A 7-day forecast for a local medical walk-in center

Menglan Lu

Introduction

This analysis is performed on a dataset which lists the number of patients that visit the walk-in center each day for the past few years. The data recorded from 1st April 2015 to 31st March 2019.

The purpose of this forecast is to predict the number of patients in the medical center in the next 7 days and to see if there are any patterns in the data that can help managers better organize employees' work.

In this Article, we first undertake preliminary analysis including numerical summaries, graphical summaries and decomposition to examine if there are any trend, seasonality and error in the data.

Based on the results of this analysis, we will try to explore a series of time series models including naïve, SES, Holt Linear, simple linear regression, ARIMAs. We will finally choose the most suitable model according to error statistics and give a 7-day forecast.

Graphical summaries

From the time plot we can see the data slightly change between years, and fluctuate around the average. And there is an obvious seasonal pattern

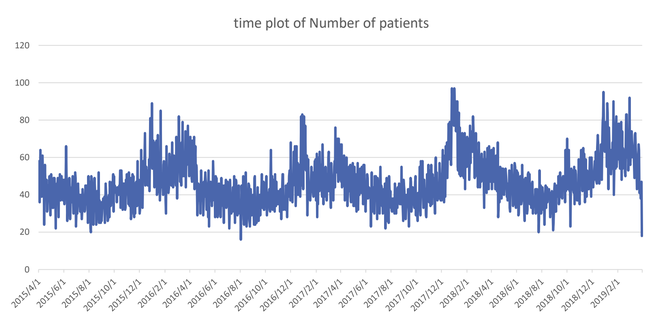


Chart 1. Time plot of number of patients.

To check the seasonal pattern, we plot seasonal chart of monthly and daily amount of patients as follow:

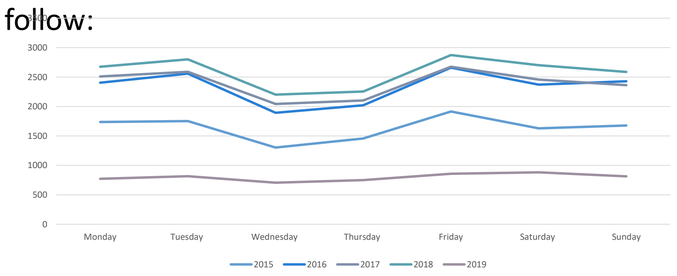


Chart 2. Seasonal chart of daily patients amount.

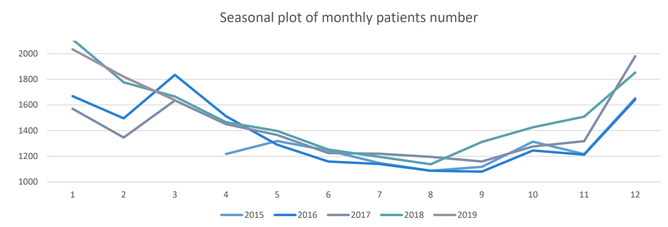


Chart 3. Seasonal plot of monthly patients amount

We can clearly see the data has a pattern on week that Wednesdays have the least patients and Fridays have the most. There are also a pattern in monthly data with summer having the least patients and then reaching its peak at winter.

To further check it, we then have a look at the autocorrelation. We plot the correlogram against month and day individually. We find there are obvious pattern on half year and week with lags larger in multiple of 6 and 7 in individual correlograms.

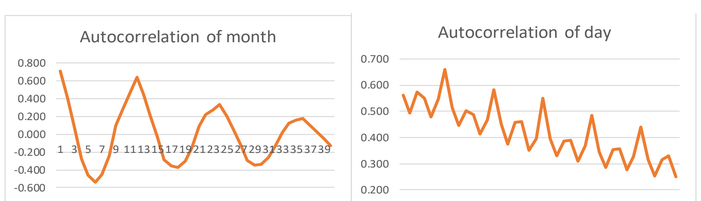


Chart 4. Correlogram of month and day

Numerical Summaries

We first take a numerical summaries, these data will show the range of the data and how much the variable changes over time.

Mean	46.72	Max	97
Median	45.00	Min	16
Mean absolute deviation	9.94	Correlation coeffient	0.27
Mean squared deviation	166.58	qaurtile_1	38
Variance	166.69	qaurtile_2	45
Standard deviation	12.91	qaurtile_3	54

Table 1. Numerical summaries

Decomposition

From previous analysis, we can see the number change slightly between each year, so we choose additive model. We can see an obvious seasonal pattern from the decomposition. In order to see the seasonal indices clearly ,we only plot the very beginning data points in chart 6

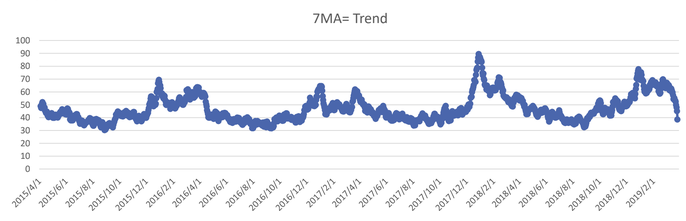


Chart 4. Trend of the data (7MA)

Chart 5. Error of the data

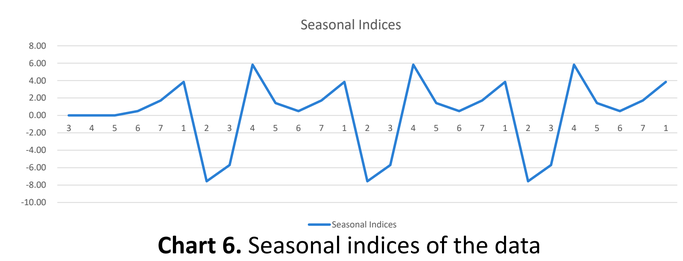


Chart 6. Seasonal indices of the data

7 day forecast

day	NF1	SES	Holt-li	simple regress	ARIMA
2019/4/1	18.00	43.47	32.89	50.57	65.02
2019/4/2	18.00	43.47	30.56	50.58	65.02
2019/4/3	18.00	43.47	28.23	50.58	65.02
2019/4/4	18.00	43.47	25.90	50.59	65.02
2019/4/5	18.00	43.47	23.57	50.59	65.02
2019/4/6	18.00	43.47	21.24	50.60	65.02
2019/4/7	18.00	43.47	18.91	50.61	65.02

Table 3. 7 day forecast summary

All Models used

Five models were used including naïve(NF1), SES, Holt Linear, simple linear regression, ARIMAs:

SES with the α=0.159;**Holt linear** with the α=0.23， β=0.16，m=1;**Simple linear regression** with b1=0.0061, b0=41.66.

With the ARIMAs model, we try several models with different order of differencing, moving average, autoregressive and seasonal component and finally choose the A(2,0,1)(1,0,1)[12] one according to the better Reference coefficient BIC and AIC.

The model of ARIMAs is shown below:

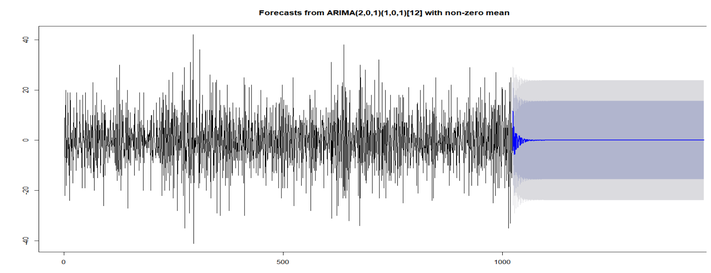
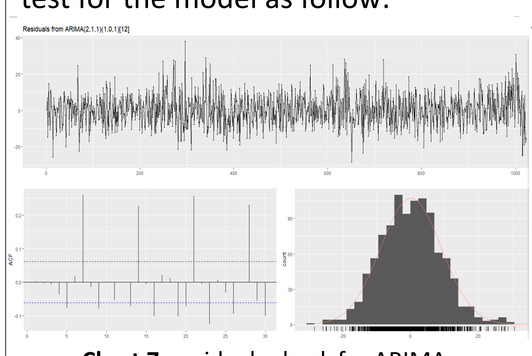


Chart 6. ARIMAS model

Then we also check the autocorrelation and do Ljung-Box test for the model as follow:



The five models and original data are plotted against date in the same chart to see if these models capture any possible pattern and trend.

Chart 7. residuals check for ARIMAs

Considering the date points is too much and we will hard to see trends or patterns, so I just plot the beginning 30 data points of testing sets. Which are show blow:

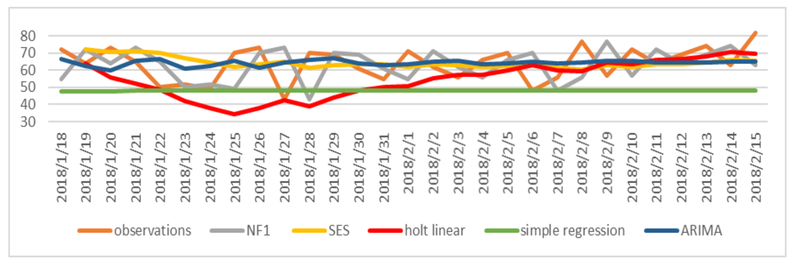


Chart 8. all models and original data

we can see SES model fitting the dataset best. We will check it using error statistics latter.

Error statistics

NF1 model is a baseline model. The model which gains a smaller MSE,MAE,MAPE could be considered as a better one. We can see SES (simple exponential smoothing) model is best again. So we Choose SES for our 7 day forecast.

	training sets		
	MSE	MAE	MAPE
NF1	145.11	9.55	22.51
SES	85	7.3	17.45
Holt-linear	96.62	7.77	18.6
simple regression	145.92	9.23	21.98
ARIMA	156.13	7.13	16.97
	test sets		
	MSE	MAE	MAPE
NF1	147.87	9.49	19.63
SES	91.12	7.61	16.11
Holt-linear	105.93	8.05	16.68
simple regression	179.02	10.56	21.79
ARIMA	366.99	16.36	39.52

Table 2. Summary of error statistics for each model

Conclusions

According to error statistics and models' chart, we choose SES model for our forecast model and estimate around 43 people would go to the health center everyday in following 7days.

However, we still have a lot of problems need to be solved. For example, the ARIMA model can only make precision forecasting for nearby data, but at the end of the model, the prediction value are always constant. And also we observed there are seasonal indices both on week and month, we can attempt to tackle them at the same time in the future.