ZHOU Kevin

Student ID : A0197122H

NUS - IE5101 - Applied Forecasting Methods

Professor : Chen Nan

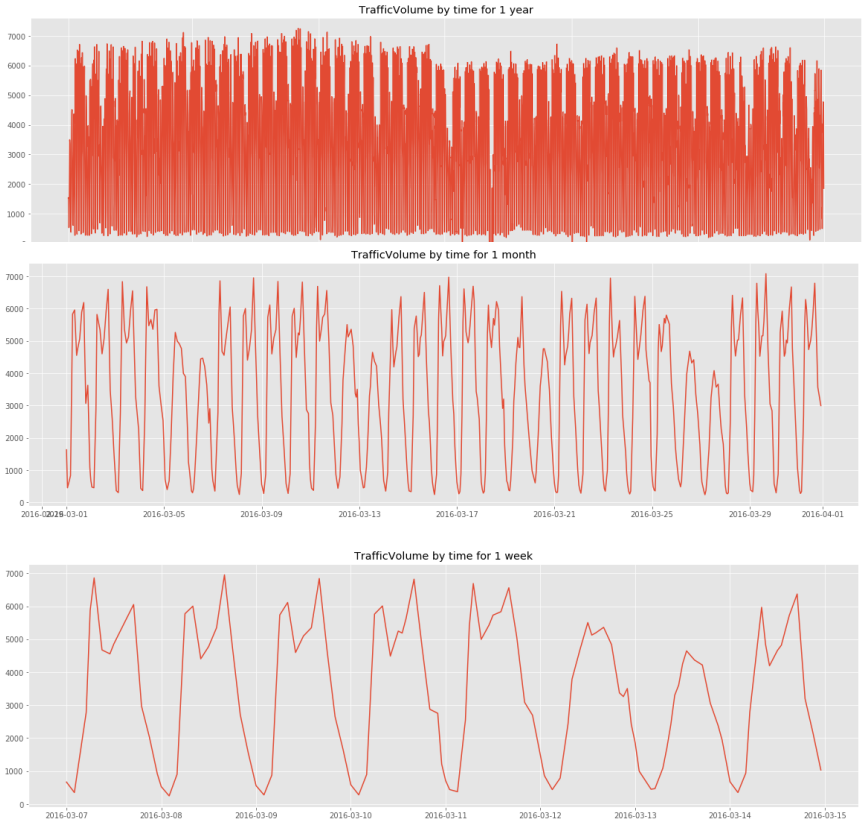
Project 2 : Forecasting Highway Car Volumes

Table of Contents

2.1 Regression on Time

2.1.1 Regression on time

In this first part our regression will only consider the variables Time and TrafficVolume for prediction.



By plotting the Traffic volume, we notice some seasonality linked to the hour (peak around 7 am to 5 pm), to the day of the week (low traffic during the weekend) and and also to the month (plateau around April and minimum around July)

Hence to compute the linear regression, we will introduce dummy variables for the **hour of the day**, the **day of the week** and the **month of the year** and add the year to account for possible trend in time.

To ease the interpretation, we have standardized the year variable: Year= Year -2012

We have opted for time series regression with no trend and constant additive seasonal variation ie $y_t = TR_t + SN_t + \varepsilon_t$ assuming :

- $TR_t = \beta_0$ with β_0 constant
- $SN_t = \beta_{S1} * x_{S1,t} + \beta_{S2} * x_{S2,t} + \ldots + \beta_{SN} * x_{SN,t}$

  with x_{Si} the dummy variables related to the seasonal variation

  More precisely $x_{Si,t} = 1$ if time period t is Si (season i) , 0 otherwise

This has yielded satisfactory results (below). All variables taken are extremely relevant (p-values close to 0) except C(monthofyear)[T.12].

A full model regression with the list of possible features below have resulted in a slightly better R^2-adjusted and smaller AIC. A stepwise forward selection regression to automatically select the best variables yielded similar results : better R^2-adjusted and smaller AIC.

OLS Regression Results

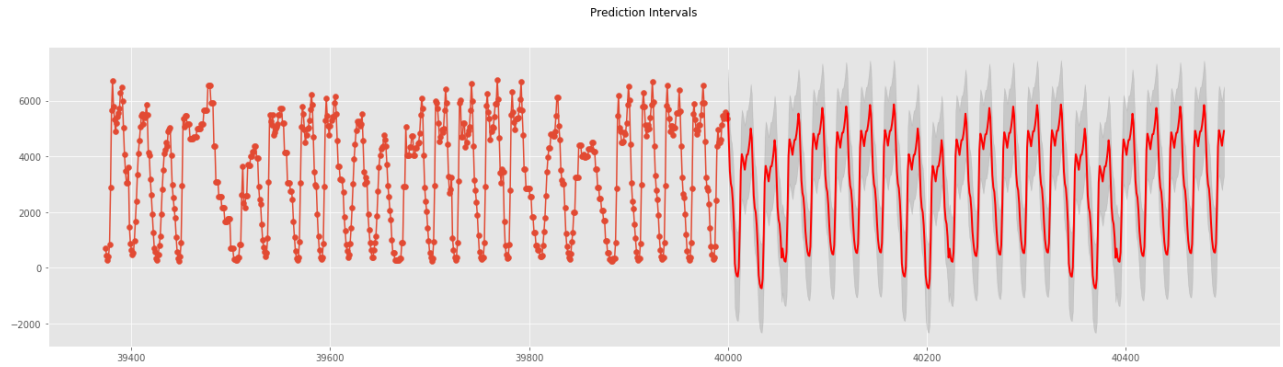
Dep. Variable:	TrafficVolume	R-squared:	0.834
Model:	OLS	Adj. R-squared:	0.834
Method:	Least Squares	F-statistic:	4911.
Date:	Sun, 20 Oct 2019	Prob (F-statistic):	0.00
Time:	18:33:10	Log-Likelihood:	-3.2467e+05
No. Observations:	40000	AIC:	6.494e+05
Df Residuals:	39958	BIC:	6.498e+05
Df Model:	41		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	635.4819	27.117	23.435	0.000	582.331	688.632
C(hour)[T.1]	-306.7032	27.784	-11.039	0.000	-361.161	-252.245
C(hour)[T.2]	-447.1942	27.833	-16.067	0.000	-501.748	-392.640
C(hour)[T.3]	-471.4459	27.896	-16.900	0.000	-526.122	-416.770
C(hour)[T.4]	-133.8727	27.648	-4.842	0.000	-188.064	-79.681
C(hour)[T.5]	1250.6530	27.819	44.957	0.000	1196.127	1305.179
C(hour)[T.6]	3327.9238	27.697	120.157	0.000	3273.638	3382.210
C(hour)[T.7]	3917.9658	27.741	141.232	0.000	3863.592	3972.339
C(hour)[T.8]	3778.9058	27.689	136.478	0.000	3724.635	3833.176
C(hour)[T.9]	3560.6685	27.976	127.275	0.000	3505.835	3615.502
C(hour)[T.10]	3359.2987	27.704	121.255	0.000	3304.997	3413.600
C(hour)[T.11]	3646.0482	28.197	129.308	0.000	3590.782	3701.314
C(hour)[T.12]	3892.3142	28.120	138.420	0.000	3837.199	3947.429
C(hour)[T.13]	3904.7812	28.353	137.723	0.000	3849.210	3960.353
C(hour)[T.14]	4114.8622	28.074	146.572	0.000	4059.837	4169.888
C(hour)[T.15]	4418.0435	28.220	156.560	0.000	4362.733	4473.354
C(hour)[T.16]	4839.3920	28.004	172.810	0.000	4784.503	4894.281
C(hour)[T.17]	4508.7693	28.220	159.770	0.000	4453.457	4564.082
C(hour)[T.18]	3433.7703	28.030	122.504	0.000	3378.831	3488.710
C(hour)[T.19]	2439.9679	28.121	86.767	0.000	2384.850	2495.085
C(hour)[T.20]	1998.4927	28.029	71.301	0.000	1943.555	2053.430
C(hour)[T.21]	1832.3378	28.007	65.424	0.000	1777.443	1887.232
C(hour)[T.22]	1368.7204	27.959	48.954	0.000	1313.920	1423.521
C(hour)[T.23]	628.2224	27.772	22.621	0.000	573.788	682.656
C(dayofweek)[T.1]	205.2955	15.142	13.558	0.000	175.617	234.974
C(dayofweek)[T.2]	256.7943	15.015	17.102	0.000	227.364	286.225
C(dayofweek)[T.3]	310.5708	15.105	20.561	0.000	280.965	340.176
C(dayofweek)[T.4]	330.2617	15.096	21.878	0.000	300.674	359.850
C(dayofweek)[T.5]	-532.9516	15.137	-35.209	0.000	-562.620	-503.283
C(dayofweek)[T.6]	-952.5392	15.080	-63.165	0.000	-982.097	-922.982
C(monthofyear)[T.2]	190.0225	21.275	8.932	0.000	148.324	231.721
C(monthofyear)[T.3]	323.0690	20.875	15.476	0.000	282.154	363.984
C(monthofyear)[T.4]	346.3163	20.168	17.172	0.000	306.786	385.846
C(monthofyear)[T.5]	297.0587	19.885	14.939	0.000	258.084	336.034
C(monthofyear)[T.6]	382.7492	20.925	18.291	0.000	341.735	423.763
C(monthofyear)[T.7]	163.0914	19.448	8.386	0.000	124.972	201.210
C(monthofyear)[T.8]	350.0345	20.069	17.442	0.000	310.699	389.370
C(monthofyear)[T.9]	272.5881	20.943	13.016	0.000	231.540	313.636
C(monthofyear)[T.10]	314.1265	20.023	15.688	0.000	274.880	353.373
C(monthofyear)[T.11]	124.6361	19.754	6.309	0.000	85.918	163.354
C(monthofyear)[T.12]	13.1563	19.474	0.676	0.499	-25.012	51.325
year	9.0160	2.451	3.679	0.000	4.213	13.819

Omnibus:	6106.812	Durbin-Watson:	0.349
Prob(Omnibus):	0.000	Jarque-Bera (JB):	16601.720
Skew:	-0.836	Prob(JB):	0.00
Kurtosis:	5.677	Cond. No.	86.9

('hour', 'C(hour)', 'dayofweek', 'C(dayofweek)', 'dayofmonth', 'C(dayofmonth)', 'dayofyear', 'C(dayofyear)','C(weekofyear)', 'weekofyear', 'C(monthofyear)', 'monthofyear','year')

Both yielded slightly better results (1.1 point higher in R^2-adjusted and 2000 lower in AIC) but have much complex models. For ease of interpretation, we will opt for the first regression model based on the AIC criteria.

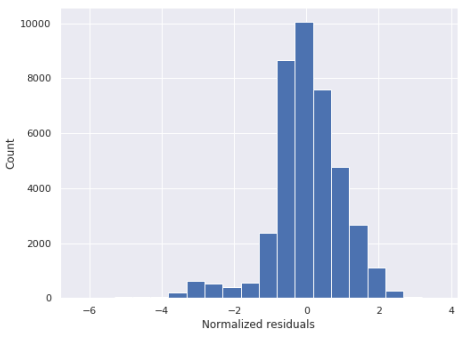


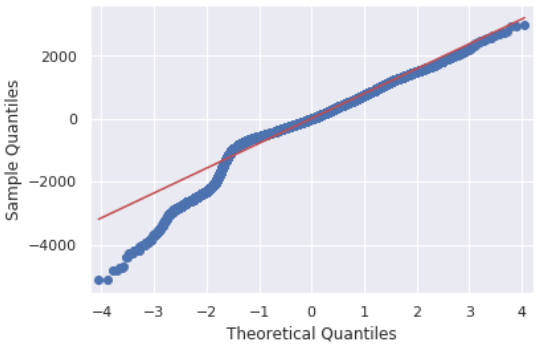
Predicting future dates gives the right seasonal trend with however some values below 0.

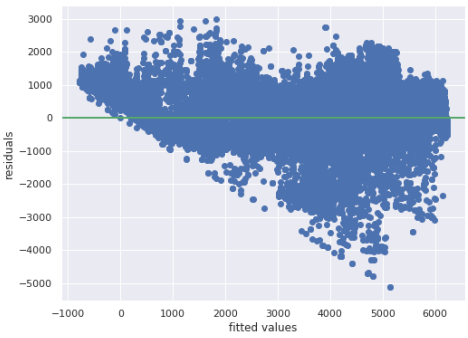
2.1.2 Diagnostic check

-Residuals plot shows a non-normally distributed set of residuals especially for the lower quantiles with non-constant growing variance of the residuals when compared to the fitted values. Assumption of homoscedasticity may not be true.

-Moreover, when fitted values are negatives, residuals are always positive. This is because the prediction intervals should always have an inferior bound of 0.







2.1.3 Model interpretation

The year is present in our model as a normal feature of linear regression. We have included 3 levels of seasonality : hour, day of the week, month of the year.

The base value is 635.48 at base hour, base day of week and base month.

- The base hour is hour 00:00. We observe that hours between 1:00 and 4:00 have negative coefficients in front of the related dummy variables and hours between 5:00 to 23:00 have positive coefficients with a culminating point at 16:00.

  C(hour)[T.16] =4839.3920 can be interpreted as : the model predicts that under the exact same context and variables (ie same features) and only changing the hour from 00:00 to 16:00 would increase the Traffic Volume by 4839 units.

- The base day is dayofweek=0, ie monday. High traffic seems to be in working days with a maximum on dayofweek=4, ie Friday. We note that variation across days are less important than variation across hours

- The base month is month=1 ie January. All coefficients are positive meaning that January can be seen as the month with the lowest traffic. There are 2 peak periods : from March to June and from August to October.

2.2 Exponential Smoothing

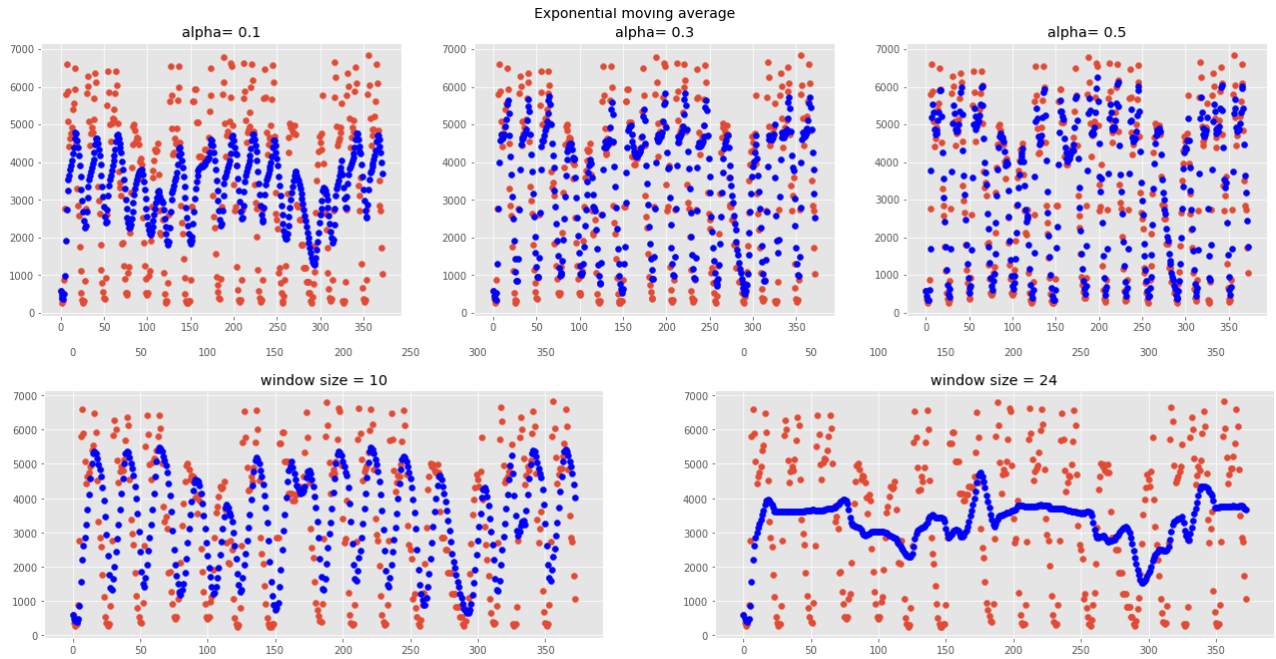
We will then apply exponential smoothing to make forecasting.

2.2.1 Naive application of Exponential smoothing

We will apply a simple exponential smoothing based on the assumption that the observations follow a constant trend modal:

$y_t = \beta_0 + \varepsilon_t$ with β_0 the constant term and ε_t the random residual term

We will calculate the optimal alpha given the SSE = $\Sigma[Y_i - L_{i-1}]^2$

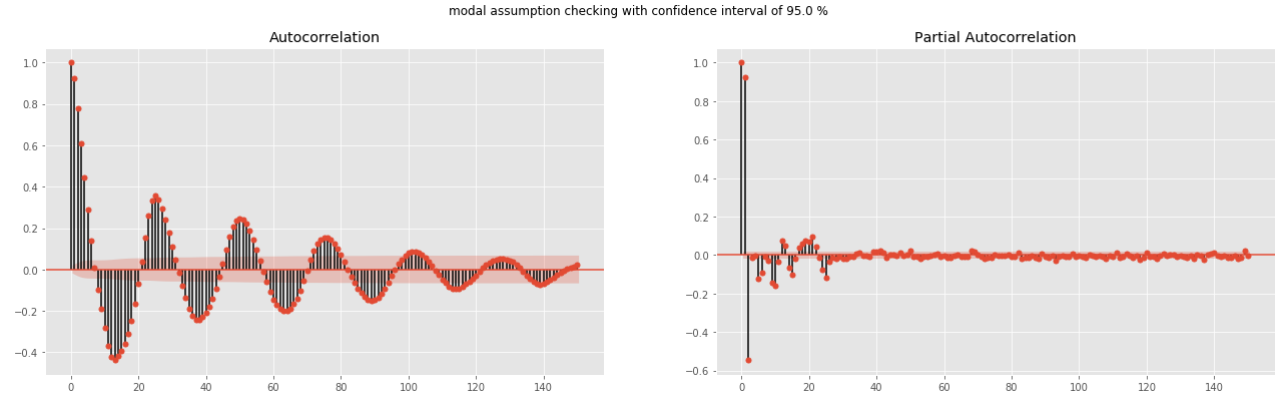


This resulted in an optimal alpha of 1.0 ie the optimal prediction of

Y_n is $L_{n-1}= \alpha Y_{n-1}+(1-\alpha) L_{n-2} =Y_{n-1}$. The best estimation is the previous value. We may need to review our assumptions when using the exponential smoothing.

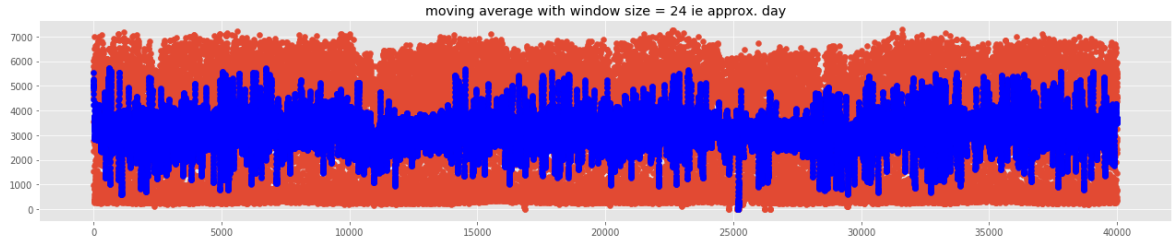
2.2.2 Verification of model choice, upsides and downsides

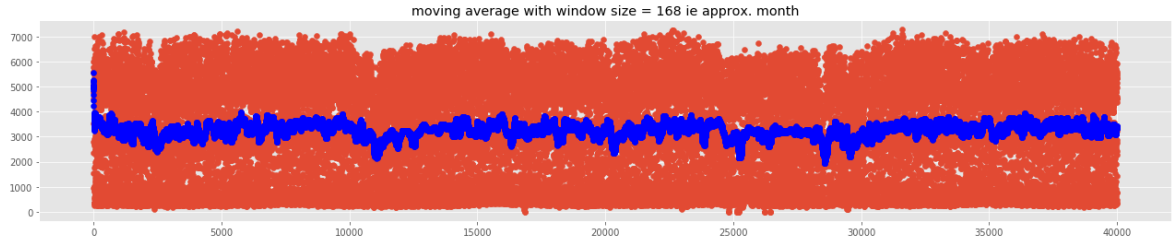
The exponential smoothing is based on constant parameters (ie TR_t, $SN_{t,}$) and constant variance over time.

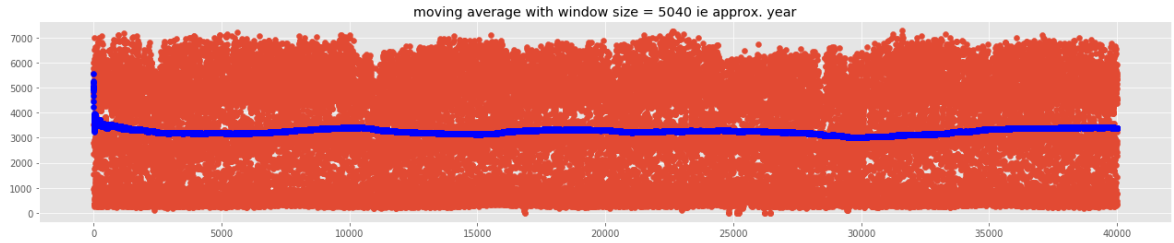


By plotting the autocorrelation plot, we do notice a sinusoidal autocorrelation function, which is a synonym for seasonality. As for the trend term, plotting the moving average by month and by year shows a relatively constant mean and thus can assume in first instance that the trend is constant

To account for the seasonal trend only, we can apply the Holt-Winters methods with a seasonal trend corrected smoothing and neglecting the linear trend term.







2.2.3 Holt-Winters exponential smoothing

- Since the hour variable is the most influential one, we can first develop a model with an hourly seasonal variation (L=24).
- $Y_t = \beta_0 + \beta_1 t + SN_t + \varepsilon_t$ with $\beta_1 t = 0$ meaning that $Y_t = \beta_0 + SN_t + \varepsilon_t$

  with the following update:

  $L_n = \alpha(Y_n - SN_{n-L}) + (1 - \alpha)(L_{n-1})$

  $SN_n = \delta(Y_n - L_n) + (1 - \delta)SN_{n-L}$
- Applying this methods would however require a uniformly sampled data with exactly one row per hour ie no duplicated or missing values for a unique hour/date which is not always the case here.

2.3 Free form forecasting

We will first apply a **linear regression** using features non related to time and dummy variables to model the seasonal variation and then apply an **ARIMA modal** (or Box-Jenkins methodologies) to modal the residuals of the previous forecast of traffic volumes.

$Y_t = TR_t + SN_t + \varepsilon_t$

with $TR_t + SN_t \sim$ Linear Regression and $\varepsilon_t \sim ARIMA(p, q)$

2.3.1 Data cleaning

Following data cleaning (**wrongly labeled/ outliers data)** have been necessary to yields satisfactory results:

-‘IsHoliday’ is only flagged for the first hour of the day. We will flag it for all hours.

We do notice also some days that are wrongly labeled that would need further investigations (ex : 2017-01-02 labeled as ‘New Year’, maybe due to an error in time zone conversion)

- ‘Rain1h’/ ‘Temp’ with abnormally high/low values are put at nan and interpolated with neighboring values.

2.3.2 Linear regression

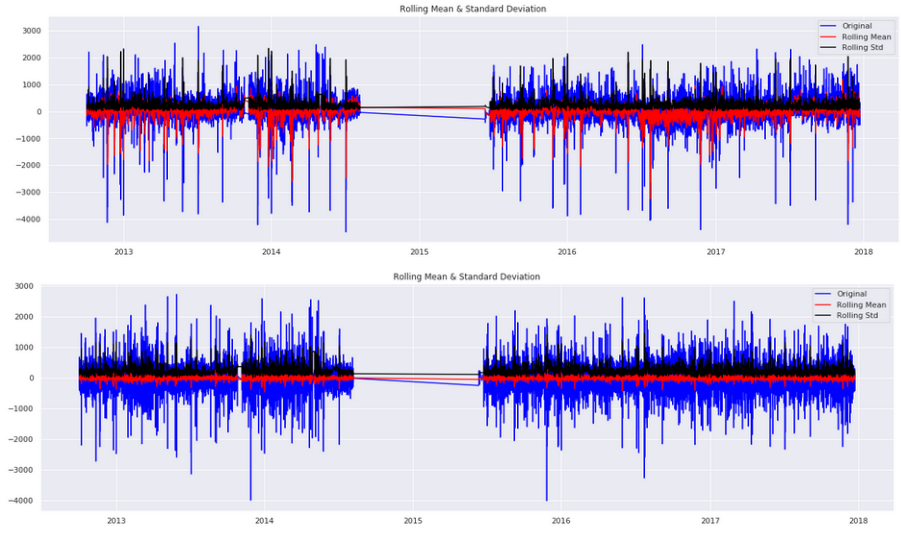
Among the different linear regression tested, the **model with interaction effects included obtained by forward selection** seems to be a good trade off between low testing errors and low model complexity.

Model and methods used	**Train errors (MAE)**	**Test error (MAE)**
1-full model	586.59	589.46
2-model obtained by forward selection	586.91	590.15
3-full model with interaction effect	**235.54**	**245.33**
4-model with interaction included obtained by forward selection	254.34	261.16

2.3.3 Linear regression with ARIMA errors

The residuals obtained have been used as entry for an ARIMA modeling. This methodology works best when data is stationary i.e. mean, variance and autocorrelation are constant over time.

Differencing ($z_t = y_t - y_{t-1}$) one time makes indeed the data more stationary (left) with a more constant average and standard deviation:



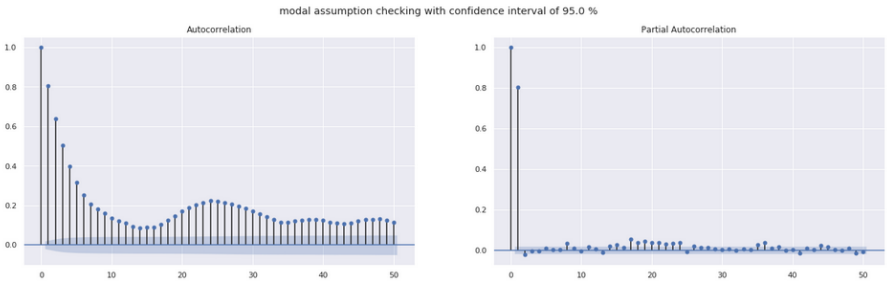
The ACP (of differenced residuals displayed below) however still reveals some seasonality effects. And lag 1 of partial autocorrelation is too negative and we might have over-differenced.

We will then try a gridsearch on both (not differenced and differenced) to look for the orders optimizing the cross validation errors.

ACP and PACP of differenced residuals :

For the AR term, by looking at the Partial autocorrelation plot, we notice that lag 1 is quite significant and lag 2 much less. We can tentatively look for values of p =0 to p=1.

For the MA term, by looking at the autocorrelation plot, we can tentatively look for values of q= 0 to q=5.

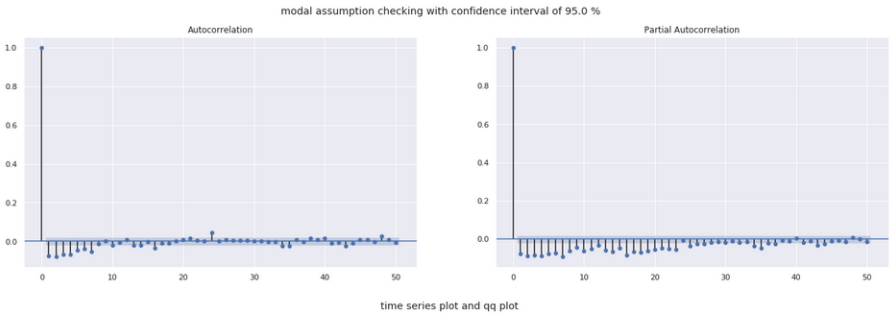


ACP and PACP of residuals

(p,d,q)=(1,0,1) gave the best results among all the possible combinations when comparing mean absolute cross validation errors.

However the errors are small (<1%) when comparing errors of ARIMA (p,d,q)=(1,0,1) and (p,d,q)=(0,0,0) (model without ARIMA)

Those results needs to be however taken with a pinch of salt since ARIMA have been performed on a datasets that was not correcly sampled.



ACP and PACP of differenced residuals

Due to time constraints and the quality of the data, I was able only to begin coding the residuals prediction with ARIMA. We would need in fact a clean, correctly sampled training and testing sets to correctly predict the right number of lags for forecasting purpose.

The first resampling step has been done and has yielded slightly better results for the linear regression model. We can explain this by assuming that some replicated rows now have less weights in the new model.

2.3.4 Further Improvements

- Further data quality improvements notably on value imputation and processing (Holidays for instance) and data resampling (define strategies to treat the nan that have appeared due to the resampling)
- We can also explore the different options of combining linear regression and ARIMA :

  -ARIMA on the time series of TrafficVolume

  -ARIMAX on the time series of TrafficVolume and features

  -SARIMA on the time series of TrafficVolume

  -SARIMAX on the time series of TrafficVolume and features

  -multiple series modeling with TrafficVolume and Temparature