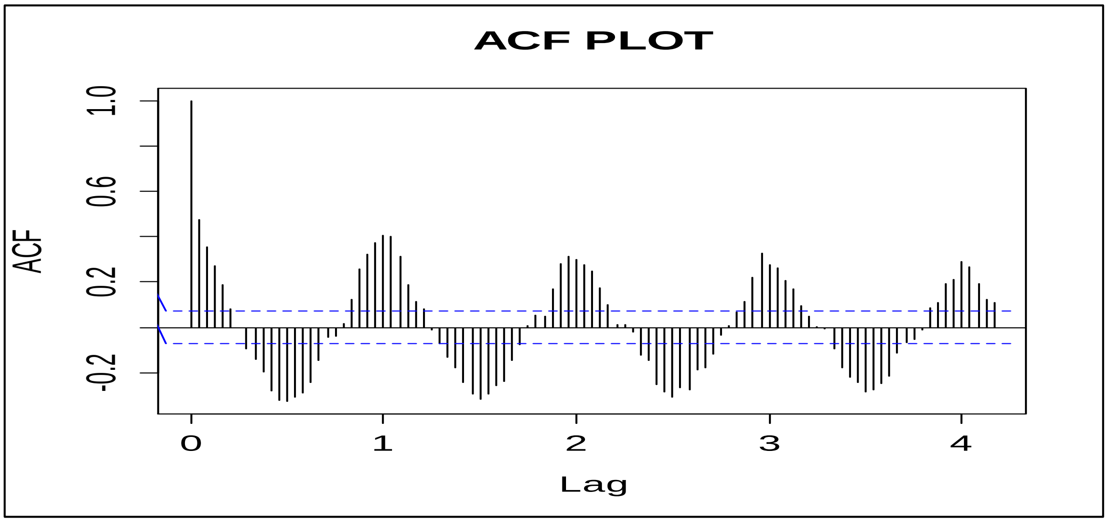

Figure 10a: ACF Plot for Patient Arrival Count Data

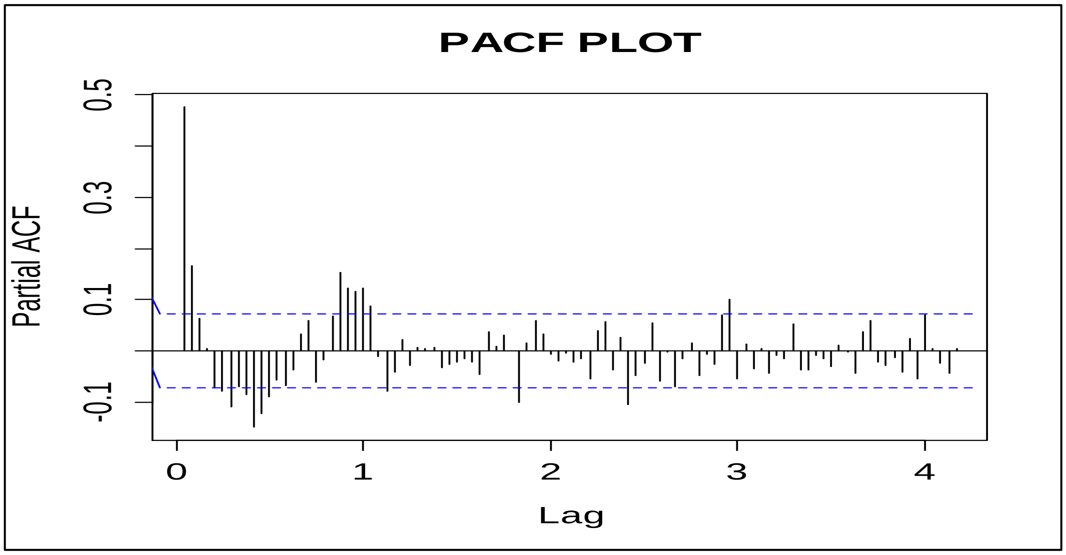

Figure 10b: PACF Plot for Patient Arrival Count Data

This multi seasonal time series will now be used to build our models.

Seasonal Autoregressive Moving Average
(Sarima) Models

The general form of the multiplicative seasonal ARIMA model can be written as (see Box et al.[3] page 333):

$$\phi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^D x_t = \theta_q(B)\Theta_Q(B^s)w_t + \delta$$

x_t Is the time series observation

Where B is the backshift operator; that is; $B^j X_t = X_{t-j},\ \ j = 0, \pm1, \pm2, \ldots$

$\theta(B)$ Is a moving average (MA) operator of the form: $1 + \theta_1 B + \ldots + \theta_q B^q$,

An autoregressive (AR) polynomial X_t is of the form $\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p \quad (\phi_p \neq 0)$.

Then the AR process can be written as $\phi(B)X_t = W_t$.

Where W_t is white noise that follows the normal distribution with mean 0 and variance σ^2. This can be written as: $\{W_t\} \sim WN(0, \sigma^2)$

Where the resulting multiplicative process will be said to be of order $(p,d,q)\times(P,D,Q)_s$. The ordinary or non-seasonal autoregressive and moving average polynomials are represented by $\phi(B)$ and $\theta(B)$ of order p and q respectively (also see Shumway &Scoffer [37] page 157) and the seasonal auto-regressive and moving average components by $\Phi_P(B^S)$ and $\Theta_Q(B^s)$ of orders P and Q, and ordinary and seasonal difference components by $\nabla^d = (1-B)^d$ and $\nabla_s^D = (1-B^S)^D$.

For preliminary analysis the data is fit as a time series with a frequency of 24 for each day. *auto.arima* function is used to fit an AR model a MA model and an ARMA model. The "best" models chosen are then used to predict up to one hour ahead.

The chosen model for the AR model is:

```
Series: Patient.arr.ct
ARIMA(47,0,0) with non-zero mean
```

For the MA model it is:

```
Series: patient.arr.ct
ARIMA(0,0,10) with non-zero mean
```

For the ARMA model is:

```
Series: patient.arr.ct
ARIMA(5,0,1) with non-zero mean
```

The plot for the forecasts and the actual plots is given below:

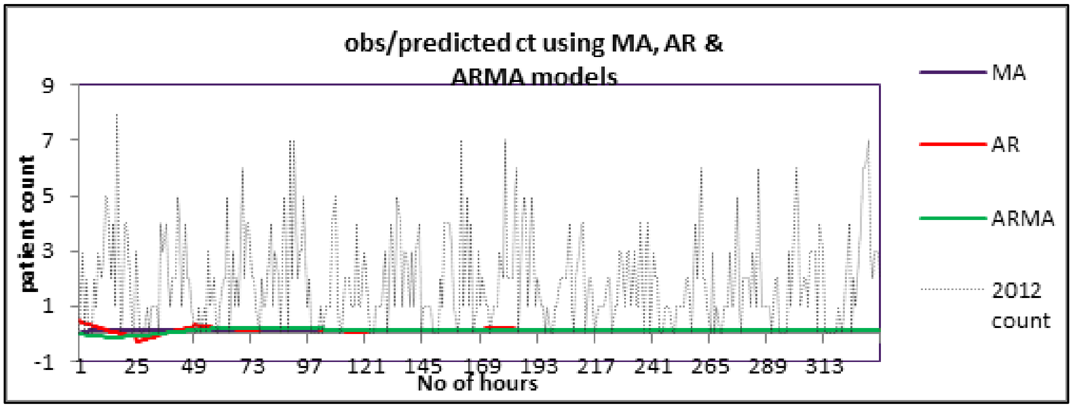

Figure 11: Graph of AR, MA and ARMA Model Predictions/Actual 2012 Count

From this plot it can be observed that these models are not able to predict our data accurately. The predictions are flat around zero. Jones et al.[22] stated that ARIMA model performed worse than the linear regression model.

Again we use the *msts* function to fit a time series as earlier described, *auto.arima* function in R (in the forecast package) is then applied. This function generates the best ARIMA model using multiple model selection criteria, and it also accommodates covariates.[17]

We fit two models, the first without covariates and the second we include the covariates as we he selected model is then used for our forecast.

Sarima Model Result

The result for the SARIMA model selected by the *auto.arima* command is :

```
Series: Patient arrival count
ARIMA(3,1,3)(0,0,2)[24] with drift

sigma^2 estimated as 3.703:  log likelihood=-54493.44
AIC=109006.9   AICc=109006.9   BIC=109088.6
```

This means the non-seasonal ARMA model has the following order; MA (3) and AR (3) with one differencing and drift, the seasonal ARMA had only MA(2) with seasonal lag of 24 (one day). The coefficients for the non-seasonal AR models are 1.53, -0.48 and-0.498 while for the MA models are, -0.113, -2.25 and 1.6094.For the seasonal MA model the coefficients are 0.169 and 0.090 respectively, this model was selected based on AIC.

The result for the selection including the covariates (hour of day, day of week and month of year) by the *auto.arima* command is:

```
Series: Patient arrival count
ARIMA(2,1,2)(0,0,1)[24] with drift

sigma^2 estimated as 3.086:  log likelihood=-52095.61
AIC=104283.2   AICc=104283.4   BIC=104659.3
```

This means the non-seasonal ARMA model has the following order; MA (2) and AR (2) with one differencing and drift, the seasonal ARMA had only MA(2) with seasonal lag of 24 (one day). For the non-seasonal components, The AR coefficients are 0.467 and 0.027 while for the MA they are -1.31 and 0.316 respectively. For the seasonal MA the coefficient is 0.0687 and the drift is 0.0001. This model was selected based on AIC.

These results are then used to forecast six months ahead to see how they would perform for long term predictions.

From the results, it was observed that including covariates in the latter SARIMA model is a significant improvement from the former ARIMA model.

Exponential Smoothing Methods

Exponential smoothing can be defined as a process for repetitively updating a forecast in light of more recent experience.[23] It assigns exponentially increasing weights to more recent observations. A time series model can be decomposed to three components; trend (T) ,cyclical component(C) ,seasonality (S) and error component[4].This method has been around since the 1950s but a modeling framework applying stochastic models, likelihood calculations, prediction intervals and model selection procedures were not developed until more recently in 1997 and 2002.[26] The state space model makes room for considerable flexibility in the specification of the parametric structure of this method.[4]

A linear innovations state space model can be defined as follows[19]

Let y_t =observation at time t

x_t = state vector

The model can be written as :

$$y_t = w'x_{t-1} + \varepsilon_t, \qquad (1.1a)$$

$$x_t = Fx_{t-1} + g\varepsilon_t, \qquad (1.1b)$$

Where $\{\varepsilon_t\}$ is a white noise series, F , g and w are coefficients. Equation (1.1a) is known as the measurement equation; it describes the relationship between the unobserved states x_{t-1} and the observation y_t . Equation (1.1b) is the transition equation. It describes the state evolution of the states over time. Using the identical errors for both models makes it an “innovation” state space model. These equations are identical to several exponential smoothing methods. One advantage that exponential smoothing models have over ARIMA models is that the trend, cyclical components and seasonality are stated explicitly in exponential smoothing models but this is not seen as easily in the ARIMA models[3]. Another useful attribute of the exponential state space model is that all the model parameters can be selected automatically without any input from the user, they are easily automated.[20]

The Holt Winters method generalizes exponential smoothing method to accommodate trend and seasonal variation.[4] There are two classes of these models: Additive and multiplicative seasonal models. A model can be described as seasonal if it displays characteristics that recurs every S period.[23] The period S is the season length.

An additive model is a model that can be expressed as:

Data=Trend+ Seasonal Effect +cyclical component + Residual

While a multiplicative model can be written as:

Data=Trend X Seasonal Effect X cyclical component X Residual

A multiplicative model can be transformed to an additive model by take the log of the data[23]

For our models we will only be considering at additive models.

The traditional Holt Winters method has been modified to handle a wider variety of seasonal patterns.[9]

The BATS model is one of such modifications. It stands for Box-cox transform, ARMA errors, Trend and Seasonal components. It comprises of the following components;

($\omega, \phi, p, q, m_1, m_2, ..., m_T$) ω indicates the Box-Cox parameter, ϕ is the damping parameter, p and q are the ARMA parameters and the seasonal periods $(m_1, ..., m_T)$.[5]

The HW method can be represented in this form, for example, BATS (1, 1, 0, 0, m_1) represents the underlying model for the traditional Holt-Winters additive single seasonal method. BATS (1, 1, 0, 0, m_1 , m_2) represents the double seasonal Holt-Winters additive seasonal described by Taylor.[39, 40]

In the TBATS model the seasonal component in the BATS model is replaced by the trigonometric seasonal formulation. It can be represented as $(\omega, \phi, p, q, \{m_1, k_1\}, \{m_2, k_2\}, ..., \{m_T, k_T\})$[9.] Due to the feature that it relies on trigonometric function, it can be used to model non-integer seasonal frequencies. Some of the advantages of the TBATS model are that, it allows for the accommodation of nested and non-nested multiple seasonal components; it handles typical nonlinear features that are often seen in real time series. Also, it accommodates any autocorrelation in the residuals.

Fitting the BATS And TBATS Models

Also like in previous methods we use 2009-2011 data to fit our model and 2012 data for validation.

BATS output.

```
BATS(0.003, {1,1}, 0.999, {24,168})
```

```
Call: bats(y = patient arrival count)
```

```
Parameters
  Lambda: 0.003109
  Alpha: 0.008765254
  Beta: -4.83746e-06
  Damping Parameter: 0.998953
  Gamma Values: 0.01069493 4.239887e-05
  AR coefficients: -0.050986
  MA coefficients: 0.091149
```

Lambda represents the Box-Cox transform which is 0.003 in this case and the smoothing parameters are alpha, beta and gamma which are 0.0088, -0.0000048, 0.011 and 0.000042 respectively .The damping parameter is 0.999 while the ARMA order is AR (1) and MA (1) with coefficients -0.051 and 0.091 respectively and finally the seasonal periods are 24 representing daily cycle and 168 representing weekly cycles, with 196 estimated parameters.

<u>TBATS output</u>.

```
TBATS(0.001, {4,3}, -, {<24,6>, <168,6>})
Call: tbats(y = patient arrival count)
Parameters
  Lambda: 0.000971
  Alpha: 0.00572104
  Gamma-1 Values: 6.403008e-07 1.90908e-06
  Gamma-2 Values: -8.633531e-06 6.324464e-07
  AR coefficients: 0.089391 -0.095853 0.031388 0.012412
  MA coefficients: -0.005474 0.126531 0.010637
```

Lambda represents the Box-Cox transform which is 0.001 in this case and the smoothing parameters are alpha and gamma values. The is no damping parameter in this model while the ARMA order is AR (4) and MA (3) with coefficients seen in the output above and finally the seasonal periods are 24 representing daily cycle and 168 representing weekly cycles, with 32 estimated parameters.

Dynamic Latent Factor Model by Matteson

There are a large number of people who can come into the emergency department at any time and each one of them as a low probability of doing so. Another observation made from the patient arrival volume is that it varies with time of the day, thus it is nonstationary. It also exhibits a seasonal pattern; it varies over weeks and months. The *Palm-Khintchine* theorem states that the arrival process that arises from a large number of independent sources, where no source contributes too much to the arrivals, is approximately a Poisson process,[7, 16] based on these we assume that the patient arrival volume has a Poisson distribution. An extension of the *Palm- Khintchine* theorem is that the suitable model for arrivals in a nonhomogeneous Poisson process (NHPP).[16]

NONHOMOGENEOUS POISSON PROCESS (NHPP)*:

A counting process $\{Y(t): t \geq 0\}$ { is said to be a

Nonhomogeneous Poisson process with intensity function $\lambda(t)$, t ≥ 0 if

i. $Y(0) = 0.$

ii. For each $t \geq 0$, $Y(t)$ has a Poisson distribution with mean

$m(t) = \int_0^t \lambda(s)ds.$

iii. For each $0 \leq t_1 < t_2 < ... < t_m$, $Y(t_1), Y(t_2) - Y(t_1), ..., Y(t_m) - Y(t_{m-1})$ are independent random variables.

Several studies have been carried out based on this assumption for modeling call center arrival rates which has similar underlying assumptions as our data.[8, 16, 27, 34] Matteson[27] proposed a model which is based on the assumption that the data has a Poisson distribution and accommodates low counts which is characteristic of our data. This model avoids use of variance stabilizing transformations. It assumes that the intensity function is a random process and that it can be forecast using previous observations. This interpretation is similar to a Cox process. A Cox process is a Poisson process with a stochastic intensity and can be referred to as a doubly stochastic Poisson process[8]. The main difference here is that while in a Cox process the random intensity depends mainly on its own history here it also depends on previous observations. The random intensity function is partitioned into stationary and nonstaionary components. We would use this method to model our data.

Notation

Our data is collected hourly, and so we assume (following the method proposed by Matteson[27]) that the latent call intensity function for these periods can be approximated to be constant, and our data was collected sequentially in time. We suppose total patient arrival follow a nonhomogeneous counting process $\{Y_t : t \in Z\}$ with discrete time index t . Underlying this is a latent, real-values nonnegative process $\{\lambda_t : t \in \mathrm{Z}\}$. It is further assumed that conditional on λ_t , Y_t follows a Poisson distribution with mean λ_t .

As seen in *figure4* the pattern of patient arrival in a given day has a distinct pattern even though the weekdays are closely similar. They considered an arrival process that has been repeatedly observed in a 24 hour time period (one day). Let $\{y_t : t = 1,...,n\} \equiv \{y_t : i = 1,...,d; j = 1,...,m\}$ denote the sequence of call arrival counts, observed over time period t denote the sequence of patient arrival counts, observed over time period t , which corresponds one-to-one with the j th sub-period of the i th day, so that $n = dm$. Their basic idea here was to model the arrival intensity λt for each unique day using some smooth curves.

λ_t Is defined as the conditional expectation of Y_t given F_{t-1} and X where X is covariate information for each model (calendar information; day-of-week and week-of year were used here) represented by $X = \{x_1,..., x_n\}$ and F_t is a $\sigma -$ field generated by $Y_1,...,Y_t$. Let $\mu_t = E(Y_t \mid X) > 0$ denote the conditional mean of Y_t given only the covariates. Let

$$(1) \qquad \lambda_t = E(Y_t \mid F_{t-1}, X) = \mu_t E(Y_t / \mu_t \mid F_{t-1}, X) = \mu_t \eta_t,$$

In which $\eta_t > 0$ is referred to as the conditional intensity inflation rate (CIIR). By construction,

$$E(\eta_t \mid X) = E(E(Y_t \mid F_{t-1}, X) \mid X) / \mu_t =_t E(Y_t / X) / \mu_t = 1.$$

The CIIR was proposed to model any remaining serial dependence in patient arrival for available covariates. This serial dependence could be due to various factors that may or may not be measureable.

MODELLING: A dynamic latent factor model with integer valued time series is combined with covariates. These covariates are introduced through simple constraints on the factor loadings. Smoothing splines is applied to estimate the model because it forces smooth evolution in the factor levels and loadings.

The factor model provides a parsimonious representation of the nonstationary pattern in intraday calls arrivals, while the time series models capture the remaining serial dependence in the patient arrival process.

DYNAMIC LATENT FACTOR: Assume m consecutive observations per day are available for d consecutive days with no omissions in the record. Let $Y = (y_{ij})$ denote the $d \times m$ matrix of observed counts for each day i over each sub- K period j . Let $\mu_{ij} = E(Y_{ij|} \mid X)$ and $M = \mu_{ij}$ denote the corresponding $d \times m$ latent intensity matrix. A K-factor model is introduced to reduce the dimension of the intensity Matrix M.

They assumed that the intraday pattern of expected patient arrivals on the log scale can be well approximated by a linear combination of (a small number) K factors or functions, denoted by f_k for $k = 1..., K$. The factors are orthogonal length- m vectors.

The intraday arrival rate model μ_i over a particular day i is given by

(2) $$\log \mu_i = L_{i1} f_1 + ... + L_{iK} f_k$$

When is much smaller than either m or d , the dimensionality of the general problem is greatly reduced. K is determined manually.

In matrix form we have

(3) $\log M = LF^T$,

in which $F = (f_1, ..., f_K)$ denotes the $m \times k$ matrix of underlying factors and L denotes the corresponding $d \times K$ matrix of factor loadings, both of which are assumed to be full column rank.

Since neither F nor L are observable, the expression (3) is not identifiable. We further require $F^T F = I$ to alleviate this ambiguity and we iteratively estimate F and L .

To further reduce dimensionality substantially, constraints are imposed with certain conditions (see paper) on the factor loading matrix L .

The constraints considered by Matteson include auxiliary information about the rows and columns of the observations Y to simplify estimation and improve out-of-sample predictions. The day-of-week and week-of-year effects are incorporated into the factor loadings by specifying appropriate constraints.

Another major assumption considered by the authors is that the nonstaionary intensity process μ_{ij} varies smoothly over the hours j of each day i . To include this smoothness into the model, Generalized Additive models (GAMs) is used in the estimation of the common factors f_k . GAMs are generalized linear models with the linear predictor partly dependent, linearly on some unknown smooth functions.[39]

Matteson recommended the use of the *gam* function in the *mgcv* library.[38]

To estimate model (2) using the *gam* function, thin plate regression splines with ten-dimensional basis, Poisson family, and the log-link functions were used. These spines are a low rank, isotropic smoother with any desirable properties.(see wood 2006). The degree of smoothness for the factors f_k and the loadings functions are automatically estimated by generalized cross validation (GVC).

<u>Adaptive Forecasting with Time Series Models</u>

Let $\hat{e}_t = Y_t / \hat{\mu}_t$ denote the multiplicative residual in period t implied by the fitted values $\hat{\mu}_t$ from a factor model as earlier described. Time series plots of the residuals even though sees stationary, reveals some serial dependence. A time series model is considered for the latent CIIR process $\eta_t = E(Y_t / \mu_t \mid F_{t-1}, X)$ to explain this dependence. We look at the ACF and PACF plots for $\hat{e}_t$.

To depict the series dependence a generalized autoregressive linear model, defined by recursion

(4) $$\eta_t = \omega + \alpha \hat{e}_{t-1} + \beta \eta_{t-1}.$$

To ensure positivity certain restrictions are employed; $\omega > 0$, $\alpha, \beta \geq 0$ and $\alpha + \beta < 1$(to guarantee stationarity of η_t).

The resulting model for Y_t when μ_t is constant is an integer-GARCH (1, 1) model. When $\hat{\mu}_t$ is a nonstationary process, the conditional intensity

$$\lambda_t = \mu_t \eta_t$$

is also nonstationary. This η_t is the stationary multiplicative deviation or inflation rate, between λ_t and μ_t. Let

$$\hat{\varepsilon}_t = Y_t / \hat{\lambda}_t$$

represent the multiplicative standardized residual process given an estimated CIIR process $\hat{\eta}_t$ the model defined by (4) adequately accounts for the observed linear dependence in $\hat{e}_t$, then the autocorrelation plot of $\hat{\varepsilon}_t$ should not be statistically significant.

Fitting the Dynamic Latent Factor Models

The data from 2009-2011 was used to fit the model, and use 2012 data for validation. The average for the three years was used after the day of the week was aligned to fit the corresponding covariates. A factor loadings model with constraints and smoothing splines with $K = 3$ was applied after using multiplicative root mean square error (RMSE) to determine the best fit for K in our models. We also added the CIIR process through time series models earlier defined. From our ACF and PACF plots we do not expect significant improvement from the CIIR process because the serial dependence in the residuals after fitting the factor model appears weak.

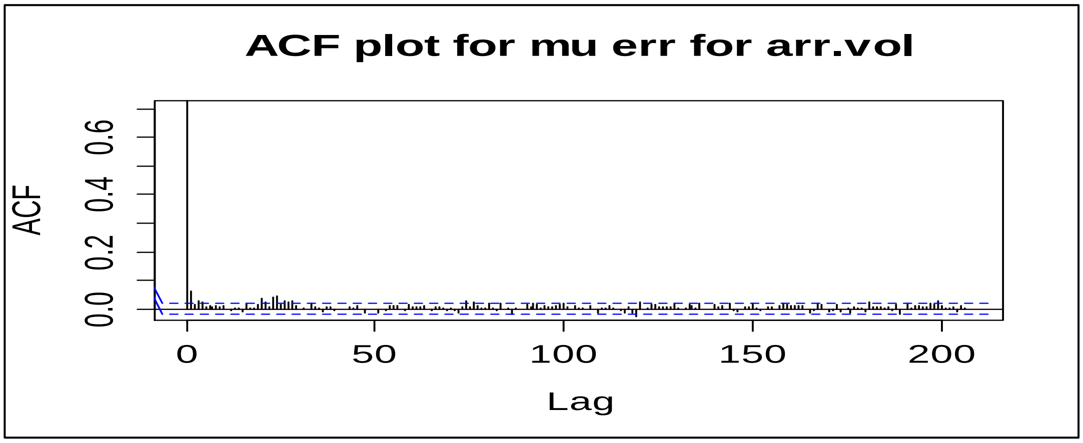

Figure 12a: ACF Plot for Factor Model Residuals for Patient Arrival Data

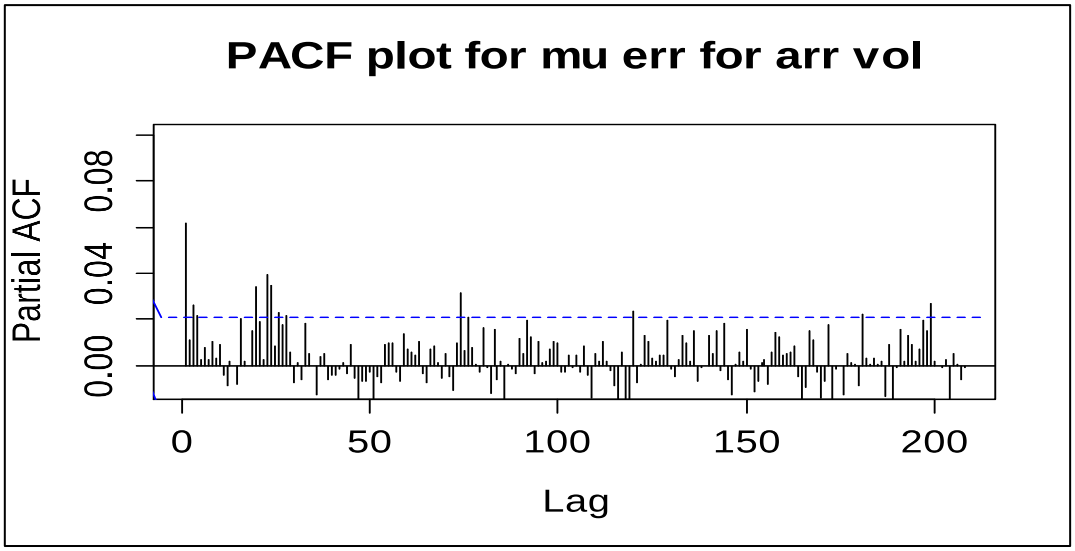

Figure 12b: PACF Plot for Factor Model Residuals for Patient Arrival Data

Results

Forecasting evaluation. There are several methods that can be used to assess the performance of our forecasting models.[18] Some of these basic methods include; mean absolute error (MAE), mean absolute percentage error (MAPE), mean square error (MSE), root mean square error (RMSE). Error is calculated by subtracting the forecasted values from the observed value, for each observation. The mean absolute error involves finding the absolute value of the errors, summing them all up and dividing by n (sample size).The mean square error involves squaring all the errors, then summing them up, finding the mean, while the RMSE involves taking the root of the MSE. The mean absolute percentage error is calculated by dividing the absolute error by the observed value for each observation, summing them up, dividing by n and multiplying by 100 to get a percent value.

The RMS and the MSE are the most commonly used of these methods[18] due to their relevance in statistical modeling. The RMSE method is on the same scale as the data so it is more preferable to the MSE but they are both more sensitive to outliers than the MAE. Another drawback *to the RMSE and MSE is that they increase as the variance associated with the frequency distribution of error in the model increases.*[42] This occurs mainly when the errors are greater than one, the reverse is the case when the errors are less than one. The mean absolute percent is calculated by dividing the absolute error by the observed value, finding the mean and multiplying by 100. The major drawback for this method is that when the observed y_i is zero this then this calculation is undefined. There are some zeros in our data and so this method is unsuitable for our data set,

Let $e_i = y_{i(observed)} - y_{i(predicted)}$

Then $MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i|$

Also $RME = \frac{1}{n}\sum_{i=1}^{n} e_i^2$, $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}$, and

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|e_i|}{y_i}\times 100$$

The out of sample forecast accuracy for the first six months of 2012 data (4368) observations is calculated using MAE, MSE and RMSE and the results are given in Table 4 below.

Table 4: Patient Arrival Count Forecast Evaluation Results

METHODS	BATS	TBATS	SARIMA	SARIMA+REG	L.REGRESSION	Factor	FACTOR MODEL+CIIR
MAE	1.16	1.15	1.60	1.21	1.33	1.33	1.33
MSE	2.41	2.36	3.67	2.50	2.88	3.05	3.04
RMSE	1.55	1.53	1.92	1.58	1.70	1.75	1.74

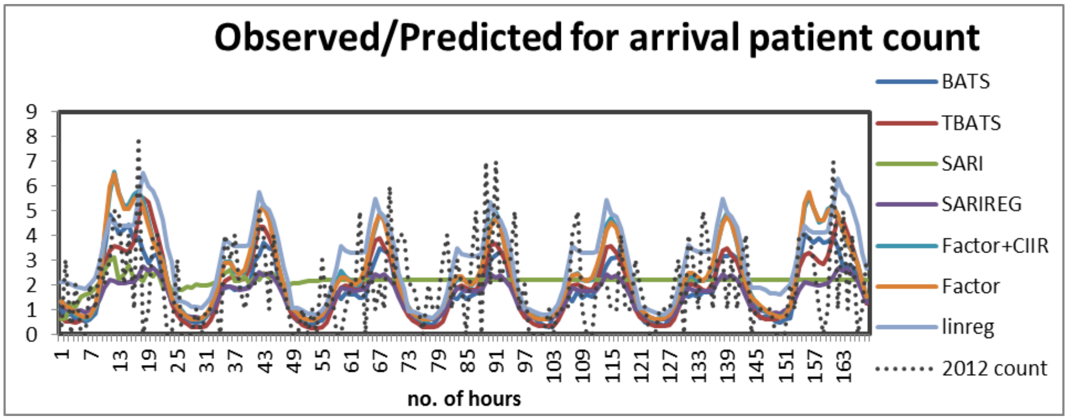

Figure 13: Graph of Predicted/Actual 2012 Patient Counts for Patient Arrival Data

Conclusion

We see from the Table 4 that the TBATS model has the smallest of all three matrices and the SARIMA model without covariates performed the worst. The BATS model performed second to the TBATS model. We also see that adding covariates to the SARIMA model improved its performance quite significantly. Most of our methods have the ability to produce reliable long term forecasts (up to one year ahead), which is needed for capacity planning. Surprisingly linear regression performed better than the SARIMA model with covariates. For our latent factor models with constraints and smoothing splines, it is observed that they are also same and this is expected because the serial dependence in the error after fitting the model is not significant. A major drawback for this model is that it does not produce confidence intervals or prediction intervals by default while the other models are capable of doing so.

SECTION II: CUMULATIVE PATIENT COUNT

Previously only the patient arrival count was considered now we would be looking at the cumulative patient count for each hour. This implies that if a patient came in at 12.35am and was discharged at 4.25 am, he would be counted for 1.am, 2.am 3.am and 4.am because he was in the ED at these times. We would be applying same methods to see if we would get similar results.

To achieve this from our arrival data a variable was created called length of stay in minutes, this is the duration of the patient’s stay in the ED. If the Length of stay is less than 60 minutes the patient is only counted for one time period which is the arrival hour, if a patient stays longer then they are counted for every hour present in the ED. The main drawback of this method is this, suppose a patient comes in at 11.38am and leaves at 12.20pm, this patient would only be counted for 11am and not for 12 noon. The system in the hospital is able to successfully generate the cumulative data but this was not available at the time of this study.

Regression Model

The covariates or explanatory variables used for fitting the regression model for the cumulative patient count are the same as used for patient arrival count. They are categorical variables for, hour of the day, day of the week and month of the year. Again for day of the week variables, Wednesday is the reference category while for hour of the day 12 midnight is the reference Category and for Month of the year December is the reference category.

There are 40 explanatory variables in total, with 1 or 0 values and using 40 degrees of freedom.

After fitting the regression model we plot the residuals to check if the conditions for regression are satisfied.

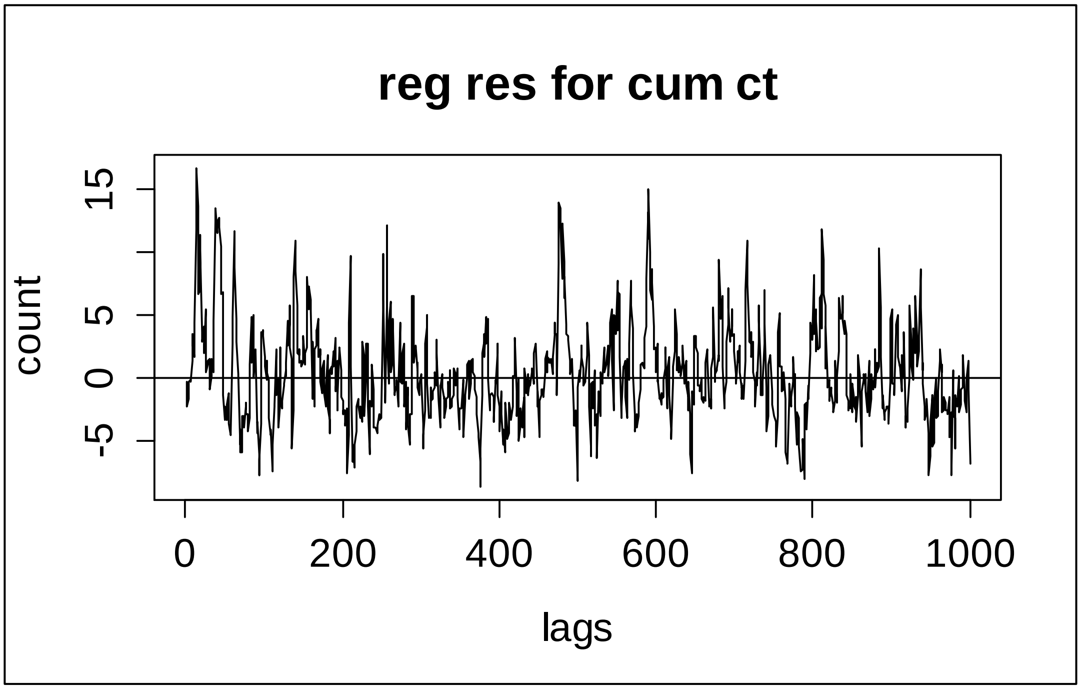

Figure 14a: Regression Residual Plot for Cumulative Patient Data

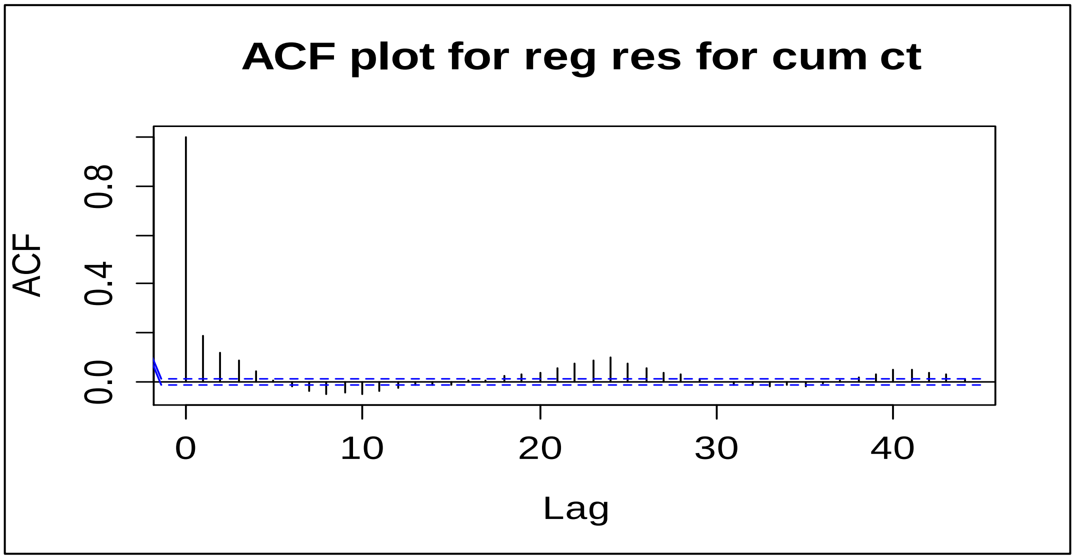

Figure 14b: ACF Plot for Cumulative Patient Data

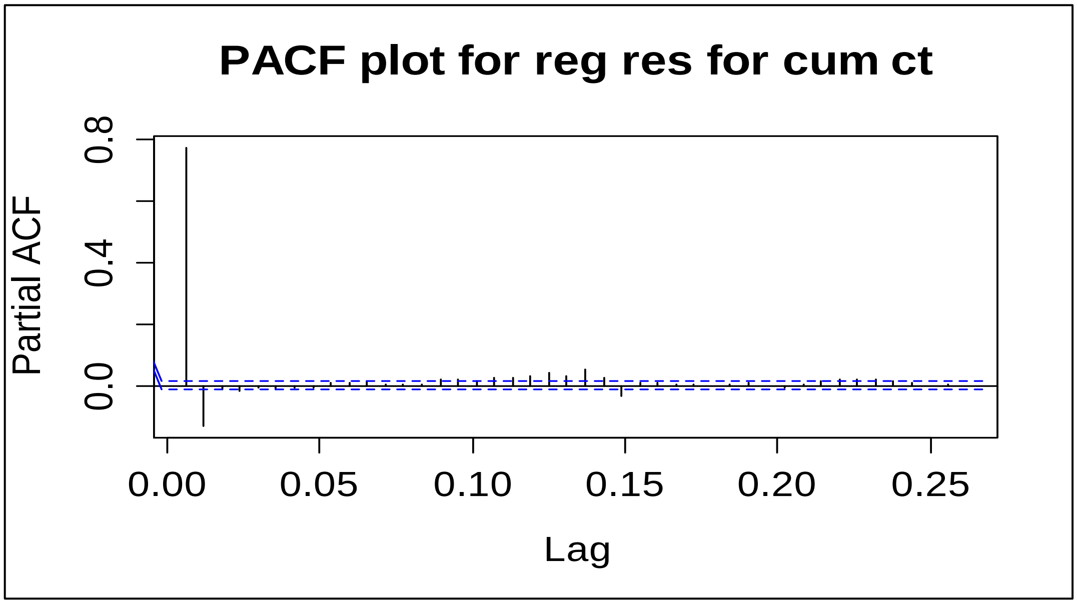

Figure 14c: PACF Plot for Cumulative Patient Data

From the ACF and PACF plot of the residuals, it can be deduced that there is still remaining serial dependence after the regression has explained 82.46% variation between cumulative patient count and the covariates. *Auto.arima* function is again applied to the residuals to model the remaining relationship, and then the residual is forecasted and added to the regression prediction.

```
Series: cum data regression model
ARIMA(2,0,2) with non-zero mean
Coefficients:
         ar1      ar2      ma1      ma2  intercept
      1.3941  -0.5040  -0.5199  -0.0741     0.1112
s.e.  0.0622   0.0433   0.0626   0.0145     0.0517

sigma^2 estimated as 5.156:  log likelihood=-58843.05
AIC=117698.1   AICc=117698.1   BIC=117747.2
```

This means the ARMA model has the following order; MA (2) and AR (2) with zero differencing and non-zero mean. The AR coefficients are 1.394 and -0.504 while for the MA they are -0.52 and 0.0741, respectively.

Time Series Plot

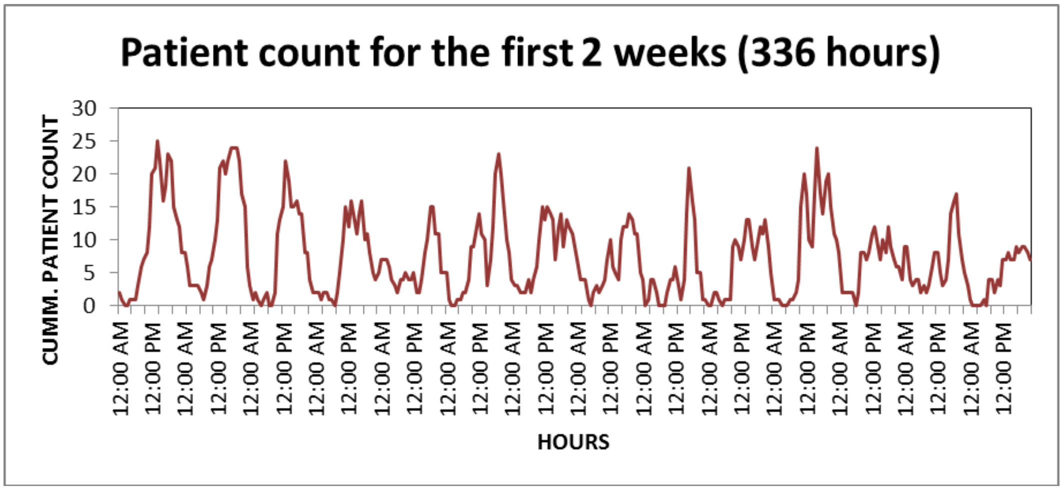

Figure 15a: Graph of Cumulative Patient Count for First Two Weeks of 2009

We plot the two series the patient count and the cumulative patient count for the first two weeks:

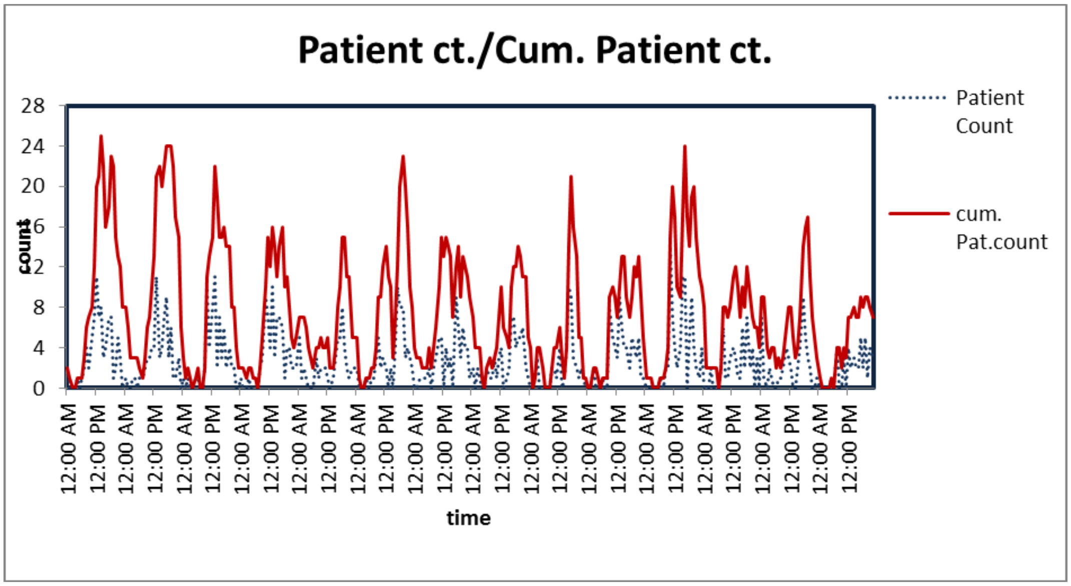

Figure 15b: Graph of Arrival/Cumulative Patient Count for First Two Weeks of 2009

It is observed that both series follow a similar pattern, but the cumulative series is appears smoother than the patient arrival count and this is expected.

We plot the ACF and the PACF for the cumulative patient count:

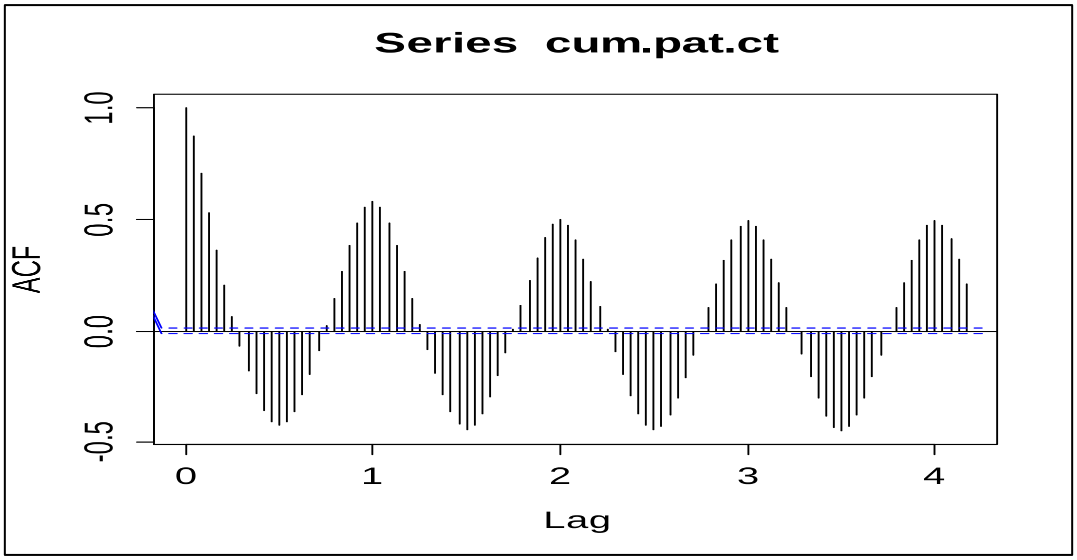

Figure 16a: ACF for Cumulative Patient Count Data

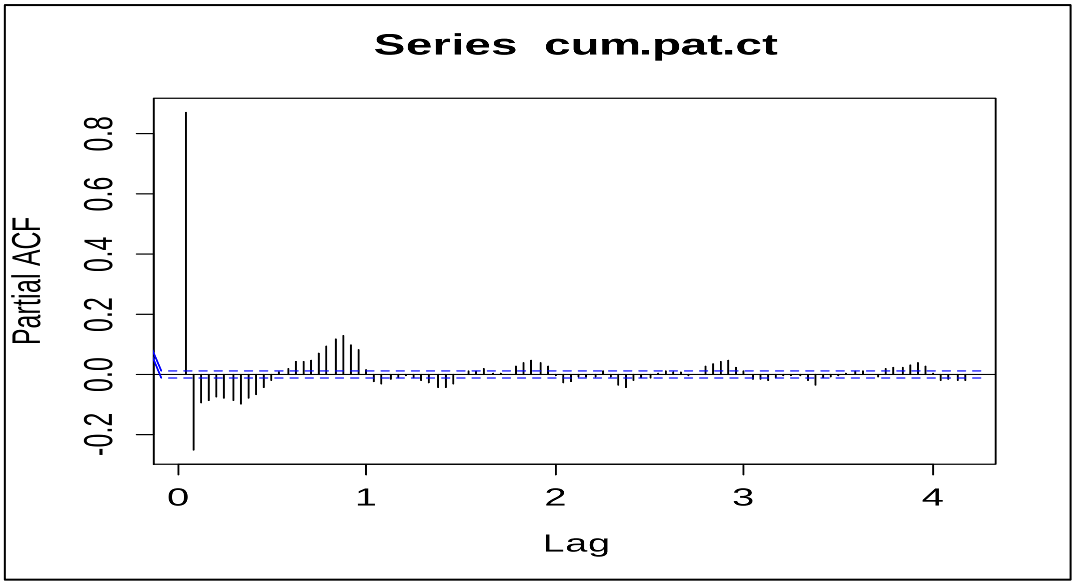

Figure 16ba: PACF for Cumulative Patient Count Data

Sarima Model

The result for the SARIMA model without covariates as described earlier is for the cumulated patient count is:

```
Series: Cumulative Patient Count
ARIMA(4,1,4)(2,0,2)[24]
sigma^2 estimated as 5.015:  log likelihood=-58474.6
AIC=114468.6   AICc=114468.6   BIC=114578.6
ar1       ar2      ar3       ar4       ma1      ma2       ma3       ma4
sar1    sar2
      0.9684  -0.8247  0.5300  -0.0597  -1.0721  0.6958  -0.5801  -
0.0381  0.3828  0.6129
s.e.  0.0731   0.0426  0.0413   0.0400   0.0729  0.0479   0.0510
0.0545  0.0615  0.0614
         sma1     sma2
      -0.3321  -0.6134
s.e.   0.0596   0.0576
```

From the above output , the non-seasonal ARMA model has the following order; AR(4) and MA (4) with one differencing and drift, the seasonal ARMA has AR(2) and MA(2) with seasonal lag of 24 (one day). For the non-seasonal components, The AR coefficients are 0.6129, 0.9684, -0.8247 and 0.53 while for the MA they are -0.06,-1.072, 0.7 and -0.58 respectively. For the seasonal components the AR coefficients are -0.038 and 0.383 while the MA coefficients are -.33 and -0.61. This model was selected based on AIC.

The corresponding SARIMA model with covariates result is:

```
Series: Cumulative Patient Count
ARIMA(4,1,5)(2,0,0)[24] with drift

Coefficients:
         ar1     ar2     ar3      ar4      ma1      ma2      ma3     ma4    ma5
sar1
      0.1859  0.5058  0.6558  -0.5935  -0.3050  -0.7197  -0.7456  0.6786  0.098
0.0429
s.e.  0.0266  0.0246  0.0190   0.0176   0.0273   0.0247   0.0218  0.0266  0.010
0.0063
        sar2  drift
      0.0029  1e-04
s.e.  0.0063  4e-04
```

From the above output , the non-seasonal ARMA model has the following order; AR (4) and MA (5) with one differencing and drift, the seasonal ARMA has AR(2) with seasonal lag of 24 (one day) and drift. For the non-seasonal components, The AR coefficients are 0.186, 0506, -0.6558 and -0.56 while for the MA they are -0.305,-0.72, -0.746, 0.679 and -0.098 respectively. For the seasonal components the AR coefficients are 0.043 and 0.003. This model was selected based on AIC.

BATS and TBATS Model

Also like in previous methods we use 2009-2011 data to fit our model and 2012 data for validation. The output for the cumulative patient count is given below:

```
BATS(0.003, {1,3}, 0.999, {24,168})
Call: bats(y = cum.patient count)
Parameters
  Lambda: 0.00347
  Alpha: 0.003780952
  Beta: -1.665636e-06
  Damping Parameter: 0.998993
  Gamma Values: -1.273514e-05 0.000489981
  AR coefficients: 0.660068
  MA coefficients: 0.067323 -0.016772 -0.014099

Sigma: 0.3418187
AIC: 306764.5
```

Lambda represents the Box-Cox transform which is 0.003 in this case and the smoothing parameters are alpha, beta and gamma which are 0.004, -0.0000017,-0.000013 and 0.00049 respectively .The damping parameter is 0.999 while the ARMA order is AR (3) and MA (1) with coefficients 0.067,-0.017 and -0.014 and for MA is 0.66 respectively and finally the seasonal periods are 24 representing daily cycle and 168 representing weekly cycles, with 198 estimated parameters.

For the TBATS model the corresponding output is;

```
TBATS(0.164, {2,1}, 0.929, {<24,6>, <168,2>})
Call: tbats(y = cum.patient count)
Parameters
  Lambda: 0.164324
  Alpha: 0.01012952
  Beta: -0.0003620897
  Damping Parameter: 0.928777
  Gamma-1 Values: 0.0001652112 0.0001819758
  Gamma-2 Values: 9.49861e-07 -0.00115996
  AR coefficients: 1.709499 -0.716389
  MA coefficients: -0.977066

Sigma: 0.4436025
AIC: 304777.9
```

Lambda represents the Box-Cox transform which is 0.164 in this case and the smoothing parameters are alpha, beta and gamma which are 0.01, -0.00037,- 0.00017 , 0.00018,0.00000095 and -0.0012 respectively .The damping parameter is 0.929 while the ARMA order is AR(2) and MA(1) with coefficients 1.71and -0.72and for AR is -0.Finally the seasonal periods are 24 representing daily cycle and 168 representing weekly cycles, with 21 estimated parameters.

Factor Latent Model

We use only the averaged count for three years after the alignment is done for the corresponding covariates. Same method is applied but with $K = 4$, and the CIIR process is also added. When the residual is plotted the serial dependent appears to be stronger than earlier observed, so we expect that the predictions including the CIIR component would be an improvement from the factor level only prediction (Figure 15a and b here).

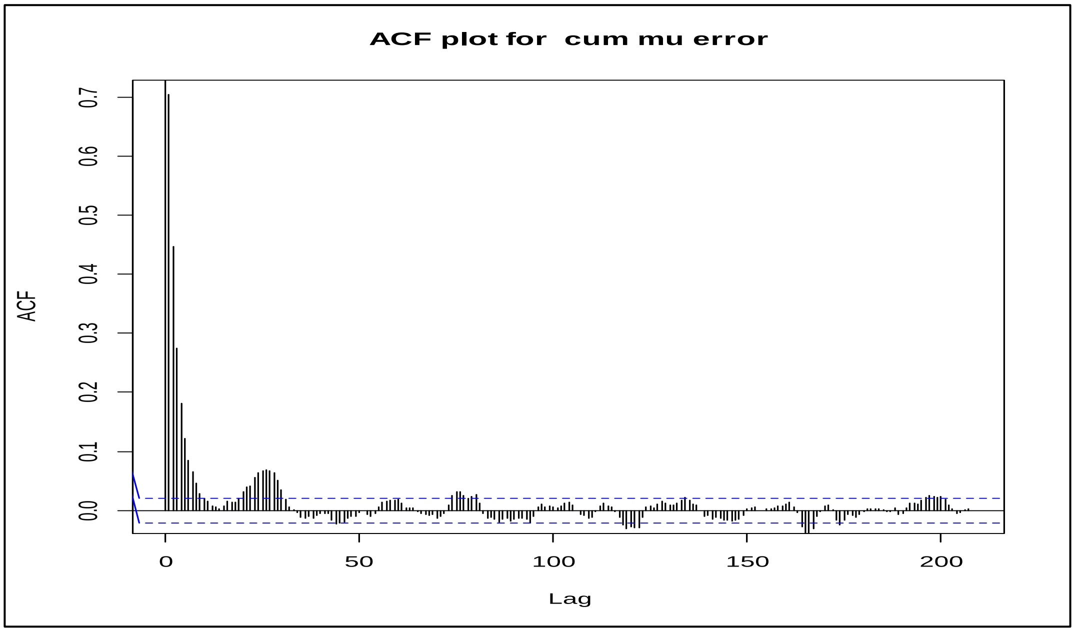

Figure 17a: ACF Plot for Factor Model Residuals for Cumulative Patient Data

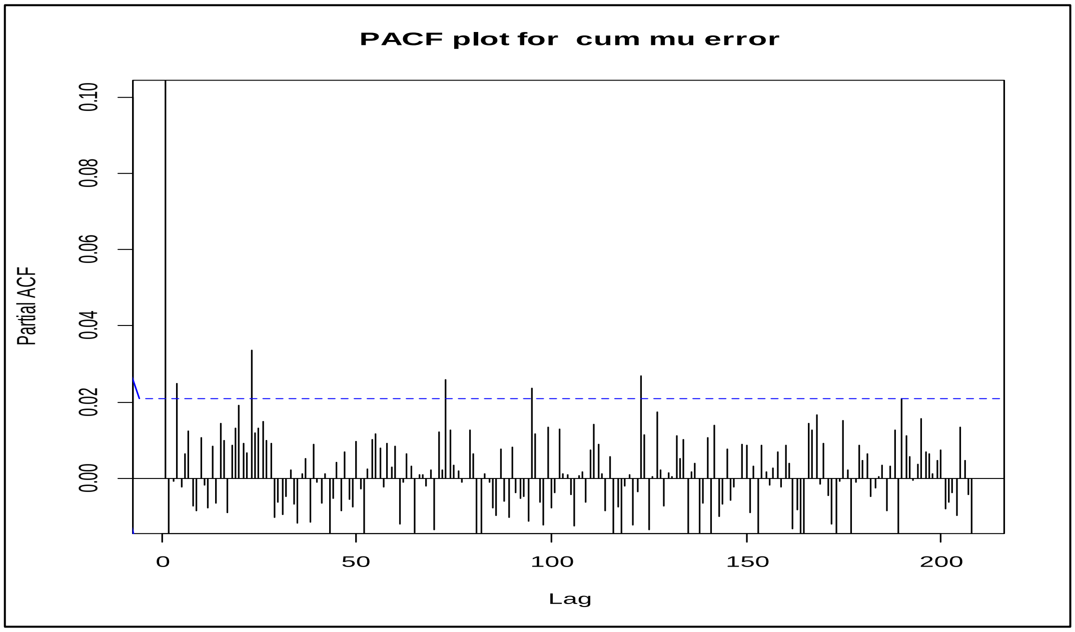

Figure 17b: PACF Plot for Factor Model Residuals for Cumulative Patient Data

Results

Forecasting evaluation. The out of sample forecast accuracy for cumulative patient count data for the first 26 weeks (4368 observations) of 2012 is calculated using MAE, MSE and RMSE and the results are given in Table 5 below.

Table 5: Cumulative Arrival Count Forecast Evaluation Results

METHODS	BATS	TBATS	SARIMA+REG	SARIMA	L.REGRESSION	Factor	FACTOR MODEL+CIIR
MAE	2.41	2.43	4.68	5.68	2.67	3.25	3.23
MSE	10.07	10.41	28.43	47.94	11.53	18.89	18.33
RMSE	3.17	3.23	5.33	6.92	3.40	4.35	4.28

Here we observe that based on all three metrics that the BATS method has the best forecast accuracy followed by the TBATS method. The SARIMA with covariates, performed worse that the SARIMA without covariates. This may be due to the ARIMA models not being suitable for long term forecasts. Also the latent factor model with the CIIR factor performed worse that the that factor model without the CIIR, this may be due to the intGARCH(1,1) not being suitable for this data.

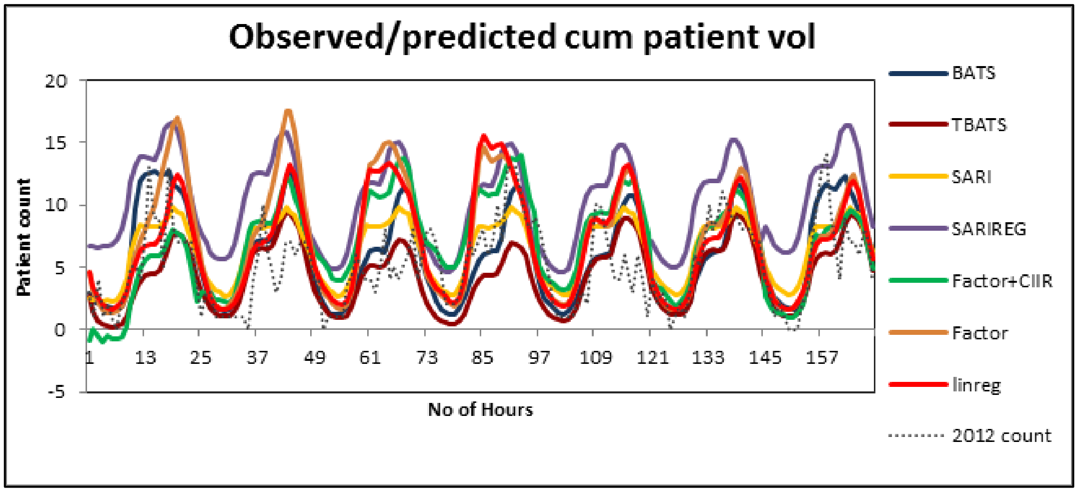

Figure 18: Graph of Predicted/Actual 2012 Patient Counts for Cumulative Patient Data

SECTION III: ANALYSIS OF URGENT ACUITY

As earlier observed, of all the patients that came into the ED from 2009-2011, 52% were of the urgent acuity category,34% the semi-urgent category, 10% were of the emergent category while the rest were resuscitation, non-urgent and unknown categories. Also it was seen that the length of stay for each patient depends on the category and this leads to a further study of the urgent category.

The average proportion of acuity for each hour by day was calculated and applied to the arrival counts predicted by each method previously. Also the methods were applied on the arrival counts data and predicted for 2012. These methods are then compared.

The average hourly urgent acuity category for three years for each day of the week is plotted below:

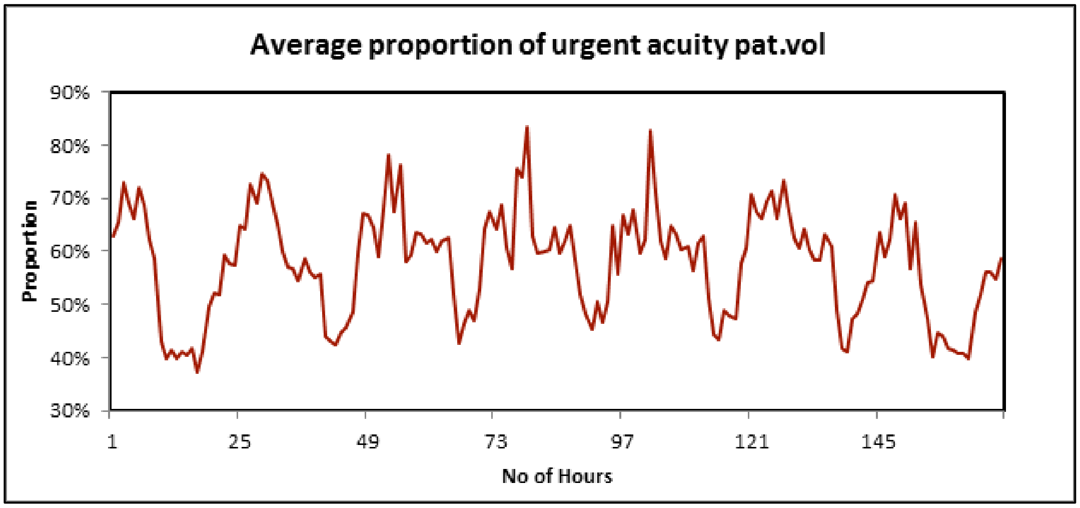

Figure 19: Mean Proportion of Urgent Acuity Patient for One Week

Regression Model

The covariates or explanatory variables used for fitting the regression model for the urgent acuity group are the same as used for patient arrival count. They are categorical variables for, hour of the day, day of the week and month of the year. Again for day of the week variables, Wednesday is the reference category while for hour of the day 12 midnight is the reference Category and for Month of the year December is the reference category.

There are 40 explanatory variables in total, with 1 or 0 values and using 40 degrees of freedom.

After fitting the regression model we plot the residuals to check if the conditions for regression are satisfied:

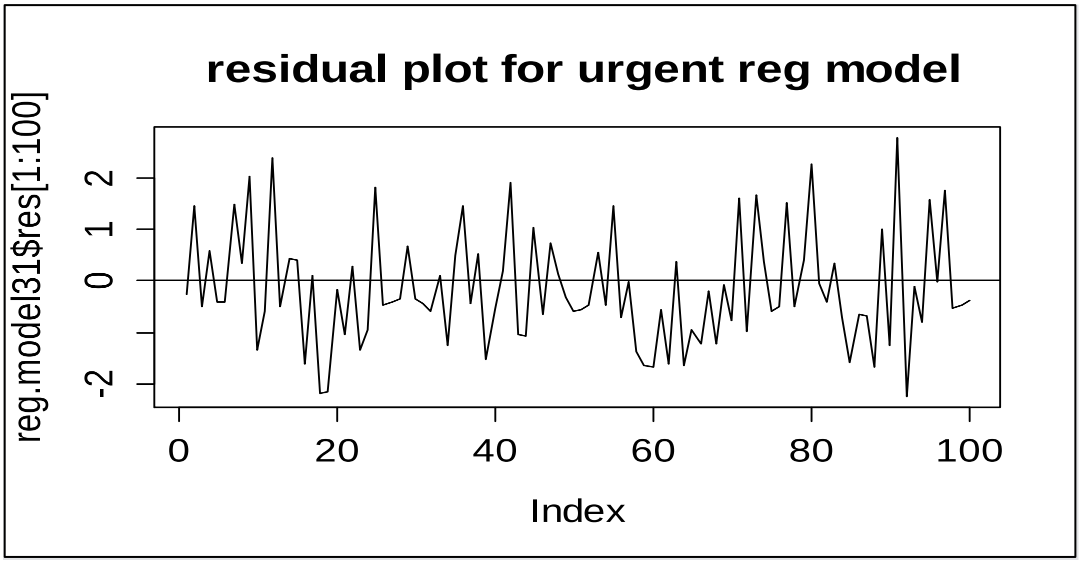

Figure 20a: Regression Residual Plot for Urgent Patient Acuity Patient Data

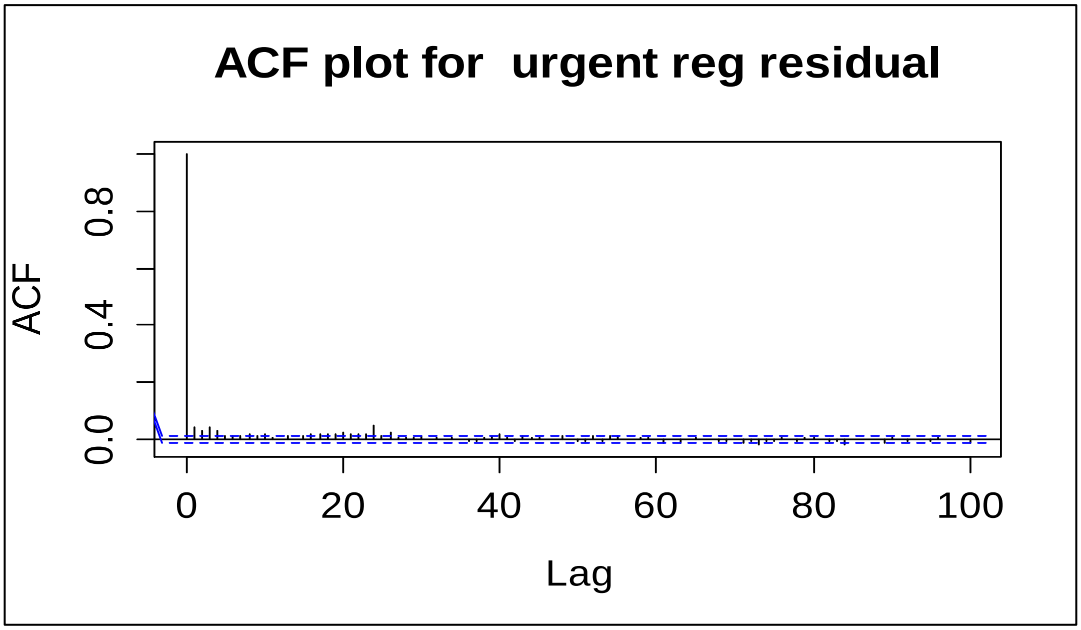

Figure 20b: ACF Plot for Regression Residual for Urgent Acuity Patient Count

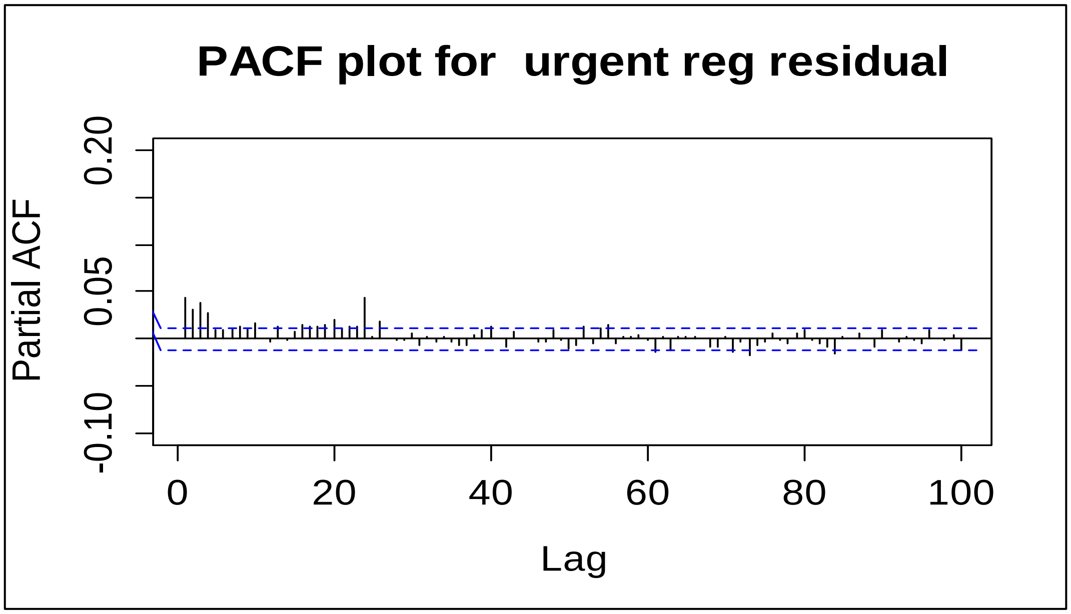

Figure 20c: PACF Plot for Regression Residual for Urgent Acuity Patient Count

From the ACF and PACF plot of the residuals, it can be deduced that the residuals of the urgent regression model is white noise, there is no indication of any serial dependence after the regression has explained about 60% of the variation between the urgent patient arrival count and the covariates. Since there is negligible information in the residuals, no further analysis is done on them.

Time Series Method

Here the urgent acuity series is plotted with the arrival count data series to compare patterns.

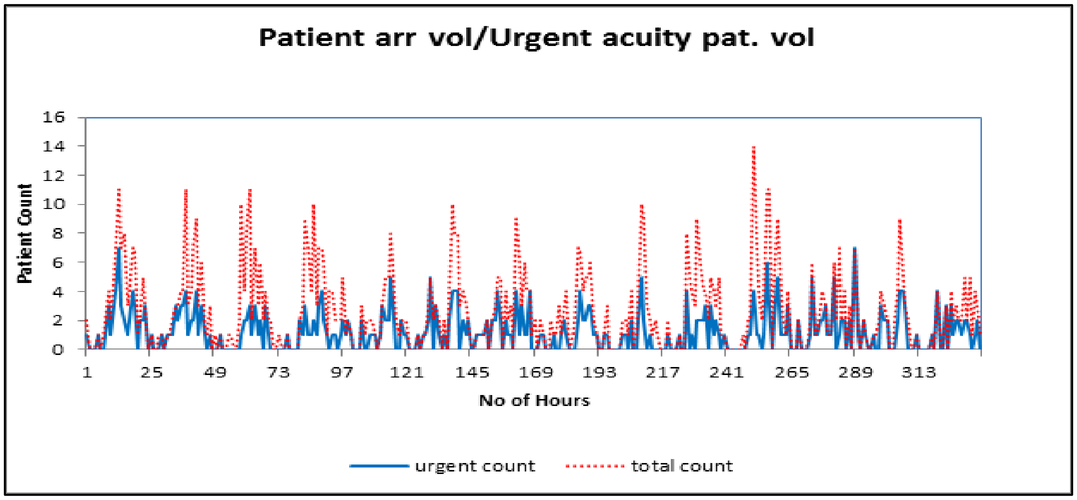

Figure 21: Graph of Urgent Acuity Patient Arrival Count/Patient Arrival Count for Two Weeks

It is observed that both series follow a similar pattern.

The ACF and PACF plots of the urgent acuity level time series models are given below:

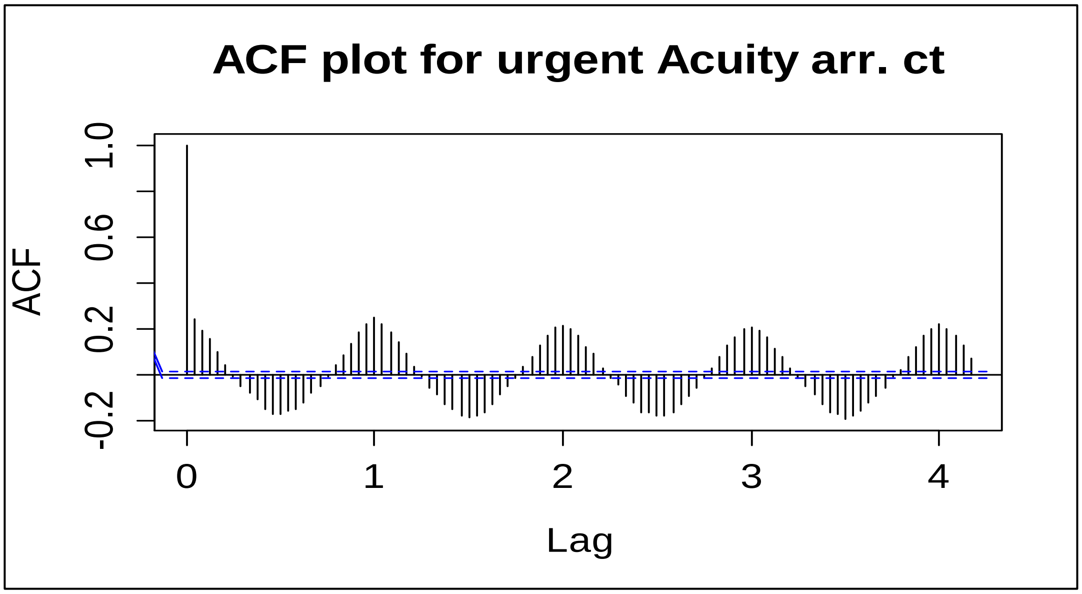

Figure 22a: ACF Plot for Urgent Acuity Patient Arrival Count Data

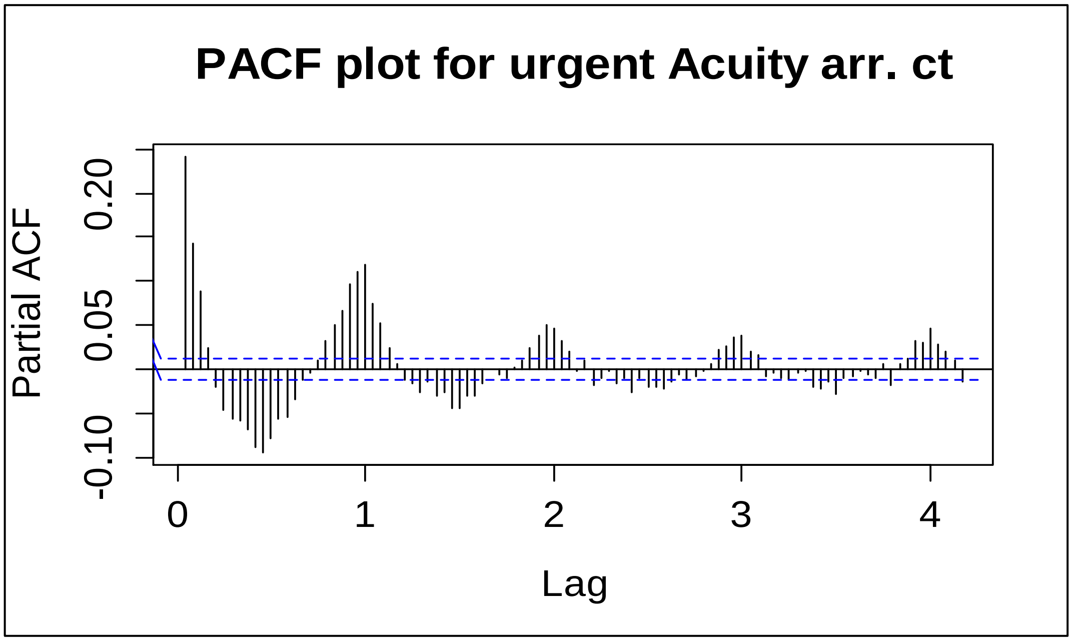

Figure 22b: PACF Plot for Urgent Acuity Patient Arrival Count Data

Also like in previous methods we use the first three years of the data to fit our model and the last year (2012) for validation. The output for the urgent acuity arrival data is given below.

Sarima Model

The result from the *auto.arima* function for the SARIMA model without covariates as described earlier is for the simulated data is:

```
Series: urgent patient count
ARIMA(3,1,4)(2,0,2)[24]

Coefficients:
         ar1     ar2     ar3      ma1      ma2     ma3     ma4    sar1
      0.1317  0.2349  0.1363  -1.0874  -0.1264  0.1001  0.1146  0.6877
sar2     sma1
0.3097   -0.6342
s.e.  0.0042  0.0071  0.0042   0.0066   0.0122  0.0121  0.0066  0.0581
0.0580   0.0573
         sma2
      -0.3285
s.e.   0.0556
```

```
sigma^2 estimated as 1.398:  log likelihood=-41686.72
AIC=83396.95   AICc=83396.96   BIC=83495.07
```

From the above output, the non-seasonal ARMA model has the following order; AR (3) and MA (4) with one differencing The seasonal ARMA has AR(2) and MA(2) with zero differencing with seasonal lag of 24 (one day) . For the non-seasonal components, The AR coefficients are 0.132, 0.235 and 0.1363, while for the MA they are -1.09, -0.13, 0.01 and 0.11, respectively. For the seasonal components the AR coefficients are 0.69 and 0.31 and the MA coefficients are -0.63 and -0.33. This model was selected based on AIC.

The output for the SARIMA model with covariates is as follows:

```
Series: urgent patient count
ARIMA(2,1,2)(2,0,2)[24] with drift

Coefficients:
          ar1     ar2      ma1      ma2     sar1      sar2      sma1    sma2
      -0.8103  0.0215  -0.1572  -0.8328  0.3912  -0.2706  -0.3624  0.2558
1e-04
s.e.   0.1596  0.0080   0.1597   0.1591  0.1815   0.1169   0.1822  0.1153
drift
  1e-04

sigma^2 estimated as 1.36:  log likelihood=-41325.85
AIC=82749.28   AICc=82749.46   BIC=83149.93
```

From the above output, the non-seasonal ARMA model has the following order; AR(2) and MA(2) with one differencing with drift. The seasonal ARMA has AR(2) and MA(2) with zero differencing with seasonal lag of 24 (one day) . For the non-seasonal components, The AR coefficients are -0.81 and 0.022 while the MA Coefficients are -0.16 and -0.83 respectively. For the seasonal components the AR coefficients are 0.39 and -0.027 also the MA coefficients are -0.36 and 0.026. This model was selected based on AIC.

BATS and TBATS Model

The BATS model that best suits our simulated data is as follows:

```
BATS(0, {0,0}, -, {24,168})

Call: bats(y = urgent patient count)

Parameters
  Lambda: 0.000114
  Alpha: 0.004796037
  Gamma Values: 1.295485e-07 -1.221693e-08
```

Lambda represents the Box-Cox transform which is 0.000114 in this case and the smoothing parameters are alpha and gamma parameters which are 0.005, 0.00000013 and -0.0000000122 respectively. There are no damping parameter and ARMA errors for this model. Finally the seasonal periods are 24 representing daily cycle and 168 representing weekly cycles, with 193 estimated parameters.

For the TBATS model the corresponding output is;

```
TBATS(0, {1,1}, 0.997, {<24,6>, <168,6>})

Call: tbats(y = urgent patient. count)
Parameters
  Lambda: 3e-06
  Alpha: 0.004296292
  Beta: -1.200812e-05
  Damping Parameter: 0.996749
  Gamma-1 Values: 6.213437e-06 3.166494e-06
  Gamma-2 Values: -1.890419e-08 3.477005e-08
  AR coefficients: 0.018446
  MA coefficients: 0.003665
```

Lambda represents the Box-Cox transform which is approximately zero in this case and the smoothing parameters are alpha, beta which is 0.0043, -0.000012 and gamma coefficients which are almost zero. The damping parameter is 0.998 while the

ARMA order is AR (1) and MA(1) with coefficients 0.018 and for MA 0.0037.Finally the seasonal periods are 24 representing daily cycle and 168 representing weekly cycles, with 28 estimated parameters.

Factor Latent Model

We use only the averaged urgent acuity count for three years after the alignment is done for the corresponding covariates. Same method is applied with $K = 4$, and the CIIR process is also added. When the residual is plotted the serial dependent appears to be stronger than earlier observed, so we expect that the predictions including the CIIR component would be an improvement from the factor level only prediction.

The ACF and PACF plots for the residuals after fitting the latent factor model are shown below:

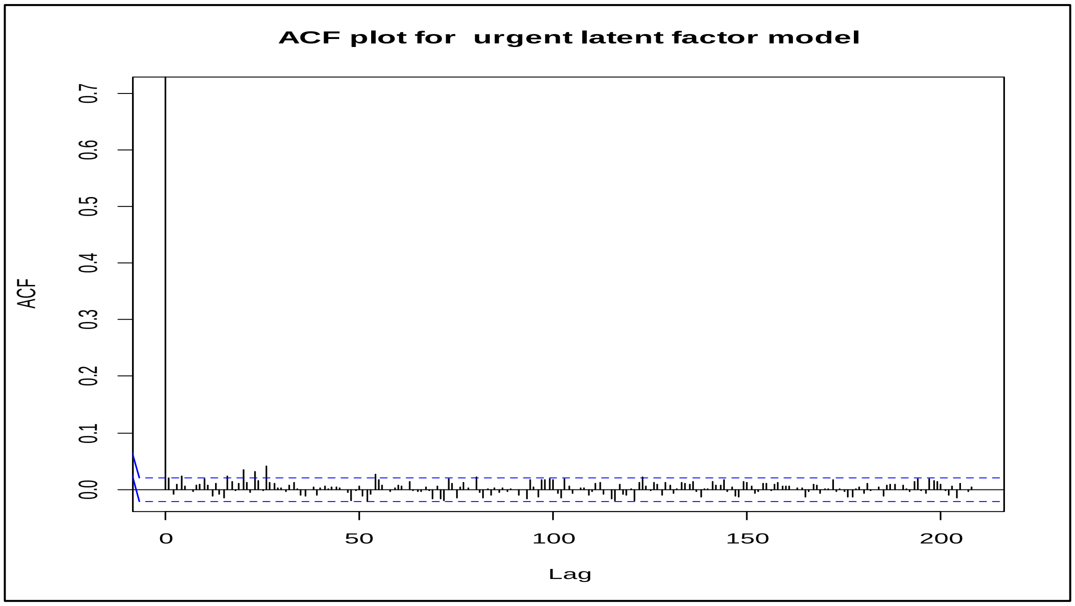

Figure 23a: ACF Plot of Factor Model Residuals for Urgent Acuity Arrival Patient Data

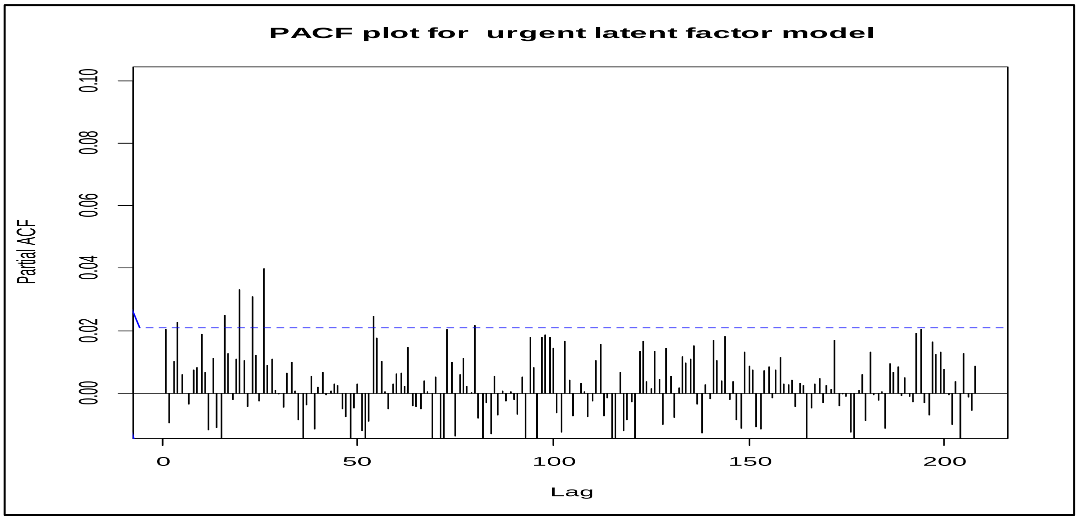

Figure 23b: PACF Plot of Factor Model Residuals for Urgent Acuity Arrival Patient Data

<u>Forecast Evaluation</u>

First the average proportions for each hour of the day calculated earlier was applied to the predicted count from patient arrival count. The out of sample forecast errors for the first 26 weeks (4368 observations) of 2012 are given below.

Table 6a: Urgent Arrival Count (using mean proportions) Forecast Evaluation Results

METHODS	BATS	TBATS	SARIMA+REG	SARIMA	L.REGRESSION	FACTOR MODEL+CIIR	Factor
MAE	0.86	0.86	1.19	0.94	0.88	0.88	0.92
MSE	1.36	1.33	2.07	1.51	1.38	1.26	1.26
RMSE	1.17	1.15	1.44	1.23	1.17	1.12	1.12

We see from the table 6a that the Factor latent models with CIIR and without CIIR both have the smallest values of all three matrices, followed by the TBATS model. The SARIMA model without covariates performed the worst followed by the SARIMA model with covariates.. We also observe again that adding covariates to the SARIMA model improved its performance quite significantly. Most of our methods have the ability to produce reliable long term forecasts (one year ahead), which is needed for capacity planning. For our latent factor models with constraints and smoothing splines, it is observed that they are also same and this is expected because the serial dependence in the error after fitting the latent factor model is not significant. A major drawback for this model is that it does not produce confidence intervals or prediction intervals by default while the other models are capable of doing so.

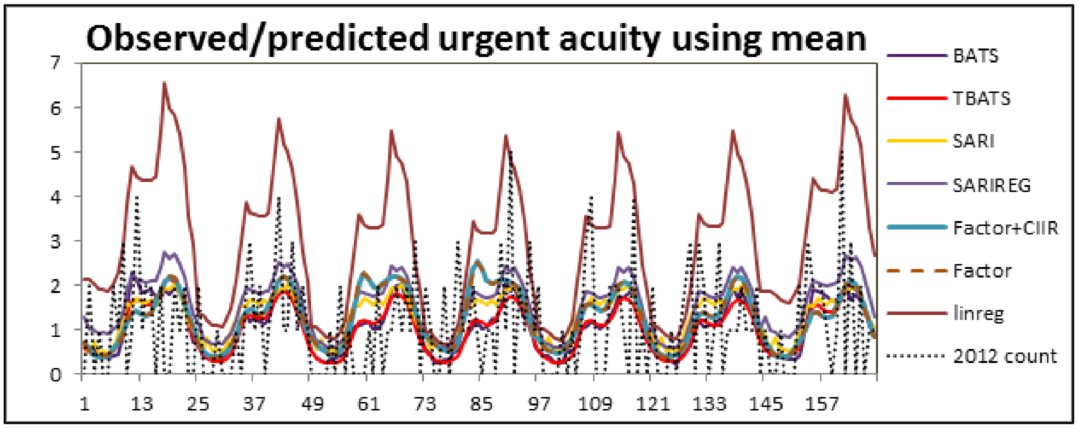

Figure 24: Predicted/Actual 2012 obs for Urgent Acuity Patient Arrival Data Using Mean Proportions

The out-of sample forecast errors for the urgent count data analysis of all the methods is given below.

Table 6b: Actual Urgent Arrival Count Forecast Evaluation Results

Methods	BATS	TBATS	SARIMA+REG	SARIMA	FACTOR MODEL+CIIR	Factor	L.REGRESSION
MAE	0.84	0.85	0.92	1.14	1.62	0.90	0.90
MSE	1.27	1.31	1.34	1.86	4.04	1.35	1.35
RMSE	1.13	1.14	1.16	1.36	2.01	1.16	1.16

It can be seen from Table 6b that the BATS model performed best of all the models followed by the TBATS model and the factor latent models. SARIMA models performed worst but adding the covariates was an improvement from the model without the covariates. This mirrors the results obtained with the arrival patient volume. This mean is a good method but it depends heavily on how good the patient volume prediction is.

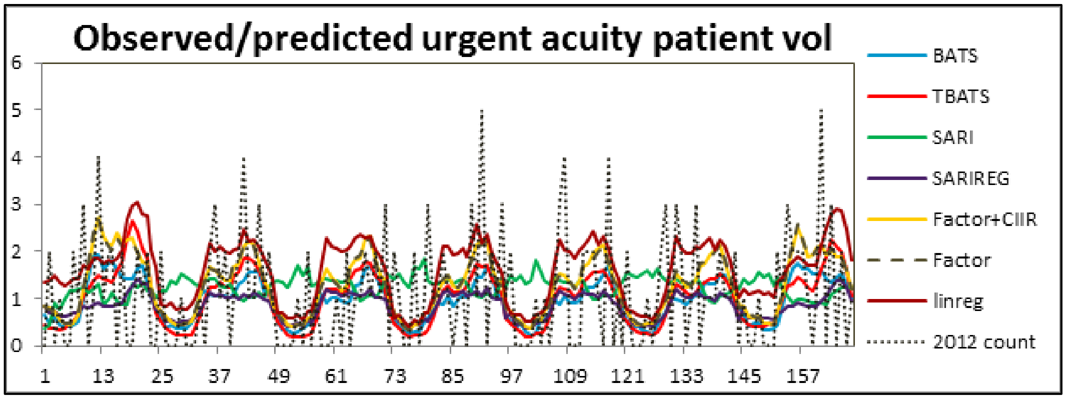

Figure 25: Predicted/Actual 2012 obs for Urgent Acuity Patient Arrival Data

SECTION IV: SIMULATED DATA WITHOUT TREND COMPONENT

Data Simulation

The purpose of this section is to simulate data that has a similar pattern with our actual patient count and to apply the methods used in the previous sections and compare with our actual results. It was earlier stated that the arrival count is a Poisson process and so to simulate the data, we would use the random Poisson distribution.[27]

It was observed that there is a daily and weekly cycle in the data; this has to be incorporated in the data also there is the error component of the data which is a ARIMA process. The error component is generated using the function *arima.sim* function in R with coefficients for the AR(2) component are 0.95 and -.45 and the MA(2) coefficients are -.84 and .29, and this is randomly generated using the random normal distribution with variance .134. We generate data for 104 weeks (two years) the first half will be used to build the model and the second half will be used for validation.

For the cycle we generate a rate defined as:

rate = 12+10* sin(2*pi*hour/24) + 2*cos(2*pi*week/52) + err ;

Finally we generate the data using:

ysim = rpois(X,rate).

We plot the patient arrival count and the simulated data to compare the patterns and we observe that the patterns are identical.

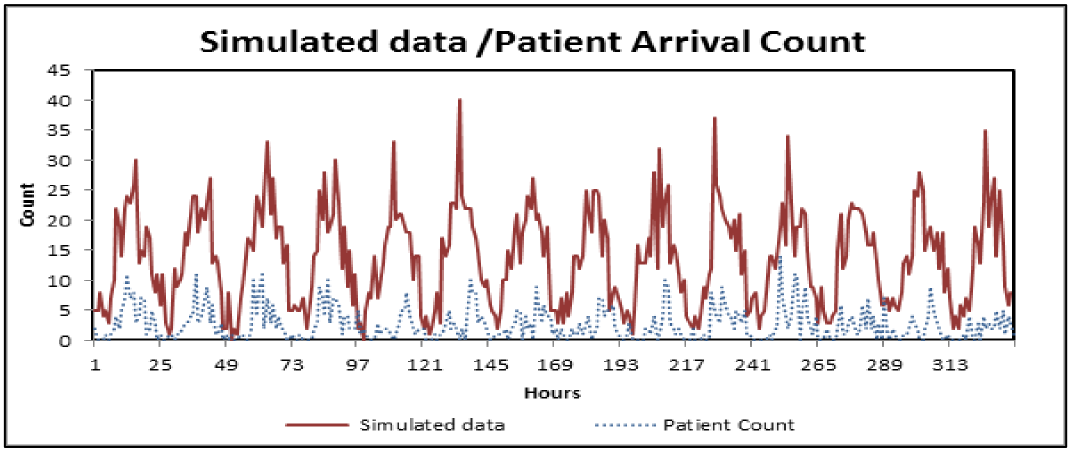

Figure 26: Plot of Simulated Data/Patient Arrival Data for First 336 obs

Regression Model

The covariates or explanatory variables used for fitting the regression model for the simulated are the same as used for patient arrival count. They are categorical variables for, hour of the day, day of the week and month of the year. Again for day of the week variables, Wednesday is the reference category while for hour of the day 12 midnight is the reference Category and for Month of the year December is the refere nce category.

There are 40 explanatory variables in total, with 1 or 0 values and using 40 degrees of freedom.

After fitting the regression model we plot the residuals to check if the conditions for regression are satisfied:

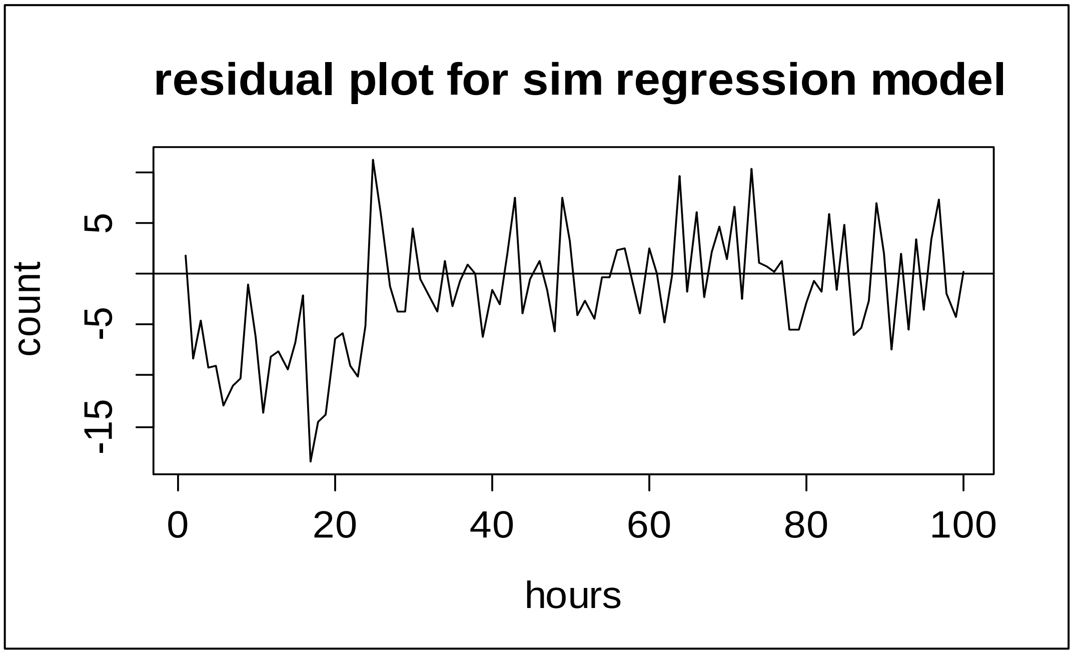

Figure 27a: Regression Residual Plot for Simulated Data without Trend

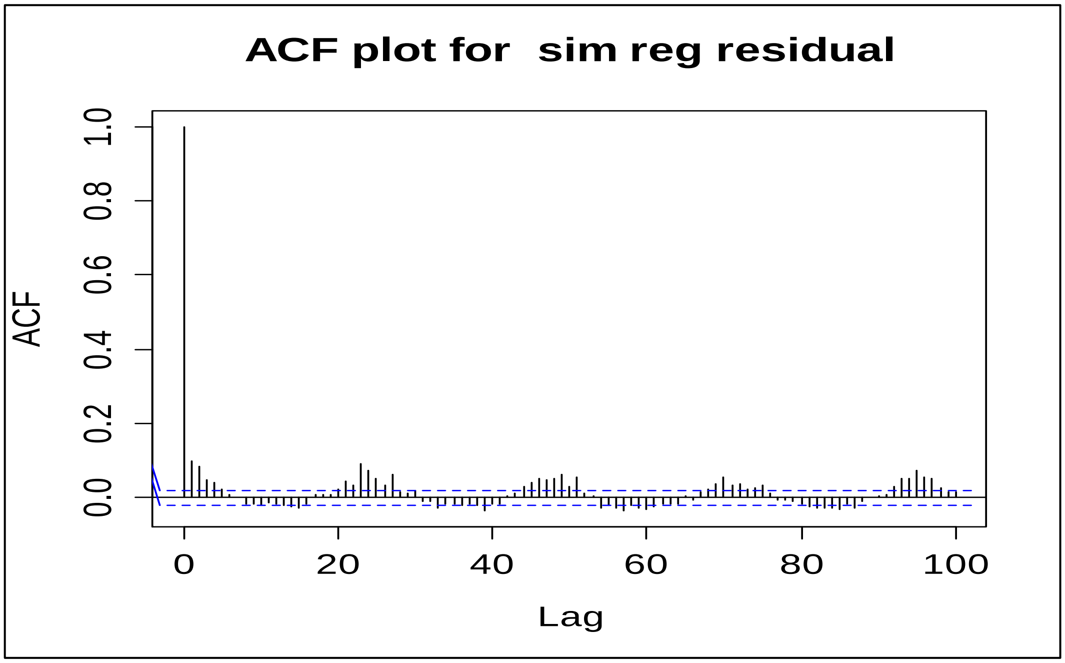

Figure 27b: ACF Plot for Regression Residual for Simulated Data without Trend

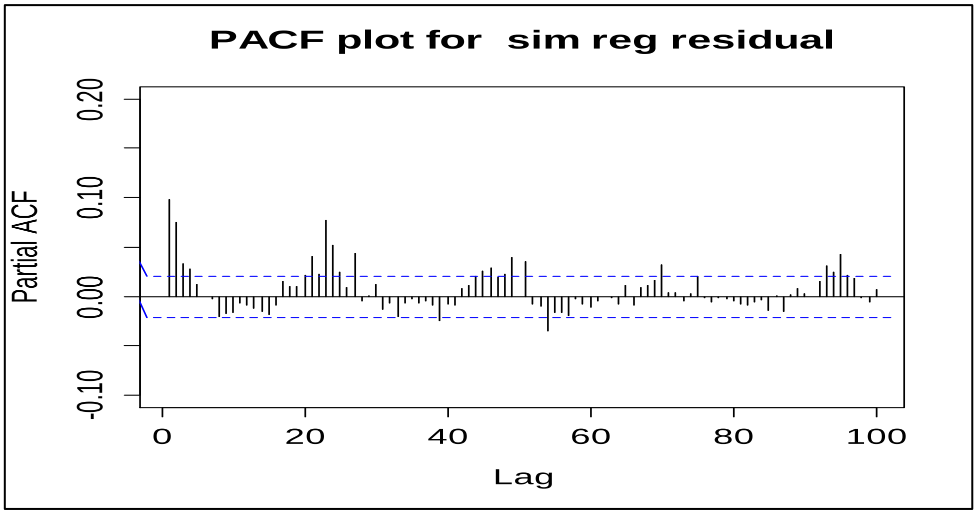

Figure 27c: PACF Plot for Regression Residual for Simulated Data without Trend

From the ACF and PACF plot of the residuals, it can be deduced that there is a very weak serial dependence after the regression has explained 90.4% variation between the simulated data and the covariates. Since there is negligible information in the residuals, no further analysis is done on them.

Time Series Method

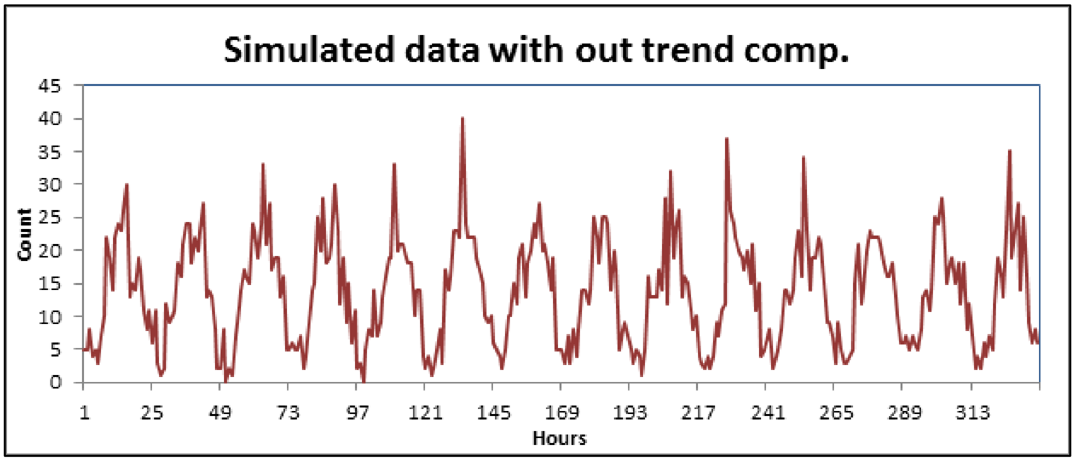

Figure 28: Time Series Plot of Simulated Data without Trend for First 336 obs

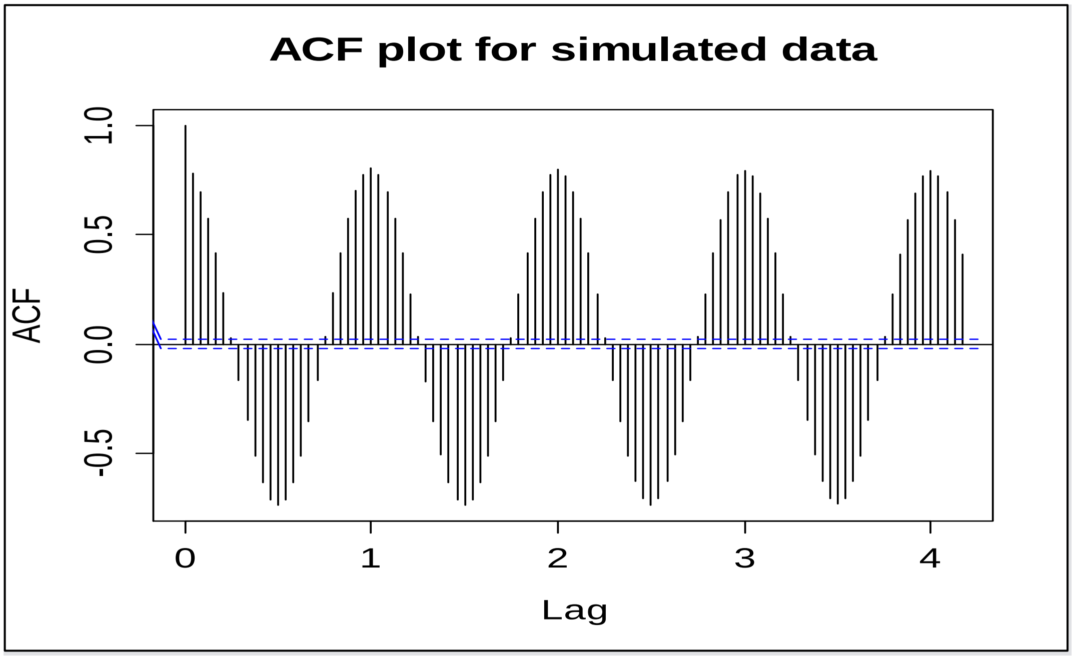

Figure 29a: ACF Plot for Simulated Data without Trend

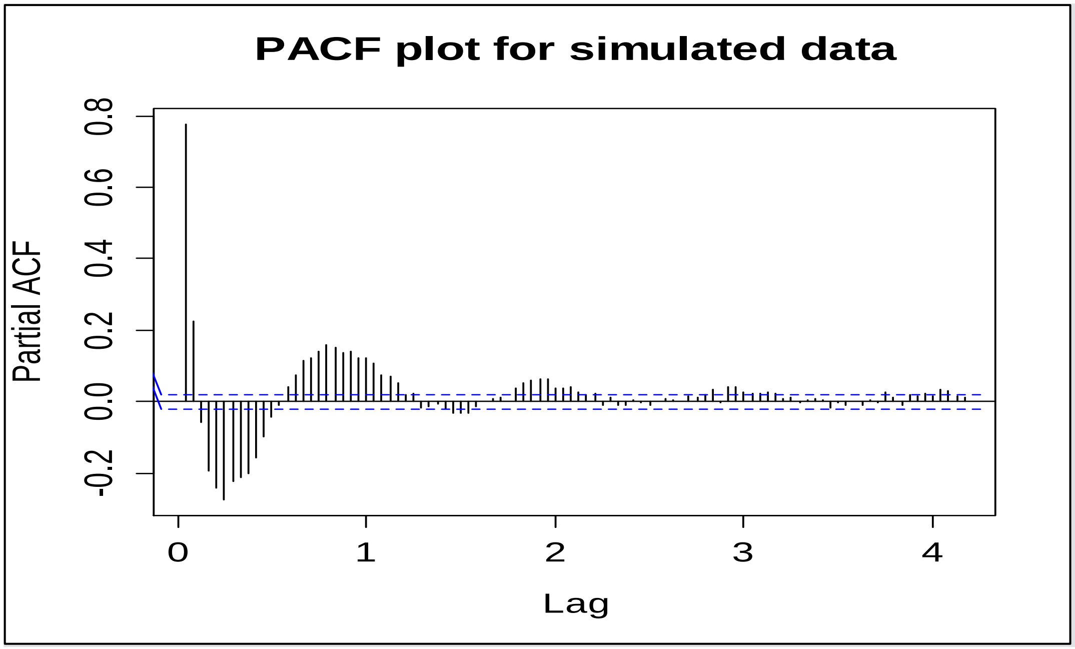

Figure 29b: PACF Plot for Simulated Data without Trend

The output for the simulated data without trend component time series models are given below.

Sarima Model

The result from the *auto.arima* function for the SARIMA model without covariates as described earlier is for the simulated data without trend is:

```
Series: simulated data1
ARIMA(3,1,2)(2,0,2)[24]

Coefficients:
         ar1      ar2      ar3      ma1     ma2    sar1    sar2     sma1
      0.1121  -0.0319  -0.0175  -1.1028  0.1104  0.4502  0.5496  -0.4052
sma2
-0.5404
s.e.  0.0250   0.0108   0.0107   0.0271  0.0270  0.0572  0.0572   0.0567
0.0550

sigma^2 estimated as 13.07:  log likelihood=-23619.3
AIC=47255.47   AICc=47255.49   BIC=47326.22

Training set error measures:
          ME           RMSE            MAE           MPE          MAPE
MASE
  0.03651052   3.6040593    2.715831   -13.789489   32.40253
0.672224
```

From the above output, the non-seasonal ARMA model has the following order; AR (3) and MA (2) with one differencing. The seasonal ARMA has AR(2) and MA(2) with zero differencing with seasonal lag of 24 (one day) . For the non-seasonal components, The AR coefficients are 0.11, -0.0319 and -0.018 while the MA Coefficients are -1.1 and 0.11 respectively. For the seasonal components the AR coefficients are 0.45 and 0.55 also the MA coefficients are -0.41 and -0.54. This model was selected based on AIC.

The corresponding SARIMA model with covariates result is:

```
Series: simulated data1
ARIMA(2,1,2)(2,0,2)[24] with drift
```

```
sigma^2 estimated as 12.79:  log likelihood=-23527.27
AIC=47146.53   AICc=47147.03   BIC=47471.99

Coefficients:
          ar1     ar2      ma1      ma2     sar1     sar2     sma1
      -0.8103  0.0215  -0.1572  -0.8328  0.3912  -0.2706  -0.3624
sma2  drift
 0.2558  1e-04
s.e.   0.1596  0.0080   0.1597   0.1591  0.1815   0.1169   0.1822  0.1153  1e-0

Training set error measures:
            ME           RMSE           MAE           MPE          MAPE
 -0.000077618   3.575836563   2.713255152 -15.438955034   33.368517394
MASE
   0.671586391
```

From the above output, the non-seasonal ARMA model has the following order; AR (2) and MA (2) with zero differencing. The seasonal ARMA has AR(2) and MA(2) with zero differencing with seasonal lag of 24 (one day). For the non-seasonal components, The AR coefficients are -0.81 and 0.0215 while the MA Coefficients are -0.16 and -0.833 respectively. For the seasonal components the AR coefficients are 0.39 and -0.271 also the MA coefficients are -.36 and 0.256 with drift. This model was selected based on AIC.

BATS and TBATS Model

The BATS model that best suits our simulated data

```
BATS(0.612, {0,0}, 0.999, {24,168})

Call: bats(y = simulated data1)

Parameters
  Lambda: 0.612368
  Alpha: 0.02963382
  Beta: 2.954125e-05
  Damping Parameter: 0.998748
  Gamma Values: 0.02549024 2.004902e-07

Sigma: 1.320385
AIC: 99975.18
```

Lambda represents the Box-Cox transform which is 0.612 in this case and the smoothing parameters are alpha, beta and gamma parameters which are 0.03, 0.00003, 0.025 and 0.0000002 respectively. The damping parameter for this model is 0.999 but there are no ARMA errors for this model. Finally the seasonal periods are 24 representing daily cycle and 168 representing weekly cycles, with 194 estimated parameters.

For the TBATS model the corresponding output is;

```
TBATS(0.673, {0,0}, 1, {<24,3>, <168,2>})

Call: tbats(y = simulated.data1)

Parameters
  Lambda: 0.67301
  Alpha: 0.003834106
  Beta: 2.570862e-05
  Damping Parameter: 1
  Gamma-1 Values: 4.845342e-07 1.03767e-05
  Gamma-2 Values: -1.571637e-08 5.779328e-06

Sigma: 1.512353
AIC: 99569.7
```

Lambda represents the Box-Cox transform which is 0.673 in this case and the smoothing parameters are alpha and beta which are 0.0038 and 0.00003 and also gamma parameters which are all close to zero. The damping parameter for this model is 1 but there are no ARMA errors for this model. Finally the seasonal periods are 24 representing daily cycle and 168 representing weekly cycles, with 22 estimated parameters.

<u>Factor Latent Model</u>

We use the first half of the data to fit the factor model with hour of the day, day of the week and week of the year covariates. Same method is applied but with

$K = 4$, and the CIIR process is also added. The later half of the data is used for validation.

The ACF and PACF plots for the residuals after fitting the latent factor model are shown below:

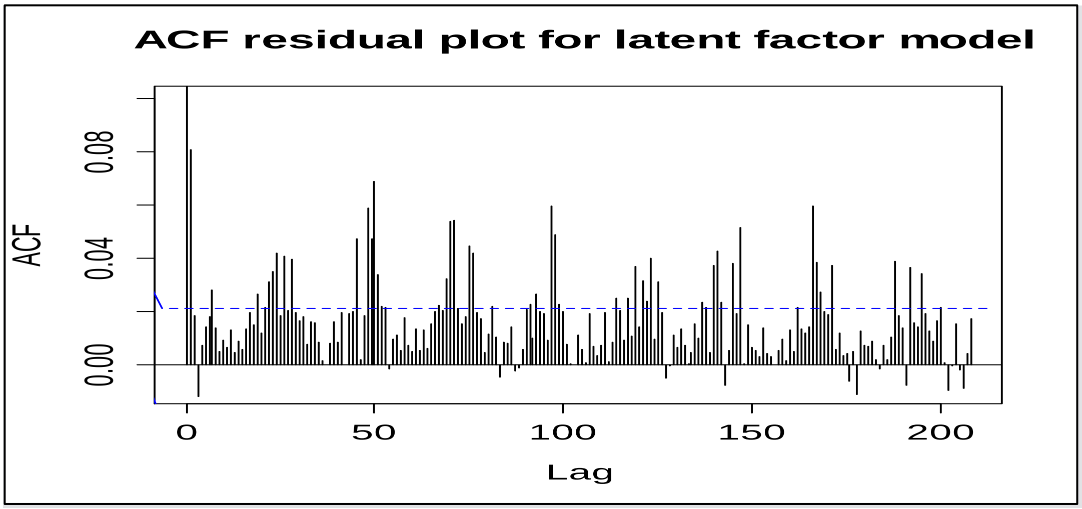

Figure 30a: ACF Plot of Factor Model Residuals for Simulated Data without Trend

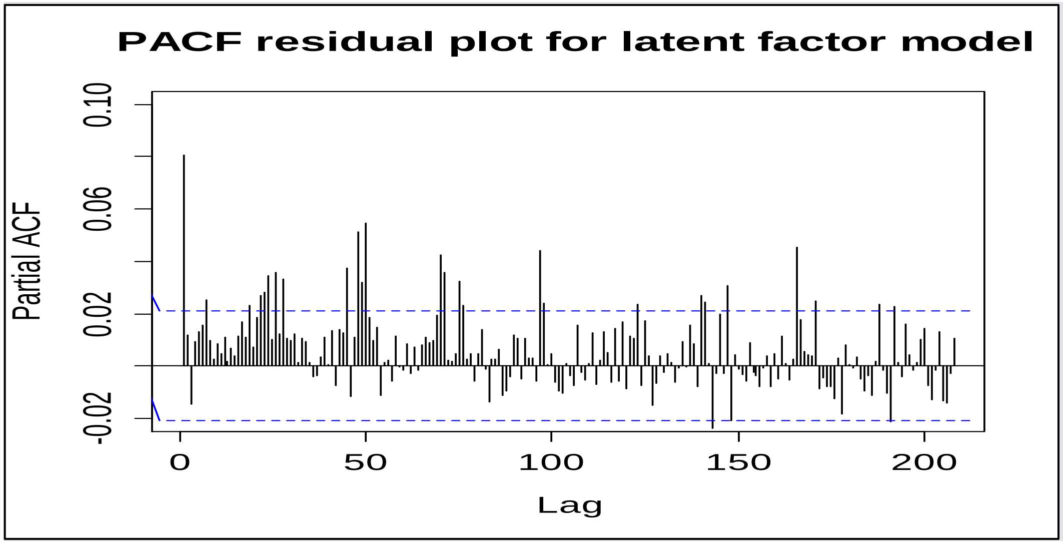

Figure 30b: PACF Plot of Factor Model Residuals for Simulated Data without Trend

The plots are similar but there is not information that can be deduced form them, We would fit a model including the CIIR process and see what improvement this might bring to our model.

Forecast Evaluation

The out of sample forecast accuracy for simulated data for the first 26 weeks (4368 observations) of 2012 is calculated using MAE, MSE and RMSE and the results are given in Table 7 below.

Table 7: Simulated Data without Trend Forecast Evaluation Results

METHODS	BATS	TBATS	SARIMA+REG	SARIMA	L.REGRESSION	FACTOR MODEL+CIIR	Factor
MAE	4.69	18.90	11.97	19.33	12.10	11.96	11.96
MSE	32.15	71.50	12.85	75.43	13.23	12.89	12.90
RMSE	5.67	8.46	3.58	8.69	3.64	3.59	3.59

We see from the table 4 that the Factor latent models with CIIR and without CIIR both have the smallest values of all three matrices, followed by the SARIMA model with covariates. The TBATS model performed the worst followed by the SARIMA model without covariates. The BATS model didn't perform as good as expected from the previous results. We also observe again that adding covariates to the SARIMA model improved its performance quite significantly. Most of our methods have the ability to produce reliable long term forecasts (one year ahead), which is needed for capacity planning. For our latent factor models with constraints and smoothing splines, it is observed that they are also same and this is expected because the serial dependence in the error after fitting the latent factor model is not significant. A major drawback for this model is that it does not produce confidence intervals or prediction intervals by default while the other models are capable of doing so.

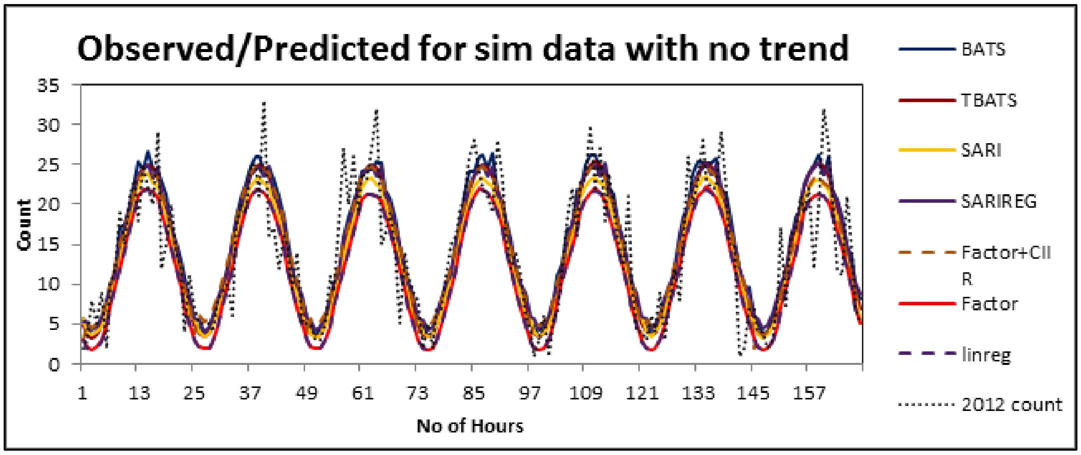

Figure 31: Predicted/Actual Simulated Data for Simulated Data without Trend

SECTION V: SIMULATED DATA WITH TREND

We observed that starting in May 2011 there was a steady decline in patient arrival volume that continued till 2012. This is a trend and so what happens when our data has a trend? Will our models be able to capture this trend?

To our simulated data we add a quadratic trend component. The data is generated as follows:

$$t = 1, 2, \ldots N$$

$ysim_t$ is our previously simulated data

$$ysim\_with\_trend = ysim_t + (t^2 / (5*10^5))$$

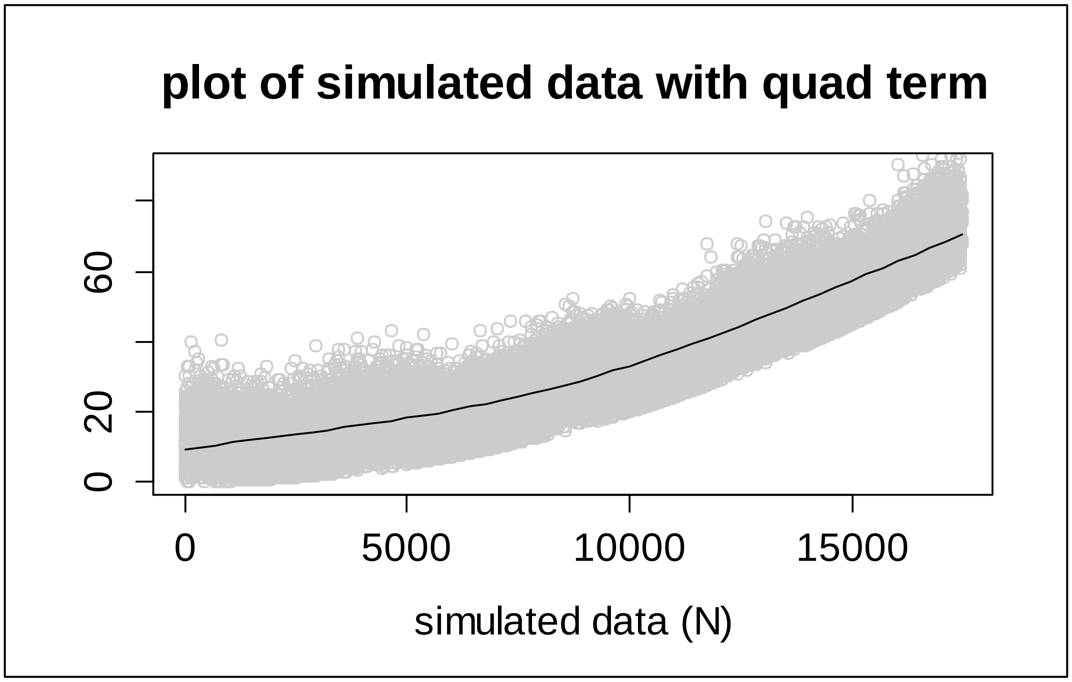

Figure 32: Plot of Simulated Data with Trend Component

Regression Model

The covariates or explanatory variables used for fitting the regression model for the simulated data with trend are the same as used for patient arrival count. They are categorical variables for, hour of the day, day of the week and month of the year. Again for day of the week variables, Wednesday is the reference category while for hour of the day 12 midnight is the reference Category and for Month of the year December is the reference category.

There are 40 explanatory variables in total, with 1 or 0 values and using 40 degrees of freedom.

After fitting the regression model we plot the residuals to check if the conditions for regression are satisfied:

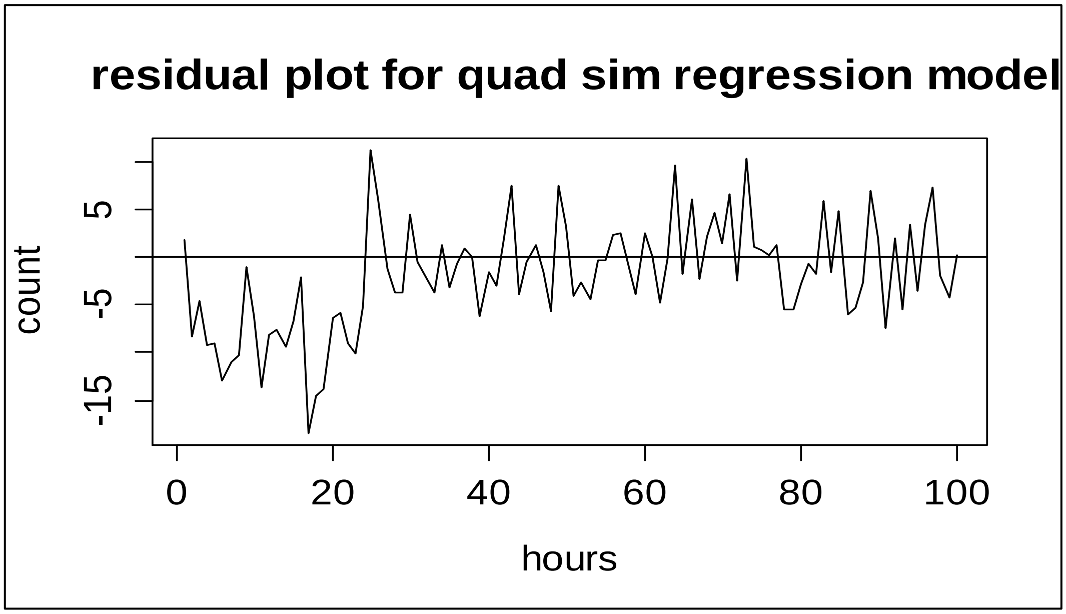

Figure 33a: Regression Residual Plot for Simulated Data with Trend

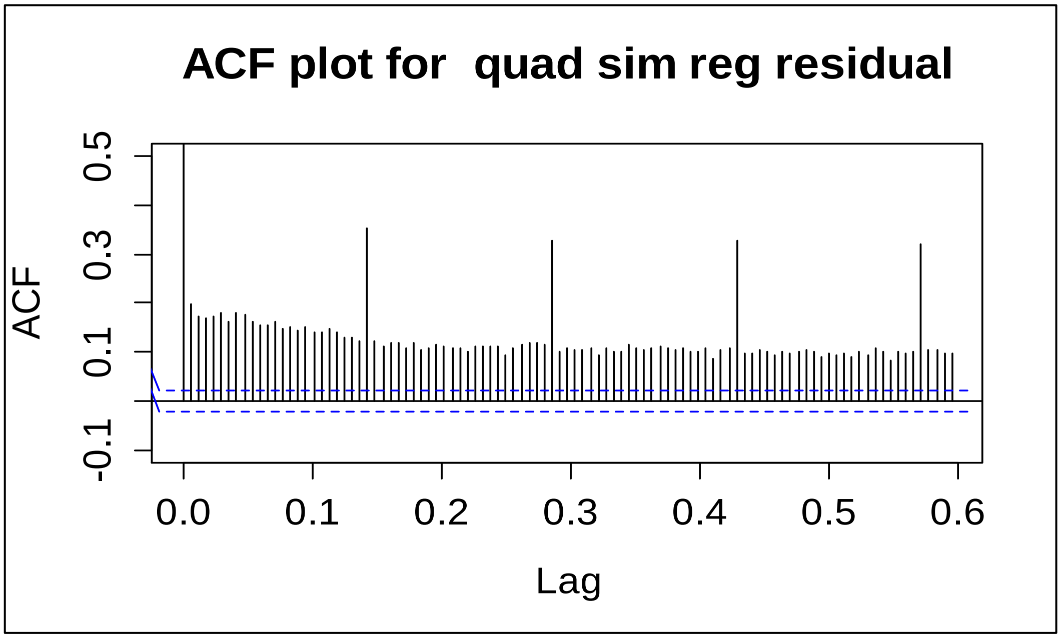

Figure 33b: ACF Plot for Regression Residual of Simulated Data with Trend

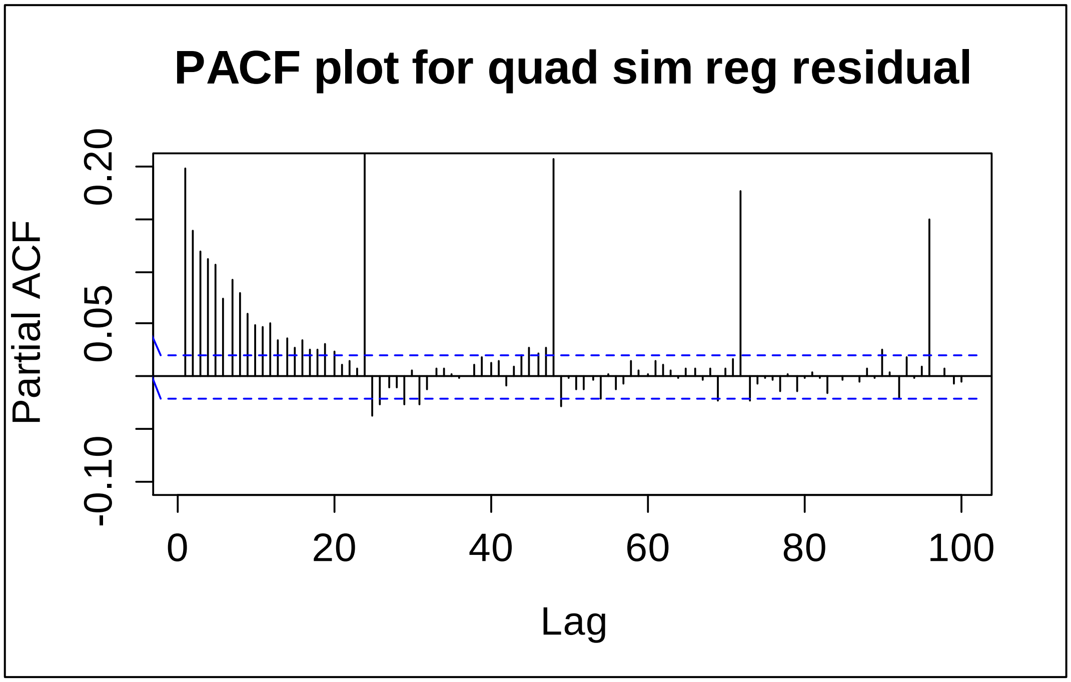

Figure 33c: PACF Plot for Regression Residual of Simulated Data with Trend

From the ACF and PACF plot of the residuals, it can be deduced that there is a serial dependence after the regression has explained 94% variation between the simulated data and the covariates. Since there is negligible information in the residuals, no further analysis is done on them. The ACF and PACF plots for the simulated data with trend component are given below:

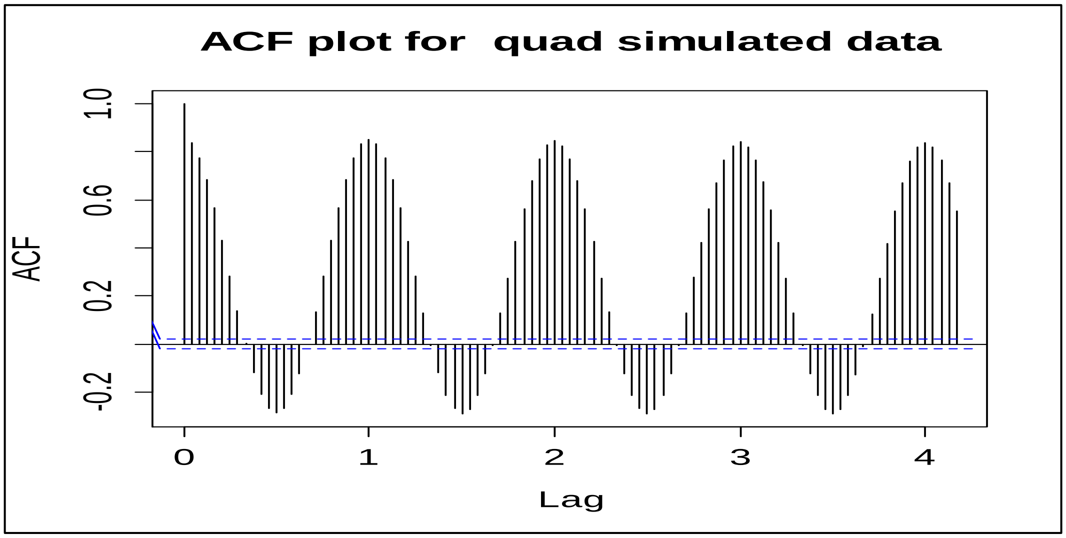

Figure 34a: ACF Plot for Simulated Data with Trend Component

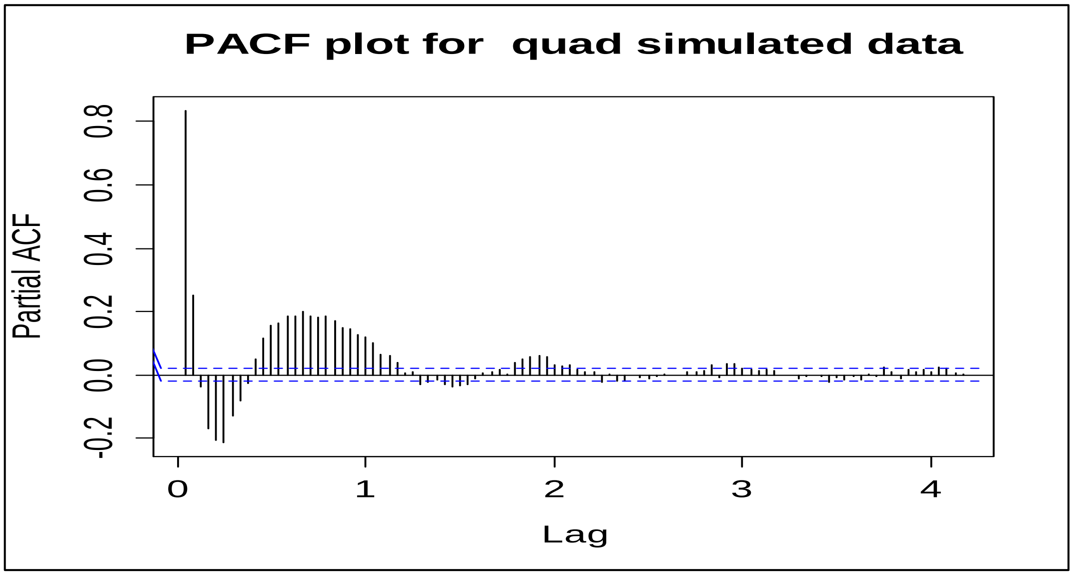

Figure 34b: PACF Plot for Simulated Data with Trend Component

The outputs for the simulated data with trend component time series models are as follows:

Sarima Model

The result from the *auto.arima* function for the SARIMA model without covariates as described earlier is for the simulated data is:

```
Series:sim quad
ARIMA(3,1,2)(2,0,2)[24]

Coefficients:
         ar1      ar2      ar3      ma1     ma2    sar1    sar2
sma1     sma2
      0.1123  -0.0319  -0.0177  -1.1028  0.1107  0.4542  0.5457  -
0.4089  -0.5364
s.e.  0.0253   0.0108   0.0107   0.0274  0.0273  0.0575  0.0575
0.0571   0.0554

sigma^2 estimated as 13.07:  log likelihood=-23620.39
AIC=47257.63   AICc=47257.66   BIC=47328.39
```

From the above output, the non-seasonal ARMA model has the following order; AR (3) and MA (2) with one differencing. The seasonal ARMA has AR(2) and MA(2) with zero differencing with seasonal lag of 24 (one day). For the non-seasonal components, The AR coefficients are 0.11, -0.0319 and -0.018 while the MA Coefficients are -1.1 and 0.11 respectively. For the seasonal components the AR coefficients are 0.45 and 0.55 also the MA coefficients are -0.41 and -0.54. This model was selected based on AIC.

The corresponding SARIMA model with covariates result is:

```
Series: sim quad
ARIMA(2,0,0)(1,0,0)[24] with non-zero mean

Coefficients:
         ar1     ar2    sar1  intercept
      0.0378  0.0065  0.0445    19.9127
s.e.  0.0108  0.0108  0.0108     0.2223

sigma^2 estimated as 13.06:  log likelihood=-23620.51

AIC=47329.02   AICc=47329.47   BIC=47640.33
```

From the above output, the non-seasonal ARMA model has the following order; AR(2) with no differencing. The seasonal ARMA has AR(1) with zero differencing with seasonal lag of 24 (one day) and non-zero mean. For the non-seasonal components, The AR coefficients are 0.038 and 0.0065. For the seasonal components the AR coefficient is 0.45. This model was selected based on AIC.

BATS and TBATS Model

To fit these models successfully, the trend option has to be specified. The BATS model that best suits our simulated data with trend is:

```
BATS(0.941, {0,0}, 1, {24,168})

Call: bats(y = sim.quad, use.trend = TRUE)

Parameters
  Lambda: 0.940917
  Alpha: 0.04138617
  Beta: 0.0009097658
  Damping Parameter: 1
  Gamma Values: 0.0012371 -1.569877e-08

Sigma: 3.025002
AIC: 101714.6
```

Lambda represents the Box-Cox transform which is 0.941 in this case and the smoothing parameters are alpha, beta and gamma parameters which are 0.041, 0.00091, 0.0012 and -0.000000016 respectively. The damping parameter for this model is 1 but there are no ARMA errors for this model. Due to the presence of trend in the data the trend option is used. Finally the seasonal periods are 24 representing daily cycle and 168 representing weekly cycles, with 194 estimated parameters.

For the TBATS model the corresponding output is;

```
TBATS(0.788, {0,0}, 1, {<24,5>, <168,5>})

Call: tbats(y = sim quad, use.trend = TRUE)

Parameters
  Lambda: 0.788415
  Alpha: 0.004433112
  Beta: 2.749054e-05
  Damping Parameter: 1
  Gamma-1 Values: 0.00211541 0.002081061
  Gamma-2 Values: -0.0001151335 0.0002147104

Sigma: 1.966455
AIC: 100807.6
```

Lambda represents the Box-Cox transform which is 0.788 in this case and the smoothing parameters are alpha, beta and gamma parameters which are 0.0044, 0.000027, 0.0021,0.002,-0.00012 and -0.00021 respectively. The damping parameter for this model is 1 but there are no ARMA errors for this model. Due to the presence of trend in the data the trend option is used. Finally the seasonal periods are 24 representing daily cycle and 168 representing weekly cycles, with 22 estimated parameters.

Factor Latent Model

We use the first half of the data to fit the factor model with hour of the day, day of the week and week of the year covariates. Same method is applied but with $K = 4$, and the CIIR process is also added.

The ACF and PACF plots for the residuals after fitting the latent factor model are shown below:

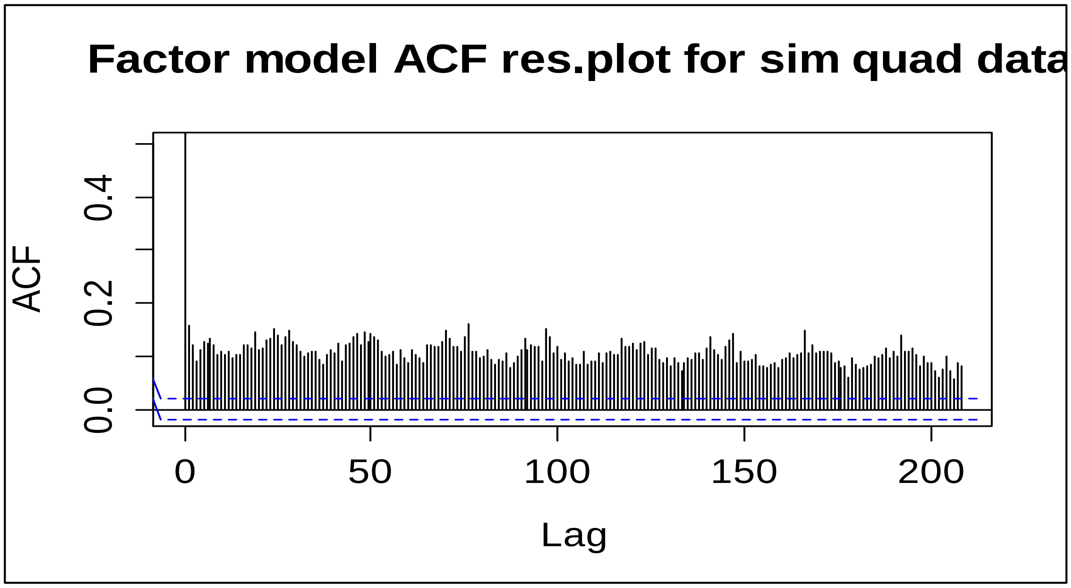

Figure 35a: ACF Plot of Factor Model Residuals for Simulated Data with Trend

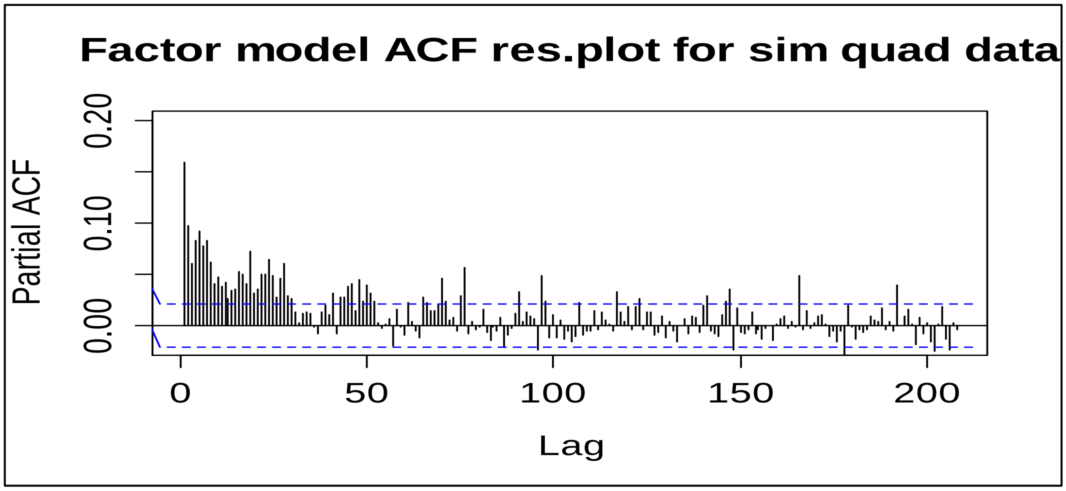

Figure 35b: PACF Plot of Factor Model Residuals for Simulated Data without Trend

The plots are similar but there is not information that can be deduced form them, we would fit a model including the CIIR process and see what improvement this might bring to our model.

Table 8: Simulated Data with Trend Forecast Evaluation Results

METHODS	BATS	TBATS	SARI	SARIREG	l.reg	Factor+CIIR	Factor
MAE	37.12	5.17	5.09	23.91	28.95	22.50	22.39
MSE	1780.20	38.10	35.75	604.17	917.01	549.14	547.59
RMSE	42.19	6.17	5.98	24.58	30.28	23.43	23.40

We observe that the SARIMA model without covariates performed the best followed by the TBATS model. BATS Model here performed worst this was due to the fact that it was able to detect the trend in the data but was not able to model the other seasonal patterns. From the plot it is observed that the factor model regression and SARIMA with covariates might over fit the data.

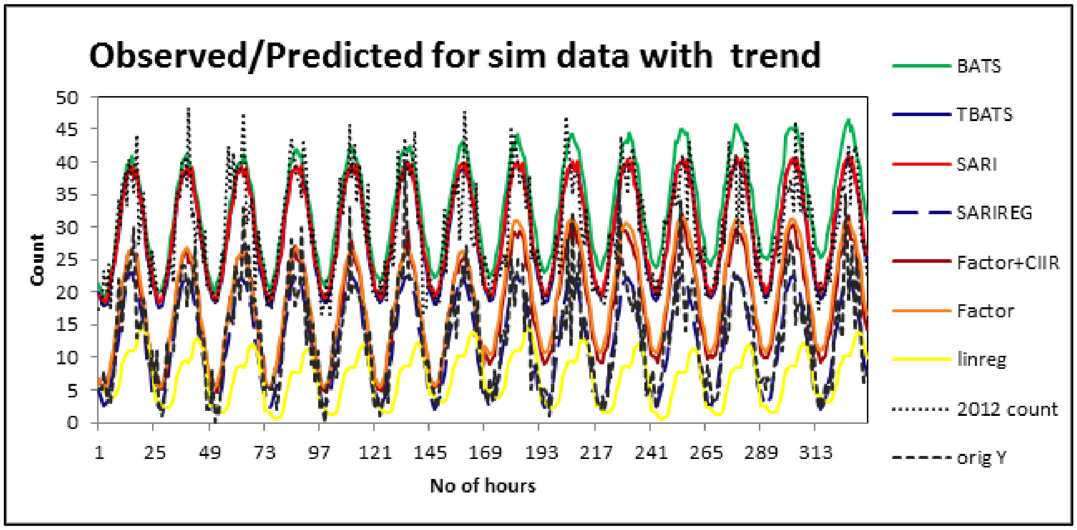

Figure 36: Predicted/Actual Simulated Data for Simulated Data with Trend

Chapter 7

SUMMARY

Here is a summary of the performance of the models utilized for our analysis.

Linear Regression

This method while it did not perform the best for any of our models it also was not the worst. Of all the methods applied, it is the easiest model to explain but the covariates have to be carefully defined. The residuals also need to be explored for any serial dependence that can still be extracted, which might improve forecast results. The adjusted R squared also plays a major role in determining how useful the residual analysis is; when the R squared is high even though there might still be serial dependence in the residuals, it might not improve our forecast.

Time Series Models

In all models, except the simulated data with trend ,the SARIMA model with covariates is an improvement from the SARIMA model without covariates. The *auto.arima* function in the *forecast* package in R has the ability to successfully capture trend according to AIC and AICc, while fitting the SARIMA model. The reason adding the covariates made the model worse might be it caused over fitting. Again the data needs to be examined carefully to determine suitable covariates.

The BATS model performed best in the cumulative data, urgent acuity data, simulated data without quadratic term and closely second to the TBATS method in the arrival data. The BATS model was pretty consistent in performing best but for the simulated data with quadratic term the trend option has to be specified, but the model over estimates the trend in the data. This seems to be the major drawback of the model.

The TBATS model did pretty well in estimating the simulated data with trend after the SARIMA model without covariates, the trend option also needs to be specified. For the arrival data it performed best and for urgent and cumulative data it performed second to the BATS model but didn't perform as good for the simulated data without the trend component.

Factor Model

The CIIR factor produced a significance improvement in only two models; the cumulative data and the simulated data with trend it was not necessary in all the other models. They performed best only in the simulated data without quadratic component. This method is not automated and requires the K to be determined manually.

In conclusion, the BATS and TBATS models performed consistently better that other models, is easily automated and does not include additional information or covariates. It also does not require residual analysis like the linear regression model and latent factor models. These models however, have a few drawbacks; they do not accommodate zeros values and so require a transformation, do not accommodate covariates and the trend option needs to be specified; *auto.arima* in R on the other hand has the ability to capture trend for a SARIMA model and covariates can be added to this model when necessary.

For the dynamic factor model, we need to align the data carefully to make sure that the factors for building the models and fitting the residuals must match. Also the number of factors *K* is decided manually before fitting the model, also residual analysis needs to be done to check for any serial dependence that can improve forecasts. The time structures for model building and forecasting should be the same. A note of caution for this model is it doesn't work well if there is any change in the pattern of our data like seen when forecasting the simulated data with quadratic component. A major drawback for this model is that it doesn't give confidence intervals for the predictions. Residual analysis is important for linear regression and factor models also the data has to be examined carefully to determine suitable covariates.

The performance of this research will be evaluated on how well we are able to answer the following questions.

- Can patient arrival volume be predicted accurately? Yes, this can be done fairly accurately.
- Using the same methods for predicting patient arrival, can cumulative patient volume also be accurately forecasted? Yes, this can also be done adequately.
- How much data is required to make the most accurate predictions? Three years of data produced the most accurate predictions.
- How accurate will six months predictions be? Six months forecasts perform comparably to one week forecasts.

- Which method(s) is most suitable for our data? BATS and TBATS were most consistently the best models and they are easily automated and do not require covariates.
- Can we predict urgent acuity patient arrival volume? Yes, this can be done satisfactorily.
- What forecast methods can handle multi seasonality? Fitting the time series with *msts* helps the models handle multiseasonalities better
- If there is a trend (steady decline or increase) in the data which forecasts method will most successfully capture it? TBATS and SARIMA were better suited for depicting trend.
- How easily can these methods be implemented in the ED? Time series methods are easily automated, residual analysis need to be done manually and this makes linear regression adds a layer of difficulty and dynamic latent factor model is not easily automated because the function is not yet automated in R and also K needs to be set manually.

The suggested procedure for analysis is as follows:

First, at least two years of data is collected to be used for analysis, though having three or more years of data to build models is likely to increase forecast accuracy.

Next, preliminary analysis like plots, descriptive statistics and other data exploration techniques should be carried out on the data to identify patterns, trends and outliers. This is vital in setting up research goals while also defining covariates.

Then, the data is divided into two parts; test and validation portion. The most recent year data is used for validation and the earlier portion is used for building the models, after which the most preferred model is then selected based on performance.

Finally, the data is now updated to include most recent observations (validation portion) and used to generate forecasts for six months ahead. It is recommended that the process be reevaluated every six months also; the performance of these models should be closely tracked.

REFERENCES

REFERENCES

1. Anonymous**The Emergency Medical Treatment and Active Labor Act**, as established under the Consolidated Omnibus Budget Reconciliation Act (COBRA) of 1985 (42 USC 1395 dd). 1194. Federal Register, 59:32086–12. *COBRA* **1985**.

2. Binsfeld, M.; Buffalo hospital Buffalo,MN. Forecasting need of Emergency department in Buffalo Hospital (interview). **2012**.

3. Box, G.; Jenkins, G.; Reinsel, G. Time series analysis: forecasting and control. **1994**.

4. Bretthauer, K. M.; Cŏté, M. J. A model for planning resource requirements in health care organizations. *Decision Sciences* **1998***, 29*, 243-270.

5. Chatfield, C. *Time-series forecasting;* Chapman and Hall/CRC: **2002**.

6. Chatfield, C. The Holt-Winters Forecasting Procedure. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **1978***, 27*, 264.

7. Cinlar, E. Superposition of point processes, Stochastic Point Processes: Statistical Analysis, Theory and Applications PAW Lewis, 549–606. **1972**.

8. Cox, D. D. R.; Isham, V. *Point processes;* Chapman & Hall/CRC: **1980**; Vol. 12.

9. De Livera, A. M.; Hyndman, R. J.; Snyder, R. D. Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association* **2011***, 106*, 1513-1527.

10. Donaldson, N.; Shapiro, S. Impact of california mandated acute care hospital nurse staffing ratios: a literature synthesis. *POLICY POLIT NURS PRACT* **2010***, 11*, 184-201.

11. Doob, J. L. Stochastic Processes and Statistics. *Proc. Natl. Acad. Sci. U. S. A.* **1934***, 20*, 376-379.

12. Faraway, J. J. *Linear models with R;* Chapman and Hall/CRC: **2004;** Vol. 63.

13. Gilboy, N.; Tanabe, P.; Travers, D.; Rosenau, A.; Eitel, D. Emergency severity index, version 4: implementation handbook. *Rockville, MD: Agency for Healthcare Research and Quality* **2005**, 1-72.

14. Gould, P. G.; Koehler, A. B.; Ord, J. K.; Snyder, R. D.; Hyndman, R. J.; Vahid-Araghi, F. Forecasting time series with multiple seasonal patterns. *Eur. J. Oper. Res.* **2008***, 191*, 207-222.

15. Habasevich, B. Defining Acuity. **2012**.

16. Henderson, S. G. In *In Should we model dependence and nonstationarity, and if so how?* Proceedings of the 37th conference on Winter simulation; Winter Simulation Conference: **2005**; , pp 120-129.

17. Hyndman, R. J. subset. ts 53. *Package 'forecast'* **2011**, 53.

18. Hyndman, R. J.; Koehler, A. B. Another look at measures of forecast accuracy. *Int. J. Forecast.* **2006***, 22*, 679-688.

19. Hyndman, R. J.; Koehler, A. B.; Snyder, R. D.; Grose, S. A state space framework for automatic forecasting using exponential smoothing methods. *Int. J. Forecast.* **2002***, 18*, 439-454.

20. Hyndman, R.; Koehler, A. B.; Ord, J. K.; Snyder, R. D. *Forecasting with exponential smoothing: the state space approach;* Springer: **2008**.

21. Institute of Medicine (US). Committee on the Future of Emergency Care in the United States Health System *Hospital-Based Emergency Care: At the Breaking Point;* National Academy Press: **2007**.

22. Jones, S. S.; Thomas, A.; Evans, R. S.; Welch, S. J.; Haug, P. J.; Snow, G. L. Forecasting daily patient volumes in the emergency department, Academic Emergency Medicine, **2008.**

23. Kalekar, P. S. Time series forecasting using Holt-Winters exponential smoothing. *Kanwal Rekhi School of Information Technology* **2004**.

24. Kam, H. J.; Sung, J. O.; Park, R. W. Prediction of Daily Patient Numbers for a Regional Emergency Medical Center using Time Series Analysis. *Healthcare informatics research* **2010***, 16*, 158-165.

25. Mahoney, J.; Buffalo hospital Emergency department registered nurse, Buffalo, MN. Interview, **2012**.

26. Makridakis, S.; Wheelwright, S. C.; Hyndman, R. J. *Forecasting methods and applications;* John Wiley & Sons: **2008**.

27. Matteson, D. S.; McLean, M. W.; Woodard, D. B.; Henderson, S. G. Forecasting emergency medical service call arrival rates. *The Annals of Applied Statistics* **2011***, 5*, 1379-1406.

28. Ord, J. K.; Koehler, A.; Snyder, R. D. Estimation and prediction for a class of dynamic nonlinear statistical models. *Journal of the American Statistical Association* **1997**, *92*, 1621-1629.

29. Page, A. *Keeping patients safe: Transforming the work environment of nurses;* National Academy Press: **2004**.

30. Rathlev, N. K.; Obendorfer, D.; White, L. F.; Rebholz, C.; Magauran, B.; Baker, W.; Ulrich, A.; Fisher, L.; Olshaker, J. Time Series Analysis of Emergency Department Length of Stay per 8-Hour Shift. *Western Journal of Emergency Medicine* **2012**, *13*, 163.

31. Rencher, A. C.; Schaalje, G. B. *Linear models in statistics;* Wiley-Interscience: **2008**.

32. Rico, F. A. University of South Florida, **2009**.

33. Schweigler, L. M.; Desmond, J. S.; McCarthy, M. L.; Bukowski, K. J.; Ionides, E. L.; Younger, J. G. Forecasting models of emergency department crowding. *Acad. Emerg. Med.* **2009**, *16*, 301-308.

34. Shen, H.; Huang, J. Z. Forecasting time series of inhomogeneous Poisson processes with application to call center workforce management. *The Annals of Applied Statistics* **2008**, 601-623.

35. Shumway, R. H.; Stoffer, D. S.; Stoffer, D. S. *Time series analysis and its applications;* Springer New York: **2000**; Vol. 549.

36. Sun, Y.; Heng, B.; Seow, Y.; Seow, E. Forecasting daily attendances at an emergency department to aid resource planning. *BMC Emergency Medicine* **2009**, *9*, 1.

37. Tandberg, D.; Qualls, C. Time series forecasts of emergency department patient volume, length of stay, and acuity. *Ann. Emerg. Med.* **1994**, *23*, 299-306.

38. Tang, N.; Stein, J.; Hsia, R. Y.; Maselli, J. H.; Gonzales, R. Trends and characteristics of US emergency department visits, 1997-2007. *JAMA: the journal of the American Medical Association* **2010**, *304*, 664-670.

39. Taylor, J. W. Triple seasonal methods for short-term electricity demand forecasting. *Eur. J. Oper. Res.* **2010**, *204*, 139-152.

40. Taylor, J. W. Short-term electricity demand forecasting using double seasonal exponential smoothing. *J. Oper. Res. Soc.* **2003**, *54*, 799-805.

41. Trzeciak, S.; Rivers, E. Emergency department overcrowding in the United States: an emerging threat to patient safety and public health. *Emergency medicine journal* **2003**, *20*, 402-405.

42. Willmott, C. J.; Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research* **2005**, *30*, 79.

43. Wood, S. mgcv: GAMs with GCV/AIC/REML smoothness estimation and GAMMs by PQL. *R package version* **2010**, 1.6-2.

44. Wood, S. N. Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **2003**, *65*, 95-114.

45. Wood, S.; Wood, M. S. The mgcv package. *www.r-project.org* **2007**.

46. Wright, P. D.; Bretthauer, K. M.; Côté, M. J. Reexamining the nurse scheduling problem: Staffing ratios and nursing shortages*. *Decision Sciences* **2006**, *37*, 39-70.

APPENDIX

TIME SERIES MODEL

```
#### R code for time series models #############################

#### Data is loaded into R #####################################

### Preliminary time series models #############################

### Data is plotted using the ts function ######################


pat.arr.ct=ts(hourly[,3],freq=24)



###we fit an AR, MA and ARMA model using the auto.arima function #########

mod.ar = auto.arima(pat.arr.ct, max.p=200, max.q=0,               max.P=0,
max.Q=0, max.order=5, start.p=2, start.q=2,                  start.P=1,
start.Q=1, stationary=FALSE, seasonal=TRUE)

mod.ma = auto.arima(pat.arr.ct, max.p=0, max.q=200, max.P=0, max.Q=0,
max.order=5, start.p=2, start.q=2,            start.P=1, start.Q=1,
stationary=FALSE, seasonal=TRUE)

mod.arma = auto.arima(pat.arr.ct, max.p=200, max.q=200,
max.P=0, max.Q=0, max.order=5, start.p=2, start.q=2,
start.P=1, start.Q=1, stationary=FALSE, seasonal=TRUE)



#### We forecast for up to one year ahead ####################

pred.ar=forecast (mod.ar,h=8736)

pred.ma=forecast (mod.ma,h=8736)

pred.arma=forecast (mod.arma,h=8736)



##### We combine all our predictions

arma.res=cbind(as.vector(pred.ma$mean)-1,as.vector(pred.ar$mean)-
1,as.vector(pred.arma$mean)-1,as.vector(predy[1:8736,3]))

colnames(arma.res)=c("MA","AR","ARMA","OBS")

################################################################

#### Data is fitted as a multi seasonal time series using the msts command

#### 24 for daily cycle
```

```
#### 168 for weekly cycle

#### 1 is added to the series due to BATS and TBATS restrictions

 patient.arr.ct=msts(hourly[,3]+1, seasonal.periods=c(24,168),
ts.frequency=24)


####ARIMA model without covariates ###########

fit.mod=auto.arima (patient.arr.ct)


####ARIMA model with covariates ###################

fit.mod.reg=auto.arima(patient.arr.ct,xreg=hourly)

### BATS and TBATS model ############################

bats.mod=bats (patient.arr.ct)

tbats.mod=tbats (patient.arr.ct)


####### We now forecast for one year ahead ####################


pred.sarim=forecast (fit.mod,h=8736)

pred.sarim=forecast (fit.mod,h=8736, xreg=hourly)

pred.bats=forecast (bats.mod,h=8736, level=c(80,95))

pred.tbats=forecast (tbats.mod,h=8736, level=c(80,95))


####### We combine all the time series predictions######

###### We subtract 1 that was added earlier ############


arr.ts.pred=cbind(as.vector(pred.bats$mean)-1,as.vector(pred.tbats$mean)-
1,as.vector(pred.sarim$mean)-1,as.vector(pred.sarim.reg$mean)-
1,as.vector(predy[1:8736,3]))

colnames(fore4)=c("BATS","TBATS","SARI","SARI+REG","OBS")
```

```
######### Calculating Residuals#######

res.arr.ts.pred = cbind(as.vector(arr.ts.pred[,1]-
arr.ts.pred[,5]),as.vector(arr.ts.pred[,2]-
arr.ts.pred[,5]),as.vector(arr.ts.pred[,3]-
arr.ts.pred[,5],as.vector(arr.ts.pred[,4]-arr.ts.pred[,5])))


######## MSE and RMSE##############

mean(res.arr.ts.pred[,1]^2); sqrt(mean(res.arr.ts.pred[,1]^2))

mean(res.arr.ts.pred[,2]^2); sqrt(mean(res.arr.ts.pred[,2]^2))

mean(res.arr.ts.pred[,3]^2); sqrt(mean(res.arr.ts.pred[,3]^2))

mean(res.arr.ts.pred[,4]^2); sqrt(mean(res.arr.ts.pred[,4]^2))


##### MAE##################

mean(abs(res.arr.ts.pred[,1]))

mean(abs(res.arr.ts.pred[,2]))

mean(abs(res.arr.ts.pred[,3]))

mean(abs(res.arr.ts.pred[,4])
```

FACTOR MODEL R CODE

```
######### The patient arrival data set is loaded into R ########

#################### It has 5 columns and 8736 observations ####

#################### Column one contains date ##################

######## Column two is day of the week ranging from 1 to 7 ####

######## Column three is week of the year ranging from 1 to 52##

### Column four contains hour of the day ranging from 1 to 24###

###### Column five is the actual y value labeled y #############


T = 24*7*52

hour = hosp[,4]

day = rep(1:(7*52), each = 24)

dofw = hosp[,2]

week = hosp[,3]

#########################################################

y=hosp[,5]

head(y)

D = length(y)/ (N); D # number of "days"

ND = length(y) # total number of observations

ND


dofwindex = as.factor(dofw)

weekindex = as.factor(week)

Y = t(matrix(y,N,D))

DoW = t(matrix(dofw,N,D))

WEEK = t(matrix(week,N,D))


########################################
```

FACTOR MODEL

```
########################################
#### The main estimation algorithm for
#### fitting the K-factor model
#### using constraints and
#### smoothing splines
########################################
K.max = 3
muhat = matrix(0,N*D,K.max)
Max.iter = 40
# Set exit level for relative reduction in deviance
dev.exit = 0.0001

for(k in 1:K.max){
  ########################################
  # Initialization:
  dim(Y);  min(Y);  min(ifelse(Y==0,0.01,Y))
  gY = log(ifelse(Y==0,0.01,Y))
  gYsvd = svd(gY)

  # coefs
  B.new = matrix(0,D,k)
  for(i in 1:k){  B.new[,i] = gYsvd$d[i]*gYsvd$u[,i]  }

  #factors
  F.new = matrix(0,N,k)
  for(i in 1:k){  F.new[,i] = gYsvd$v[,i]  }
  ########################################
  # Begin iterative algorithm
  iter = 1
  dev.new = Inf

  while(iter < Max.iter){
    tic = proc.time()[3]

    dev.old = dev.new
    F.old = F.new
    B.old = B.new

    ########################################
    X.temp = matrix(0,ND,k)
    for(kk in 1:k){ X.temp[,kk] = rep(F.old[,kk],D) }

    xnam <- paste(paste("s(as.numeric(weekindex),by = X.temp[,", 1:k,
sep=""), "],bs='cc')", sep = "")
    fmla <- as.formula(paste("y ~ -1 + X.temp:dofwindex +", paste(xnam,
collapse= "+")))

    fit6 = gam(fmla, family = poisson)
    B.tempD = matrix(as.vector(fit6$coefficients[1:(7*k)]), 7, k,
byrow=TRUE)

    # Extracting fitted values
    n = 52 # number of weeks in the year
    S = NULL
    for(s in 1:k){

      raw <- fit6$model[fit6$smooth[[s]]$term]

      xx <- seq(min(raw), max(raw), length = n)

      by <- rep(1, n)

      dat <- data.frame(x = xx, by = by)

      names(dat) <- c(fit6$smooth[[s]]$term, fit6$smooth[[s]]$by)

      Xmat <- PredictMat(fit6$smooth[[s]], dat)

      first <- fit6$smooth[[s]]$first.para

      last <- fit6$smooth[[s]]$last.para

      p <- fit6$coefficients[first:last]

      S.temp <- Xmat %*% p

      S = c(S,S.temp)

    }


    B.tempW = matrix(as.vector(S), 52, k, byrow=FALSE)


    ########################################

    B.temp = matrix(0, D, k, byrow=TRUE)

    rm(fit6)


    # 7 days in the week

    for(j in 1:7){

      for(ell in 1:k){

        B.temp[which(DoW[1:D,1] == levels(dofwindex)[j]),ell] =
B.temp[which(DoW[1:D,1] == levels(dofwindex)[j]),ell] +
as.numeric(B.tempD[j,ell])

      }

    }
    for(j in 2:53){
      for(ell in 1:k){
        B.temp[which(WEEK[1:D,1] == levels(weekindex)[j]),ell] = 
B.temp[which(WEEK[1:D,1] == levels(weekindex)[j]),ell] + 
as.numeric(B.tempW[(j-1),ell])
      }
    }

    Z.temp = matrix(0,ND,k)
    for(kk in 1:k){ Z.temp[,kk] = rep(B.temp[,kk],each=N) }

    ########################################

    znam <- paste(paste("s(hour,by = Z.temp[,", 1:k, sep=""), "])", sep = 
"")
    fmla <- as.formula(paste("y ~ -1 +", paste(znam, collapse= "+")))

    fit4 = gam(fmla, family = poisson)

    # Extracting fitted values
    n = 24 # 24 hours per day
    S = NULL

    for(s in 1:k){
      raw <- fit4$model[fit4$smooth[[s]]$term]
      xx <- seq(min(raw), max(raw), length = n)
      by <- rep(1, n)
      dat <- data.frame(x = xx, by = by)
      names(dat) <- c(fit4$smooth[[s]]$term, fit4$smooth[[s]]$by)
```

```
  Xmat <- PredictMat(fit4$smooth[[s]], dat)
  first <- fit4$smooth[[s]]$first.para
  last <- fit4$smooth[[s]]$last.para
  p <- fit4$coefficients[first:last]
  S.temp <- Xmat %*% p
  S = c(S,S.temp)
}

F.temp = matrix(as.vector(S),N,k, byrow=FALSE)

# Save most recent fit before orthogonalization
fit.final = fit4
rm(fit4)

########################################
# Orthogonalize Factors F
G.temp = B.temp %*% t(F.temp)
Gsvd = svd(G.temp)

B.new = matrix(0, D, k)
for(i in 1:k){  B.new[,i] = Gsvd$d[i]*Gsvd$u[,i]  }
F.new = matrix(0,N,k)
for(i in 1:k){  F.new[,i] = Gsvd$v[,i]  }

dev.new = fit.final$deviance
if(0 < dev.old - dev.new & dev.old - dev.new < dev.exit) iter = Inf

toc = proc.time()[3] - tic ; toc
```

```
    # optional print statements
    print(c(iter, toc/60))
    iter = iter + 1
    print(fit.final$deviance)
    flush.console()

  }

  muhat[,k] = fit.final$fitted

  # optional print statements
  #print(k)
  #print(summary(F.old - F.new)) ;
  #print(max(abs(F.old - F.new)))
  #print(summary(B.old - B.new)) ;
  #print(max(abs(B.old - B.new)))
  #print(round(crossprod(F.old,F.new),4))
  #print(diag(round(crossprod(F.old,F.new),4)))

}

# fitted values in vector form (same length as y) for k = K.max
index = seq(1,24,by=1)
mu.hat = numeric(ND)

for(i in 1:D){

  mu.hat[((i-1)*N+1):(i*N)] = as.vector(exp(F.new[index,]%*%B.new[i,]))
```

```
}


# multiplicative residual

Et = y/mu.hat


# a couple residual plots

par(mfrow=c(1,1))

ts.plot(Y[1:500])

acf(y,lag.max=100,main="ACF plot for Y")

pacf(y,lag.max=100,main="PACF plot for Y")

ts.plot(Et[1:500],main="error time series plot") ; abline(h = 1)

acf(Et, ylim=c(-0.01,0.7), lag.max = 96*2+16,main="ACF plot for mu err for
arr.vol")

pacf(Et, ylim=c(-0.01,0.1), lag.max = 96*2+16,main="PACF plot for mu err
for arr vol")

abline(v = c(96.6, 192.6), lty = 2, col = 2)

acf(Et, ylim=c(-0.02,0.1), lag.max = 50, type = "partial")

abline(v = c(96.6, 192.6), lty = 2, col = 2)

##############################################################

# if some missing days were removed use 'misshour' below

# to reinitilize the conditional likelihoods below

misshour = c(1, ifelse(diff(day) > 1 , 1, 0))

sum(misshour)


##############################################################

######## For conditional ML estimation of              #######

######## Int-GARCH(1,1)                                #######

##############################################################

"condPoissonInt11" = function(parms, y, mu, misshour, llik){
  alpha = parms[1]
  beta  = parms[2]
  omega = 1 - alpha - beta
  N = length(y)
  lambda = numeric(N)
  eta = numeric(N)
  epsilon = y/mu
  eta[1] = 1
  lambda[1] = 1
  loglik = 0 # -sum(lfactorial(y))
  for(i in 2:N){
    eta[i] = omega + alpha*epsilon[(i-1)] + beta*ifelse(misshour[(i-1)] ==
1, 1, eta[(i-1)])
    lambda[i] = mu[i]*eta[i]
    #    if(lambda[i] <= 0){print(c(i,lambda[i],alpha,beta))}
    temp = -lambda[i] + y[i]*log(lambda[i]) - (lfactorial(y[i]))
    loglik = loglik + ifelse(misshour[i] == 1, 0, temp)
  }
  if(llik==TRUE){-loglik}
  else{eta}
}
################################################################
theta.0 = c(0.05, 0.5)

condPoissonInt11(parms = theta.0, y = y, mu = mu.hat, misshour = misshour,
llik = TRUE)

outInt11 = optim(par=theta.0, fn = condPoissonInt11, y = y, mu=mu.hat,
llik=TRUE,
                 misshour= misshour, method = "L-BFGS-B", lower =
c(0.000,0.000),
                 upper = c(0.2,0.9), hessian=T, control =
                   list(trace = TRUE, ndeps = rep.int(0.000001, 2),
                        maxit = 200L, factr = 1e+31, pgtol = 0))


# parameter estimates
igparInt11 = outInt11$par ; igparInt11 ; 1 - sum(igparInt11)


# approximate SEs
igseInt11 = sqrt(diag(solve(outInt11$hessian))) ; igseInt11


# CIIR
etaInt11 = condPoissonInt11(parms = outInt11$par,y= y, mu=mu.hat,
misshour= misshour, llik=FALSE)


# Mltiplicative residuals
e = y/mu.hat


# Fitted values
lambdaInt11 =mu.hat*etaInt11
length(lambdaInt11)
pred.factor=cbind(as.vector(lambdaInt11),as.vector(mu.hat),as.vector(predy
[1:8736,3]))
colnames(pred.factor)=c("lamda","mu.hat","OBS")


######### Residuals are calculated#######
res.factor=cbind(as.vector(pred.factor[,1]-
pred.factor[,3]),as.vector(pred.factor[,2]-pred.factor[,3]))
######## MSE and RMSE##############
```

```
mean(res.factor[,1]^2); sqrt(mean(res.factor[,1]^2))

mean(res.factor[,2]^2); sqrt(mean(res.factor[,2]^2))


##### MAE##################

mean(abs(res.factor[,1]))

mean(abs(res.factor[,2]))
```

REGRESSION CODE

```
########   Patient arrival Data is loaded into R ##############


pat.arr.reg=read.csv("C:\\Users\\utchay\\Dropbox\\reg1.csv", header=T)

predy=read.csv("C:\\Users\\utchay\\Dropbox\\hourly2reg.csv", header=T)


######## Fitting regression model ####################


arr.reg=lm(patient.count~0+ ., data=pat.arr.reg)

summary(arr.reg)


######### checking residuals plots  ###################


plot(arr.reg$res[1:1000],type="l", main=" residual plot for regression
model", ylab="count", xlab="lags")

abline(h=0)


res.arr.reg=ts(arr.reg$res,start=1, freq=1)

acf(res.arr.reg, lag.max=100,main="ACF plot for reg residual",ylab="
count")

pacf(res.arr.reg, lag.max=100,main="PACF plot for reg residual",ylim=c(-
.1,.2))


########## Fitting an ARIMA model for the regression residuals
############


reg.res.mod=auto.arima(res.arr.reg,ic="aicc",d=0,D=0,max.p=10,max.q=10)

reg.res.mod
```

```
########## Predicting up to one year ahead
################################


pred.res.reg=forecast(reg.res.mod,h=8736,level=c(80,95))




########### (Adding the time series residuals prediction to the regression
predictions ###############

fore.reg=cbind(as.vector(pred.res.reg$mean[1:8736]),as.vector(arr.reg$fit[
1:8736]))

colnames(fore.reg)=c("reg", "res")

fore.reg$pred=as.vector(fore.reg[,1] + fore.reg[,2])

dim(fore.reg)

reg.pred=apply(fore.reg,1,sum)

dim(reg.pred)


###### Extracting residuals #################

reg.pred.res=cbind(as.vector(reg.pred[4368]),as.vector(predy[1:4368,3]))

reg.res= apply(reg.pred.res,1,sum)


mean(reg.res^2); sqrt(mean(reg.res^2))


##### MAE##################

                            mean(abs(reg.res))
```