Forecasting for Ambulance Services in Jakarta

GLEB KHAMIN

INTRODUCTION

Ambulance service providers in Jakarta are seeking a planning tool for scheduling and rostering their crews. For efficient medical treatment and transportation to patients, ambulance services require an accurate forecast of demand. The aim of this project is to analyse the provided data on the daily number of patients (01/04/2015 - 31/05/2019) to see if there is any pattern to it, and to forecast the number of patients for the first week of June 2019. The following steps will be taken:

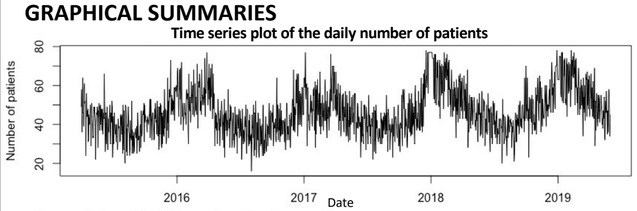
(i) Analysing and pre-processing the provided data
(ii) Converting the data into daily time series format and splitting it into training (70%) and testing (30%) sets to facilitate model training and validation
(iii) Training the 10 selected models on the train set, and subsequently testing their accuracy on the test set
(iv) Evaluating the performance of these models using statistical measures, such as MAE, RMSE, and MAPE
(v) Selecting the model with the best performance, and using it to forecast the number of patients for the first week in June 2019.

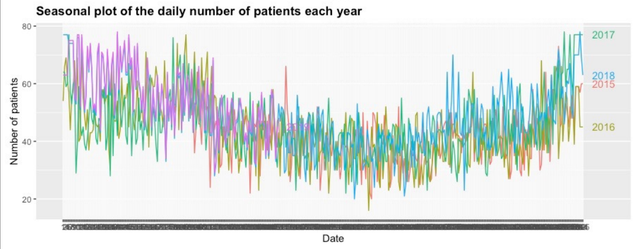
NUMERICAL SUMMARIES

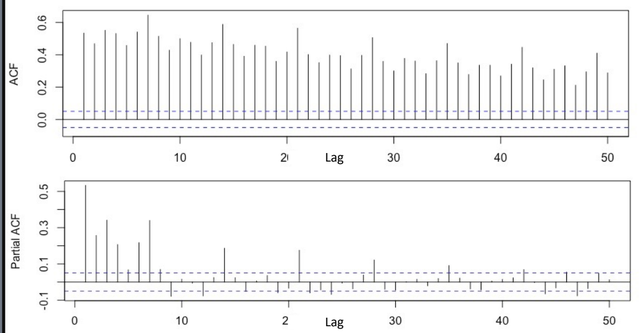
- 1521 observations from 1 April 2015 to 31 May 2019
- No missing values were found in the dataset
- 37 outliers were identified. First, the outliers were replaced with NA values. After that, the Last observation carried forward (LOCF) imputation method was used to replace the NA values with the most recent non-NA value that it could find.
- Since the p-value obtained from the Dickey-Fuller test was very low (below the 0.01 significance level), we reject the null hypothesis that the data is non-stationary.

Min	Q1	Median	Mean	Q3	Max	IQR	SD
16	38	45	46.83	54	104	16	12.86

GRAPHICAL SUMMARIES







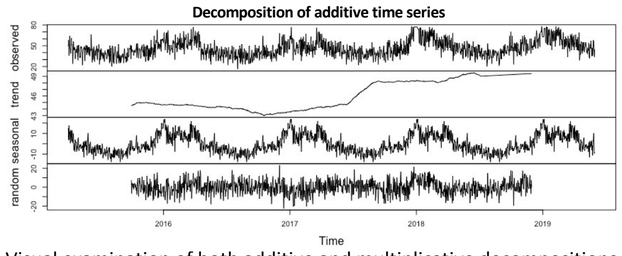
The seasonality plot shows strong seasonality in the time series, particularly from June to August. However, determining the exact seasonal period from the seasonal plot alone proves challenging. Therefore, decomposition will be used for a more precise understanding of seasonality.

GRAPHICAL SUMMARIES (CONTINUED)

The ACF plot indicates significant autocorrelation up to lag 50, whereas the PACF plot shows significant partial autocorrelation up to lag 4. The ACF and PACF plots also depict seasonality in the time series, which peaks at lag 7, lag 14, lag 21 and so on.

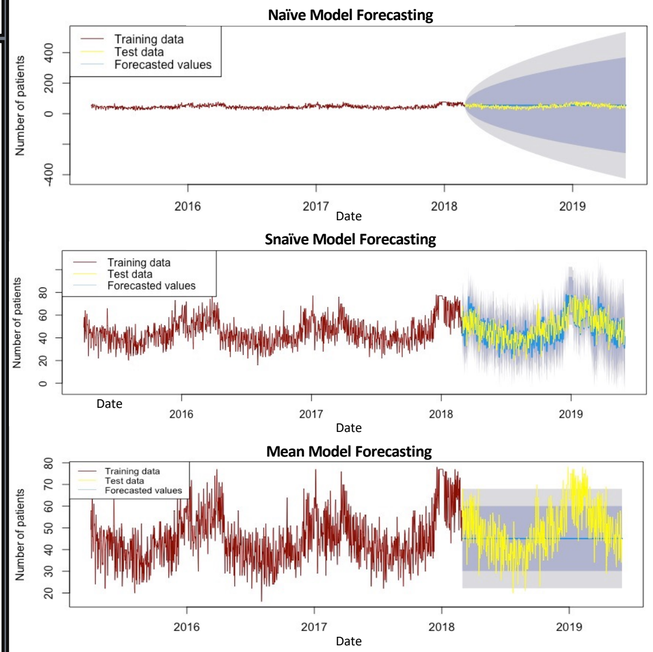
DECOMPOSITION OF THE DATA

We carried out a decomposition of the time series into its fundamental components: trend, seasonality, and residuals. A clear upward trend is noticeable from 2015 to 2019. Additionally, a distinct seasonal pattern emerges, with patient numbers typically escalating at the start of each year, likely within the first quarter, and decreasing from June to August.



Visual examination of both additive and multiplicative decompositions reveals that a seasonal component is constant over time, therefore, an additive decomposition model is to be used for our analysis.

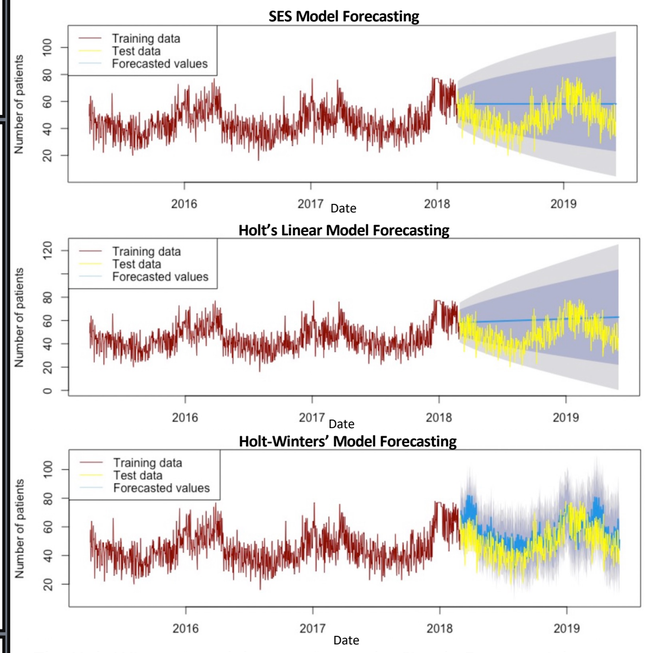
NAÏVE, SNAÏVE & MEAN MODELS



- Naive model was chosen to be the baseline model
- Snaive model shows the best error statistics scores compared to Naïve and Mean Models
- Snaive model accounts for seasonality while other two models do not
- Snaive model is robust at predicting short term as well as long term
- Snaive model is the best model out of these three models.

Model	MAE	RMSE	MAPE
Naïve	10.67	12.83	25.49
Snaïve	7.82*	10.48*	16.98*
Mean	9.59	12.32	19.30

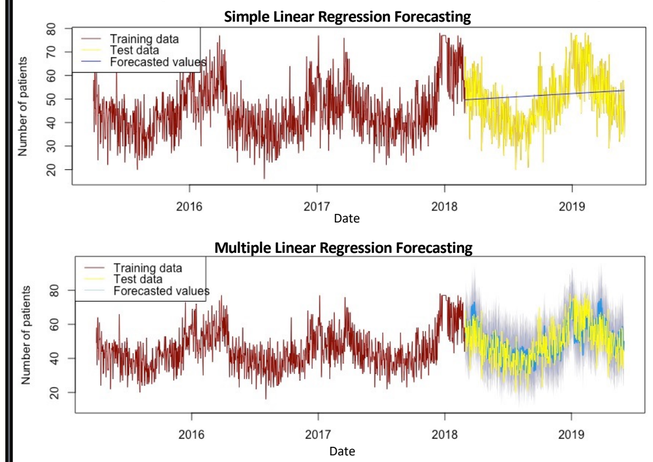
EXTRAPOLATION MODELS (SES, HOLT LINEAR & HOLT-WINTERS)



- The Holt-Winters' model outperforms the Simple Exponential Smoothing model and Holt's Linear Model, providing the most accurate fit to the data with the lowest MAE, RMSE and MAPE
- The Holt-Winters' model also captures the dynamics of the data, including its trend and seasonality.

Model	MAE	RMSE	MAPE
SES	12.27	14.53	29.89
Holt's Linear	13.52	15.86	33.11
Holt-Winters'	9.60*	12.05*	22.44*

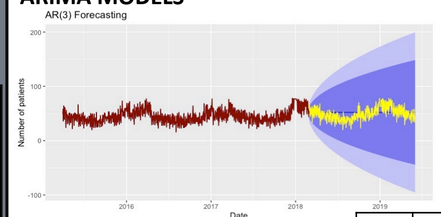
REGRESSION (SIMPLE & MULTIPLE REGRESSION MODELS)



- The SLR model performs a regression on the daily number of patients based on the time trend
- The MLR model, on the other hand, regresses the number of patients on both the time trend and seasonality
- The MLR model is the most robust model out of the two as it outperforms the SLR model in terms of MAE, RMSE and MAPE, and also captures the seasonality of the time series.

Model	MAE	RMSE	MAPE
SLR	9.41	11.49	21.56
MLR	7.51*	9.30*	16.71*

ARIMA MODELS



- When forecasting with Auto ARIMA, the ARIMA(p, q, d) parameters were automatically selected to be (3, 1, 1);
- Since the ACF plot decays geometrically and the PACF plot cuts off after lag 4, the AR(3), AR(4) and AR(5) models have been used on the first differenced series, and the most robust model AR(3) has been identified.

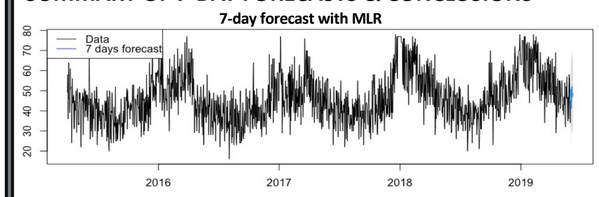
Model	MAE	RMSE	MAPE
Auto ARIMA	11.70	13.92	28.37
AR(3)	9.70*	11.84*	22.36*
AR(4)	9.74	11.88	22.52
AR(5)	9.94	12.07	23.20

SUMMARY OF ERROR STATISTICS

Model	MAE	RMSE	MAPE
Naive Method	10.67	12.83	25.49
Snaive Method	7.82	10.48	16.98
Mean Model	9.59	12.32	19.30
SES	12.27	14.53	29.89
Holt's Linear	13.52	15.86	33.11
Holt-Winters'	9.60	12.05	22.44
SLR	9.41	11.49	21.56
MLR	7.51*	9.30*	16.71*
AR(3)	9.70	11.84	22.36
MA(7)	7.57	9.57	17.40

The Multiple Linear Regression model with trend and season, the Snaive model, and the Moving Average model MA(7) outperform other models in terms of error statistics. Although all three captured the fluctuations and dynamics of the daily number of ambulance patients effectively, the MLR model demonstrates superior robustness compared to the other two when considering MAE, RMSE and MAPE. With its outstanding error statistics, the MLR model will be used as the preferred model for forecasting the number of ambulance patients for the first week of June 2019.

SUMMARY OF 7-DAY FORECASTS & CONCLUSIONS



	days_to_forecast	Point Forecast	Lo 95	Hi 95
2019.4137	2019-06-01	50.60563	32.09463	69.11663
2019.4164	2019-06-02	41.35563	22.84463	59.86663
2019.4192	2019-06-03	49.85563	31.34463	68.36663
2019.4219	2019-06-04	51.85563	33.34463	70.36663
2019.4247	2019-06-05	46.85563	28.34463	65.36663
2019.4274	2019-06-06	48.60563	30.09463	67.11663
2019.4301	2019-06-07	47.85563	29.34463	66.36663

- The daily number of patients over 1521 days, from 1 April 2015 to 31 May 2019, has been analysed using time series analysis. The analysis shows an upward trend in daily patients from 2015 to 2019, accompanied by a consistent seasonal component of an additive nature across these years. The Multiple Linear Regression model, which displayed the best performance with an RMSE of 9.30, has been used to forecast for the first 7 days of June 2019.
- The use of neural methods could be considered for future work as they can eliminate the need for feature engineering, data scaling, and making the data stationary by differencing. Recurrent neural networks are especially promising as they can handle sequential data, such as time series, without making assumptions about the independence of individual data points.