Time Series Outlier Detection

Hitesha Mukherjee (MS2016007), Srinivasa Raghavan (MS2017008), Anoop Toffy (MT2016016)
International Institute of Information Technology, Bangalore

(Draft Paper)

***Abstract*—Outlier detection in time series is critical in system security and health care applications. This involves building mathematical model of time series that are robust to outliers which can prevent false alarm and determine the location and type of outlier in the given time series. It is observed that because of the presence of outliers in time series, the parameter estimates of the mathematical model built for the series could be biased or even inappropriate. The conventional STL decomposition technique could also be subjected to these biases. These skewed models may result in masking effect and false alarms while detecting outliers. In our study we have considered Yahoo Webscope data set [1] which contains network traffic data from Yahoo Fantasy sports. We have used iterative technique for estimating model parameters and outlier effect jointly which prevents bias in model. We also consider time series involving change point, we suggest a new technique that invloves moving average smoothing followed by multivariate adaptive regression splines (MARS) [3] fitting to identify change points, further the outlier detection was carried out for time series subdivided based on change points.**

***Index Terms*—Machine learning, Time series analysis, Outlier detection, ARIMA, MARS, Change points, Joint parameter estimation, Outlier classification.**

1. Introduction

A. Time series analysis

A Time-Series is a sequence of data points corresponding to a set of observations made at a particular time instance e.g.- Rainfall data in a year, solar radiation data, speech signal, network traffic data, ECG and EEG data etc. To model or analyze these signals, there is a need to understand the underlying pattern. These patterns may include components such as trends, seasonality, stochasticity, long term cyclicity and autocorrelation etc.

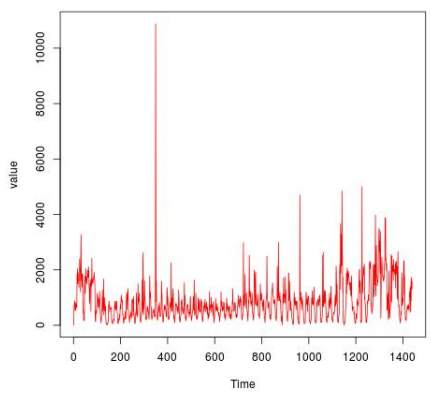

Fig. 1: Time series (example)

The trend component exhibits long pattern, producing irregular effects which are positive, negative, linear or nonlinear involving low, medium and high frequency components, typically observed over several time frames e.g.- decreasing trend of rainfall after monsoon. Seasonality is the regularly repeating fluctuations in time series. The seasonality may be present due to the climatic conditions, calendar, social habit or inherent system characteristics. The modelling might involve additive model or multiplicative model.

The time series may also be associated with deterministic and stochastic noisy component. The random noise will have zero mean and unit variance property as per Gauss Markov condition. If the present or future values are dependent on the past inputs, the property is termed as autocorrelation. Stationarity is the time-invariant property which result in constant mean, covariances between observations at different time window of the time series. The degree of stationarity is measured in terms of autocorrelation function (ACF) and partial auto correlation function (PACF). Pure noise and some autoregressive features are the components which are stationary but trend, cyclicity and seasonality are non-stationary. Once the components of trend, cyclicity and seasonality are removed from the data, the remaining stationary series is referred to as residue. If the residual time series is not stationary, the process is iterated from earlier step to properly model seasonality and trends. Classical decomposition, differencing, locally weighted scatterplot smoothening (loess) decomposition techniques could be used for stationarizing the non-stationary time series. Hypothesis testing such as Augmented Dickey Fuller test could be performed to check whether the residue is stationary or not.

The residue obtained which is stationary can be modelled as any one of the appropriate Auto Regression (AR) process, Moving Average (MA) process, Auto Regressive and Moving Average (ARMA) process. Once predictions from these models are removed from stationary part of time series, it should result in pure white noise. Otherwise the model parameters are updated till the residue is pure white noise. Once these underlying patterns are determined, the time series can be represented by an appropriate mathematical model. The autoregressive integrated moving average (ARIMA) can be used to model the nonstationary time series data. If the order of the model (p+q) increases, the model complexity increases which results in overfitting. Akaike Information Criterion (AIC), AIC corrected (AICc) and Bayesian Information criterion (BIC) determines effectiveness of these models. Lower the AIC and BIC values implies better the model.

B. Outliers

Time series may sometime exhibit eccentric pattern that deviate from normal operating behavior because of the presence of outliers. The data points which are lying outside the regular operating region which results in such deviating characteristics are referred to as outliers. (As shown in fig. 2)

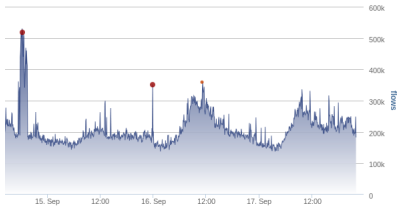

Fig. 2: Outliers (example)

C. Types of outliers

Let $\{Y_t\}$ be a time series following a general ARIMA process,

$$Y_t = \frac{\theta(B)}{\alpha(B)\phi(B)} a_i \quad 1, ...n \tag{1}$$

Where n is the number of observation in the time series; $\theta(B), \phi(B)$ and $\alpha(B)$ are polynomials in B, all roots of $\theta(B)$ and$\phi(B)$ are outside the unit circle; and all roots of $\alpha(B)$ are on the unit circle. To describe time series subject to the influence of a nonrepetative event, the following model is considered:

$$Y_t^* = Y_t + w\frac{A(B)}{G(B)H(B)} I_t(t_1) \tag{2}$$

where Y_t follows a general ARIMA model described in Eq. 1.Here $I_t(t_1)$ is a indicator function for the occurence of the outlier impact, t_1 is possible unknown location of the outlier and w and $A(B)/\ G(B)H(B)$ denote the magnitude and dynamic pattern of the outlier effect. The approach is to classify outlier effect into four types by imposing a special structure to the $A(B)/\ G(B)H(B)$. The type include innovational outlier(IO), and additive outlier (AO), a level shift (LS) and a temporal change (TC). Their definitons are below:

$$IO : \frac{A(B)}{G(B)H(B)} \approx \frac{\theta(B)}{\alpha(B)\phi(B)} \tag{3}$$

$$AO : \frac{A(B)}{G(B)H(B)} = 1 \tag{4}$$

$$TC : \frac{A(B)}{G(B)H(B)} = \frac{1}{1-\delta B} \tag{5}$$

$$LS : \frac{A(B)}{G(B)H(B)} = \frac{1}{1-B} \tag{6}$$

2. Methodology

In this section we give a detailed iterative algorithm used for outlier detection. Section A describes the iterative algorithm that is described in Chen and Lie paper [2]. In section B we propose a method to handle changes points using MARS modelling and figuring out potential change points.

A. Joint Estimation of Model Parameters and Outlier Effects in Time Series

This section summarizes the joint estimation modelling approach following the work of Chung Chen et al [2]. Let Y_t be a time series for modelling ARMA process. If there are 'm' outliers at the time stamps $t_1, t_2, ...t_m$. The model Y_t^* is the model jointly considering the outlier effects and Model paramters. 'e' is the estimated residuals. Here ω and $L_j(B)$ represents the magnitude and dynamic nature of the outlier effect. This methodology involves iteratively updating model parametes and effects of the outlier using the following equations

$$Y_T^* = \Sigma_{j=1}^m \omega_j L_j(B) I_t(t_j) + \frac{\theta(B)}{\phi(B)\alpha(B)} a_t \tag{7}$$

$$e_t = \Sigma_{j=1}^m \omega_j \pi(B) L_j(B) I_t(t_j) + a_t \tag{8}$$

Here α, ϕ, θ are functions of lag order 'B' which models AR, MA and ARMA model. $\pi(B) = \phi(B) * \alpha(B)/\theta(B)/$. Residuals are estimated as $e_t = \pi(B) * Y_t$

The least square estimate of ω from each type of outlier is estimated along with residual standard deviation to obtain the standard outlier statistics τ as described below from Chen and Lie paper [2].

$$\tau_{IO}(t_1) = \omega_{IO}(t1)/\sigma_a \tag{9}$$

$$\tau_{AO}(t_1) = \{\omega_{AO}(t_1)/\sigma_a\}(\Sigma_{t=t_1}^n x_{2t}^2)^{1/2} \tag{10}$$

$$\tau_{LS}(t_1) = \{\omega_{LS}(t_1)/\sigma_a\}(\Sigma_{t=t_1}^n x_{3t}^2)^{1/2} \tag{11}$$

$$\tau_{TC}(t_1) = \{\omega_{TC}(t_1)/\sigma_a\}(\Sigma_{t=t_1}^n x_{4t}^2)^{1/2} \tag{12}$$

Based on these statistics joint estimation is performed in the following 3 stages.

1) Stage 1 :Maximum likelihood estimates for the original time series is obtained. For each time stamp $t_1, t_2, \ldots, t_n$, the statistics τ is computed for all 4 types of outliers. If it is greater than critical value, C there is an outlier at that time stamp. Remove the outlier effect and store location of outlier. This process is continued till all outliers effects are removed and location is stored. Update model parameters if outlier were found with τ statistics. The stage-1 process is repeated if outlier is observed after model updates. Repeat the procedure till no outliers are observed after updating the model. The locations of 'm' outliers from $t_1, t_2, \ldots, t_n$ is noted.

2) Stage 2 :Once location of outlier $t_1, t_2, \ldots, t_n$ is known, from the residue of eqn 8 , the outlier effect ω_j can be estimated. With the τ_j computed on the recent estimates ω_j, if the statistic is less than the critical value, C the outlier is deleted at that time point and repeat for m-1 outliers. Remove the significant outliers with the adjusted series and recent values of ω_j estimated. Compute the maximum likelihood estimates of the model parameters based on the adjusted series obtained. If the relative change of the residual standard error from the previous estimate is greater than epsilon repeat the above Stage-1.
3) Stage 3 :With latest model parameter estimates, obtain residuals after original series is filtered. Use residuals and iterate through stage-1 and stage-2 to obtain final model parameter estimates and outlier effects.

B. Using Multivariate Adaptive Regression Splines and recursive partitioning for dataset containing change points

When moving average smoothening is applied on time series involving change points, the change in trend corresponds to change point anomaly. The same can be observed for a time series in A4 Benchmark dataset.

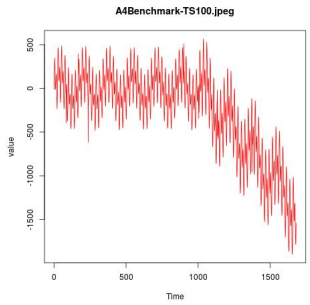

Fig. 3: Change points (example)

Enhanced Adaptive Regression Through Hinges (EARTH) is a implementation of MARS in R. This function can fit linear model between predictor and target by segmenting predictor variables. Integrating all these linear chunks, the technique can be used to model complex non-linearity. Hinge functions is used in MARS which play important role in the partition of data into disjoint regions. Hinge functions or constant or product of hinge functions can be used to create basis functions. The MARS builds a model as a weighted sum of these basis functions. The MARS adds basis functions in pairs with least residual error at every cut points in forward pass stage. Cut points/knots/terms are selected automatically by MARS during forward pass. To overcome overfitting problem, it prunes the unwanted terms in backward pass stage. The pruning is done until best fit for sub models is obtained with optimum cut points.

For a given multivariate data, EARTH package in R returns the location of cut points. From a moving average smoothened univariate time series, EARTH package can be used to find cut points which corresponds to change points in original time series.

3. Dataset and Results

A. Description of our Dataset

The dataset we have is the Yahoo Dataset [1], we have received this data from Yahoo. This dataset contains 371 files which includes four sets of Csv files.

- A1Benchmark/real_(int).csv
- A2Benchmark/synthetic_(int).csv
- A3Benchmark/A3Benchmark-TS(int).csv
- A4Benchmark/A4Benchmark-TS(int).csv

A1Benchmark is based on the real production traffic to some of the Yahoo Servers. The other 3 benchmarks are based on synthetic time-series. A2 and A3 Benchmarks include outliers, while the A4Benchmark includes change-point anomalies. The synthetic data set contains time-series with random seasonality, trend and noise. The outliers in the synthetic dataset are inserted at random positions. Note that the timestamps of the A1Benchmark represents 1-hour worth of data. The A3Benchmark only contains outliers while the A4Benchmark also contains the anomalies that are marked as change-points. The synthetic time-series have varying noise and trends with three pre-specified seasonality's. The anomalies in the synthetic time-series are inserted at random positions.

B. Observations

The Time series outlier detection on Yahoo time series data were analysed with STL decomposition, joint estimation and MARS techniques. The Outlier detection using joint estimation approach on the time series was used for A1, A2 and A3 Benchmark dataset. The outliers and it effects on time series, one from each of the three Benchmark set is as shown in figure 4, 5, 6, 7, 8 and 9 respectively. For a time series in A4 Benchmark data set, the time series is divided into subseries based on the number of cut points obtained with EARTH package. Further, the outlier detection is carried out for each sub series figure 13, figure 15 and the results are as shown in figure 14 and figure 16.

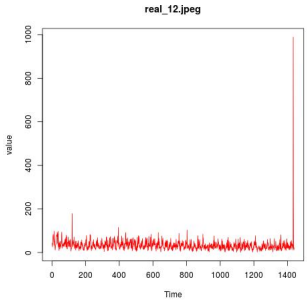

Fig. 4: A1 Benchmark time series

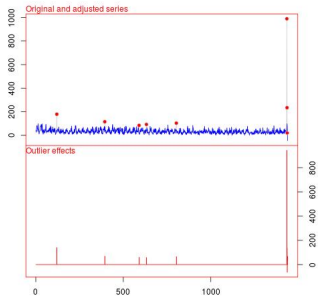

Fig. 5: A1 Benchmark time series and outliers

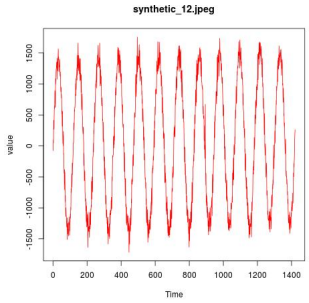

Fig. 6: A2 Benchmark time series

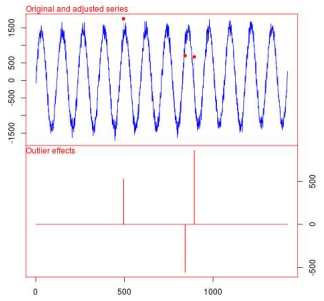

Fig. 7: A2 Benchmark time series and outliers

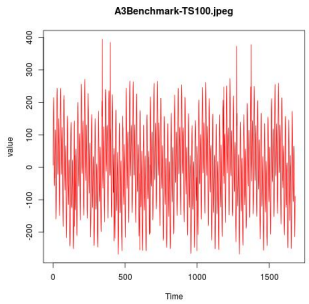

Fig. 8: A3 Benchmark time series

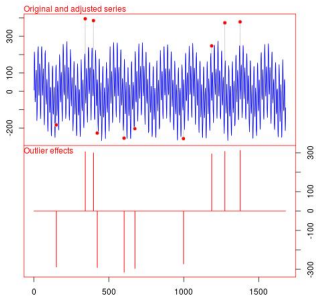

Fig. 9: A3 Benchmark time series and Outliers

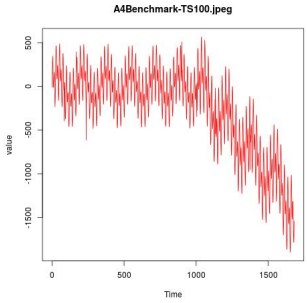

Fig. 10: A4 Benchmark time series

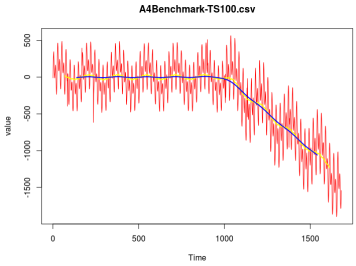

Fig. 11: A4 Benchmark time series (Smoothened)

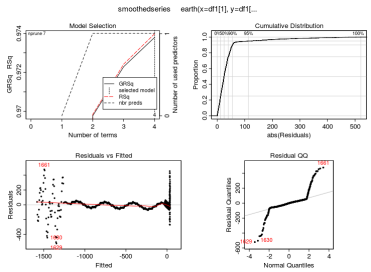

Fig. 12: A4 Benchmark time series (EARTH fit)

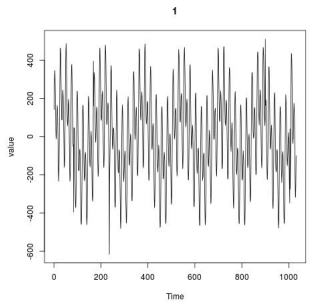

Fig. 13: A4 Benchmark time series (Split One)

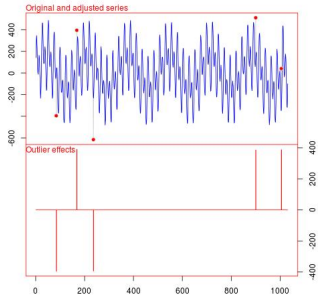

Fig. 14: A4 Benchmark time series (Split One) Outlier

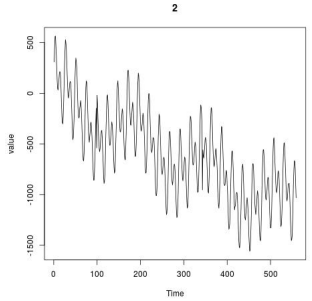

Fig. 15: A4 Benchmark time series (Split Two)

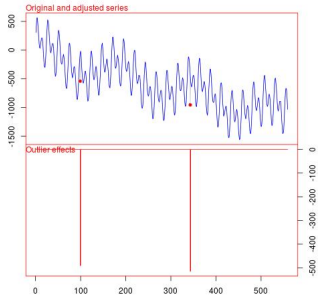

Fig. 16: A4 Benchmark time series (Split Two) Outlier

ACKNOWLEDGMENT

We would like to thanks Yahoo Labs for providing us with the necessary data set [1]. We also would like to convey our sincere thanks to our Professor G. Srinivasaraghavan for guiding us and motivating us throughout this work.

APPENDIX A
CODE

- **Github link** : https://goo.gl/yRdE46

APPENDIX B
MISCELLANEOUS

AUTO REGRESSIVE (AR) TIME SERIES

In this model, the output variable is linear combination on present and past value of inputs in addition to white noise. It is all pole infinite impulse response filter applied on white noise.

$$X(t) = \mu + \epsilon(t) + \Sigma\alpha(i)X(t-i); i = 1, 2, \ldots, p \quad (13)$$

Where X(t) is time series, ϵ is stochastic noise. The above equation is pth order Auto Regressive model: AR(p)

MOVING AVERAGE(MA) TIME SERIES

In this model, the output variable is linear combination on present and past value of noise or error. It is finite impulse response filter applied on white noise.

$$X(t) = \mu + \epsilon(t) + \Sigma\alpha(i)\epsilon(t-i); i = 1, 2, \ldots\ldots, q \quad (14)$$

Where X(t) is time series, ϵ is stochastic noise. The above equation is qth order Moving Average model: MA(q)

AUTO REGRESSIVE MOVING AVERAGE (ARMA) TIME SERIES

To combine the characteristics of AR time series and MA time series model we can use ARMA(p,q) model.

$$\begin{aligned} X(t) &= \mu + \Sigma\alpha(i)X(t-i) + \Sigma\beta(j)\epsilon(t-j) + \epsilon(t); i \\ &= 1, 2, \ldots\ldots, p; j \\ &= 1, 2, \ldots\ldots, q. \end{aligned} \quad (15)$$

STL DECOMPOSITION

STL Decomposition: STL is a very versatile and robust method for decomposing time series. STL is an acronym for “Seasonal and Trend decomposition using Loess”, while Loess is a method for estimating nonlinear relationships. The STL method was developed by Cleveland et al.. STL has several advantages over the classical decomposition method STL will handle any type of seasonality, not only monthly and quarterly data. The seasonal component is allowed to change over time, and the rate of change can be controlled by the user. The smoothness of the trend-cycle can also be controlled by the user. It can be robust to outliers (i.e., the user can specify a robust decomposition). So occasional unusual observations will not affect the estimates of the trend-cycle and seasonal components. They will, however, affect the remainder component.

A time series decomposition is a mathematical procedure which transform a time series into multiple different time series. The original time series is often computed (decompose) into 3 sub-time series:

- Seasonal: patterns that repeat with fixed period of time. A website might receive more visit during weekends. This is a seasonality of 7 days.
- Trend: The underlying trend of the metrics. A website which gains popularity should have a general trend who go up.
- Cycle: Cycle component consists of decreasing or increasing patterns that are not seasonal. Usually, trend and cycle components are grouped together. Trend-cycle component is estimated using moving averages.
- Residual: Finally, part of the series that can't be attributed to seasonal, cycle, or trend components is referred to as residual or error.

The process of extracting these components is referred to as decomposition. First, we calculate seasonal component of the data using stl(). We have used STL Algorithm for decomposition along with Outliers package for plotting the outliers in A1Benchmark and A2Benchmark dataset. STL is a flexible function for decomposing and forecasting the series. It calculates the seasonal component of the series using smoothing, and adjusts the original series by subtracting seasonality.

Stationary Series

One of the basic requirement for time series modelling is to first check whether the data is stationary or non-stationary. The data is said to be stationary if it satisfies the following property

- The mean of the series should not be a function of time but should be a constant.
- The variance of the series should not be a function of time.
- The co-variance of the ith term and (i + m)th term should not be a function of time.

So, for the time series modelling, if the data is not stationary, it has to be first converted into stationary data. To verify whether data is stationary or not, we can use Dickey Fuller Test of Stationarity. For the Dickey-Fuller Test, the p value has to be less than 0.05. If it satisfies the above condition, it is considered as stationary series, else it is a non-stationary series. To convert non-stationary into stationary, differencing should be applied before proceeding. Next step identifies if the data is stationary or not, by applying Dickey Fuller Test and verifying with Plot. Without any differencing, the result of Dickey Fuller Test, Alternative Hypothesis: stationary p-value = 0.01, Alternative Hypothesis: Stationary Since p-value ¡ 0.05, it implies that data is stationary, also seen from Figure below. We have also plotted PACF and ACF.

Autocorrelation Function(ACF)

It is a correlation function which rather than showing correlation between two different variables, shows the correlation between different values of same variable i.e Xi and Xi+k. ACF is used to identify the q value for MA component of ARIMA/Multivariate ARIMA. To calculate the MA term of the model, the lag at which the ACF cuts is considered. It displays the sharp cut-off at the negative correlation.

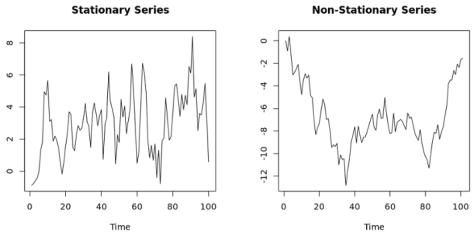

Fig. 17: Stationary and non-stationary series

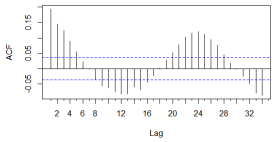

Fig. 18: Autocorrelation Function(ACF)

Partial Autocorrelation Function(PACF)

It is also a correlation function, but it gives the partial correlation of time series with its own lagged values, controlling for the values of the time series at all shorter lags. It contrasts with the autocorrelation function, which does not control the other lags. PACF is used to identify the p value for the AR component of ARIMA/Multivariate- ARIMA. The lag at which the PACF cuts off is the indicated number of AR terms. It displays a sharp cut off at the positive autocorrelation.

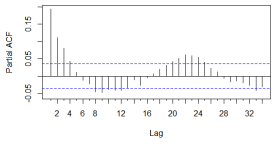

Fig. 19: Partial Autocorrelation Function(PACF)

STL Decomposition and ARIMA:

- The process of extracting Seasonal Component, Trend Component and residue is referred to as Decomposition.
- We calculated seasonal component of the data using stl(). STL is a flexible function for decomposing and forecasting the series. It calculates the seasonal component of the series using smoothing, and adjusts the original series by subtracting seasonality. After decomposing the data we have plotted graphs for all of them.

- We use ts function to specify data in the time series format. As for the frequency parameter in ts() object, we are specifying periodicity of the data, i.e., number of observations per period.
- The task decomposing the series and removing the seasonality can be accomplished by simply subtracting the seasonal component from the original series. seasadj() is a convenient method inside the forecast package. We now have a de-seasonalized series.
- Fitting an ARIMA model requires the series to be Stationary. A series is said to be stationary when its mean, variance, and autocovariance are time invariant.
- The Augmented Dickey-Fuller (ADF) test is a formal statistical test for stationarity. The null hypothesis assumes that the series is non-stationary. ADF procedure tests whether the change in Y can be explained by lagged value and a linear trend. If contribution of the lagged value to the change in Y is non-significant and there is a presence of a trend component, the series is non-stationary and null hypothesis will not be rejected.
- Some of the files in our dataset are stationary is a stationary, since p value obtained is less than 0.05. No differencing technique is needed.
- We determine the ACF and PACF for our dataset. We plot them for our dataset.
- After determining the ACF and PACF and differencing factor. Now let's fit a model. The forecast package allows the user to explicitly specify the order of the model using the arima() function, or automatically generate a set of optimal (p, d, q) using auto.arima().
- Two of the most widely used criteria are Akaike information criteria (AIC) and Baysian information criteria (BIC). These criteria are closely related and can be interpreted as an estimate of how much information would be lost if a given model is chosen. When comparing models, one wants to minimize AIC and BIC.
- While auto.arima() can be very useful, it is still important to complete steps 1-5 in order to understand the series and interpret model results. Note that auto.arima()also allows the user to specify maximum order for (p, d, q), which is set to 5 by default.

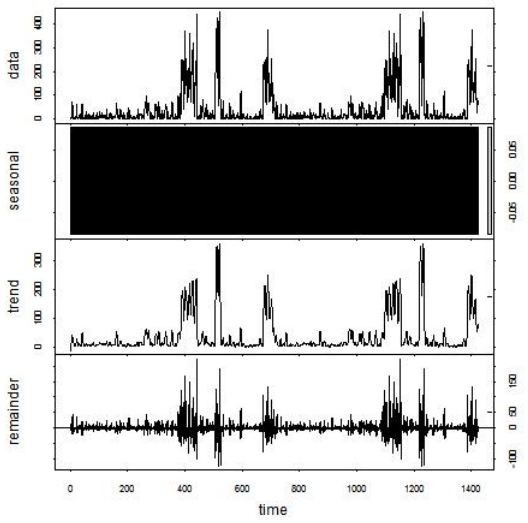

Fig. 20: STL decomposition

REFERENCES

[1] S5 - A Labeled Anomaly Detection Dataset, version 1.0(16M) *A Benchmark Dataset for Time Series Anomaly Detection*. Yahoo Webscope Program https://webscope.sandbox.yahoo.com/catalog.php?datatype=s&did=70

[2] Chung Chen and Lon-Mu Liu *"Joint Estimation of Model Parameters and Outlier Effects in Time Series"* Journal of the American Statistical Association, Vol. 88, No. 421 (Mar., 1993), pp. 284-297 https://www.researchgate.net/publication/243768707_Joint_Estimation_of_Model_Parameters_and_Outlier_Effects_in_Time_Series

[3] Friedman, Jerome H *Multivariate adaptive regression splines*, The annals of statistics, 1–67, 1991, JSTOR