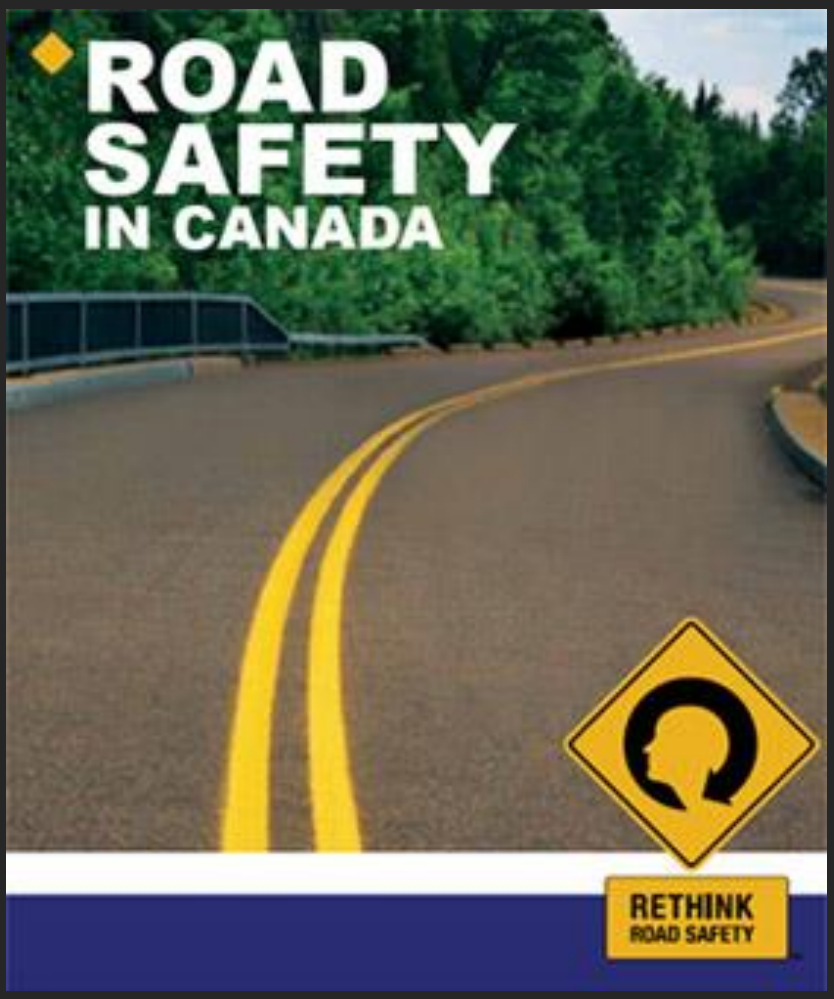
FORECASTING BIKE TRAFFIC FOR BETTER TRAFFIC MANAGEMENT IN OTTAWA CITY

FCAS Project



Authored by - Study Group B8
RHISHABH GARG - 61910352
GIDEON JAMES DRAVIAM - 61910680
UJJWAL KEJRIWAL - 61910587
MRIDUL KRISHN MISHRA - 61910569
JASMEET SINGH - 61910021
KUHELI JATI - 61910343

1. EXECUTIVE SUMMARY

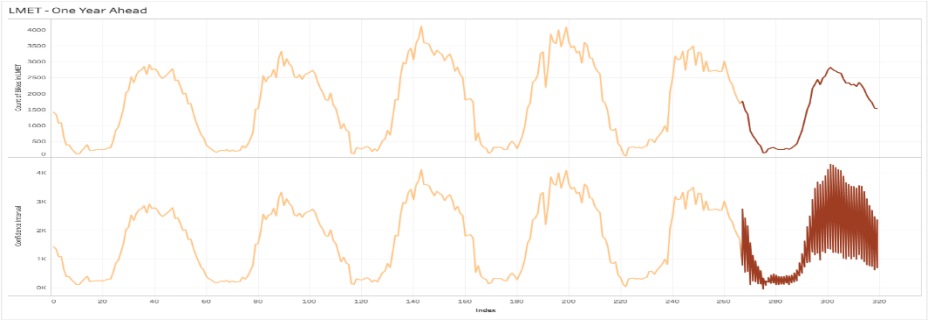
Problem Description:- The business problem involves the forecasting of Ottawa city bike traffic to understand the requirement of increasing bike corridors in city. The traffic accidents in Ottawa is on rise with a 7-year peak reaching in 2017. As per Ottawa police, the traffic accidents have increased by 23%. Additionally, the registered two-wheeler vehicles went up by 15000 with 5000 additional drivers.

The analysis of data across 6 junctions in Ottawa helped to solve the below two problems-

- Identifying the high traffic areas for bikers and small light-weighted vehicles, and planning for construction of special biker lanes
- Hiring of traffic management personnel based on peak times

We check for the traffic during different months of the year for the last 5 years (4 years training + 1 year of validation) to forecast for the next coming years. The forecasting horizon is 1st Nov'17 to 31st Oct'18.

Performance Metrics:- Based on running various models on the training and validation data set, we arrive at the conclusion that "Holt's multiplicative" is the best fitted model. The model is performing better on all the metrics like MAPE, SSE etc.



			Benchmark	
			SNaive	MHW
LMET	Year 1	RMSE	500.788341	324.4784
		SSE	13291815	5580170
		MSE	250788.962	105286.2
		MAP	21%	18%
	Year 2	RMSE	434.237359	342.4266
		SSE	20176143	12546391
		MSE	188562.084	117256
		MAP	24%	17%

Recommendations: -

Short Term: - Months of June and July see spikes in traffic for some select days. It can be used for the efficient allocation of traffic personnel

Long Term: - Creating corridor extensions for the junctions of OBVW, OGLD, SOMO.

2. TECHNICAL SUMMARY

DATA CLEANING REQUIRED

- We have taken two-wheeler traffic data captured in six different traffic counters spread across Ottawa city. The data has been collected from the transportation department of Ottawa city

- Analysis is done over a five-year period starting from November 2013 to November 2018 where first 4 years are used for training while last year for validation
- There were quite a few missing values in the data, especially during the winter months when presumably the counters went kaput. For these missing values, we took the average of the recent weeks and completed the dataset

DATA AGGREGATION

- Due to business objective and high data fluctuation, we aggregated the data at weekly level
- To accomplish this, we analysed the data of all the 7 days of each week and took the maximum traffic amongst each of these 7 days as the capacity needed each week
- So basically, we have used the approach of upper capping or maximum shielding to ensure that the corridors have sufficient space and they aren't running over capacity on any day of the week
- Post aggregation, we used this max. weekly traffic data to train our model using 4 years, validate it over 1 year and then forecast it for 1 year after choosing the model with least error and best fit

CATEGORICAL VARIABLES

- We have 'WEEK' variable as our categorical variable which essentially is the week number ranging from 1 to 53 for each of the five years. But Excel doesn't allow a categorical variable to take more than 30 distinct values while we had 53 distinct values representing 53 weeks of a year
- To deal with this, we made a separate variable 'WINTER WEEK' which represented the data of the weeks that fell under the winter season. This way, 'WEEK' had 30 distinct weeks under its umbrella while 'WINTER WEEK' had 23 thus satisfying the condition

3. FORECASTING METHODS USED

		Smoothening					Regression			
	Seasonal Naïve	Moving Average	Simple Exponential Smoothening	Double Exponential Smoothening	Holt's Winters Additive	Holt's Winter's Multiplicative	Linear (T + Weekly Index)	Linear (T + T^2 + Weekly Index)	Logarithmic (T + Weekly Index)	AR Model
Level	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Trend	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓
Season	✓	✗	✗	✗	✓	✓	✓	✓	✓	✓
Auto-Correlation	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
Business Objective (No Running Forecast)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗

The methods that could be used for Smoothening are:

1. Moving Averages
2. Simple Exponential Smoothening
3. Double Exponential Smoothening
4. Holt's Winters Additive
5. Holt's Winters Multiplicative

Methods that could be used for regression are:

1. Linear (T + Weekly Index)
2. Linear (T + T^2 + Weekly Index)
3. Logarithmic (T + Weekly Index)
4. AR Model

In the data, we have both trend and seasonality. Therefore, we've used models that capture these factors.

Among smoothening methods, only Holt's Winters method captures both trend and seasonality because of which this method is used for forecasting. To find out whether the data has additive or multiplicative seasonality, we tried both methods and majorly, Holt's Winters Multiplicative was the most appropriate model. Importantly, this type of forecasting is one-time static forecast and not a continuous one which makes it a better method.

We performed regression as it considers the external factors and focuses on seasons more granularly. We tried 3 variants: Linear – T, Linear – T^2, and Multiplicative – log(T). Majorly, Logarithmic – T showed the best results.

4. MODEL EVALUATION

1. We partitioned the data twice. First, 4 years were taken as training period, and 1 as validation. In the second partition, 2 years were taken as validation period.
2. Validation performance was also evaluated with 2 years, to negate isolated instances which can hamper the performance metric of the model

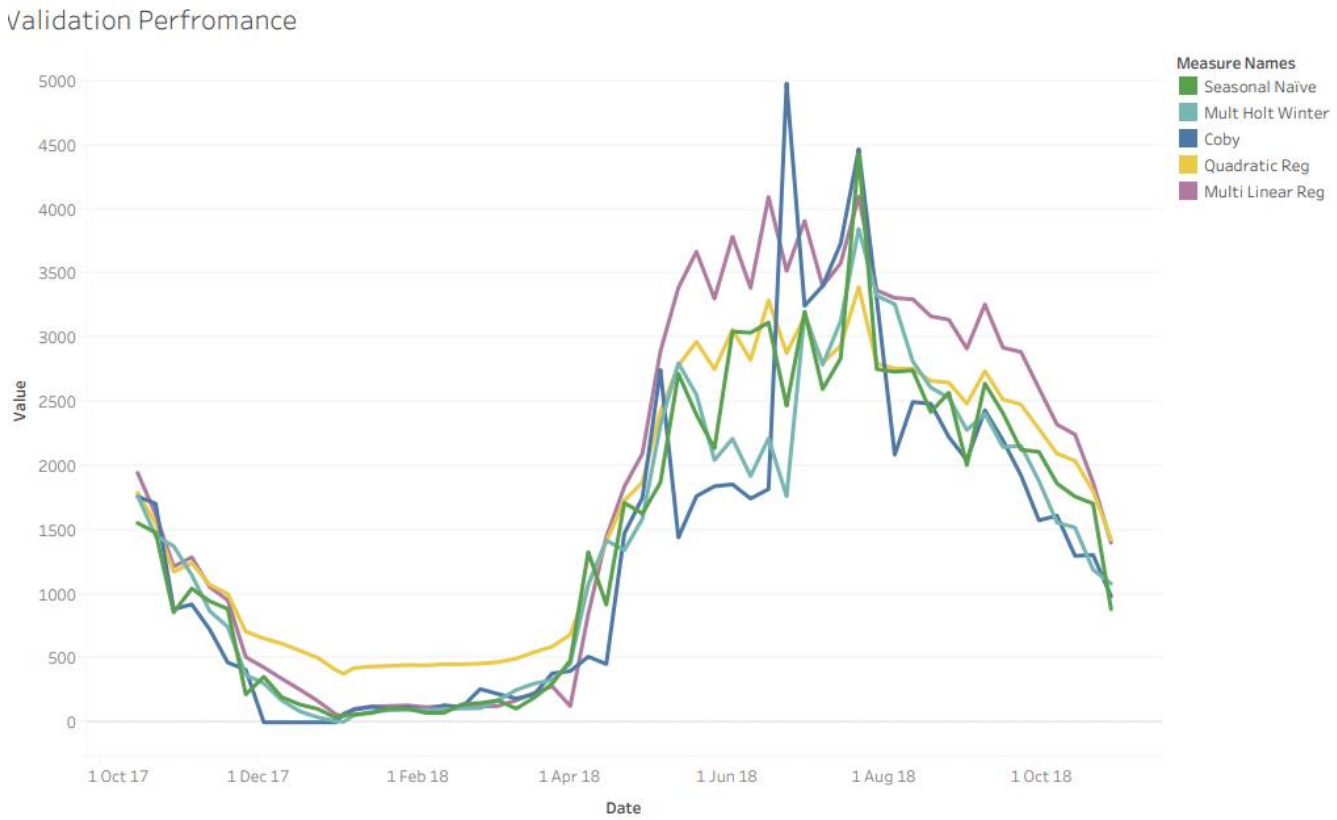
3. Seasonal Naïve of 1 year was taken as the benchmark model
4. Metrics of Interest which we specified: We need to have the model performing its best at the peaks, hence, the errors SSE, MSE, RMSE and MAPE at the peaks in the season are considered
5. To maintain congruency of the series, and the scale, we considered MAPE as the primary metric Our data contains trend, seasonality. We removed noise by aggregating data to week level, and ran three smoothening methods, and three regression methods.

The validation performance results for the 6 counter data are presented below

			Benchmark	Smoothening			Regression		
			SNaive	DE	MHW	AHW	LTS	LQTS	MLTS
COBY	Year 1	RMSE	593.290155	639.6202	582.7585	585.331	607.4637	668.5677	818.9306
		SSE	18655640	21683041	17999195	18158458	19557643	23690088	35544312
		MSE	351993.208	409114	339607.5	342612.4	369012.1	446982.8	670647.4
		MAP	29%	29%	26%	31%	60%	90%	41%
LMET	Year 1	RMSE	500.788341	729.7026	324.4784	325.1053	2094.987	726.775	729.7026
		SSE	13291815	28220693	5580170	5601754	2.33E+08	27994702	28220693
		MSE	250788.962	532465.9	105286.2	105693.5	4388972	528201.9	532465.9
		MAP	21%	68%	18%	27%	90%	59%	68%
OBVW	Year 1	RMSE	383.102282	296.2126	310.9741	250.7224	411.4875	389.2967	248.7749
		SSE	7778670	4650322	5125359	3331671	8974065	8032252	3280114
		MSE	146767.359	87741.92	96704.89	62861.72	169322	151551.9	61888.94
		MAP	41%	50%	32%	50%	289%	422%	25%
OGLD	Year 1	RMSE	494.11	553.6205	498.8572	464.151	291.304	319.878	2513.97
		SSE	12939437	16244267	13189503	11418117	4497474	5423062	3.35E+08
		MSE	244140.32	306495.6	248858.6	215436.2	84858	102321.9	6320045
		MAP	28%	62%	56%	43%	57%	77%	157%
ORPY	Year 1	RMSE	757.394292	537.2081	456.4612	372.3085	375.5374	502.4588	2408.939
		SSE	30403244	15295403	11042910	7346522	7474501	8256754	3.08E+08
		MSE	573646.113	288592.5	208356.8	138613.6	141028.3	231421.2	5802989
		MAP	45%	32%	47%	111%	81%	103%	84%
SOMO	Year 1	RMSE	123.659606	130.7609	125.8483	115.5498	128.5035	120.0769	616.6638
		SSE	810460	906216.4	839402.6	707643.6	875197.6	764177.9	20154534
		MSE	15291.6981	17098.42	15837.78	13351.77	16513.16	14418.45	380274.2
		MAP	28%	32%	24%	26%	25%	35%	87%

			Benchmark	Smoothening			Regression		
			SNaive	DE	MHW	AHW	LTS	LQTS	MLTS
COBY	Year 2	RMSE	531.762657	545.5412	468.6041	457.3917	468.6492	505.1234	614.4169
		SSE	30256553	31844825	23496112	22385162	23500636	27301008	40393373
		MSE	282771.523	297615.2	219589.8	209207.1	219632.1	255149.6	377508.2
		MAP	38%	39%	25%	27%	60%	75%	32%
LMET	Year 2	RMSE	434.237359	728.1715	342.4266	308.4142	2255.269	544.2611	728.1715
		SSE	20176143	56735007	12546391	10177770	5.44E+08	31695552	56735007
		MSE	188562.084	530233.7	117256	95119.34	5086239	296220.1	530233.7
		MAP	24%	72%	17%	27%	95%	48%	72%
OBVW	Year 2	RMSE	349.61708	276.0973	276.6242	223.6592	323.2137	307.9929	217.8533
		SSE	13078835	8156582	8187739	5352510	11177976	10149981	5078228
		MSE	122232.103	76229.74	76520.92	50023.45	104467.1	94859.64	47460.08
		MAP	143%	110%	44%	71%	314%	390%	34%
OGLD	Year 2	RMSE	381.334094	411.6047	374.2524	340.1815	232.8658	247.7017	1901.777
		SSE	15559479	18127774	14986942	12382407	5802235	6565105	3.87E+08
		MSE	145415.692	169418.4	140064.9	115723.4	54226.49	61356.12	3616757
		MAP	33%	50%	38%	33%	63%	74%	115%
ORPY	Year 2	RMSE	697.407777	599.4078	565.0089	441.4673	431.1039	511.1325	2277.937
		SSE	52042404	38443997	34158153	20853587	19886014	22564867	5.55E+08
		MSE	486377.608	359289.7	319235.1	194893.3	185850.6	205450.1	5188999
		MAP	165%	72%	64%	141%	99%	121%	88%
SOMO	Year 2	RMSE	118.095795	116.6549	102.7716	95.68615	103.935	97.24142	631.6043
		SSE	1492288	1456096	1130133	979674.8	1155866	1011781	42684869
		MSE	13946.6168	13608.37	10561.99	9155.84	10802.49	9455.893	398924
		MAP	33%	33%	21%	24%	22%	28%	92%

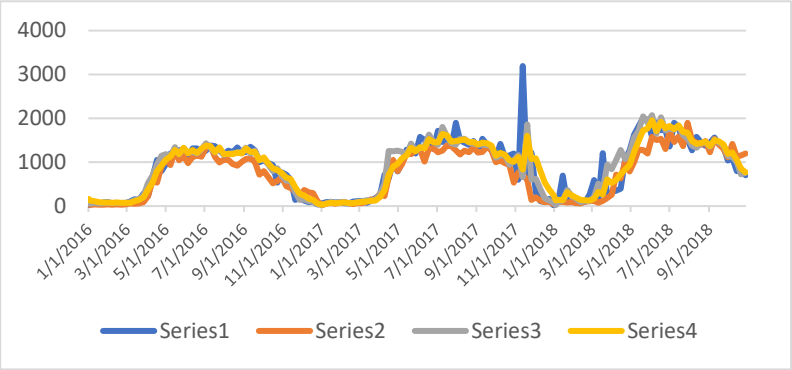
*Graphs for all these models and their residuals are given in the appendix



he trends of Seasonal Naïve, Mult Holt Winter, Coby, Quadratic Reg and Multi Linear Reg for Date. Color shows details about Seasonal Naïve, Mult Holt Winter, Coby, Quadratic Reg and Multi Linear Reg. The view is filtered on Date, which ranges from 09/10/2017 to 28/10/2018.

OBSERVATIONS FROM THE RESULTS

- For COBY, LMET, and SOMO, Holt's Winter's with multiplicative seasonality gives the best MAPE.
- For OBVW, Linear Regression with multiplicative seasonality worked best.
- For ORPY, running the model on two-year data, shows Multiplicative Holt's Winter as best model.
- For OGLD, we observe a high value outlier in the November month of 2017. This datapoint biases the model leading to high error. For this reason, the naïve forecast comes out as the best fit on overall data. To check the efficiency of all the models at the peaks, we plot the forecasts and focus on peaks.



Measuring the errors here,

2 Year	SN	DE	MHW	AHW	LTS	LQTS	LMTS
MAPE	0.9920069	0.741506	0.67081	0.584742	0.618747	0.657863	6.78443

we see that the Additive Holt's Winter's smoothening method performs the best.

We optimize the α, β and γ values, and hence our models have some over-fitting on the training data.

Key Takeaways :

1. Model selection was done based on the performance compared to each other on predicting the peaks
2. We chose the model which over-forecasted and had minimum error as the cost of under-forecasting is high (life)

EXTERNAL FACTORS NOT CONSIDERED IN THE ANALYSIS

1. Weather
2. Holidays
3. Traffic in nearby counters
4. Vehicle Growth

5. BUSINESS RECOMMENDATIONS

Based on the data analysis, and our values forecasted, we divide our recommendations on the basis of short-term implementation, and long-term solution.

SHORT TERM : We analyze which days in particular months witness the maximum two-wheeler traffic.

- The months to consider are June and July
- The days indicated by blocks colored green for each junction, should have additional traffic management personnel to direct the traffic efficiently
- Set up temporary lanes to accommodate the spike in traffic

Months To Consider : **June & July**							
	Sun	Mon	Tue	Wed	Thu	Fri	Sat
COBY	-	-	-	-	-	-	-
LMET	-	-	-	-	-	-	-
OBVW	-	-	-			-	-
OGLD	-		-			-	
ORPY	-	-	-	-	-	-	-
SOMO	-	-	-	-			

LONG TERM :

We look at the 99th percentile data historically and in forecasts to estimate if we are crossing the capacity:

	Historical 99th Percentile	Forecasted 99th Percentile	Difference	Decision
COBY	2846	2859	0.45%	No
LMET	3478	2786	-19.91%	No
OBVW	1653	2055	24.32%	Yes
OGLD	1648	1695	2.85%	Maybe
ORPY	3946	3741	-5.20%	No
SOMO	1100	1195	8.64%	Yes

Approximately 8 fatal accidents occurred in 2017.

Corridor Cost = 3x 5km x $2M/km = $30M

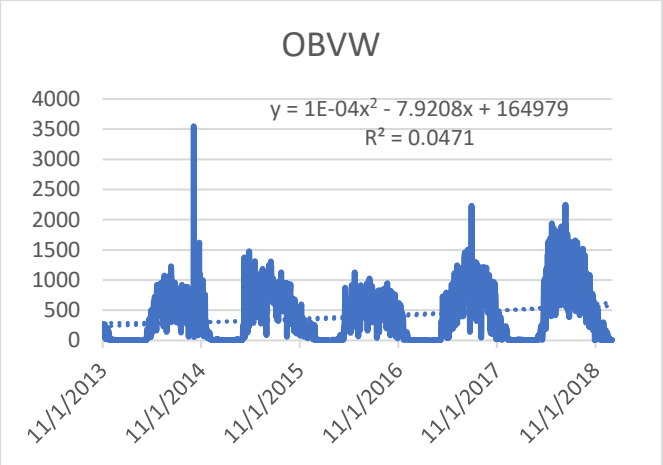
Life Saved = 8 lives x $9M x 40% = $29M (assuming 40% decrease in life loss)

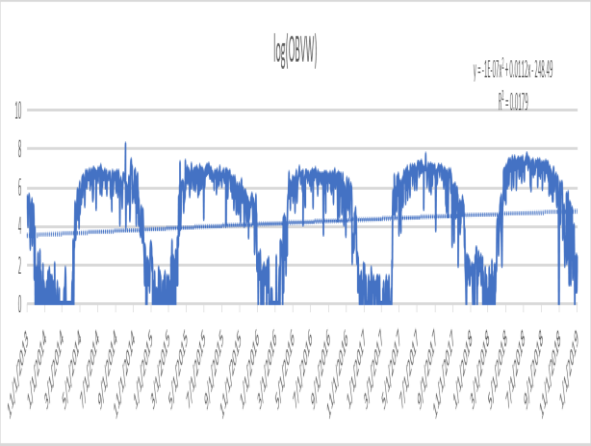
Monetarily, building the extension as a permanent solution is the way to go forward.

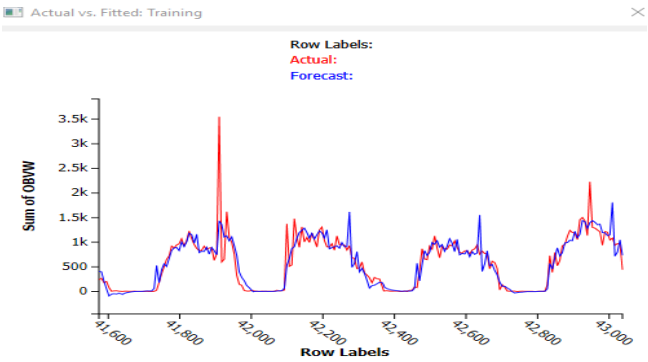
6. Appendix

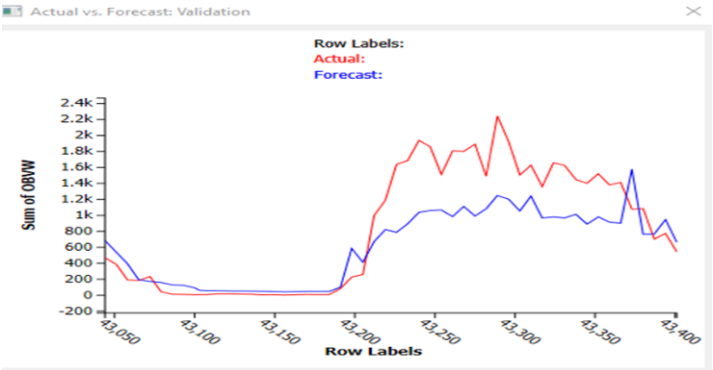
The graphs of the models run on each 6 junction, plus their residuals, and performance metrics are attached here

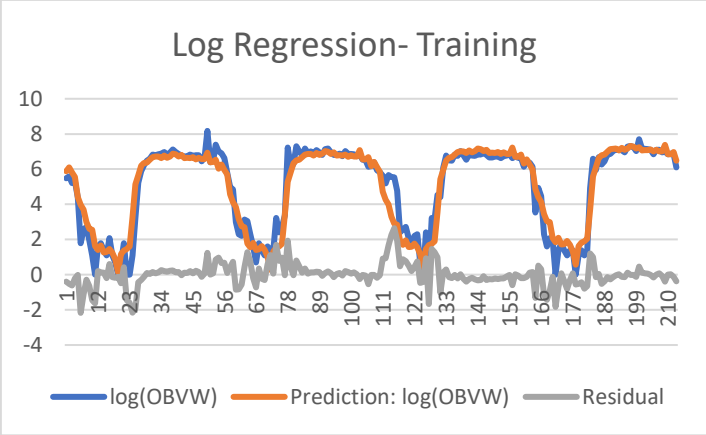
1. OBVW

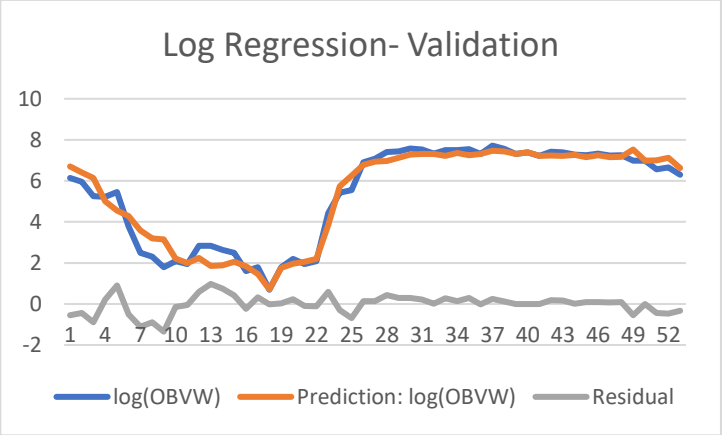




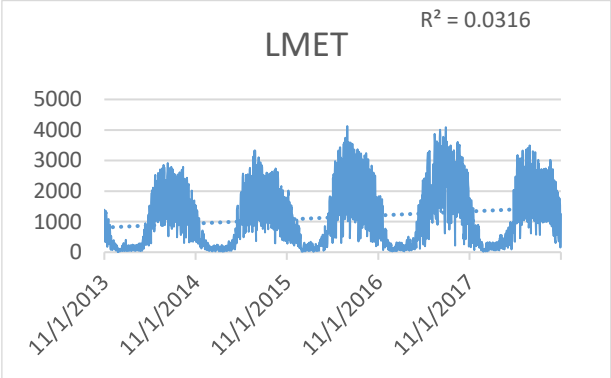


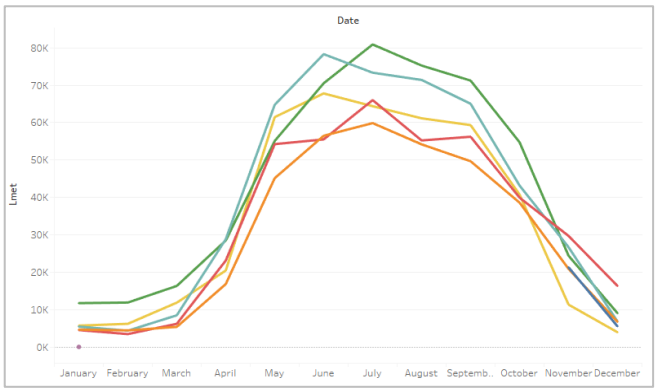






2. LMET





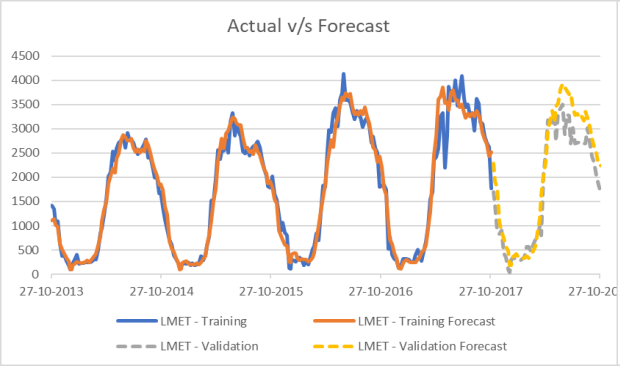
Error Measures: Training

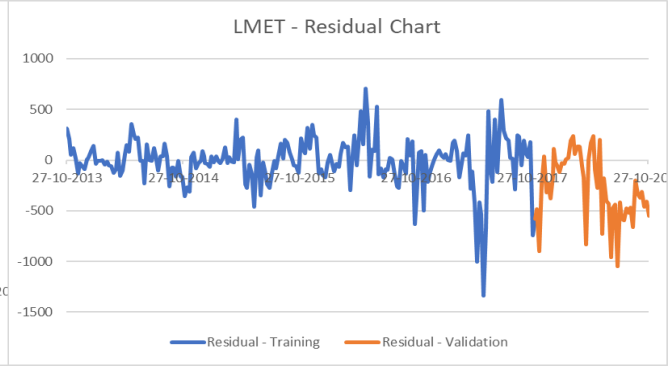
Record ID	Value
SSE	11604411
MSE	54480.803
MAPE	14.262816
MAD	154.66086
CFE	-2058.6493
MFE	-9.6650203
TSE	-13.310733

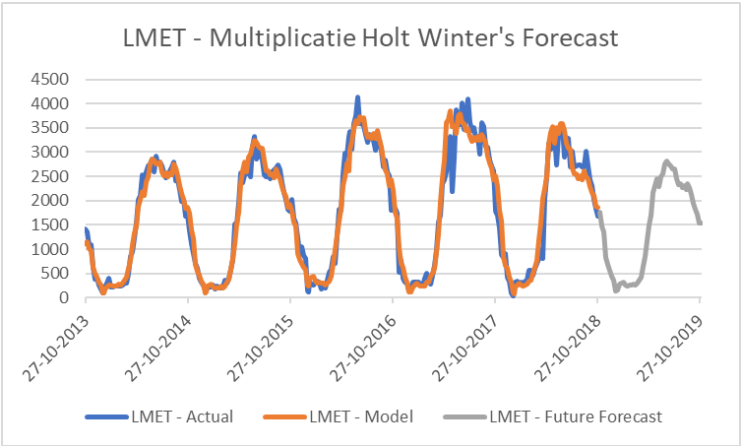
Error Measures: Validation

Record ID	Value
SSE	9287074.3
MSE	175227.82
MAPE	37.786342
MAD	330.66857
CFE	-14438.67
MFE	-272.42773
TSE	-43.665081

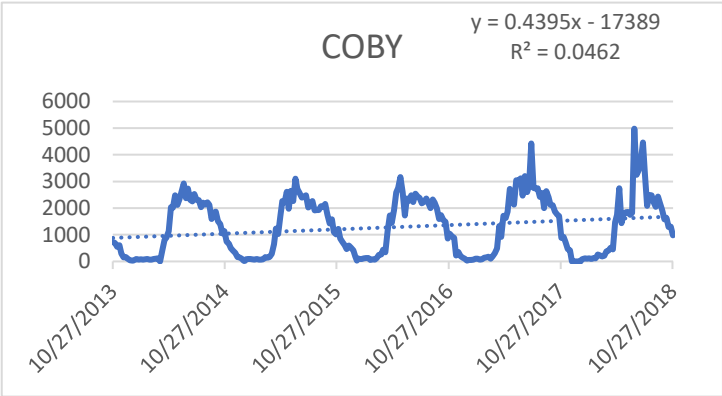
MultHoltWinters

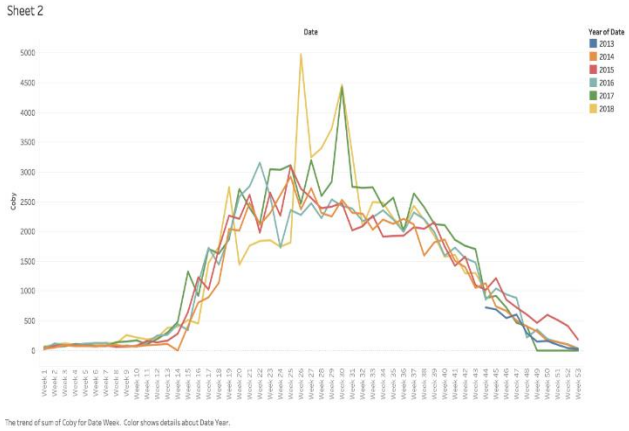






3. COBY





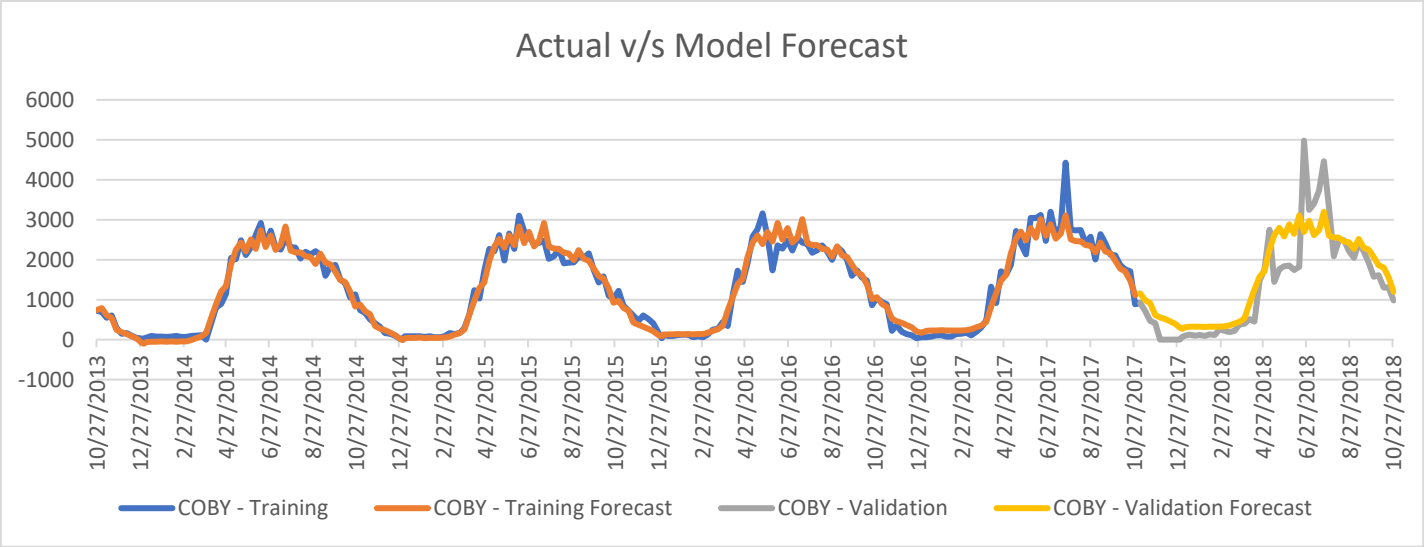
MultHoltWinters

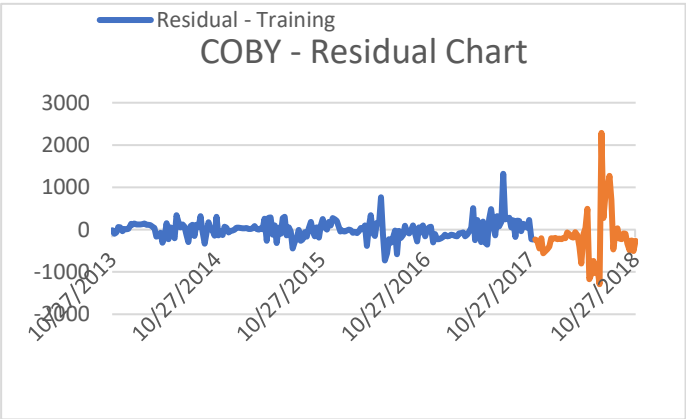
Error Measures: Training

Record ID	Value
SSE	11531244.4
MSE	53884.3194
MAPE	19.7653422
MAD	145.510333
CFE	-1977.5276
MFE	-9.240783
TSE	-13.59029

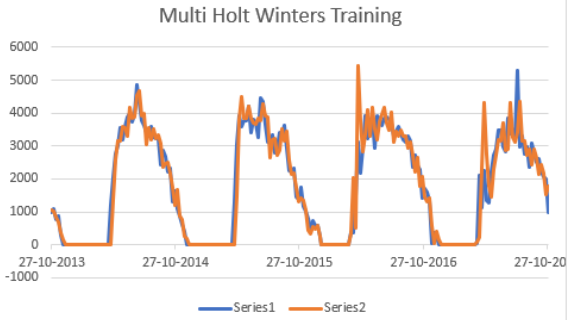
Error Measures: Validation

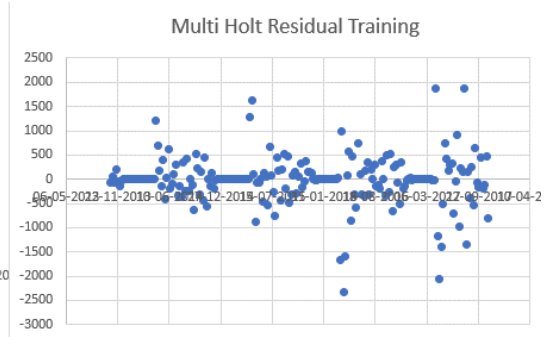
Record ID	Value
SSE	18033536.9
MSE	346798.786
MAPE	28.0115228
MAD	362.721246
CFE	-2335.7821
MFE	-44.918887
TSE	-6.4396066





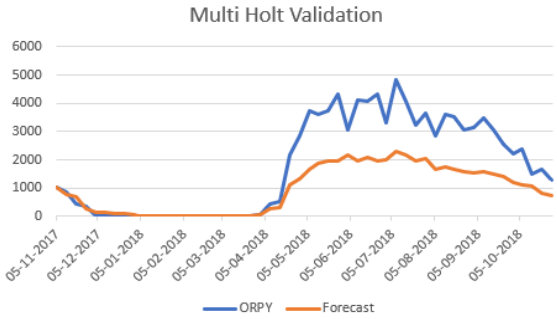
4. ORPY

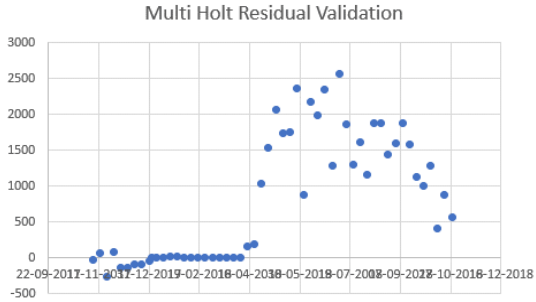




Error Measures: Training

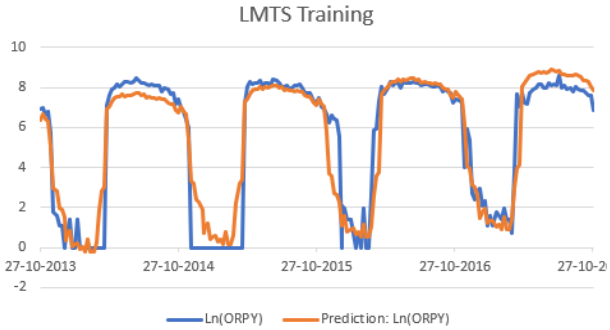
Record ID	Value
SSE	52296363
MSE	245522.83
MAPE	55.968703
MAD	281.59171
CFE	-2764.9525
MFE	-12.980997
TSE	-9.8190124

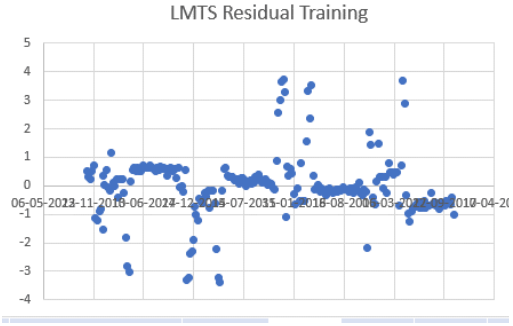




Error Measures: Validation

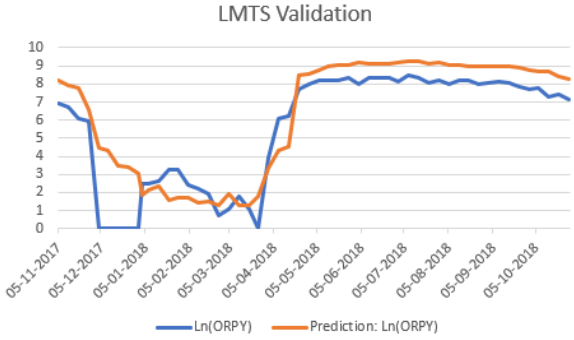
Record ID	Value
SSE	74993660
MSE	1414974.7
MAPE	53.294586
MAD	842.41313
CFE	43093.689
MFE	813.08848
TSE	51.155055

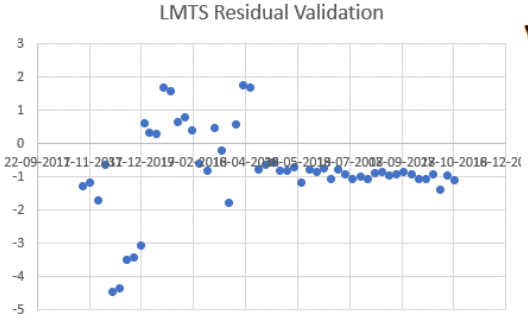




Training: Prediction Summary

Metric	Value
SSE	253.386579
MSE	1.18960835
RMSE	1.09069169
MAD	0.70780879
R2	0.89548581





Validation: Prediction Summary

Metric	Value
SSE	119.748962
MSE	2.25941437
RMSE	1.50313485
MAD	1.19279686
R2	0.76747653