Time series forecasting of hospital Inpatients and Day case waiting list using ARIMA, TBATS and Neural Network Models

MSc Research Project
Data Analytics

Raj Kumar Tamatta
x17106966

School of Computing
National College of Ireland

Supervisor: Noel Cosgrave

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Raj Kumar Tamatta
Student ID:	x17106966
Programme:	Data Analytics
Year:	2018
Module:	MSc Research Project
Lecturer:	Noel Cosgrave
Submission Due Date:	13/08/2018
Project Title:	Time series forecasting of hospital Inpatients and Day case waiting list using ARIMA, TBATS and Neural Network Models
Word Count:	5650

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	17th September 2018

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Time series forecasting of hospital Inpatients and Day case waiting list using ARIMA, TBATS and Neural Network Models

Raj Kumar Tamatta
x17106966
MSc Research Project in Data Analytics

17th September 2018

Abstract

This research aims to provide the forecasting of patients waiting list in different time band over all Ireland's Hospitals. The NTPF collect and manage waiting list data of patients from all hospitals of Ireland. The data is used to build forecasting model, first data is decomposed to identify the pattern of on time series like trending, seasonality, variation and de-seasonalize to remove irregularity and noise from time series data. Based on related work and literature reviews, all relevant time series algorithms are evaluated to choose best time series model. The ARIMA, TBATS and AR-NN are suitable algorithm for this research which is further compare with least RMSE value. The performance of each algorithm is depended on different time series data. The ARIMA, TBATS and AR-NN model were applied on five different time series forecasting cases like overall waiting list of patients, waiting list of patients in different 0-3, 3-6, 6-12 and 12 + time band. The performance of ARIMA model found to be better in most of the case, where time series data has constant seasonal pattern, TBATS model is better where data having irregular seasonal pattern. As this is linear time series problem the performance of AR-NN neural network model not great. In overall comparison of all three models, ARIMA is the best model with least RMSE value.

keyword: ARIMA, TBATS, Seasonal Decomposition, Time series, Artificial Neural Network, Hospitals waiting list.

1 Introduction

The aim of this research is to forecast the waiting list of patients in Ireland hospital over different specialty by examine data published by national treatment purchase fund (NTPF). In this research, monthly patients waiting list data from Jan-2014 to Jun-2018 has been utilized. Based on historical data, time series prediction has been performed, which gives the knowledge of expected number of patients waiting list in future. Health service is the top priority for all nations, they have department of health a regulation body of government, who ensure to provide best health care for their citizen. Different government bodies and NGOs are always trying to improve effective health service from

all aspect like medication improvement, treatment scheduling, managing efficient follow of patient waiting list as per criticality of treatment, etc. The outcome of this research will help to understand the number of patient waiting for treatment in different specialty in future. This research has used, time series model Autoregressive Integrated Moving Average (ARIMA), TBATS, Autoregressive Neural Network (AR-NN) for forecasting of patient waiting list.
The effective operations of treatment means, getting right treatment in right time. Due to unexpected increase in the volume of patients waiting list, bring pressure on hospital management and administration for scheduling plan to treat each and individual patients. Medicine, equipments and technologies are the major requirement for treatment. Some medicine and equipment are very expensive and unavailable locally, where hospital needs some time to organize all necessary elements. The forecasting model help to understand the medical need for future, which helps the hospitals administration and government bodies to pre-plan the availability of all the necessary resources for treatment like bed, medicine, medical equipment, specialist doctor, nurse, other manpower, etc. To identify the future medical demand, this research is conducted to forecast the monthly waiting list of patients in various specialty by applying time series forecasting models.

1.1 Background

The department of Health of Ireland has mission to keep people healthy, provide healthcare, deliver improved high-quality service by developing best value resources. The project Ireland 2040 is the great initiative of government of Ireland to develop better country for all, which will showcase best aspire of country. Health service is major programme of investment, that reflect to best healthcare outcomes, which is to be achieved by reorienting primary and community health care services by providing high quality acute and emergency care with appropriate acute hospital setting. (https://health.gov.ie/) . The current major problem is numerous waiting list of patients in different hospital in various time band.

1.1.1 National waiting list of patients

The national treatment purchase fund (NTPF) is the corporate body, that closely work with acute public hospital, private nursing homes across Ireland, the HSE, department of health. NTPF currently gathering and examine data in respect of Inpatient, Day Case, Planned Procedure (IDPP) and Outpatient (OP) Waiting Lists. NTPF provide set of guidelines to ensure standard and people friendly approach of management and patient waiting list scheduling for treatment consistently within each hospital and across a hospital group. The motivation behind this protocol is to ensure the safe and effective access and treatment of patient on time in a fair manner. The bottom up, top down collaborative approach is involved to develop this protocol for stakeholder engagement and central focus on patient. In this approach Department of health (DOH), Special Delivery Unit (SDU), Acute Hospitals Division, Clinical Care Programmes (CCP), each hospital group and individual hospitals has been involved. The Minimum data set (MDS) developed in 2017, which intended to integrate IDPP waiting list management protocol by operational changes. The proven lean six sigma (LSS) was utilized to develop protocol for healthcare methodology, data from all the relevant stakeholder were collected and analyse, to ensure the voice of customer (VOC) is captured. The reason behind to implement lean six sigma (LSS) approach to provide improve clinical process, identify and remove waste practice, enable staff to improve their quality, safety and efficiency of the work by examine their practice. [1] /

1.1.2 Time Series Forecasting of Patient Waiting list

The time series forecasting is powerful statistical and machine learning tool, which examine the historical data of related domain and predict the outlook of business. The time series prediction are hugely used in finance, marketing, sales and other government agencies. In sales and finance industry different time series model were used to predict the price of stock, which helps stock buyer to invest in right stock. In a similar way, government agency uses time series model to forecast the environmental condition like rain, air quality, agriculture etc. which helps government to take remedial actions for disaster management. Time series forecasting has great influence in everyones life, which helps in decision making process to prevent from future cause. In this research, time series prediction model is being used to forecast the waiting list of patients in overall and various major specialty, which aim to support planning of patients waiting list based on forecasted value. This forecasting model gives an idea of increasing number of patients and medical demand. This research is conducted by utilizing some popular time series model like ARIMA, TBATS and AR-NN.

1.2 Scope and Objective

1.2.1 Scope

The historical data of patients waiting list is analyzing to identify the existing correlation, determine the time series forecasting model, and to predict the future number of patients waiting list.

1.2.2 Objective

In this research, the waiting list of patients data from different hospital across Ireland were used, which is managed by NTPF and available in their website from Jan 2014 to June 2018. The data are further collated to examine correlation, perform time series analysis to build patient waiting list forecasting model.

1.3 Research Question

The aim of this research is:
With what accuracy different time series model forecasting the number of patients waiting for treatment across Ireland in different time band?
Which time series model will outperform as compare with another model to forecast patients waiting list?

[1] NTPF Report website `http://www.ntpf.ie/home/pdf/National/`

Table 1: Description of variable.

Data	Description
Date	Date of archival.
Hospital_Group	This is referring to the categorical variable, having hospital group name.
Hospital_HIPE	This is also referring to the categorical variable have HIPE (Hospital In-patient Enquiry) number of hospital across of Ireland.
Hospital_Name	This is referring to name of hospital across of Ireland in different group.
Specialty_HIPE	This is referring to the HIPE number of various specialty.
Specialty_Name	This referring to the Specialty name.
Case_Type	This referring to categorical variable having value of case type.
Adult_Child	This variable referring to binary variable, which categorize the patient into Child and Adult.
Age_Profile	This is categorical variable refer to age of patient which is categorize into different age band.
Time_Band	This also a categorical variable, which show the different waiting time band in a month.

1.4 Data Description

The NTPF closely working with various hospital across Ireland, which collect and manage waiting list of patients datasets and publish in their website for public. This is monthly time series data, which is available from Jan 2014 to Jun 2016 in csv format consist of 211074 rows having 54 observations for this research.

1.4.1 Dataset

This research is conducted on Inpatient/Day case waiting list:
`http://www.ntpf.ie/home/inpatient_group.htm`

1.4.2 Column of Data

The data consist of following attributes:

In this research, total number of patient waiting list is dependent variable, which is under goes to develop time series forecasting model in different time series over different specialty.

1.5 Research Overview

This research paper cover following section.
Background: In this section, the overview of research has been provided including motivation, background and definition of research variable.
Literature Review: The examination of previous and recent related work in similar domain, which included analysis and forecasting of time series data, and supporting statistical and machine learning technique for this research.
Experimental Setup: In this part of research paper, the entire related experimental work and description of test implementation were clearly described.
Result and Evaluation: This presentation of finding result from experimental setup.

Conclusion: In conclusion overall argument of research project is describe and provide justification of further research in this area.

2 Related Work

The enormous number of IPDC patients are in waiting list for a long time in a different hospitals for their medical treatment, there are number of related work carried out to determine the impact of unexpected huge waiting time on patients and hospital operations. There are various machine learning techniques are used to forecast the expected volume of patients in different department of hospital, which helps hospital management to be prepared with necessary plan and equipment to carried out smooth operation.

- The impact of waiting time on satisfaction of patients.
- Machine learning technique for waiting time forecasting.

2.1 The impact of waiting time on satisfaction of patients

The waiting time has directly impact on the satisfaction of patients, which is treated as most important factor of quality of service. The research was carried out in Mulago assessment center Uganda, where real time data and 401 patients views were captured base on 5 interview questions, it observed that patients spent 95 percent of their time in waiting and 5 percent with the health staffs. Conrad (2013). As availability and accessibility are the most important factor for effective health care service. In Canada, the access of family physician in time were challenging, which lead longer waiting time for patients to get appointment. Due to longer waiting time, patients left to use emergency department Ansell, Crispo, Simard and Bjerre (2017). The long waiting time in health care center, attract the attention of public due to its adverse effect on patients satisfaction. The cross-section time study and questionnaire interview conducted in major teaching hospital of China on endocrinology outpatients. The patients are dissatisfied more by waiting time than sociocultural atmosphere Xie and Or (2017).

2.2 Time Series Algorithm for patients waiting list forecasting

A several techniques and methodologies have been proposed in the literature reviews of times series forecasting in different medical research, education, business, government areas. The various approaches were developed for forecasting using various statistical and machine learning models with better accuracy.

2.2.1 Time series and Smoothing Models

The time series model is being popular to forecast monthly and daily flow of patients in Emergency Department, the autoregressive moving average and exponential smoothing techniques are the most commonly used (Gopakumar, Tran, Luo, Phung and Venkatesh; 2016). The classical ARIMA model were used to forecast the bed occupancy on daily basis in emergency department of European Hospitals. The model forecast 32 days in advance with RMS error of 3 percent with seasonality term (Jones, Joy and Pearson; 2002). The ARIMA model further used for short term forecasting the number of patients in emergency department, the study carried out by collecting hourly beds occupancy data from July 2005 to June 2006 from three tertiary care hospitals. The three models were used for each area, hourly historical average, seasonal autoregressive integrated moving average (ARIMA) and sinusoidal with an autoregressive structure term error (AR). Where Akaikes Information Criterion (AIC) and root mean square (RMS) error. The seasonal autoregressive integrated moving average (ARIMA) outperformed than others models, which able to prediction bed occupancy in 4 12 hours in Advance (Schweigler, Desmond, McCarthy, Bukowski, Ionides and Younger; 2009). During SARS outbreak in Singapore in Tan Tock Seng Hospital in 2003, the ARIMA model were applied to predict the occupancy of beds by using data from 14th March 2003 to 31st May 2003. It was found that ARIMA (1,0,3) model able predict the bed occupancy in three day prior accurately with 5.7 percent to 8.6 percent MAPE (mean absolute percentage error). The forecasting model was implemented in hospital to use as administrator tools for planning beds capacity in real time (Earnest, Chen, Ng and Sin; 2005). The time series forecasting models were using to predict future epidemic, which was developed in hospital of china for planning resource of patients. The Autoregressive integrated moving average (ARIMA) and exponential smoothing (ETS) models were applied to predict the number of epidemic case, the ARIMA (0,1,0) (1,1,1)12 model has selected as best performing with AIC=1342.2 and BIC = 1350.3, the ETS(M,N,M) model has AIC=1678.6 and BIC=1715.4 but ETS has minimum RMSE, hence ETS were selected for sample-simulation forecasting (Zeng, Li, Huang, Xia, Wang, Zhang, Tang and Zhou; 2016). The time series were used to predict the patients of influenza, the incidence data of highly pathogenic avian influenza (H5N1) breakout in Egypt and applied ARIMA and Random Forest model. The study is made to compare the performance of both models ARIMA and Random Forest, the application of Random Forest applied various area of health studies, in this analysis Random Forest provides good result over ARIMA model (Kane, Price, Scotch and Rabinowitz; 2014). Overall, in many cases of time series problem ARIMA model found to be better model other to predict the patients in hospital.

2.2.2 Simulation Models

The simulation model was commonly used to analyses the behavior of complex framework. In earlier, the discrete event stochastic stimulation models were used to investigate the impact of emergency admissions on the requirement of beds regular basis in acute care (Bagust, Place and Posnett; 1999). The simulation model proposed to build, based on data of 5 Israeli hospitals from emergency department. Their techniques examine the flow of patients into 8 different cluster along with elements of time. It is observed that, the follow of patient better characterized by their patient type rather than specific visited hospitals (El-Darzi, Vasilakis, Chaussalet and Millard; 1998). In 2007, Yeh and Lin were used simulation method to categorized the patients in hospital emergency department and reducing the waiting time by using genetic algorithm (Arisha and Abo-Hamad; 2013).

2.2.3 Regression Models for Forecasting

The regression models was used to analyses the relationship between the forecasting features with variable in data. The linear regression was used to forecast the patients admission in emergency department by encoding monthly variation over 6 months of horizon, that outperformed quadratic and autoregressive model with 1.79 percent of MAPE (Boyle, Wallis, Jessup, Crilly, Lind, Miller and Fitzgerald; 2008). Another approach was

used to prediction model for length of stay in hospital at emergency department by using clustering and principle component analysis (PCA) to determine the significant predictors from data of patients from Pediatric Emergency unit France and performed linear regression (Combes, Kadri and Chaabane; 2014). The nonlinear approach for forecasting patients admission using regression tree has performed superior over a neural network models. The randomized algorithm for non-linear regression RT-KGERS provides more accuracy with more features but the limitation of regression tree was stability, that provide different output in different runs (Garcia and Chan; 2012).
Non-linear regression model is better for dynamic changing of patients flow, the regression model like random forest, kNN, SVR were used to categorize outflow of patients. The kNN is most effective method to exploit the repetitive patterns, which was used for pattern recognition from data (Cover and Hart; 1967). The kNN (k-Nearest Neighbours) algorithm was used for Histogram time series that measure dissimilarities in a sequence of histograms and calculate for forecasting, which was applied on financial data (Arroyo and Maté; 2009). The k-NN non-parametric regression algorithm developed in MATLAB for short-terms traffic flow in Shanghai urban expressway using historical dataset, search algorithm along with prediction plan. The accuracy of proposed model was more that 90 percent found reliable for short-term prediction model (Zhang, Liu, Yang, Wei and Dong; 2013). Overall performance of k-NN regression model were better in above research for short-term prediction, but not yet applied to patients wait in hospital.
Support vector regression (SVR) model also use as powerful and popular algorithm for time series forecasting, as kernel function in SVR map features into high-dimension space to perform linear regression. The application of SVR are successfully used for financial market, electricity and business forecasting. The SVR algorithm was reliable approach to predict time series data in non-linear framework (Sapankevych and Sankar; 2009).
There are many other time series prediction techniques, apart from auto-regression model. k-NN regression algorithm were used to gauge the patterns from historical data. RF and SVR are powerful models, that efficiently handle non-linearity with minimum tuning.

2.2.4 Neural Network model for forecasting.

The neural network has better performance over large dataset, the neural network algorithm was used to uber request in special occasions, the model was chosen base on three criteria length, number and correlation of time series data, all these three components are high which suggest that neural network model was right choice than general time-series model. The recurrent neural network Long Short-Term Memory (LSTM) model was use for forecasting which has given 26.66 sMAPE value (Laptev, Yosinski, Li and Smyl; 2017). In time series stock forecasting, number of models were used for better accuracy, the dual-stage attention-based recurrent neural network (DA-RNN) were used on SML 2010 dataset and NASDAQ 100 Stock dataset. DA-RNN consists of encoder and decoder mechanism which encode input with input attention and decode with temporal attention, which can capture long-range dependencies information of time series. DA-RNN achieved best performance by 0.42 to 0.009 RMSE (Qin, Song, Chen, Cheng, Jiang and Cottrell; 2017). The Artificial neural network was used to predict the wind speed in anemometric tower which is place in 50-meter height at two different locations of coastal region, the ANN model is evaluated with Holt-Winter (HW). The data from Modern-Era Retrospective analysis for Research and Application Version-2 (MERRA-2) at same height of tower. As ANN hybrid model has better performance of than other models with RMSE of 0.91 and MSE is 0.62 m/s. The ANN hybrid model consider to be better model for short term wind forecasting (Ferreira, Santos and Lucio; 2018). The nonlinear autoregressive neural network model (NARNN) where use to forecast the new admission of patients in hospital, where model is evaluated with time series ARIMA and Hybrid model, in this case neural network model outperform as compare to ARIMA and Hybrid model. The root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) is better least in neural network model (Zhou, Zhao, Wu, Cheng and Huang; 2018). As there are several researches carried out using artificial neural network and recurrent neural network to forecast value with better accuracy. The neural network model outperformed for various area of forecasting problems like stock market, medical, environmental etc.

3 Methodology

The implementation of this research follows CRIPS-DM methodology, CRIPS-DM stand for cross-industry process for data mining. This methodology provides robust project planning and structural pipeline for implementation of data mining project. In most of the data mining project need large amount of time to spend on useful experiment to get quality output (Huber, Wiemer, Schneider and Ihlenfeldt; 2018). This project involved number of experimental tests to obtain quality output for business objective. The CRIPS-DM provides better communication and planning foundation for data analytics project and its deployment in test environment.

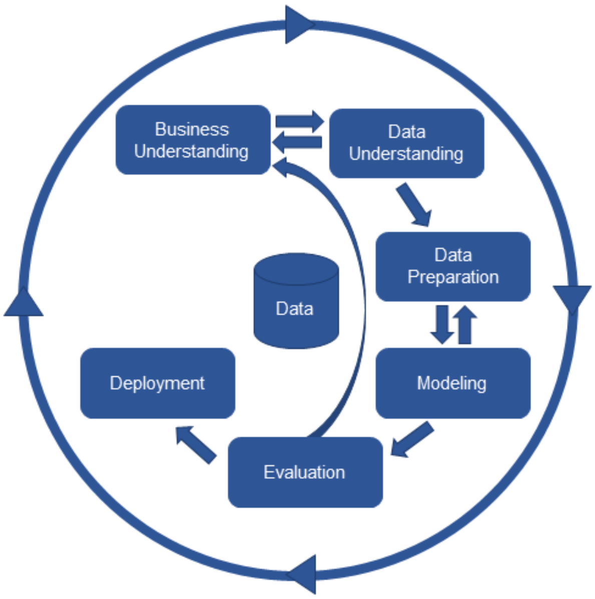


Figure 1: CRIPS-DM Approach

The CRIPS-DM methodology involving following steps:

Business Understanding: To start any project, defining objective of research is very important to gain more domain knowledge and understanding of business requirements, which make easier to deliver right solution.

Data Understanding: In data mining project, understanding of data is the basic building block to develop right data model.
Data Preparation: The data preparation involves extracting data from different as per business needs, clean and transform data into appropriate format to apply machine learning models.
Modelling: This is the experimental setup to prepare actual solution models base on business and data understanding.
Evaluation: Once the model build, its is very important that how models are performing in test phase, decide which algorithm can do better work.
Deployment: Once the project is ready, the proposed algorithm is deploy into production for actual functioning, where result play vital role in decision making.

4 Implementation

4.1 Business Understanding

This research is aiming to provide the forecasting of patient waiting list over Irelands hospital. The forecasting model can work as administrative tool for different major hospital or NTPF who closely working with hospital, for managing resource, develop protocol for management practice to minimize the waiting list.

4.2 Data Understanding

To conduct data mining project, understanding of data is second step of CRIPS DM methodology. The data are initially collected from NTPF official website as mentioned in data description. The separate yearly data load into R to form single data set.

4.3 Data Preparation

The preparation involved the act of preparing data as per business requirement for model. Once data import into R. Data is examining to check outlier and convert data into required form. This is times series analysis of patient waiting list, so data need to sort, select and transform into time series format. Reshape2 package was used to format data to apply pivot to get aggregated month wise time series data into different time band. Data are converted into time series data using tseries package and prepared data ready for examine.

4.4 Data Exploration

Below figure show the overall waiting list of patients in different time band.
From the above figure, it is observed that 12+ months waiting number of patients are keep increasing along with total number of patients in waiting list.

4.5 Data Decomposition

The data decomposition is building block of time series analysis to identify the seasonality trend and cycle, which capture the historical pattern from the time series data. All these

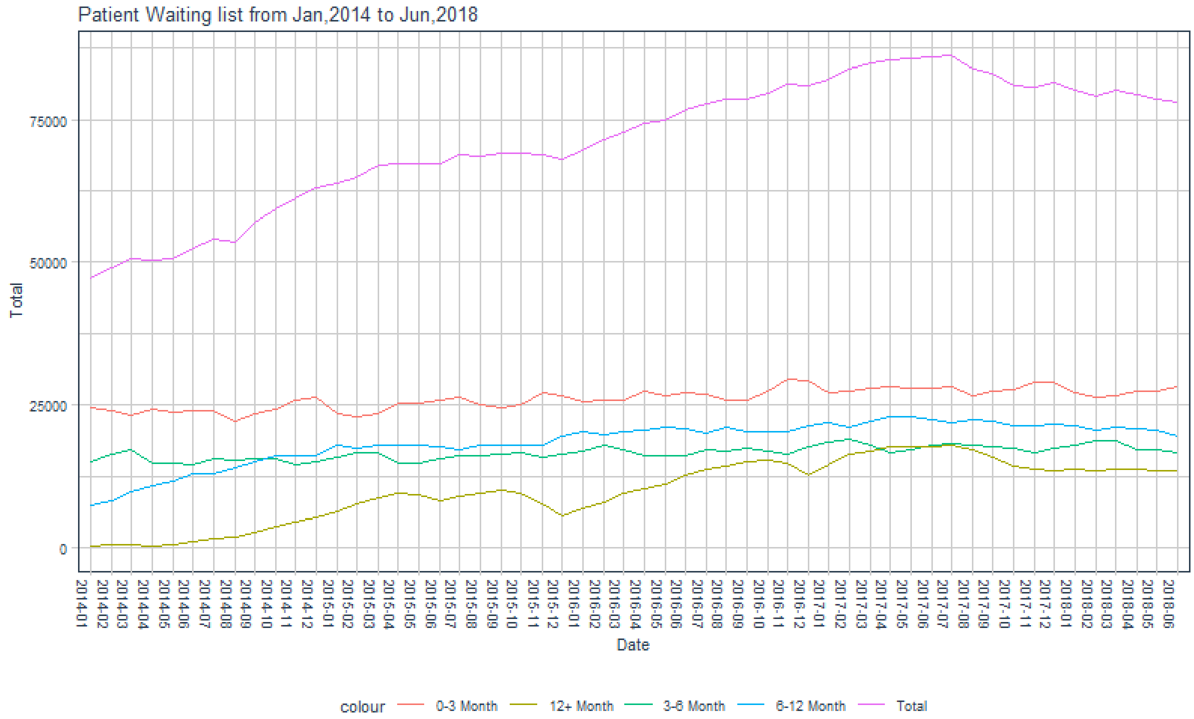


Figure 2: Overall Trending of Patients wating list in different time-tand

components will help to understand the behavior of data as well as to build the foundation model for time series.

4.5.1 Seasonal Component

Seasonal component refers to data fluctuations over different calendar cycles.

4.5.2 Trend Component

The trend component provides the overall trend of time series data.

4.5.3 Cycle Component

Cycle Component helps to understand the decreasing and increasing pattern that are not seasonal.
The final part shows the residual or error of time series data that cant be attribute to above mention component of data.

4.5.4 STL Decomposition

Below figures shows the additive decomposition of series data for overall waiting list of patients with different time band. There are two type of decomposition additive and multiplicative, to choose additive decomposition data magnitude has seasonal fluctuation and have variation around trend cycle, level of time series doesnt have variance. For choosing multiplicative when more variation in seasonal pattern and trend cycle appear proportional to the time series level. In this research, additive decomposition is more appropriate as there is season fluctuation and variation of trend cycle is not varying with the level of time series.

STL decomposition technique were used in this research, STL decomposition is versatile and robust method for decomposition of time series data. STL stand for Seasonal and Trend decomposition using Loess, which can estimate the non-linear relationship. STL decomposition has more advantage than classical, SEATS and X11 decomposition like handle any type of seasonality, control rate of change in seasonal component, robust to outlier. [3].

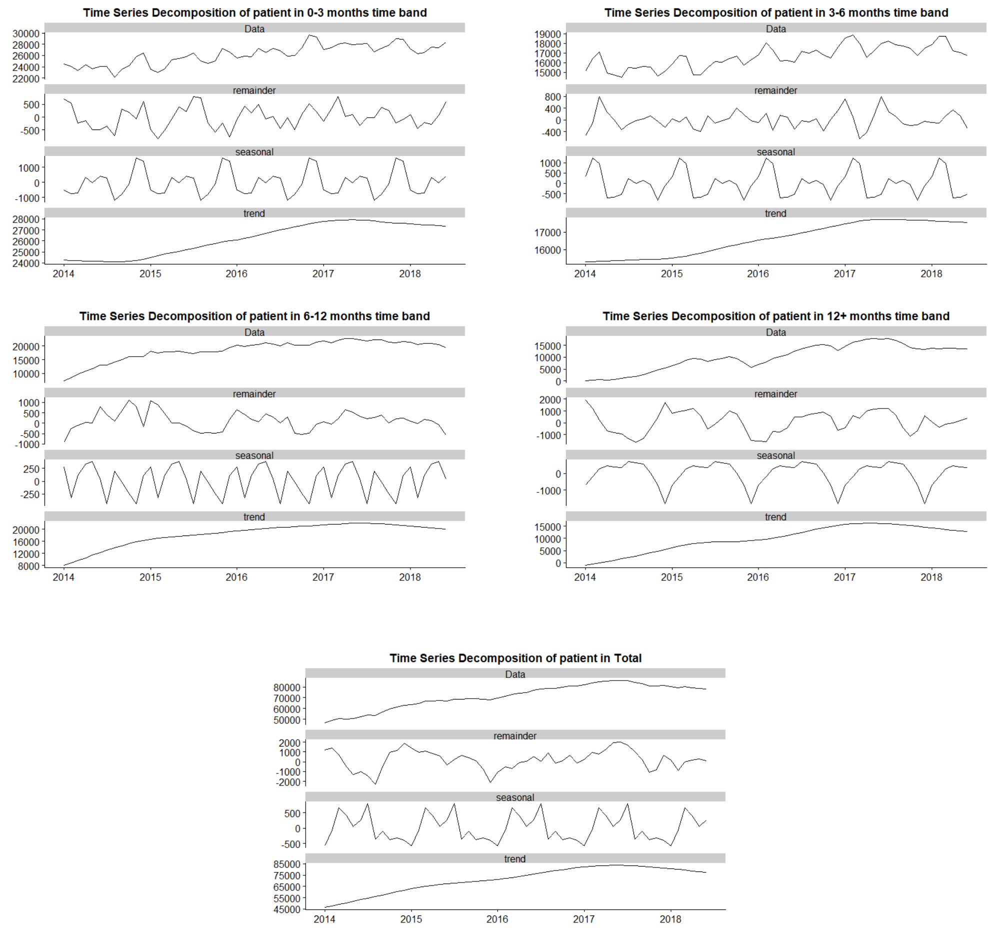


Figure 3: Decomposition of Hospital waiting list time series data

In above figures of decomposition, it is observed that there is seasonality exist on time series data in all time band.

4.5.5 Stationarity

The test of stationarity is another important part of time series analysis to apply fitting ARIMA model. when variance, autocovariance and mean of data are time invariant, it

[3] otexts`https://otexts.org/fpp2/stl.html`

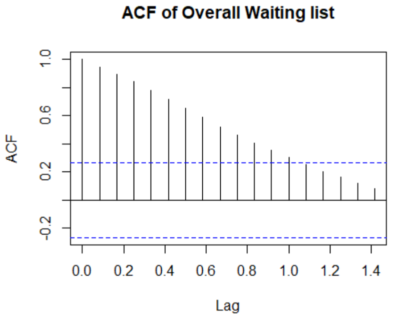
is said to be stationary time series data. The Augmented Dickey-Fuller (ADF) test is used to test the stationarity of time series data, in which null hypothesis assumed that non-stationarity of series. In ADF test procedure, how change in y value explained by lag value and linear trend. If their non-significant change in Y by lagged value and presence of trend component, the null hypothesis is rejecting to prove series is non-stationary.

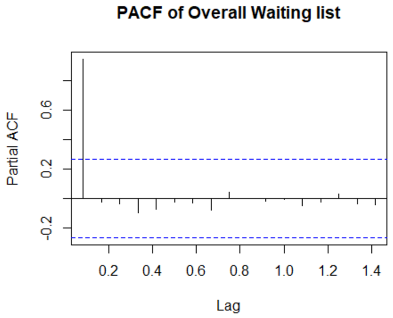
The ADF test is performed below are the p-value of different time series test.

1. Test1: For Overall Waiting list (p-value) = 0.99
2. Test2: For waiting list of 0-3 months time band (p-value) = 0.1803
3. Test3: For waiting list of 3-6 months time band (p-value) = 0.06713
4. Test4: For waiting list of 6-12 months time band (p-value) = 0.7001
5. Test5: For waiting list of 12+ months time band (p-value) = 0.8331

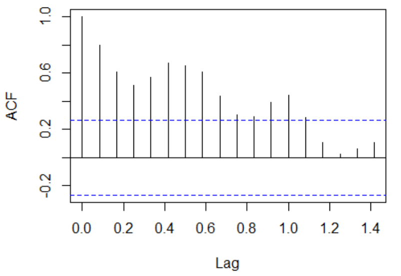
4.5.6 Autocorrelation and choosing Model Order

Autocorrelation plot is also another useful tool to determining stationarity of time series data. Autocorrelation also helpful to choose ARIMA order parameter. The ACF plot flash the correlation between series of data and its lags. In ACF plot also help to determine MA (q) order of ARIMA Model. Similarly, PACF is also an important factor to determine the correlation between variable and its lags, which is not explained by previous lag. PACF plot help to determine the order of AR (p) of ARIMA Model.
Below figure show the ACF and PACF of different time series data in various time band of waiting list.

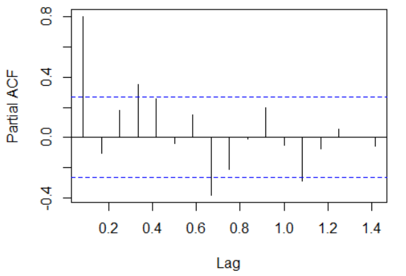




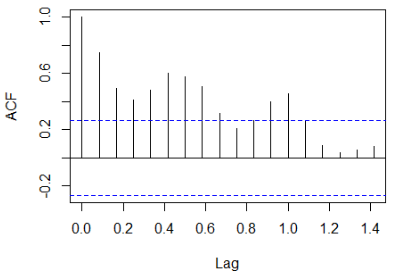
ACF of 0-3 months time band Waiting list



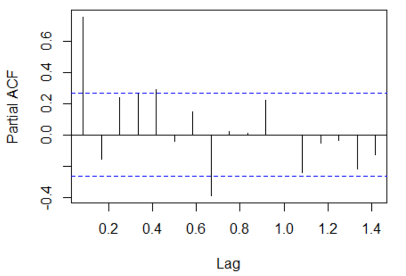
PACF of 0-3 months time band Waiting list



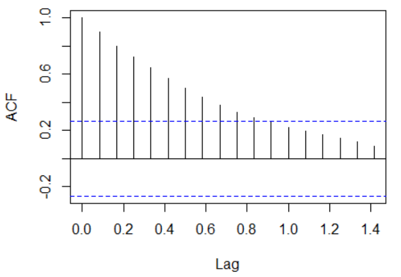
ACF of 3-6 months time band Waiting list



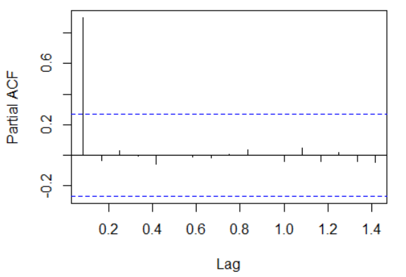
PACF of 3-6 months time band Waiting list

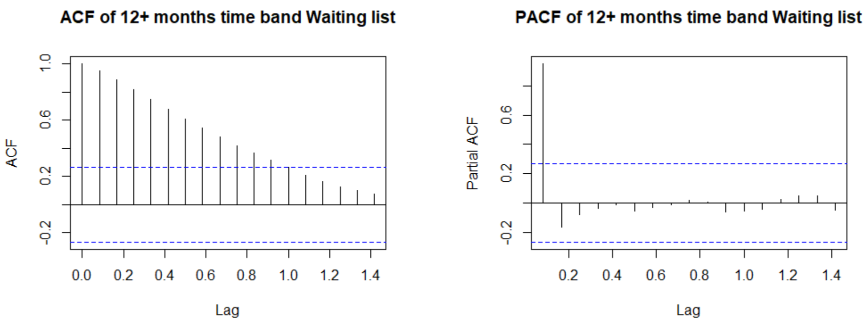


ACF of 6-12 months time band Waiting list



PACF of 6-12 months time band Waiting list

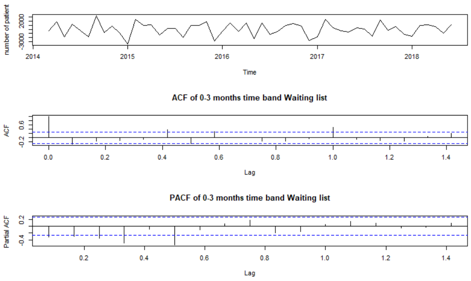



ACF of 12+ months time band Waiting list

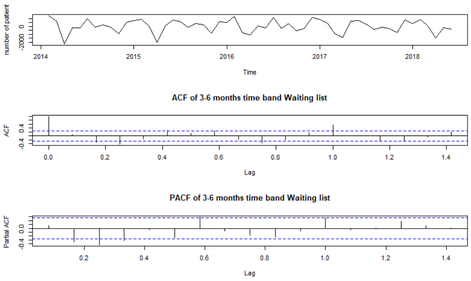
PACF of 12+ months time band Waiting list

Figure 4:

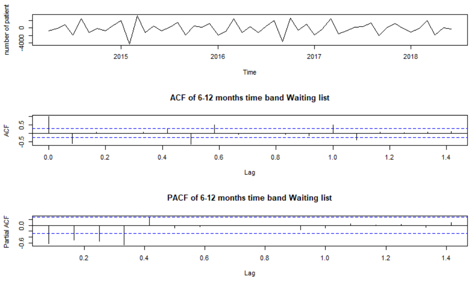
The ACF and PACF of time series difference d = 1 is plotted to identify the AR (p) and MA (p) values for different time series case.


ACF of 0-3 months time band Waiting list

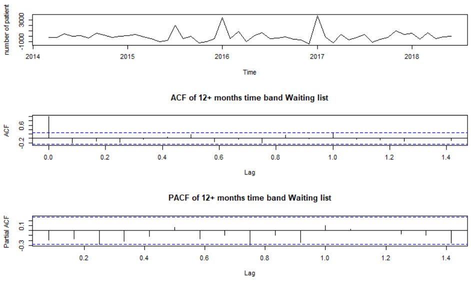
PACF of 0-3 months time band Waiting list


ACF of 3-6 months time band Waiting list

PACF of 3-6 months time band Waiting list


ACF of 6-12 months time band Waiting list

PACF of 6-12 months time band Waiting list


ACF of 12+ months time band Waiting list

PACF of 12+ months time band Waiting list

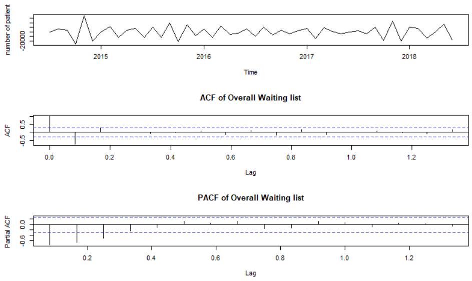


Figure 5:

4.5.7 Seasonal Adjustment

After examining the seasonality of data, it is determining that, this time series data has seasonal component. Seasonal adjustment is another necessary component in time series analysis to estimate the reasonable forecasting. As it is found that there is no theoretical definition that explain the demand of seasonal adjustment, but it was practically tested to considered into seasonal adjustment of data. There is some internal work carried out in 'Rapport vedr. ssonkorrigering i Danmarks Statistik' in 2001 at statistic Denmark [4]. In this research, seasonal adjustment is performed in other to remove irregular pattern, disturbing factor and pre-adjust the series.

5 Modeling

There are several time series models were applied for forecasting of patient in hospital, stock market, sales and many more. After testing number of time series model in this research, the three best performing algorithm ARIMA, TBATS and AR-NN were selected to carry out entire research and evaluate individual performance.

5.0.1 Autoregressive Integrated Moving Average (ARIMA)

The ARIMA model used to forecast future value in time series by examine historical data. Auto Regressive (AR) and Moving Average (MA) are the two main component of ARIMA model. The Auto Regressive (AR) and Moving Average (MA) work together to build ARIMA models (Rahimi and Khashei, 2018). Below is the equation of ARIMA model show in equation (1).

$$\emptyset(B)\nabla^{d}(y_t - \mu) = \theta(B)\alpha_t \tag{1}$$

Where y_t stand for actual value in time t, α_t refer as white noise which is independently and identically distributed with zero mean and constant variance of $\sigma 2$. $\emptyset(B) = 1 - \sum_{i=1}^{p} \varphi_i B^i, \theta(B) = 1 - \sum_{j=1}^{q} \theta_j B^j$ are the polynomial in B with p and q degree $\emptyset_i$(i = 1,2,...p) and $\emptyset_j(j = 1, 2, ...q)$ are the parameter of the model, $\nabla = (1 - B)$, where B is backward shift operator, p and q refer as backward shift operator are integer, d is integer value stand for differencing. Box and Jenkins (1976) methodology were used for ARIMA modeling which contain three iterative steps, identification of model, estimation of parameter and diagnostic checking. All these steps are mention below in detail.

[4] ref `https://www.dst.dk/-/media/Kontorer/13-Forskning-og-Metode/seasonal_001-pdf.pdf`

1. **Identification:** This step involves searching actual value of p, d and q. P is value of auto regressive, q is the value of moving average and d is the number of differencing. The Autocorrelation function and partial correlation function are tool to identify value of auto regressive (AR) and moving average (MA) in a sample data.

2. **Estimation:** After picking a determine ARIMA (p, d, q), the parameters which were identied in the last stage, ought to be evaluated by the ordinary least squared (OLS) approach.

3. **Diagnosis:** This stage involves the test of model adequacy by diagnostic checking, checking process involves the check the accuracy of model, weather model able to predict value correctly or not, the model can be evaluating by measuring AIC, BIC, RMSE, MAPE value.

All these steps are repetitive process to get adequate structure of ARIMA mode, to eliminate manual work, Grid Search algorithm were used to find bet p, d and q value for ARIMA model, which is finally evaluate with least AIC, MAE and RMSE value. Grid search algorithm is hyperparameter optimization in machine learning model, which involve exhaustive search of large combination of hyperparameter to determine best result (Elmasdotter and Nyströmer; 2018).

5.1 TBATS Model for times series

TBATS model is time series model, which is using to demonstrating multiple complex seasonality. TBATS model was introduced by De Livera in 2011. The TBATS model is the space of exponential smoothing model, which allows automatic Box-Cox transformation along with ARMA error. The TBATS stand for Trigonometric Box-Cox ARMA Trend Seasonality which is like BATS model without trigonometric regressor which is use to handle complex seasonality (Jain; 2018).

5.2 Artificial Neural Network model for time series

The ANN is non-linear model for time series forecasting and successful alternative to ARIMA, in many problems artificial neural network used with statistical time series model like ARIMA and exponential model to form hybrid model. The hybrid model was not best in every problem, in some problem neural network performed better. The multi-layer feed forward and recurrent neural architecture were used to build ANN time series model (Tamatta; 2018). Below is the equation for ANN model.

$$y_t \;=\; \emptyset_0 + \sum_{j=1}^{q} \emptyset_j g\left(\theta_{0j} \;+\; \sum_{i=1}^{p} \theta_{ij} y_{t-i}\right) + \varepsilon_t \tag{2}$$

Where, $\emptyset_j (j = 1, 2, 3q)$, $\theta_{ij}(I = 0, 1, 3, p; j = 1, 2, , q)$ are the weights. $\emptyset_0$, θ_{0j} are term as bias, and ε_t represent the white noise. The ANN model is effective non-linear model to generate time series (Khandelwal, Adhikari and Verma; 2015).

6 Evaluation

In this research, there are several models applied from simple to complex to find better forecasting model. Based on many git hub project and research, simple model like Navie,

Patient Waiting		OverAll		0-3 Months		3-6 Months		6-12 Months		12+ Months
Model	RMSE	MPE	RMSE	MPE	RMSE	MPE	RMSE	MPE	RMSE	MPE
ARIMA	771.685	0.154	406.510	0.184	300.818	0.382	402.437	0.060	402.437	0.060
TBATS	842.335	0.073	480.423	0.268	313.668	0.111	533.671	0.699	777.629	2.974
NNETAR	872.377	-0.036839	467.0437	-0.049832	334.957	0.060	415.870	0.165	764.470	-13.537

Mean, sNaive and drift were applied to test time series data as these models are rejected due to presence of seasonality in time series data set. Some more model like time series linear regression, dynamic linear regression, moving average and exponential smoothing the performance of moving average at order 3 and exponential smoothing is better than others. The more complex algorithms were used as time series has seasonality trends with the support of above described literature reviews, three models are shortlisted to perform different time series analysis. The models are ARIMA with Grid Search, TBATS and ANN. The performance of models judge based on RMSE and MAPE. In many case ARIMA outperform than other models and in some case TBATS better prediction.

This research consists of five different time series cases, below table provides the RMSE and MPE of three different models in different time series cases.
Above table shows the performance of individual model in different time band.

This plot shows the comparison of fore-casted value with actuals value of overall waiting list of patients in Hospital. As we can see TBATS forecast value closer to actual.

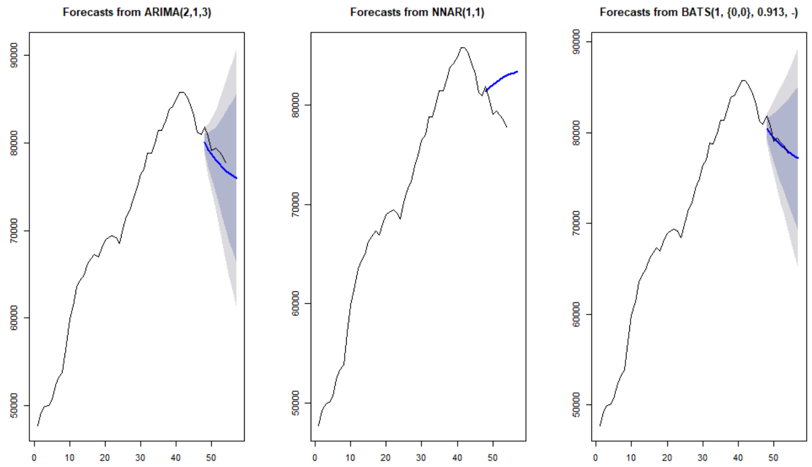


Figure 6: Overall waiting list forecasting

Below figures shows the comparison of fore-casted value with actual value of 0-3 months time band waiting list of patients in different Hospitals. The performance of ARIMA (2,1,5) is better that others.

Below figures shows the comparison of fore-casted value with actual value of 3-6 months time band waiting list of patients in different Hospitals. The performance of NNAR is better that others.

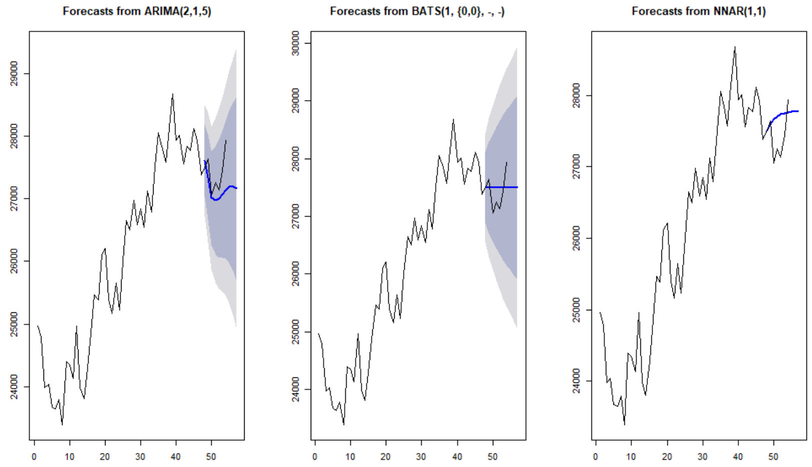


Figure 7: 0-3 months waiting list forecasting

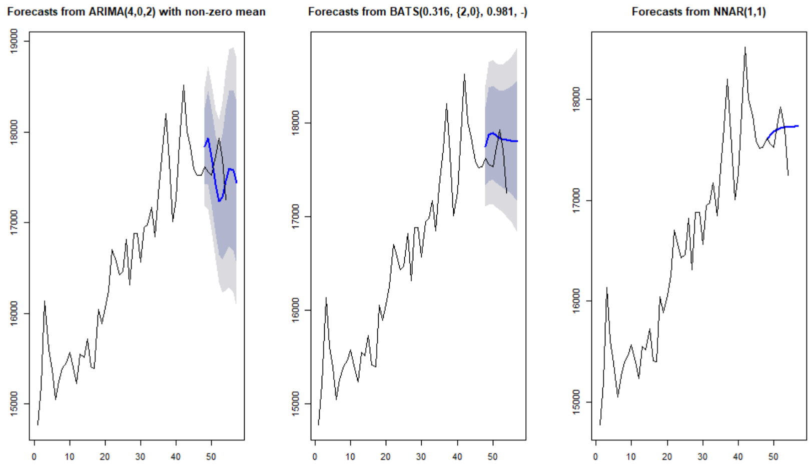


Figure 8: 3-6 months waiting list forecasting

Below figures shows the comparison of fore-casted value with actual value of 6-12 months time band waiting list of patients different in Hospitals. The performance of ARIMA (5,1,3) is better that others.

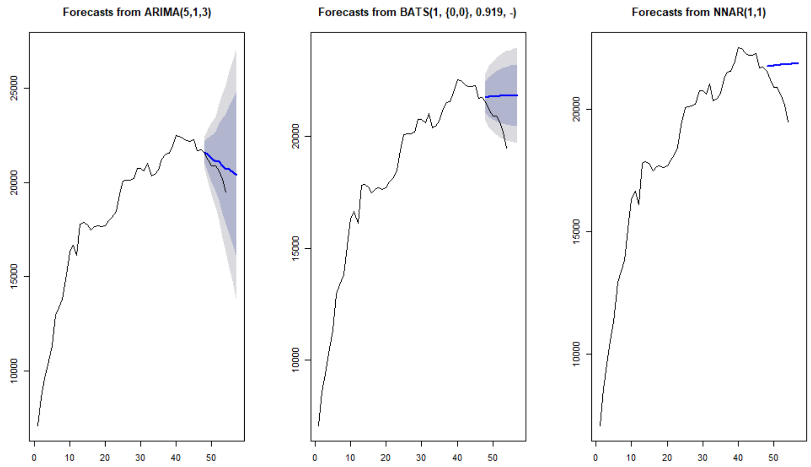


Figure 9: 6-12 months waiting list forecasting

Below figures shows the comparison of fore-casted value with actual value of 12+ months time band waiting list of patients different in Hospitals. The performance of ARIMA (1,0,5) is better that others.

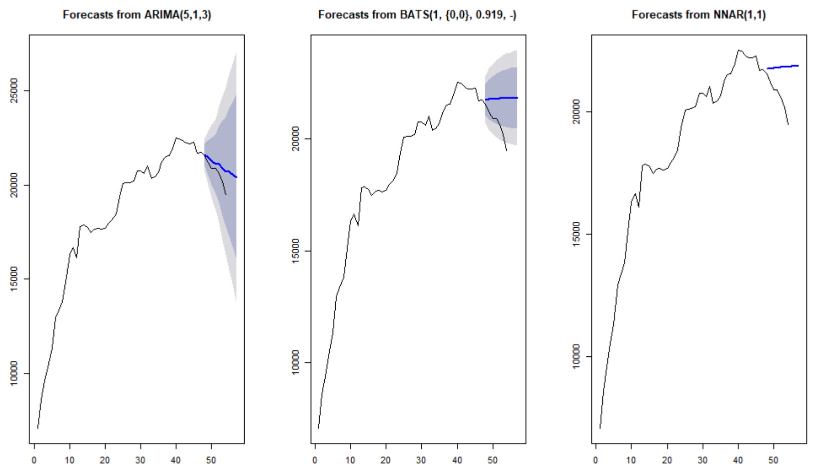


Figure 10: 12+ months waiting list forecasting

6.1 Discussion

This project follows CRIPS-DM methodology, the purpose of this research to serve as academic research project for Master thesis. This project will not be deploying in real time production environment. This research suggests that above discuss models can use solve similar real-life problems.

7 Conclusion and Future Work

In this research project number of time series models were applied and evaluated, the performance of ARIMA model is better than TBATS and NNETAR with least great RMSE value and accuracy of models further improved by de-seasonalized the data. This research conducted with standalone time series models, due to limitation of data points, hybrid model with neural network and popular LSTM recurrent neural network not applied. As, it is observed that the performance of neural network model NNETAR is not good. The data used in research is monthly time series data, which is available from Jan,2014 to Jun 2018, which is not enough to build better complex model. The availability of daily patients waiting list time series data can make this research better and allows us to use complex neural network model. However, the time series data are keep increasing over a time, in future due to availability of more data point researcher can built better models by using hybrid and deep learning algorithms. This research encourages NTPF, Hospital administration and concern government regulate body like ministry of health to estimate the number of patients in different hospitals in different time bands. This research also supports to build the models for demand of medicine, hospitals staff, doctor, equipment, beds, hospital building plans etc. Which further support government for budgeting and planning in health care sector.

References

Ansell, D., Crispo, J. A., Simard, B. and Bjerre, L. M. (2017). Interventions to reduce wait times for primary care appointments: a systematic review, *BMC health services research* **17**(1): 295.

Arisha, A. and Abo-Hamad, W. (2013). Towards operations excellence: optimising staff scheduling for new emergency department.

Arroyo, J. and Maté, C. (2009). Forecasting histogram time series with k-nearest neighbours methods, *International Journal of Forecasting* **25**(1): 192–207.

Bagust, A., Place, M. and Posnett, J. W. (1999). Dynamics of bed use in accommodating emergency admissions: stochastic simulation model, *Bmj* **319**(7203): 155–158.

Boyle, J., Wallis, M., Jessup, M., Crilly, J., Lind, J., Miller, P. and Fitzgerald, G. (2008). Regression forecasting of patient admission data, *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, IEEE, pp. 3819–3822.

Combes, C., Kadri, F. and Chaabane, S. (2014). Predicting hospital length of stay using regression models: application to emergency department, *10ème Conférence Francophone de Modélisation, Optimisation et Simulation-MOSIM14*.

Conrad, M. (2013). Patient waiting time and associated factors at the assessment center, general out-patient department mulago hospital uganda.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification, *IEEE transactions on information theory* **13**(1): 21–27.

Earnest, A., Chen, M. I., Ng, D. and Sin, L. Y. (2005). Using autoregressive integrated moving average (arima) models to predict and monitor the number of beds occupied during a sars outbreak in a tertiary hospital in singapore, *BMC Health Services Research* **5**(1): 36.

El-Darzi, E., Vasilakis, C., Chaussalet, T. and Millard, P. (1998). A simulation modelling approach to evaluating length of stay, occupancy, emptiness and bed blocking in a hospital geriatric department, *Health care management science* **1**(2): 143.

Elmasdotter, A. and Nyströmer, C. (2018). A comparative study between lstm and arima for sales forecasting in retail.

Ferreira, M., Santos, A. and Lucio, P. (2018). Short-term forecast of wind speed through mathematical models.

Garcia, K. A. and Chan, P. K. (2012). Estimating hospital admissions with a randomized regression approach, *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, Vol. 1, IEEE, pp. 179–184.

Gopakumar, S., Tran, T., Luo, W., Phung, D. and Venkatesh, S. (2016). Forecasting daily patient outflow from a ward having no real-time clinical data, *JMIR medical informatics* **4**(3).

Huber, S., Wiemer, H., Schneider, D. and Ihlenfeldt, S. (2018). Dmme: Data mining methodology for engineering applications–a holistic extension to the crisp-dm model.

Jain, G. (2018). Time-series analysis for wind speed forecasting, *Malaya Journal of Matematik (MJM)* (1, 2018): 55–61.

Jones, S. A., Joy, M. P. and Pearson, J. (2002). Forecasting demand of emergency care, *Health care management science* **5**(4): 297–305.

Kane, M. J., Price, N., Scotch, M. and Rabinowitz, P. (2014). Comparison of arima and random forest time series models for prediction of avian influenza h5n1 outbreaks, *BMC bioinformatics* **15**(1): 276.

Khandelwal, I., Adhikari, R. and Verma, G. (2015). Time series forecasting using hybrid arima and ann models based on dwt decomposition, *Procedia Computer Science* **48**: 173–179.

Laptev, N., Yosinski, J., Li, L. E. and Smyl, S. (2017). Time-series extreme event forecasting with neural networks at uber, *International Conference on Machine Learning*, number 34, pp. 1–5.

Qin, Y., Song, D., Chen, H., Cheng, W., Jiang, G. and Cottrell, G. (2017). A dual-stage attention-based recurrent neural network for time series prediction, *arXiv preprint arXiv:1704.02971* .

Sapankevych, N. I. and Sankar, R. (2009). Time series prediction using support vector machines: a survey, *IEEE Computational Intelligence Magazine* **4**(2).

Schweigler, L. M., Desmond, J. S., McCarthy, M. L., Bukowski, K. J., Ionides, E. L. and Younger, J. G. (2009). Forecasting models of emergency department crowding, *Academic Emergency Medicine* **16**(4): 301–308.

Tamatta, R. (2018). Time series forecasting of hospital inpatient and day case waiting list using hybrid arima and neural network model.

Xie, Z. and Or, C. (2017). Associations between waiting times, service times, and patient satisfaction in an endocrinology outpatient department: A time study and questionnaire survey, *INQUIRY: The Journal of Health Care Organization, Provision, and Financing* **54**: 0046958017739527.

Zeng, Q., Li, D., Huang, G., Xia, J., Wang, X., Zhang, Y., Tang, W. and Zhou, H. (2016). Time series analysis of temporal trends in the pertussis incidence in mainland china from 2005 to 2016, *Scientific reports* **6**: 32367.

Zhang, L., Liu, Q., Yang, W., Wei, N. and Dong, D. (2013). An improved k-nearest neighbor model for short-term traffic flow prediction, *Procedia-Social and Behavioral Sciences* **96**: 653–662.

Zhou, L., Zhao, P., Wu, D., Cheng, C. and Huang, H. (2018). Time series model for forecasting the number of new admission inpatients, *BMC medical informatics and decision making* **18**(1): 39.