ASSIGNMENT SUBMISSION FORM

Treat this as the first page of your assignment

Course Name: FCAS Project Work

Assignment Title: Forecast daily demand for a week in the top region in terms of origination of bookings

Submitted by: Group

Group Member Name	PG ID
Ankit Kansal	61410768
Shruti Jain	61410519
Rahul Gupta	61410608
Garrett Butler	61419107
Vikram Deshpande	61410485

1. Executive Summary

1.1 Problem description

After visualizing the data in the first cut, we picked up the issue to forecast regional level demand. In the Indian context, where radio cabs have flooded the market, it has become very necessary to plan and forecast the demand at a regional level. Yourcabs works on taxi aggregator model. Hence, to devise a real time schedule for vendors based on the regions that they operate is very critical. This will help not only to manage capacity allocation but also to develop new vendor relationships.

1.2 Brief description of the data, its source, key characteristics, and chart(s)

The data given to us contains the historical demand numbers at a latitude and longitude level both at origin and destination. We have picked up the demand origins in the light of business issue mentioned above. Further, the data contains the mobile, phone and online booking as three different demand patterns. We have strived to forecast the aggregate demand by combining each of the three series independently. Figure 3 provides the visualizations of the series used.

1.3 High level description of the final forecasting method and performance on meaningful metrics (compared to benchmark)

We used a wide variety of methods ranging from Naïve to Neural net and ensembles to arrive at the final forecasts. We started with the Naïve method for each of the series and used it as a benchmark for any future models. Our final recommendation uses a mix of the following based on the patters for each of the series: Regression +AR (1) [Total Demand, Mobile bookings], Ensemble (Lag2+Neural Net) [Phone bookings], Holt Winters [Online bookings], Ensemble [demand ratio for region 1].

To arrive on a conclusion for the best methods, we benchmarked each method's output from the naïve forecast, MAPE figures, ACF plot for residuals, and visualization of residuals.

1.4 Conclusions and recommendations

The daily demand when forecasted using the aggregate of the bookings from the three channels performs better than the forecast of the total demand. This demand when used with the forecasted proportion of the demand from the top-booking region can give us the estimated demand from that region. In view of the increasing overall trend of demand, *yourcabs* can use this information to manage vendor relationships and capacity allocation for better serving its customers.

2. Technical Summary

2.1 Forecasting Problem

a) To forecast the daily number of total bookings for *yourcabs* from the city of Bangalore and then use it to forecast the demand from the top region in terms of bookings.

b) **Forecast Horizon and granularity** - one week and daily data, which we think will provide suitable time to *yourcabs* to manage regional capacity and manage its vendor relationships.

c) **Percentage demand in an area** – Demand in any particular area is a function of–

   a. Intrinsic demand in the area due to the income levels/ population/ and other such factors

   b. Overall performance of *yourcabs* which reflects awareness about the company and customer willingness to try its services

   We thus define R1 as the ratio of demand D1 to the total demand D witnessed by *yourcabs.*

d) Total Demand – Total demand D is again a sum of demand via phone bookings (P), online site (O) and mobile site (M). Note that we are assuming that the bookings which are done through neither mobile and online booking are done through phone. We have segregated the forecasting exercise into channels of booking to account for the changing technological landscape. This will help us use appropriate forecasting models for series components and give us more accurate forecasts. We believe that the aggregate of these 3 series will give us better forecasts than the total demand series in itself. However, we will compare the performance of the two forecasting techniques.

Thus, five time series of interest are P, O, M, R1, and D.

And metrics of interest D1 is obtained as –

D1 = intrinsic factors**yourcabs* performance = R1*(P+O+M)

2.2 Preprocessing the Data:

Defining Regions: We first divided the city into nine major demand zones. Below is a chart of the nine major regions based on the total demand pattern. Each region is defined at 2.5 km in radius. We identified that Airport, Majestic and Marathalli each account for around 8% of total demand. We selected region 1 (airport) as a base region to forecast region level demand as a ratio of total demand.

Data visualization on top of the Map of Bangalore was made possible by using Tableau software

a) Because our goal was to forecast demand one week ahead on a daily basis, we first aggregated from hourly demand into daily demand. We then used Tibco Spotfire's visualization tools to see if we there are any trends or seasonality in the overall demand. We were able to see some seasonality in both monthly and weekly demand. **(Figure 4)**

b) In general, there seemed to be a lull in the mid-week data (Tuesday and Wednesday) and more demand from Thursday to Monday. Also the monthly data seemed to have a peak around the late summer months (August and September) but this might also be affected as a result of having incomplete data for the later part of 2013. We thus broke up the data into dummy variables for both day of week and month of year, which resulted in a total of 17 dummy variables. (when using neural networks and linear regression)

c) We saw spot an upward trend, which seemed to fit a sine/cosine cycle for the overall demand.

d) We then further broke up the demand into the three types of channels by which bookings were made - mobile, online and phone and visualized each on Spotfire. **(Figure 5)**

e) Mobile and online bookings displayed an upward trend while phone bookings seem to plateau off.

f) Handling missing data - We then replaced all missing data fields with zero. Though this seemed to create some outliers as a result, most of the missing data occurred within the first half of the first year, when the off take was relatively slow and the points actually represented meaningful data points. Thus we decided to keep it as is and not worry about applying other methods such as averaging with nearest neighbors.

g) Data range - we decided to only use the data from 11/1/2011 to 11/15/2013 because there were several instances where bookings were made but had yet to be fulfilled and we didn't want that to alter our measuring methods on the training and validation data.

h) Training and Validation periods - We decided to partition the data with the validation data only containing the last 28 days of data because we were only forecasting ahead one week.)

2.3 Reference Forecast

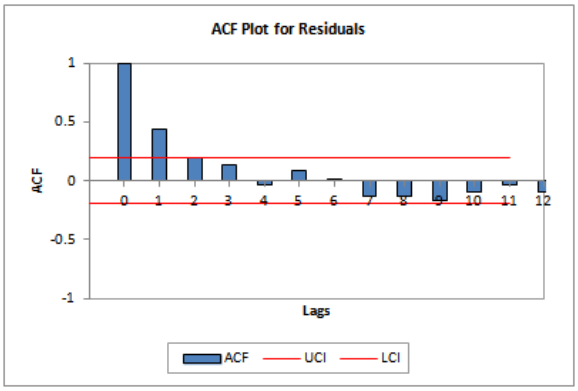
A seasonal Naïve forecast was used as the benchmark forecasting technique. A season was defined as the day of a week, e.g, for next Sunday, we would simply use a naïve forecast of last Sunday's Demand.

2.4 Forecasting methods used

2.4.1 Forecast of Total Demand

We used linear regression with sine and cosine inputs and applied the model on weekly average demand to eliminate seasonality. Then a closer look at the residuals highlighted the need to incorporate AR(1) layer into the model. The obtained weekly forecast was then split into daily forecast by using weekly demand trend of the previous month. Plot of actual and forecasted data is shown in the Figure 6 in Appendix. Forecasted daily values are presented in Appendix 5.

Weekly average Demand (t) = a +b(t) + c*Sin(2*Pi*t/52.18) + d*Cos(2*Pi*t/52.18) +AR(1) Layer



AR plot of residuals after linear regression shows the necessity to incorporate an AR (1) layer into the final model

Method	MAPE
Naïve	23%
Regression + AR(1)	**17%**
Exponential Smoothing	28%
Neural	28%

Week Ending	Sun	Mon	Tue	Wed	Thu	Fri	Sat
22-Oct-13	*127%*	*94%*	*96%*	*85%*	*92%*	*90%*	*117%*
29-Oct-13	*86%*	*107%*	*85%*	*94%*	*98%*	*118%*	*113%*
5-Nov-13	*78%*	*98%*	*92%*	*113%*	*136%*	*113%*	*71%*
12-Nov-13	*69%*	*94%*	*109%*	*107%*	*115%*	*95%*	*111%*

Sample distribution of weekly demand for the validation month. For example, the demand on Saturday of week ending 22 Oct 2013 was 117% of the average demand in the week.

2.4.2 Forecasting the number of bookings through online mode

The past data of online bookings showed a steep increasing trend over the given duration. The data also had many local maxima and minima. The Holt Winter's model was used to forecast the data. However, the 'level' of the forecast data was more than the actual data. A second level forecast using the residuals did not reduce the 'level' of the data. Therefore, an ensemble approach was adopted. The weighted average of the output from both Holt Winter's and Linear regression model was used to forecast the number of bookings that were made on-line.The below graphs shows the 'level' of data when (a) only Holt Winter's model was used and (b) Ensemble model was used.

Naïve model:

The raw data for the online booking series exhibited different patterns on different days of the week. Therefore, we had to use the value on the same day of the previous week as the Naïve forecast. The statistical measures for the Naïve model and the Ensemble model are tabulated below:

Model	RMSE	MAPE
Ensemble	35.02	48.00
Naïve (weekly)	44.11	43.90

2.4.3 Forecasting the number of bookings through mobile site

The data series for mobile site bookings starts from 24th Feb 2013 and shows an increasing trend. We started with a naïve forecast (with the last 4 weeks as validation data) of the data series and then used different models like simple linear regression and exponential smoothing to forecast the demand. On

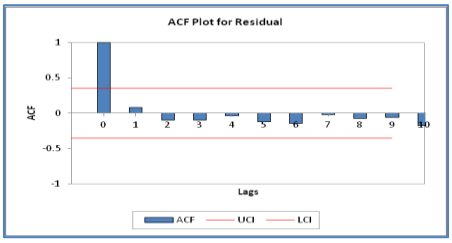
comparing the performance of various models, we finally used the below method to forecast weekly average demand:

Demand (t) = a + b * t + c * t^2 + d * Demand (t-1)

The regression model is as follows:

Input variables	Coefficient
Constant term	-0.36816841
Week	0.44542211
Week square	-0.01011591
Lag 1	0.58167869

Forecasting method	MAPE
Seasonal naïve	66.07%
Regression + AR(1)	35%



The ACF plot for the residuals shows that there is no correlation between the residuals:

2.4.4 Demand through phone bookings site (M)

The data series for phone bookings shows a stabilizing trend.

Ensemble RMSE	**23.59835**
Ensemble MAPE	**26.09166**
Naïve RMSE	**33.26356**
Naïve MAPE	**30.91721**

We observed a negative correlation between Linear Regression (Lag2) and NN and eventually use an ensemble with 70% LR and 30% NN.

2.4.5 Demand in a region

Weighted RMSE	**10.43509**
Weighted MAPE	**89.12875**
Naïve RMSE	**13.72693**
Niave MAPE	**110.4779**

We applied an ensemble approach of 50% LR and 50% NN to the time series of Ratio R1. Correspondingly, demand forecast for the airport region was obtained as D1 = R1 * (M+P+O) , as plotted in the appendix

Figure 1 The prposed forecasting approach beats the seasonal naive

2.5 Performance Comparisons

For the performance metrics we compared the naïve MAPE and RMSE to whichever model had the lowest combination of values. The reason we focused on MAPE was because most of us have prior experience using this and it was a quick and easy way to compare between models. The reason we choose RMSE was because this made the most business sense and this figure can easily translate to people who aren't familiar with statistics. Refer figure 10.

By segregating the forecasting exercise into channels of booking, we had the flexibility to apply different models specific to the underlying series. Demand forecast obtained by aggregating forecasts on mobile, phone and online booking channel was better than the forecast obtained by applying a regression + AR(1) model on total demand.

3. Conclusion

The total demand is a factor of the variation of mobile, phone and online booking. Since, all of the series have different demand patterns we cannot directly forecast the total demand of the company based on the total series. Plot of the aggregate series (sum of individual forecasts of mobile, phone and online) is a better fit plot (based on MAPE, residuals) as compared to the forecast series of overall total demand.

Further, our hypothesis that the total demand from a region will be comprised of two factors – intrinsic demand from that region and extrinsic drivers of the company brand name stands true. Hence, we have forecasted the ratio of the region 1 demand with respect to the total demand.

4. Recommendations

From our analysis, we observe that the demand of online and mobile are on a growth trajectory while the demand from phone is almost constant. Hence, the demand drivers in the near future will be from online and mobile components. We recommend that your cabs should invest in more IT infrastructure from a futuristic point of view.

Also, the demand from region 1 is growing a ratio of the total demand from its current levels of 8%. We recommend that the company should focus on gaining more fleet from the airport. This increasing rend combined with our forecasts will help plan the fleet and manage resources, build strong and better vendor relationships at a region which serves as the peak demand.

5. Appendix

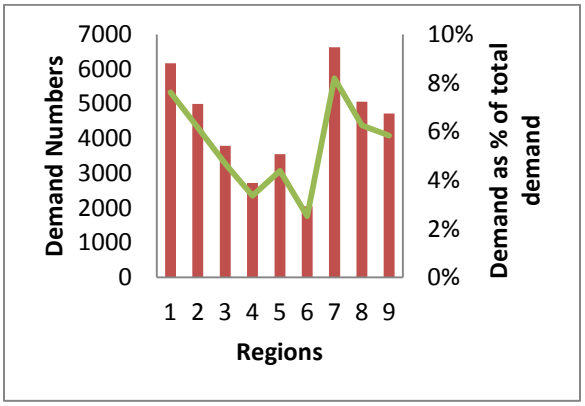


Figure 1 % demand in top nine regions

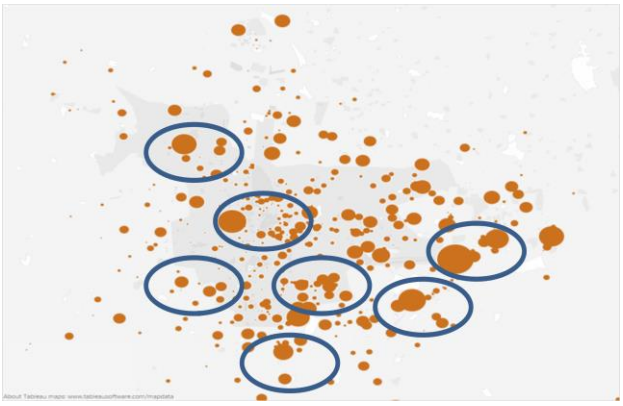


Figure 2 region definitions - 2.5 km radius, visualization through Tableau

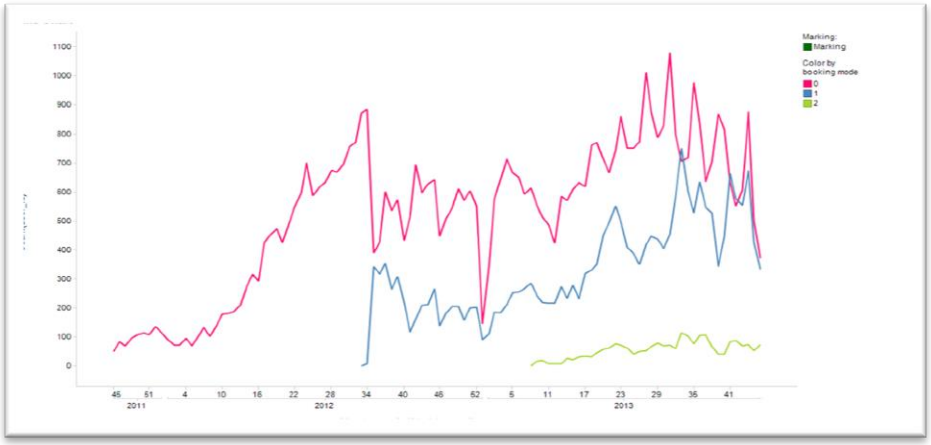


Figure 2 Total demand split by channel type; rise in mobile and online bookings given the changing technological landscape

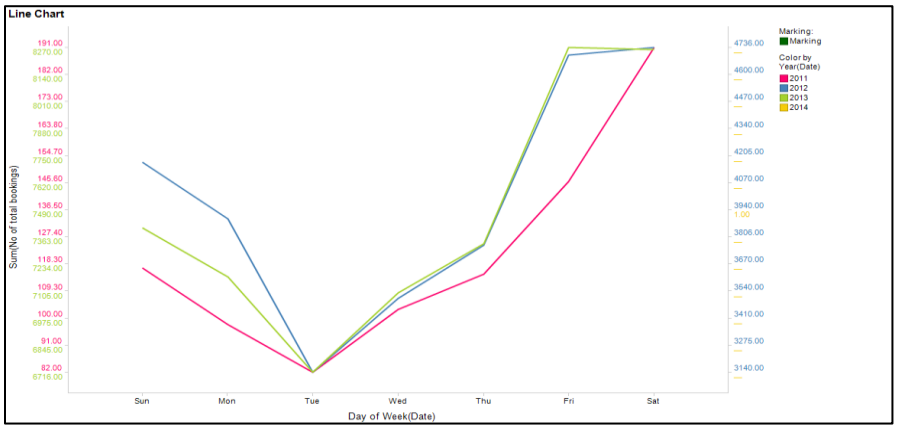


Figure 3 Weekly seasonality in total demand; An average Tuesday has 15-20% less demand than an average Saturday

		Forecasted Total demand	*Ratio of daily demand to weekly average*
Sun	17-Nov-13	145	*84%*
Mon	18-Nov-13	175	*101%*
Tue	19-Nov-13	166	*96%*
Wed	20-Nov-13	173	*100%*
Thu	21-Nov-13	200	*116%*
Fri	22-Nov-13	175	*101%*
Sat	23-Nov-13	175	*101%*
	Average Forecast obtained	172.53	

Figure 4 Forecast of total demand

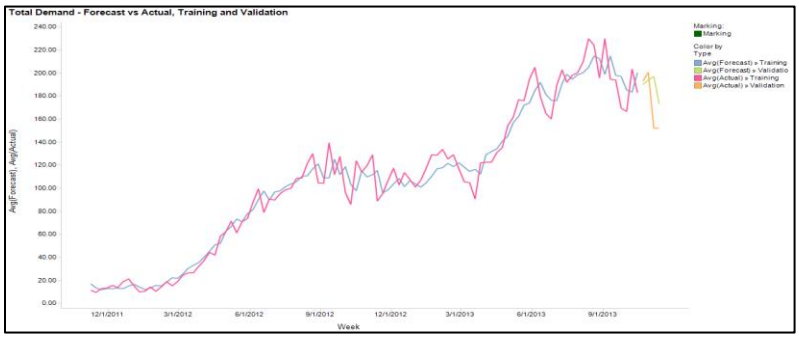


Figure 5 Total demand Actual vs Forecast

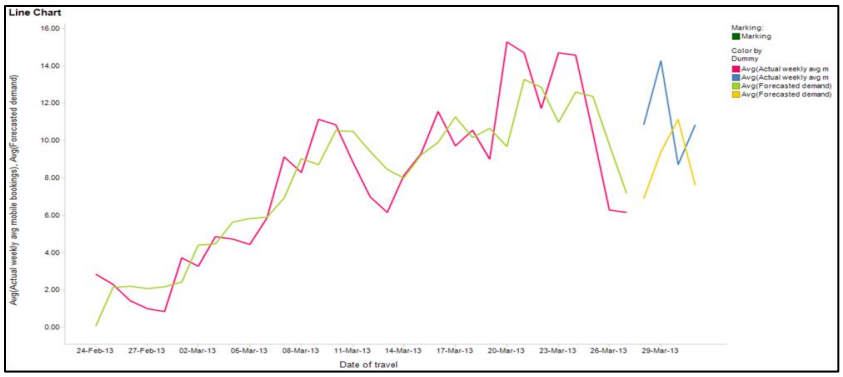


Figure 6 Mobile bookings - Forecast vs Actual on training and validation data

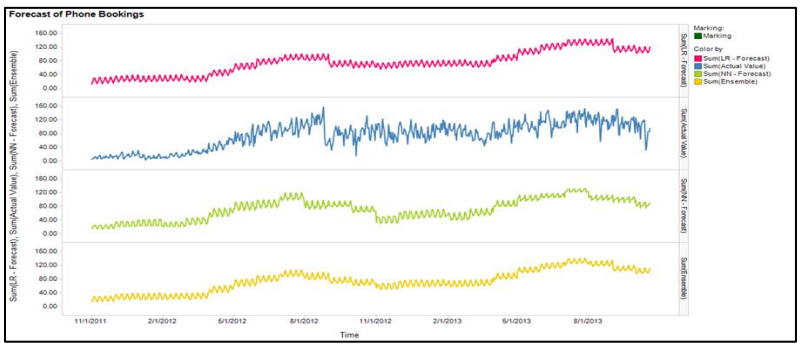


Figure 7 Forecast for Phone bookings, LR, NN and ensemble

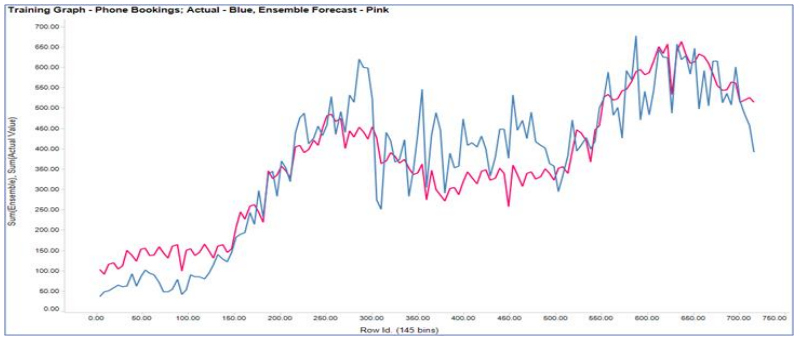


Figure 8 Phone bookings, Forecast vs Actual

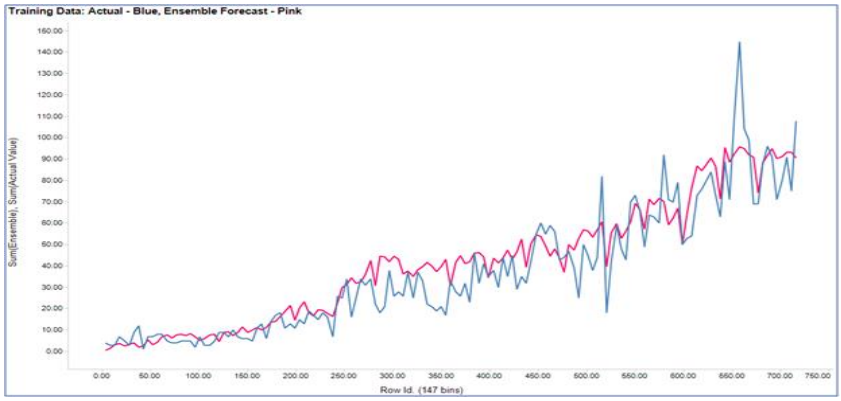
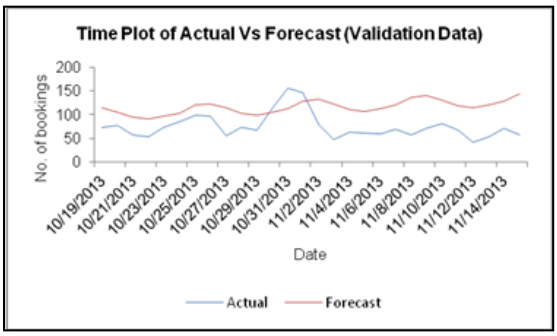


Figure 9 Demand in the region Bookings, Forecast vs Actual



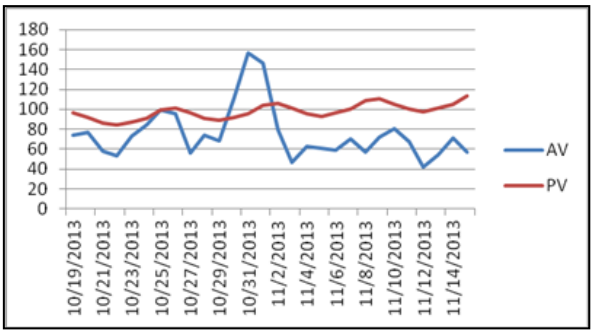


Figure 10 Ensemble

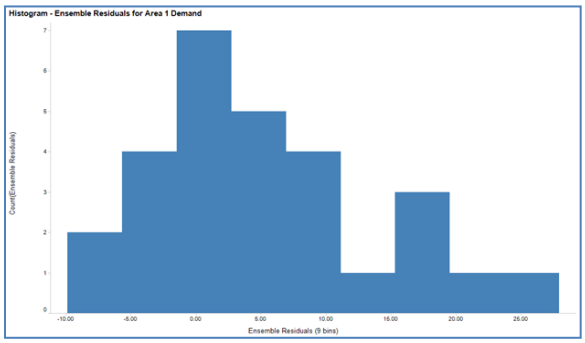


Figure 11 Plot of residuals for Regional demand bookings

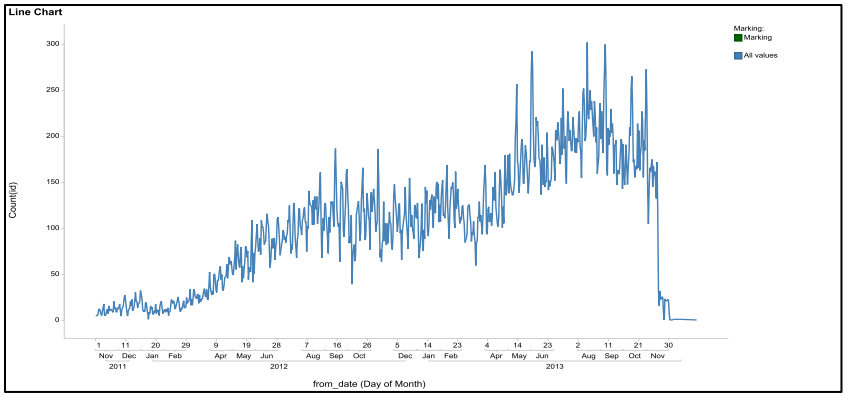


Figure 8: Total demand pattern (Series 4)

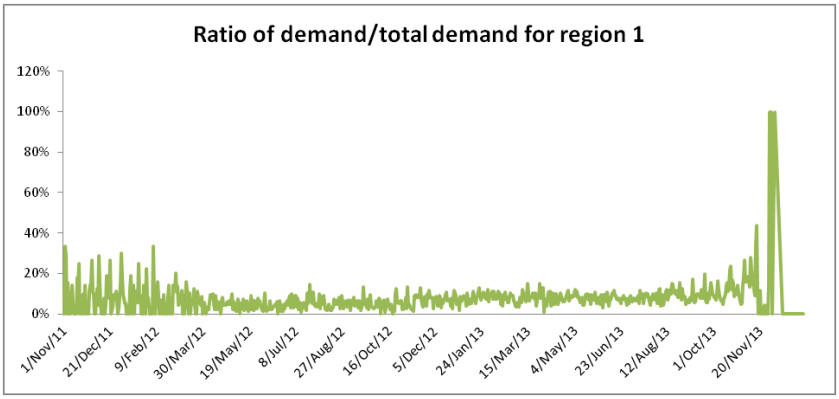
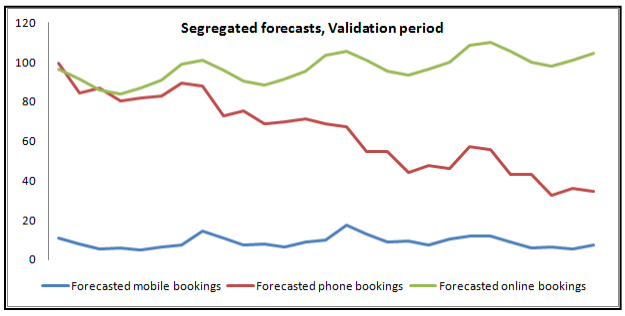
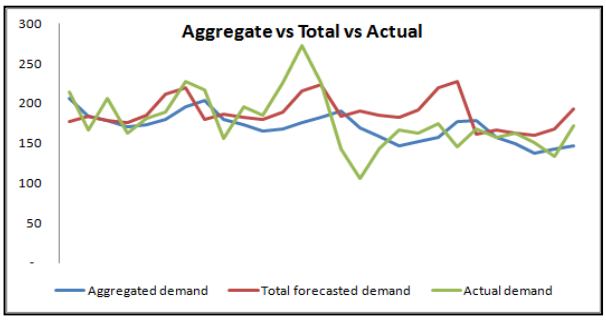


Figure 9: Ratio of airport demand to total demand (Series 5)

Comparison of total demand and aggregate demand forecast





MAPE	
Aggregate	14%
Total	17%