Federal State Autonomous Educational Institution of Higher Education

ITMO University

Methods and models for multivariate data analysis

Report on learning practice # 4

Stationarity of the processes

Performed by:

Nemat Allah Aloush

(J4134c)

Saint-Petersburg

2021

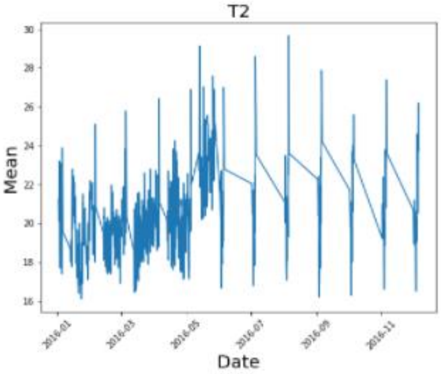
Contents

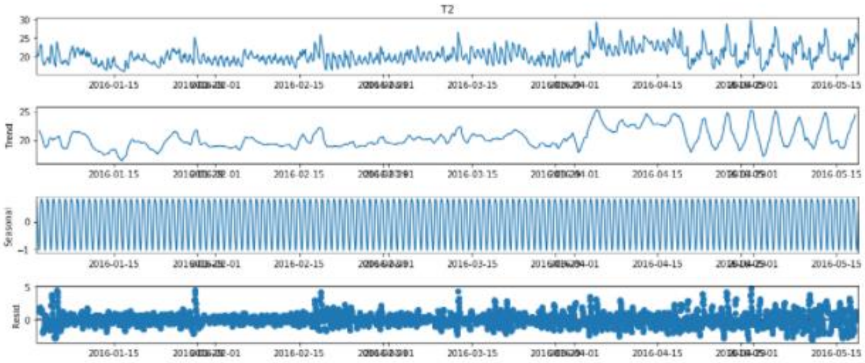
1. Substantiation of chosen sampling.

For this lab work, five continuous variables were chosen to investigate them. The chosen variables are: 'date','T2', 'T5','T6','T_out'. Where 'T2'and'T_out' are targets.

2. Stationary analysis.

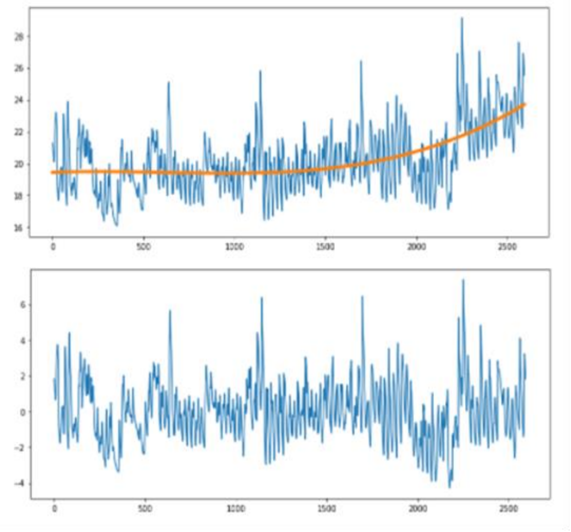
After observing the following stationarity figure for (T2), I had chosen to work on only the left half of the data.

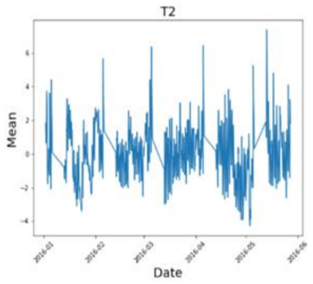


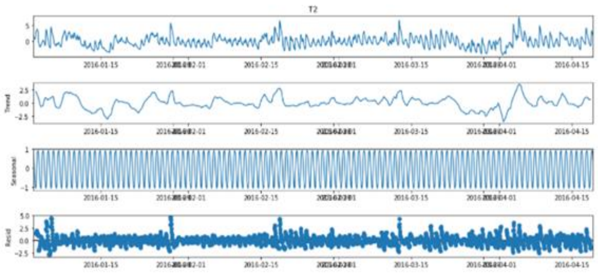


After taking only one half of the data, I draw each variable, and there was trend in each of them. That is why I deleted the trend for each of them, then reinvestigated the stationarity for each, to make sure we are working on suitable data.

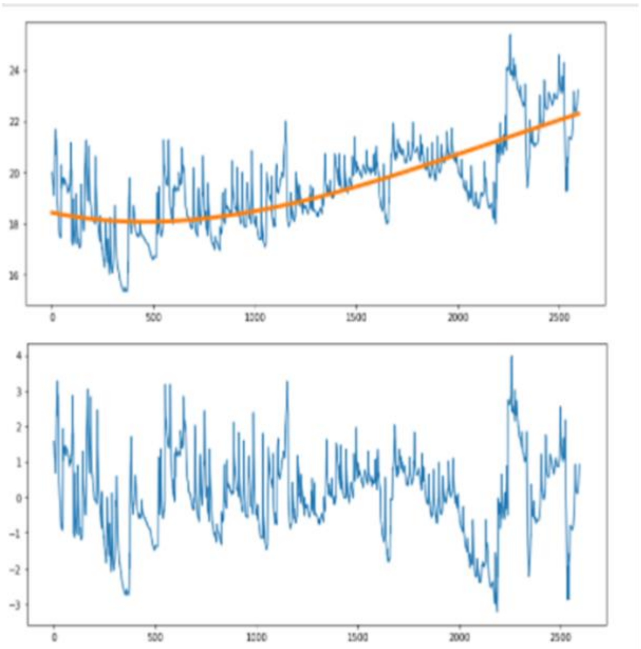
- T2 variable, on the left of following figure we can see the variable (T2) before and after removing the trend, on the right we can see the stationarity for (T2) after removing the trend:

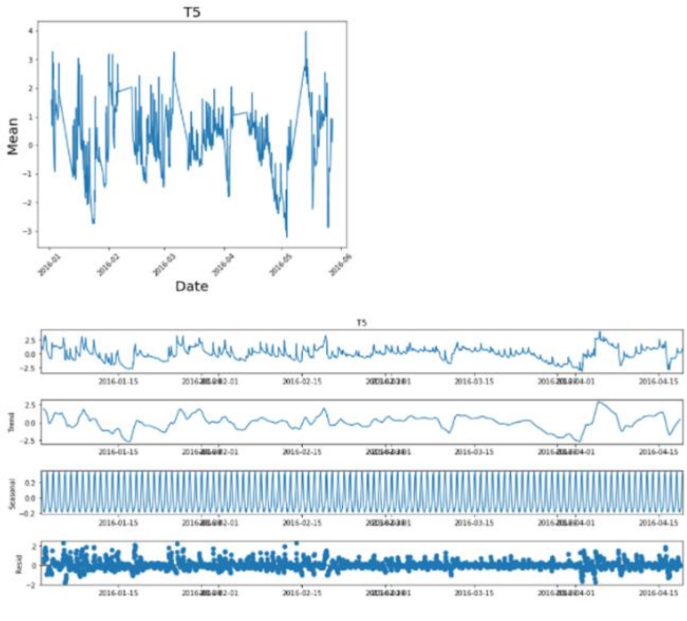




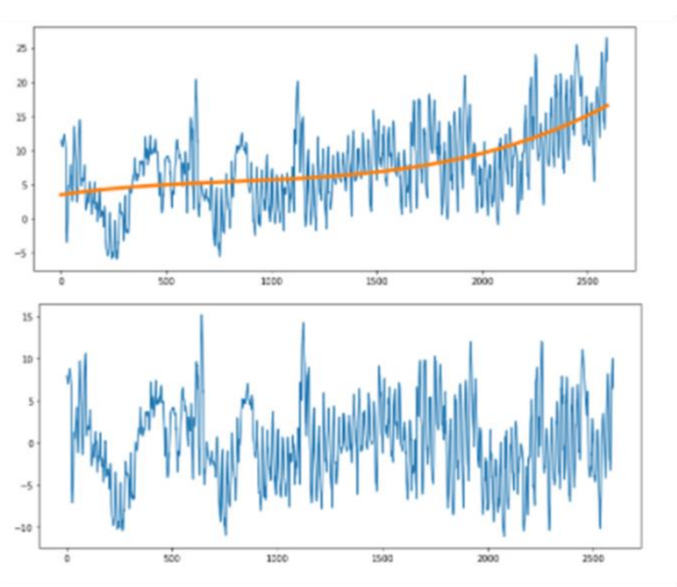


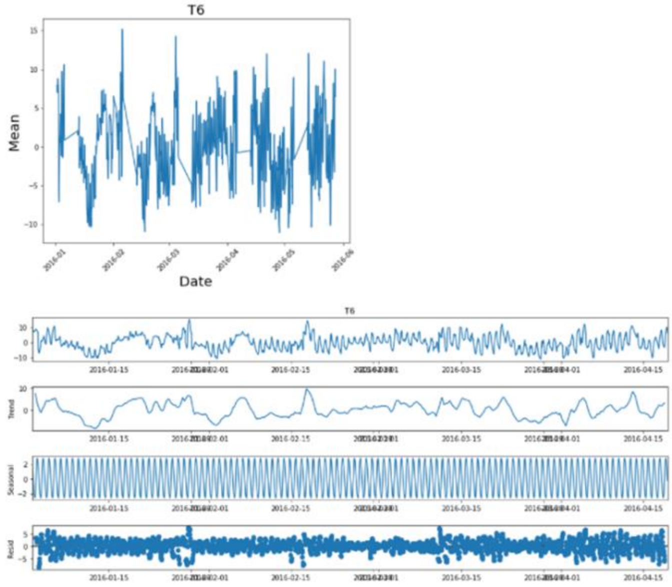
- T5 variable, on the left of following figure we can see the variable (T5) before and after removing the trend, on the right we can see the stationarity for (T5) after removing the trend:



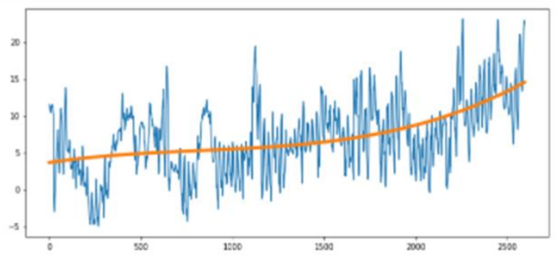


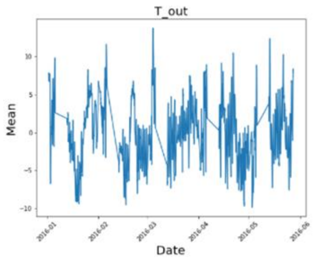
- T6 variable, on the left of following figure we can see the variable (T6) before and after removing the trend, on the right we can see the stationarity for (T6) after removing the trend:

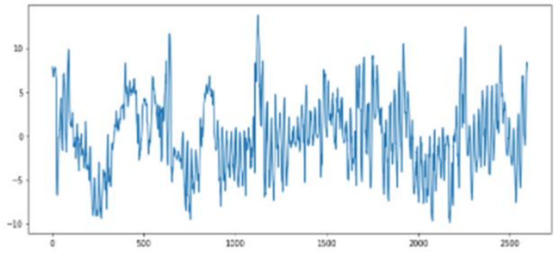


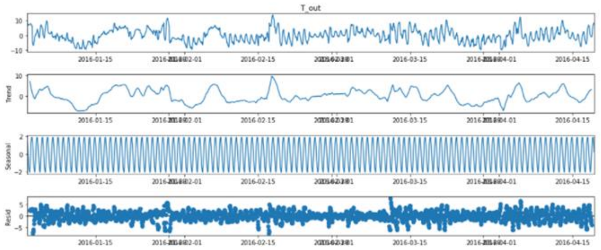


- T_out variable, on the left of following figure we can see the variable (T_out) before and after removing the trend, on the right we can see the stationarity for (T_out) after removing the trend:







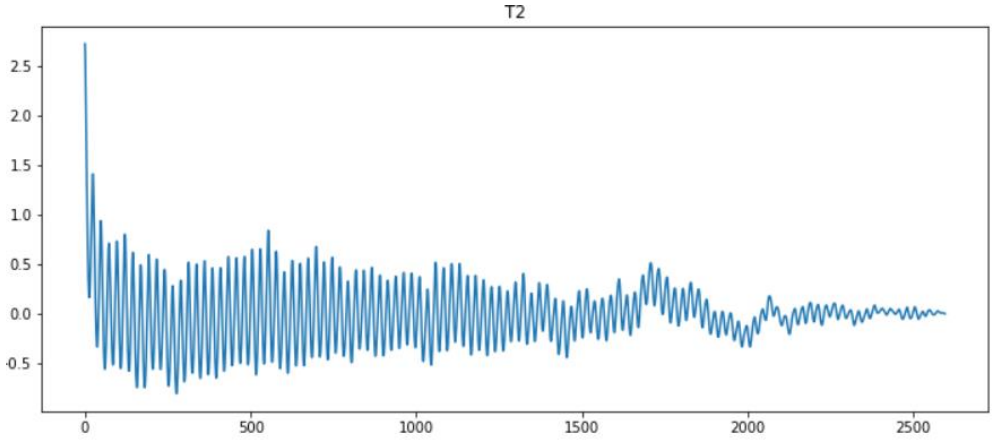


In the following table we can see the results of Augmented Dickey-Fuller test for each of the variables before and after removing the trend.

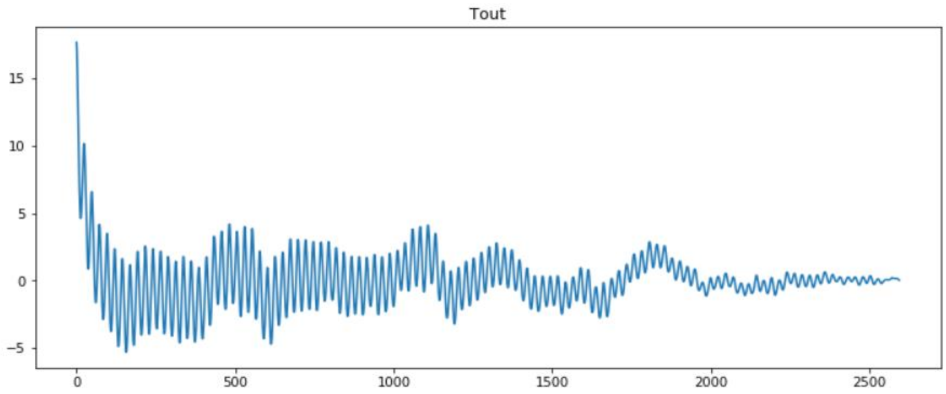
Variable		Statistic value	p-value
T2	Before Taking Half of the Data and Removing the Trend	-3.887122	0.002132
	After	-6.274650	0.000000
T5	Before	-2.711019	0.072151
	After	-5.496873	0.000002
T6	Before	-3.830888	0.002607
	After	-5.671630	0.000001
T_out	Before	-3.777386	0.003147
	After	-5.465004	0.000002

3. Covariance or correlation function analysis.

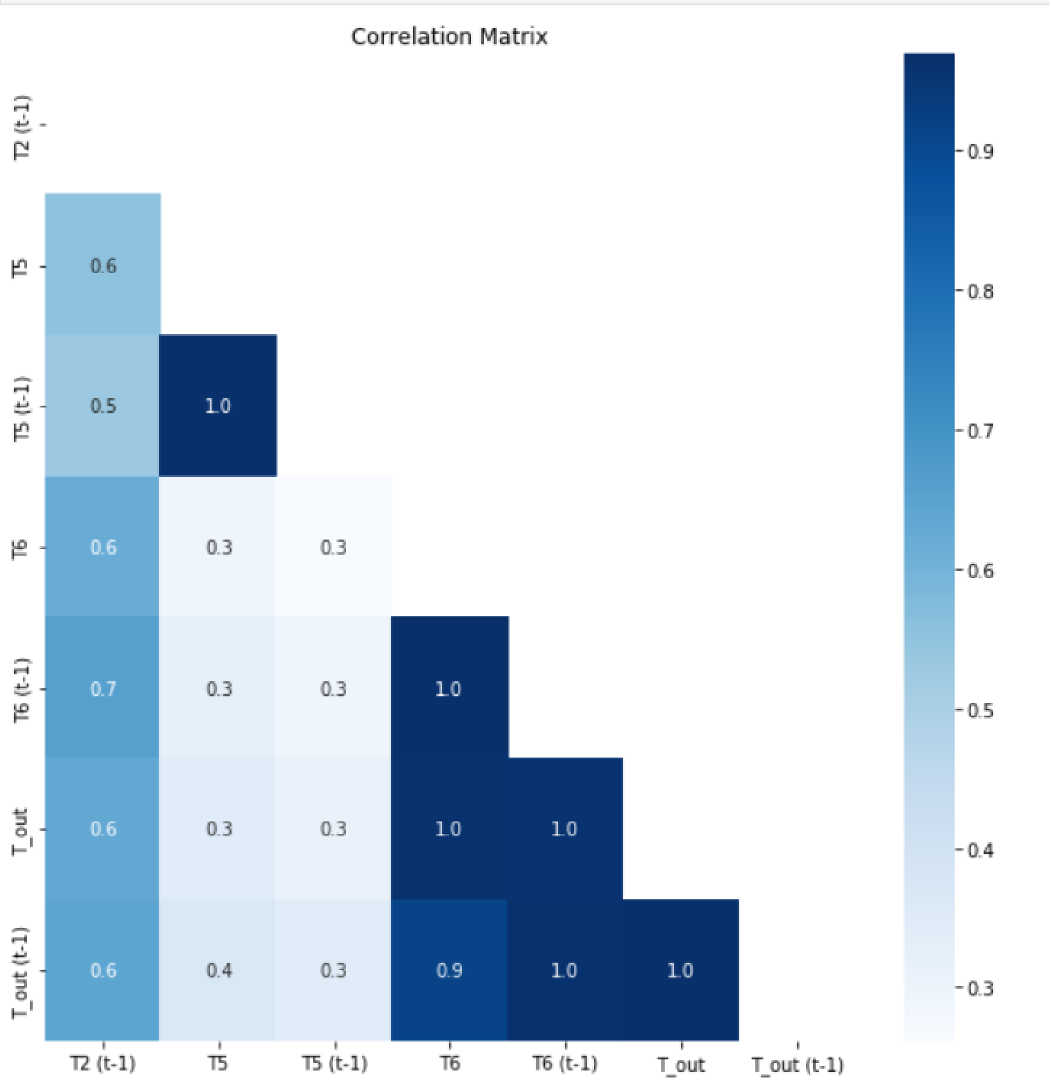
In the following figure, we can find Covariance for Target : T2.



In the following figure, we can find Covariance for Target : T_out.



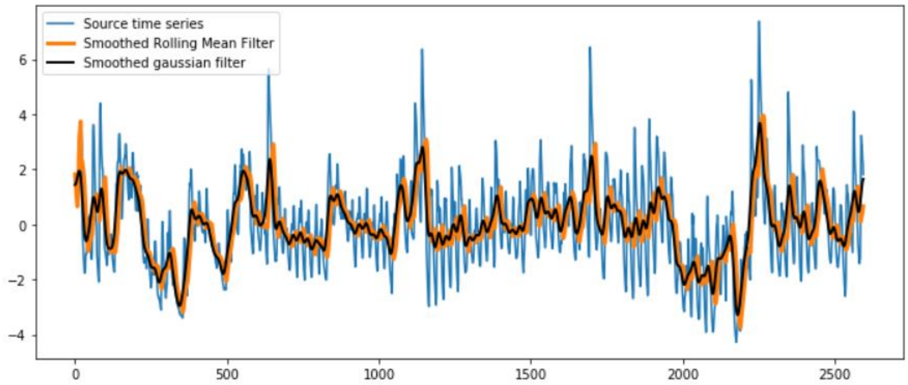
In the following figure, we can find the mutual correlation among predictors and targets, first I calculated for each variable V(t-1) from the V(t) we have. Then plotted the correlation matrix between all variables V(T) and V(t-1).



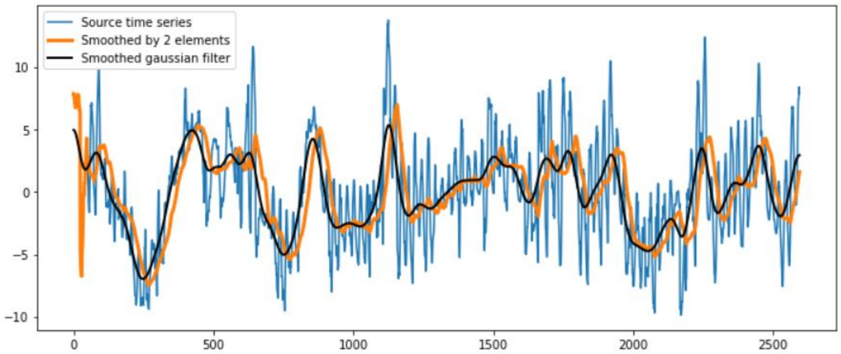
4. Noise filtration.

For each of the two targets (T2, T_out), two filters were applied: Rolling Mean filter and Gaussian Filter. For each filter several parameters were used to try filtering and one of them was chosen eventually.

In the following figure, we can find the original target data (T2) and the flittered data using both filters.

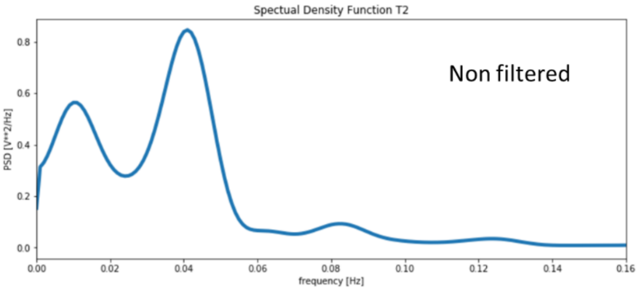


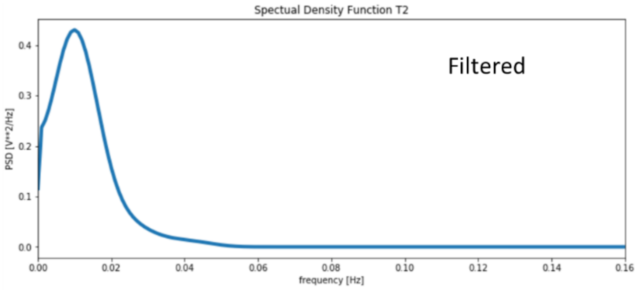
In the following figure, we can find the original target data (T_out) and the flittered data using both filters.



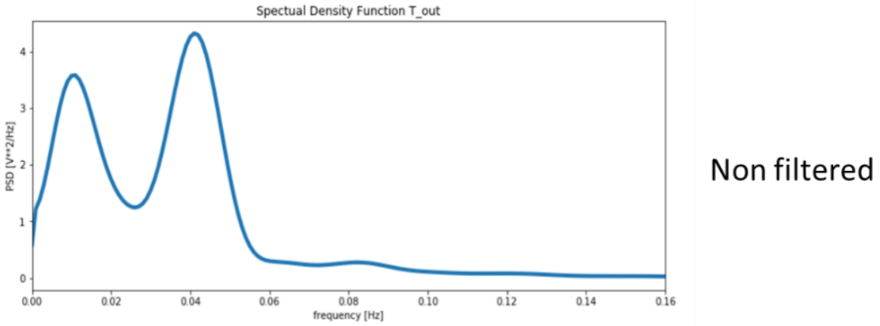
5. Estimation of spectral density function.

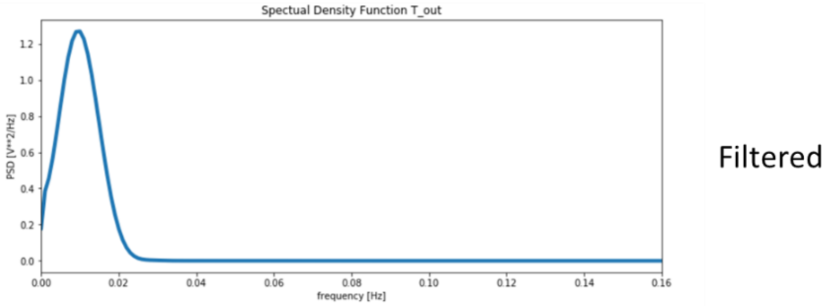
In the following figure, we can find the spectral density function for (T2) before and after filtering. The filtered data is from the previous step, which is the original data filtered by Gaussian Filter with sigma value equal to eight.





In the following figure, we can find the spectral density function for (T_out) before and after filtering. The filtered data is from the previous step, which is the original data filtered by Gaussian Filter with sigma value equal to twenty.



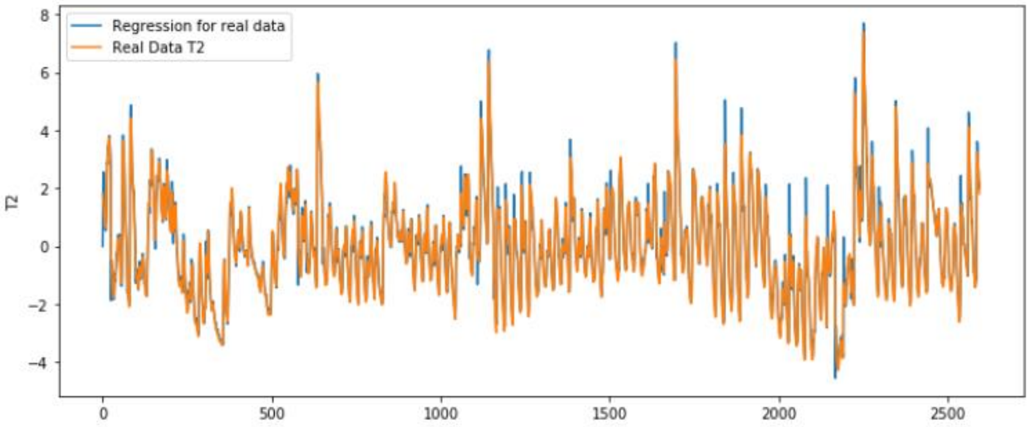


For both targets, we can notice the difference between filtered and non-filtered spectral function. In addition, for both variables, we can see that in the filtered spectral function, the frequencies starting from 0.02 are having smaller values than the non-filtered data, which indicates that prediction for longer period of time can be achieved.

6. Auto-regression model.

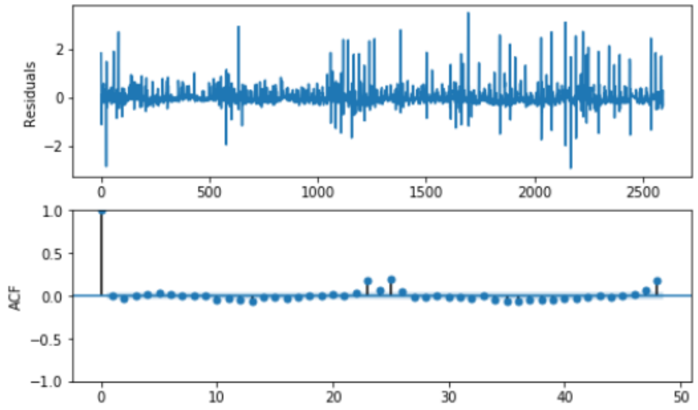
6.1 Train a SARIMA model with the values of variable (T2)

- In the following figure, the SARIMA and AIC model regression model for the non-filtered (T2) target variable.



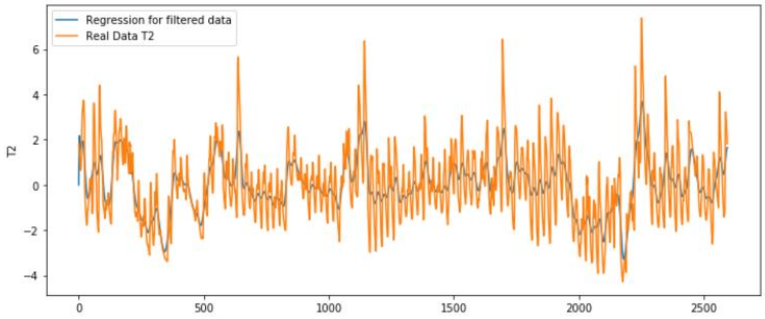
In the previous figure, we can see that the prediction for the existed data is good. The difference between statistical models and machine learning models, is that statistical models are not so affected by the overfitting, so we can simply compare the result for train sample and if statistical model represents (cover) the training data well (as our case) then the model has a strict structure and we can use it for forecasting the future.

- The orders of the best model: SARIMAX (3, 1, 3)x(0, 1, 0), 2
- In the following figure we can see that the residuals for this model are stationarity. Meanwhile, the p-value from Shapiro-Wilk test is very small 0.0 << 0.05, which indicates that the residuals do not follow a normal distribution.

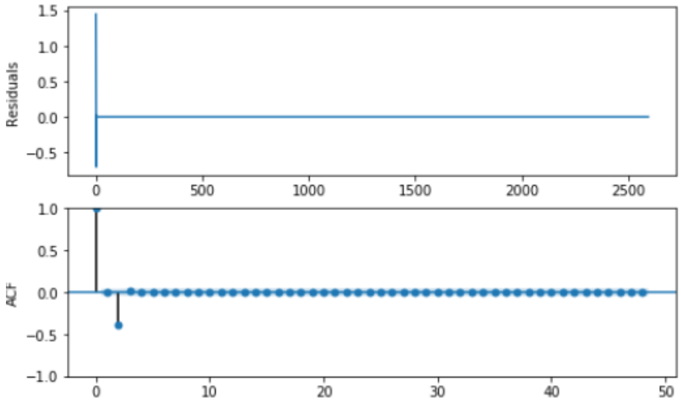


6.2 Train a SARIMA model with the filtered values of variable (T2)

- In the following figure, the SARIMA and AIC model regression model for the filtered (T2) target variable.
  The filtered data is from the previous step, which is the original data filtered by Gaussian Filter with sigma value equal to eight.

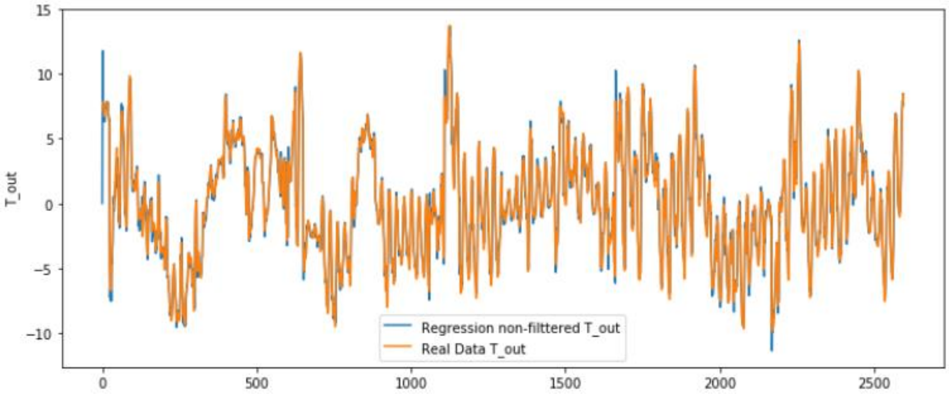


- The orders of the best model: SARIMAX(5, 1, 5)x(0, 1, 0), 2
- In the following figure we can see that the residuals for this model are stationarity. Meanwhile, the p-value from Shapiro-Wilk test is very small 0.0 << 0.05, which indicates that the residuals do not follow a normal distribution.

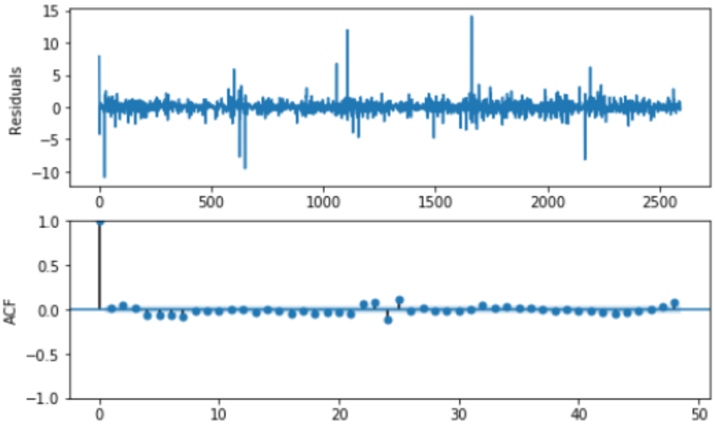


6.3 Train a SARIMA model with the values of variable (T_out)

- In the following figure, the SARIMA and AIC model regression model for the non-filtered (T_out) target variable.

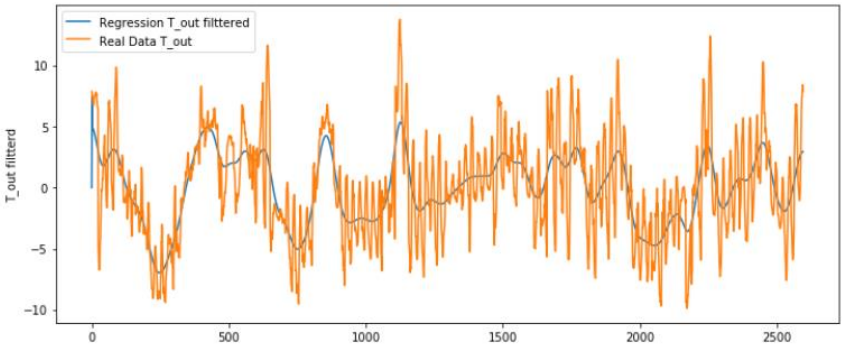


- The orders of the best model: SARIMAX (2, 1, 5) x (0, 1, 0), 2
- In the following figure we can see that the residuals for this model are stationarity. Meanwhile, the p-value from Shapiro-Wilk test is very small 0. << 0.05, which indicates that the residuals do not follow a normal distribution.

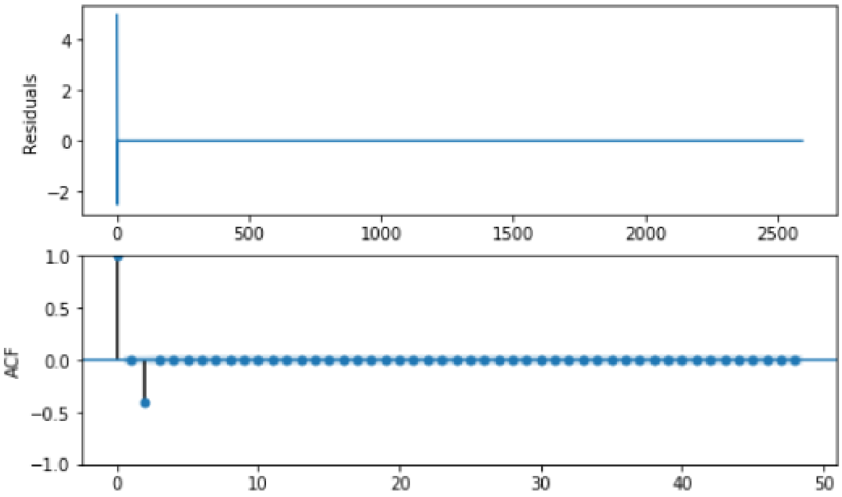


6.4 Train a SARIMA model with the filtered values of variable (T_out)

- In the following figure, the SARIMA and AIC model regression model for the filtered (T_out) target variable.
  The filtered data is from the previous step, which is the original data filtered by Gaussian Filter with sigma value equal to eight.



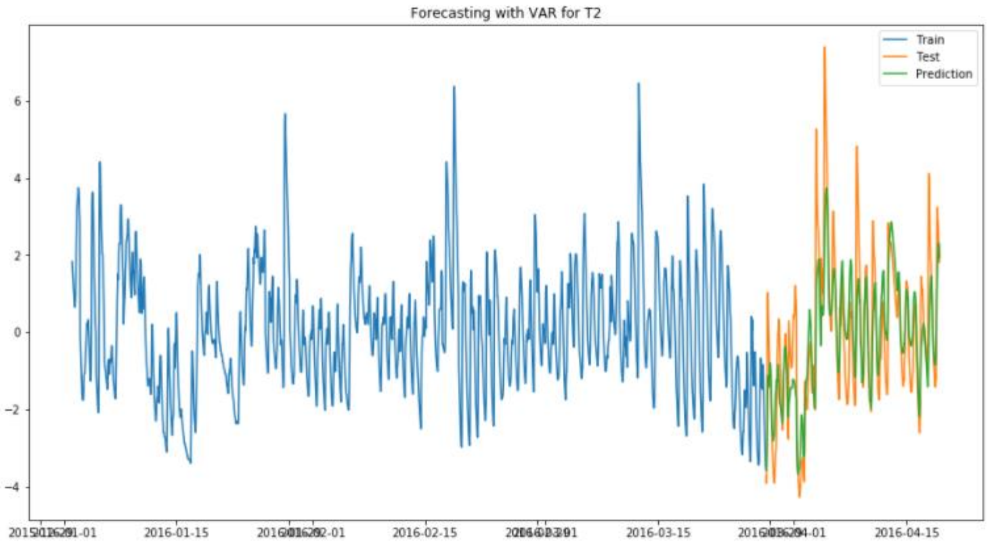
- The orders of the best model: SARIMAX(5, 1, 5)x(0, 1, 0), 2
- In the following figure we can see that the residuals for this model are stationarity. Meanwhile, the p-value from Shapiro-Wilk test is very small 0.0 << 0.05, which indicates that the residuals do not follow a normal distribution.

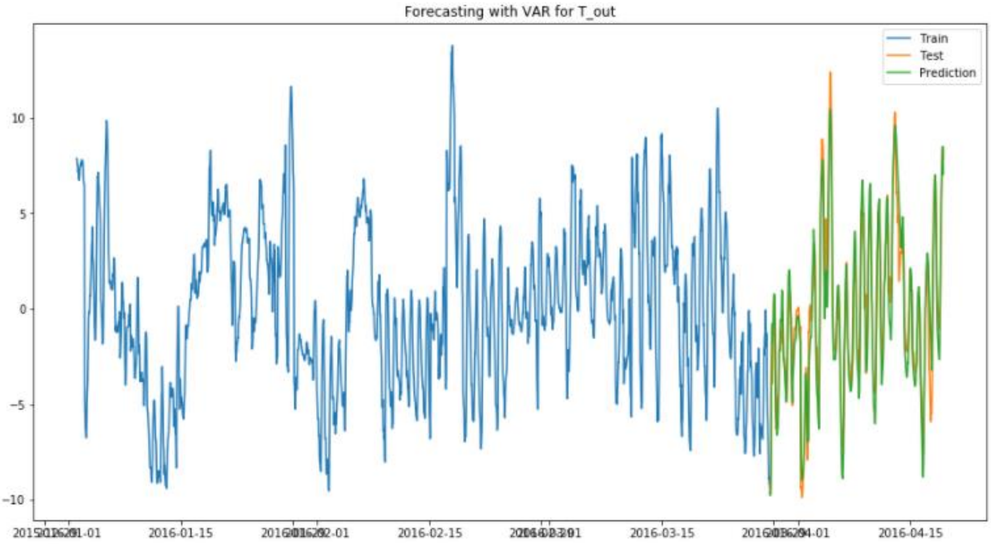


7. Model in a form of linear dynamical system.

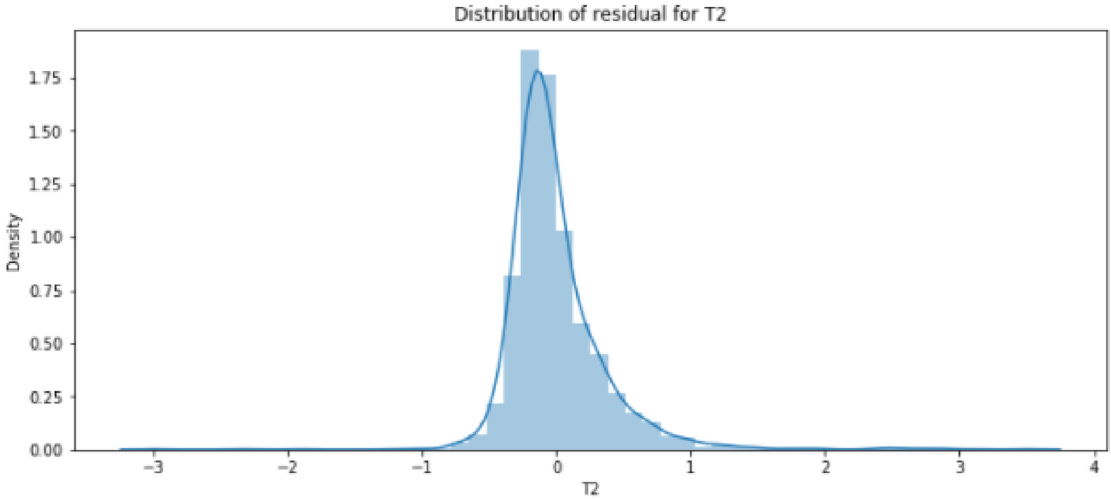
The chosen model for this task is VAR model - vector autoregression model. VAR is a statistical model used to capture the relationship between multiple quantities as they change over time. The data (Here I use the data after editing it to be stationarity) was partitioned to train and test data and fitted a model for the predictors and target variables.

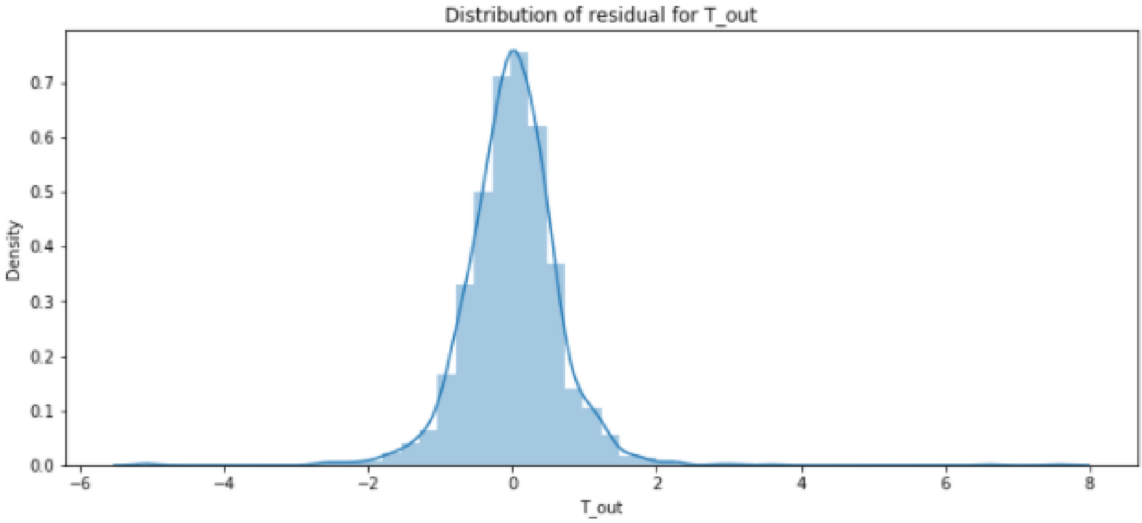
In the following figure, we can find the forecast model for T_out and T2 targets.





In the following figure, we can see the distribution of the residuals for each target value. In addition, in the following table, there are the value of the mean of the residuals and the results of Shapiro test to check the normality distribution for the residuals.





	T2	T_out
Residuals mean	-7.597670819121553e-17	-2.2686003009208017e-17
Residuals std	0.3901230095097995	0.6635648360006856
ShapiroResult-statistic	0.7801107168197632	0.9036952257156372
ShapiroResult- pvalue	0.0	2.7280790241937106e-34

We can see that the mean value is actually close to zero. And according to p-value from Shapiro test the residuals do not really follow a normal distribution, as it was the case for SARIMA models. Which means here that the VAR model do not fit the data.

8.Source code

Please find my code in the following GitHub link:

https://github.com/neematAllosh/MultiderivativeDataAnalysis/blob/master/lab_04.ipynb