ANOMALY DETECTION TECHNIQUES FOR SENSOR TIME SERIES DATA USING STL+IQR AND TWITTER+GESD METHODS

ABSTRACT

Anomaly detection is a subfield of machine learning where a model is produced which pays special mind to any variations from the norm in the data. Anomaly detection has a wide variety of uses in the different domains. For these reasons, there is a lot of research in this area. In this paper will study about anomaly detection and some techniques which are used to identify the anomalous data points in the time series.

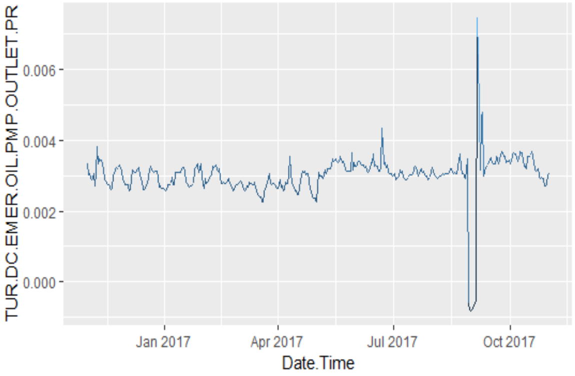
INTRODUCTION

METHODLOGY

There are lots of packages and libraries to detect anomaly data points in both the statistical software R and python. There are many different types of techniques in these packages. Keeping this paper in mind, only some algorithms will be studied. And the most suitable one is anomalize package in R for anomaly detection. Two different types of methods will be used in this package. First with the use of inter quartile range with seasonal and trend decomposition using loess and second seasonal hybrid esd with generalized extreme studentized deviate test. Then by comparing both the packages, will see how many data points by both methods declared anonymous data points.

DATASET AND TOOLS

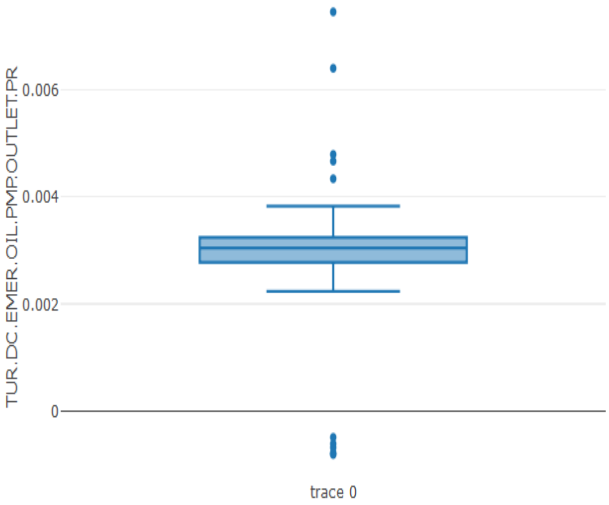
For this research paper a dataset has been used which is taken from the thermal power plant's turbine machine. The data was collected from a turbine through a sensor. However, there is a lot of sensor data in the dataset but only one sensor will be used for this paper because whatever result will be valid for all the sensors. The depiction of the plot shown below is the behaviour of a sensor.



The statistical software R is used for this paper. Along with so many packages have been used such as anomalize, ggplot2, tidyverse, magrittr, tibbletime and reshape. The Package anomalize used for this paper uses STL + IQR and Twitter + GESD. Therefore only these algorithms will be discussed.

DESCRIPTIVE ANALYSIS

The more information about the data, the better it can be analysed. Our purpose is to identify the data points that are not doing the usual behaviour like other data points. The statistical summary of the data can be used for this purpose. Information about average, median, minimum, maximum and standard deviation can be beneficial for us.



Descriptive analysis can be done easily with the above boxplot. The numeric values of the box plot above can also be obtained by using the summary function in R.

Measures	Values
Minimum	-0.0008177
Q_1	0.0027755
Median	0.0030441
Mean	0.0030047
Q_3	0.0032396
Maximum	0.0074586

TIME SERIES MODELING

Time has a different contribution in business. It is time to achieve success in business by analysing it. Well, data can be categorized in many type, but there is data what is called time series which involve time. So, Forecasting can be done using Time Series technique where time is involve in the data. There are so many methods for Modelling Time Series, but before starting modelling, it is very important to analyse the time series such that modelling from which to be done well. So first of all, Time Series should be stationary for forecasting. It can be said that the time series is the stationary when the mean of the time series is constant or it is not the time function. In such a way, variance and covariance should also be constant. Only if these three things satisfy then only we can say that time series is stationary. It is not necessary that the time series should be already stationary. Dicky Fuller Test can be used to find out that the time series is already a stationary or not. So when the time series is not stationary, it can be made stationary by using differencing or transforming techniques. Once the time series is stationary, the forecasting can be done. Methods such as Simple Naive, Simple average, Moving Average, Single Exponential Smoothing, Holt's linear trend method, Holt's winter seasonal method, ARIMA can be used for forecasting. The Important point is that the anomalies can also be detected using the forecasting technique.

TIME SERIES DECOMPOSITION

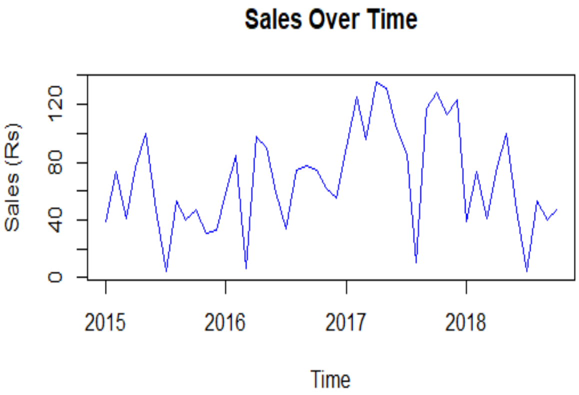
Time series can display a wide variety of data patterns and have to classify some patterns and behaviours that can be seen in the time series. Sometimes it is useful to try to split a time series into several components, each of the underlying patterns represents one of the categories. Often this is done to help understand the time series better, but its use can also be used to improve the forecasting. There are many methods that help in time series decomposition. Some of these are as follows: classical decomposition, x-12-arima, STL, SH-ESD etc. in this paper will study only about STL i.e. seasonal and trend decomposition using loess method and SH-ESD i.e. seasonal hybrid esd.

SEASONAL AND TREND DECOMPOSITION USING LOESS

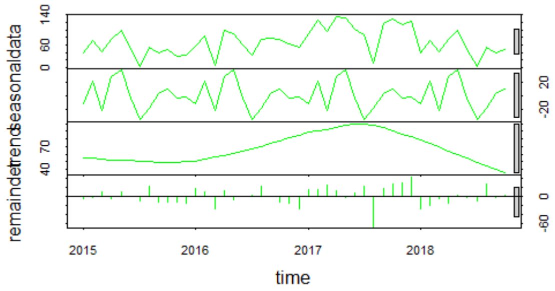
STL is a short term for "seasonal and trend decomposition using lace", while Loes (locally weighted regression and scatterplot smoothing) is a way to estimate nonlinear relationships. The basic idea is that a time series can be broken into three components: seasonal, trend and reminder from t = 1 to N measured data points as shown below.

$$Y_t = Trend_t + Seasonal_t + Remainder_t$$

It is done through two loops. In the external loop, the strength weight on the basis of the remainder is allocated to each data point. This takes into consideration lessening or wiping out the impacts of anomalies. The inward loop interactively refreshes the pattern and seasonal segments. This is finished by subtracting the current estimate of the pattern from the raw series. Time series is divided into cycle-sub-squares. The cycle-subseries are loess smoothed and afterwards passed intensive a low-pass filter. The seasonal components are the smoothed cycle-subseries minus the result from the low-pass filter. The seasonal components are subtracted from the raw data. The result is loess smoothed, which becomes the trend. The rest is the reminder. To explain how this method works, we will use a dataset that is already in R.



After using STL methods, we get the plot shown below.



INTER QUARTILE RANGE

Interquartile range is a method to extract anomalous data points from the data. It uses the distribution to identify the outliers. Below formula is used to find Inter quartile range.

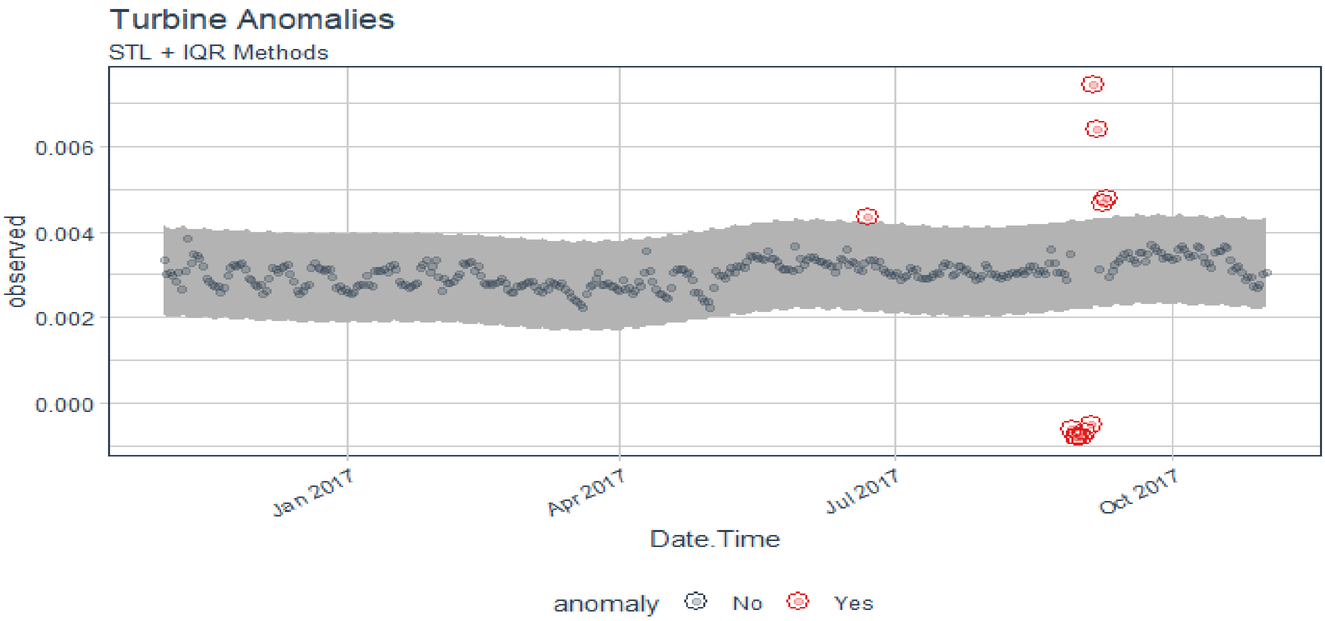
$$IQR=Q_3-Q_1$$

Where, Q_3 is third quartile and Q_1 is first quartile.

In IQR, the first quartile, as denoted by Q1, is a value in the data set that holds 25% of the value below it. The third quartile, Q3 is shown, there is a value in the data set that holds 25% of the value above it. The IQR method isn't subject to any loop hence, it is quicker and more effortlessly scaled than the GESD technique. Be that as it may, this cannot be precise in recognizing irregularities since high use inconsistencies can contact the IQR centre line (median).

STL+IQR

In order to know about the anomalous data points in the time series, we will first have to paraphrase the time series. For this, we can use the STL methods. After decomposing, we will be in the position to detect Anomalous data points. And for this will use interquartile range. So have to use both the techniques STL+IQR to achieve the goal. After applying this method we get the below plot.



SEASONAL HYBRID ESD

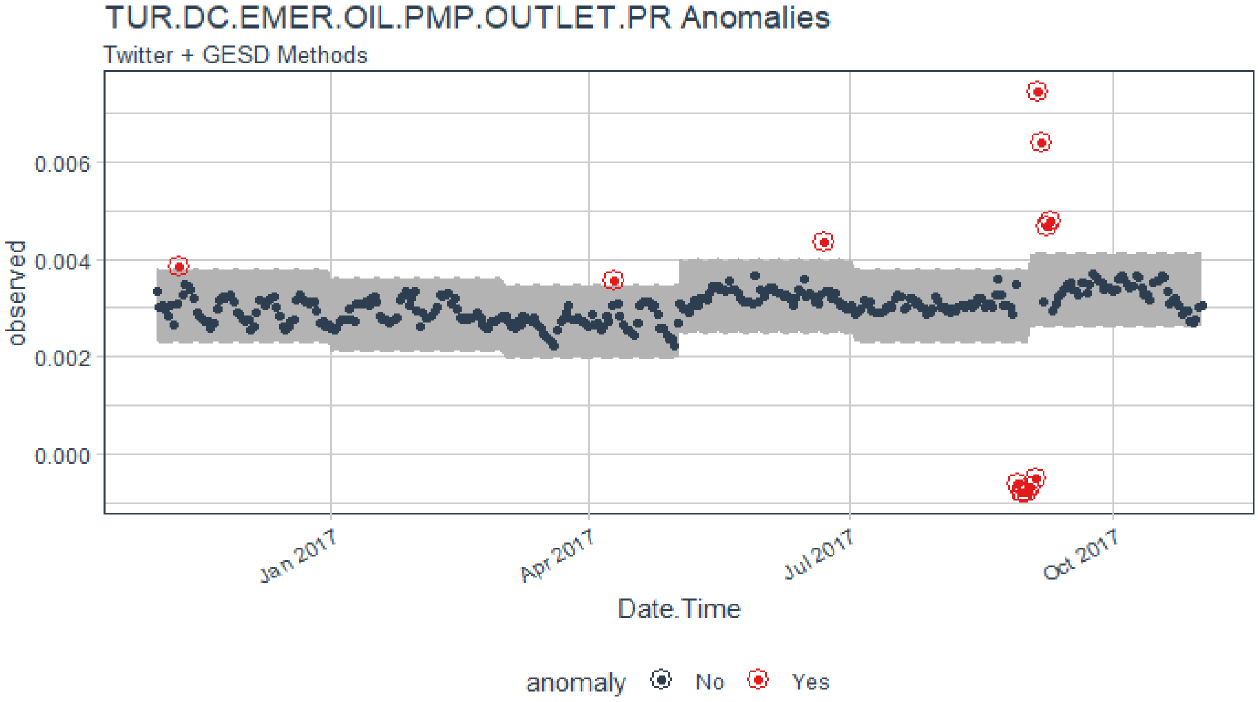
S-ESD strategy utilizes time series disintegration to decide the Seasonal segment of a given time series. S-ESD then applies ESD on the subsequent time series to recognize the peculiarities. Seasonal hybrid ESD (S-H-ESD) builds on generalized ESD test to detect discrepancies. S-H-ESD can be used to detect both global and local anomalies. It is achieved by employing time series decomposition and using strong statistical metrics, e.g., median with ESD.

GENERALIZED EXTREME STUDENTIZED DEVIATE TEST

It includes an iterative assessment of the Generalized Extreme Studentized Deviate test, which continuously assesses abnormalities, evacuating the most exceedingly bad guilty parties and recalculating the test measurement and basic esteem. The basic qualities continuously contract as more high use focuses are expelled. The alpha parameter changes the width of the basic qualities. As a matter of course, alpha = 0.05. The GESD strategy is iterative, and thusly more costly that the IQR technique. The fundamental advantage is that GESD is less impervious to high use focuses since the circulation of the Information is dynamically broke down as oddities are expelled

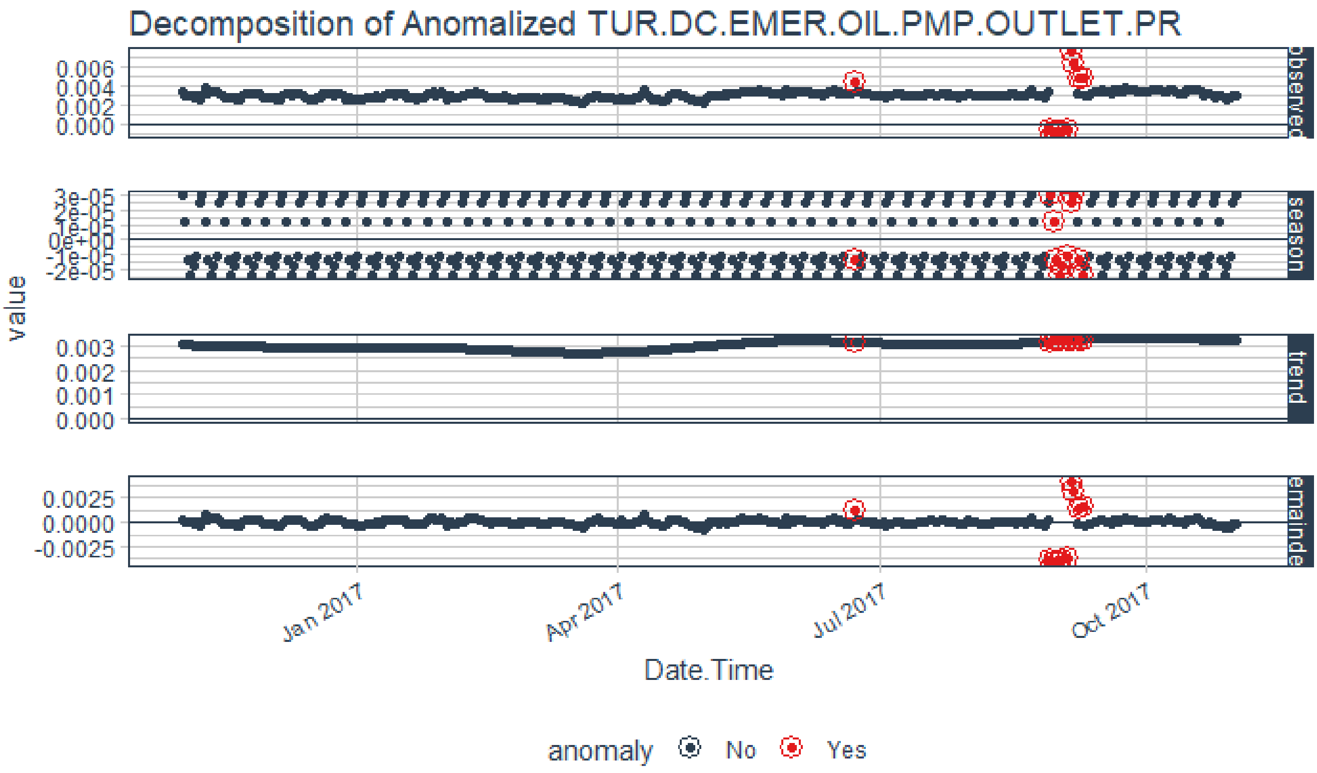
TWITTER+GESD

As has been done in the first method STL+IQR, this is exactly what Twitter + GESD has to do. The basic idea behind SH-ESD + is to use a modified version of STL decomposition to extract residual to apply a modified version of the ESD test series, and then to detect discrepancies in this residual. Twitter uses SEASONAL HYBRID ESD technique. The time series will be decomposed using SEASONAL HYBRID ESD, then will use the GESD method to detect the anomalous data points and this method is named as Twitter + GESD method in Anomalies package. After using this methods, the result will be found below.



DECOMPOSITION OF ANOMALIES DATA POINTS

Before locating Anomalies Data Points, it is also important to know how they looks after the time series data decomposes. As it is known, the time series data is decomposed in Seasonal, Trend and Remainder. The same has been done in the given plot too.



CONCLUSION

OTHER PACKAGES AND METHODS

Only packages used above and not only have their methods, which are capable of detecting anomalous data points, but there are also more packages in use, which can be used to find anomalous data points. Few of them are:

Package	Methods
Twitter's AnomalyDetection Package	Seasonal Hybrid ESD (S-H-ESD)
anomalyDetection Package	Mahalanobis distance, factor analysis, Horn's parallel analysis, block inspection, principle components analysis
tsoutliers package	Chen and Liu procedure
anomalous-acm	Works by computing a vector of features on each time series then applying robust principal component decomposition on
rainbow package	bagplots and boxplots
kmodR package	k-means proposed by Chawla and Gionis in 2013

ACKNOWLEDGEMENT

REFERENCES

[1] Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I. (1990). Stl: A seasonal-trend decomposition procedure based on loess. Journal of Official Statistics, 6(1):3–73.

[2] STL Algorithm Explained: STL Part II
Retrieved from http://www.gardner.fyi/blog/STL-Part-II/

[3] Rosner, B., (May 1983), "Percentage Points for a Generalized ESD Many-Outlier Procedure" , Technometrics, 25(2), pp. 165-172

[4] Owen S. Vallis, Jordan Hochenbaum and Arun Kejariwal (2014). A Novel Technique for Long-Term Anomaly Detection in the Cloud. Twitter Inc.

[5] Owen S. Vallis, Jordan Hochenbaum and Arun Kejariwal (2014). AnomalyDetection: Anomaly Detection Using Seasonal Hybrid Extreme Studentized Deviate Test. R package version 1.0.

[6] Anomalize Methods Retrieved from https://cran.r-project.org/web/packages/anomalize/vignettes/anomalize_methods.html

[7] Decomposition Models Retrieved from

https://onlinecourses.science.psu.edu/stat510/node/69/

[8] Algorithms for Time Series Anomaly Detection Retrieved from https://stats.stackexchange.com/questions/137094/algorithms-for-time-series-anomaly-detection/137214

[9] How to correct outliers once detected for time series data forecasting? Cross Validated, Retrieved from https://stats.stackexchange.com

[10] Alex T.C. Lau (November/December 2015). GESD - A Robust and Effective Technique for Dealing with Multiple Outliers. ASTM Standardization News. Retrieved from www.astm.org/sn

[11] Seasonal Decomposition of Time Series by Loess—An Experiment Retrieved from https://align-alytics.com/seasonal-decomposition-of-time-series-by-loessan-experiment/

[12] Extracting Seasonality and Trend from Data: Decomposition Using R Retrieved from https://anomaly.io/seasonal-trend-decomposition-in-r/

[13] Time series decomposition forecasting principles and practice book by rob Hyndman and George athanasopoulos, slides by peter fuleky oct. 2014 Retrieved from https://blog.datascienceheroes.com/anomaly-detection-in-r/

[14] Ways to Detect and Remove the Outliers Retrieved from https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba