SEQUENTIAL CHANGE-POINT DECTECTION

Kushal Bhyregowda, Shubhvaratha Dutta, Parth Parth Mahendrabhai patel, Venkata Surya Teja Sistla,
Arizona State University

1 Introduction

Statistical control charts have long been used to monitor the quality of the process. Traditional Statistical process control methods assume the process data are independent and normally distributed. In contrast, the data in real-world applications is usually dependent and not always normally distributed. Also, traditional methods are not effective in identifying small changes in the distribution. If the process failure costs are high, then detecting these changes as early as possible becomes important.[1] So detecting abrupt changes quickly in a sequence of observations $X1 \dots \dots Xn$ has attracted researchers in statistics and various other communities for decades.[2].

Sequential change-point detection has many applications such as heart arrhythmia in the electrocardiogram, monitoring air and water pollution in a city, ecosystem disturbances, intrusion in networks etc. [3]. While numerous algorithms are developed to detect changes in univariant time series only a few algorithms are developed for detecting changes in multivariate time series.[3]. The detection of changes in multivariate time series is more difficult for several reasons. Firstly, it's hard to establish a concise definition of change. Secondly, it is highly susceptible to the presence of noise in one or more time series.[3]

In this project, we are going to discuss the methodology used for sequential change detection and an algorithm and how is it applied for detecting changes in financial time series: monitor changes in the real-time price of bitcoin and other cryptocurrencies prices such as Ethereum, Litecoin, and ripple. [4]. Also, we are going to discuss the implementation of the MKS model to the COVID 19 time series.

2 Methodology

There are two main types of change detection methods, batch and sequential. In the quality control literature, these are respectively knowns as Phase-I and Phase-II analysis. To understand the methodology of Sequential Change Point Detection we must first understand Phase-I analysis –

2.1 Batch change detection (Phase I) -

In this phase, we consider a fixed length sequence containing the n observations x1…, xn which may or may not contain a change point. For analysis, we assume that the sequence contains at most one change point. If no change point exists, the observations are independent and identically distributed according to some distribution F0. If a change point exists at some time $\boldsymbol{\tau}$, then the observations have a distribution F0 prior to this point, and a distribution F1 afterwards, where F0 $\neq$ F1. For a specified point k in this sequence, we can use a standard two-sample hypothesis test to assess whether a change point occurs at $\tau = k$, with the null hypothesis being that there is no change and that all n observations are identically distributed. The two-sample hypothesis can be defined as-

$$H_0 : X_i \sim F_0(x; \theta_0),\ i = 1, \ldots , n,$$

$$H_1 : X_i \sim \begin{cases} F_0(x; \theta_0), & i = 1, 2, \ldots , k, \\ F_1(x; \theta_1), & i = k + 1, k + 2, \ldots , n, \end{cases}$$

where θi represent the potentially unknown parameters of each distribution.

A suitable test statistic D_{kn} is chosen for the above two-sample hypothesis test based on information we have about the distribution of the observations, and the type of change which they may undergo. For example, Fisher's exact test (FET) statistic can be used for observations following Bernoulli distribution, or we can use Mann-Whitney and Mood statistics for detecting location and scale changes respectively for the sequences where there is no prior information is available about the distribution.

Since the location of changepoint in the sequence is not known, we evaluate D_{kn} at every value $1 < k < n$ in the sequence, and the maximum value is used for the analysis. In other words, our finite length sequence data is split into two contiguous subsequences in every different possible ways so that the two-sample hypothesis test can be applied at each of the split points. So, the test statistic is:

$$D_n = \max_{k=2,\ldots,n-1} D_{k,n} = \max_{\mathrm{k=2,\ldots,n\text{-}1}} \left| \frac{\tilde{D}_{k,n} - \mu_{\tilde{D}_{k,n}}}{\sigma_{\tilde{D}_{k,n}}} \right|$$

if $D_n > h_n$ → Reject Null → Changepoint Detected
otherwise, Fail to Reject Null → No changepoint detected

where, h_n = appropriately chosen threshold value

The threshold value h_n is chosen such that we can bound the Type-1 error (α) rate which is a part of standard statistical hypothesis testing. This is done to keep the false alarm low and can be determined in the cpm (Change Point Model) package.

Finally, the best estimate of the change point location will be immediately following the value of k which maximized D_n:

$$\hat{\tau} = \arg \max_{k} D_{k,n}$$

2.2 Sequential change detection (Phase II) –

Using online available CPM package/framework, the two-sample hypothesis testing approach used in the batch case can be extended to sequential change detection where new observations are received over time, and multiple change points may be present.

Let x_t denote the tth new observation that has been received, where $t \in \{1,2,\ldots.\}$. Whenever a new observation x_t is received, the CPM approach treats $x_1,\ldots..,x_t$ as a fixed length sequence and computes D_t using the previous batch methodology. Here, we are using the notation D_t rather than D_n to highlight the sequential nature of the procedure.

Again, like the previous section change is flagged if Dt > ht for some appropriately chosen threshold. If no change is detected, the next observation xt+1 is received, then Dt+1 is computed and compared to ht+1, and so on. Whenever a ch, therefore, is detected, the change detector is simply restarted from the following observation in the sequence. The procedallowerefore consists of a repeated sequence of hypothesis tests. In general, the Dk,n statistics have the properties which allows Dt+1 to be computed from Dt without incurring too much computational cost overhead. Therefore, this process may seem computationally expensive, but it is computationally efficient.

Next key step in this approach is then determining the sequence of threshold values {ht}. These false alarms constitute flagging that a change has occurred when no change has taken place. In the sequential setting, h_t is chosen so that the probability of incurring a Type-1 error (α) is constant over time, so that under the null hypothesis of no change:

$$P\ (D_1 > h_1) = \alpha$$

$$P\ (D_t > h_t\ |D_{t-1} \leq h_{t-1}, \ldots, D_1 \leq h_1) = \alpha, \quad t > 1$$

In this case, assuming that no change occurs, the average number of observations received before a false positive detection occurs is equal to $1/\alpha$ = the average run length, or ARL_0.

In general, Monte Carlo simulation is used to compute the required sequences of ht values corresponding to a given choice of α. This is a computationally expensive procedure, but it only needs to be carried out a single time, and the values can then be stored in a look-up table. The cpm package contains pre-computed sequences of thresholds which correspond to a variety of choices of α.

2.3 Different CPM Packages available –

As discussed earlier, to extend Phase-1 analysis to our Phase-II Sequential, we can use different CPM packages available online -

- *The Student-t, Bartlett and GLR statistics* for detecting changes in a Gaussian sequence of random variables. The first two monitor for changes in either the mean or variance respectively, while the latter can detect changes in both
- *The Exponential statistic* for detecting a parameter change in a sequence of Exponentially random variables.
- *The GLR Adjusted and Exponential Adjusted statistics* which are identical to the GLR and Exponential statistics, except for using the finite sample correction described in Ross (2014) which can lead to more powerful change detection.
- *The Fisher's exact test (FET) statistic* for detecting a change in a sequence of Bernoulli random variables.
- *The Mann-Whitney and Mood statistics* for detecting location and scale changes respectively in sequences of random variables, where no assumptions are made about the distribution.
- *The Lepage, Kolmogorov-Smirnov, and Cramer-von-Mises statistics* for detecting more general distributional changes where again no assumptions are made about the sequence distribution.

2.4 MKS (Mann-Kendall-Sneyers) Test Methodology –

The MKS Test is a non-parametric test. We will see its application on COVID-19 time series data to understand how sequential change point detection method works. For this method we first complete, three major steps –

- ***Step 1: Determine test statistics (Sk)***

$$Sk = \sum_{i=1}^{k} mi \ , \quad where \ (k = 1, 2, 3, . . . , N) \qquad (1)$$

m_i = total number of elements xj preceding xi (j < i) where xj < xi (i = 1, 2, …, *N)*

Mean, $E(S_k) = k(k-1)/4$

$VAR(Sk) = k(k-1)(2k-5)/72$

- ***Step 2: Determine two sequences (U_f and U_b)***

  Forward Sequence, $U_f = (Sk - E(Sk))/\sqrt{(VAR(Sk))}$

  Backward Sequence, U_b = Obtained by reversing time series data X and deriving Xr. We derived the intermediate sequence U_{fr} by applying Eq. (1) to Xr. Lastly, we derived the backward sequence Ub (dashed line in Fig. 1) by first reversing the sequence of values in Ufr and then adding a negative sign to these values.

- ***Step 3: Detect change points***

  To detect change points accurately we employ a statistical filter—the points of intersection falling beyond the 95% confidence Intervals (CIs), which correspond to Z-scores = ±1.96, are rejected.

  Upward shift = if a point of intersection is between the Z-scores of 0 and 1.96,

  Downward shift = if the point is between the Z-scores of − 1.96 and 0, the change is downward.

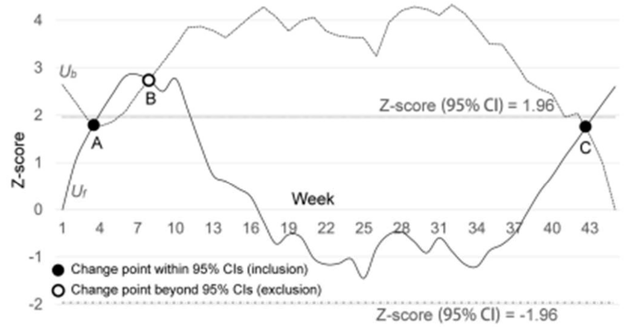


Fig. 1 MKS test of new weekly cases in Virginia with the forward sequence (solid line) and the backward sequence (dashed line). The black dot is the identified change point, and the white dot is the excluded change point

Figure 1 :- MKS test case study for Virginia

This is the weekly reported COVID case data study of Virginia, we can see that the forward sequence is shown in a solid line whereas the backward sequence is plotted in a dashed line. To detect the change point, we can put a statistical filter so if a point intersection falls beyond the 95% confidence interval (CIS) which corresponds to the Z-scores = ±*1.96,* is rejected. The Z-score value between 0 to 1.96 is known as an upward shift whereas the Z-score value between 0 to -1.96 is known as a downward shift. So, in the figure, we can see 3 points A, B, C. where Point A (week 4) and Point C(week 43) are within 95% of CIS so it is known as identical change point whereas, point B(week 8) is beyond the 95% of CIS so it is known as excluded change point. We can say that points A and point C showed an upward change.

3 Application 1: Monitoring Financial Time Series

Times series analysis provides key insights while analyzing various factors in finance. The distributions of a financial time series change with respect to time with the input of new data. In terms of the finance market, the change points can predict a boom or crash in the market. The changes in the finance market have no prior indication and occur suddenly due to the noisy and non-stationary structure of the data. While detecting these change points, decreasing the false alarm rate is crucial.

In the following application, the daily returns of Bitcoin are distributed, and change-point detection strategies are applied to analyze its market. Later the entire crypto market is analyzed for change points using price returns of Bitcoin, Ethereum, Ripple, and Litecoin.

(A)Univariate Change point detection of Bitcoin Price Returns: -

(1) Methodology

For ease of analysis, we convert Bitcoin price returns to Logarithmic Bitcoin price returns. The distribution is tested for change points using the statistic Cramer-von-mises at ARLO= 1000,900,800 in the time interval September 2014 to July 2021 Following graphs are results obtained for Cramer-von-mises test at ARL0=1000 for log return, log means, and log variance plots of Bitcoin returns.

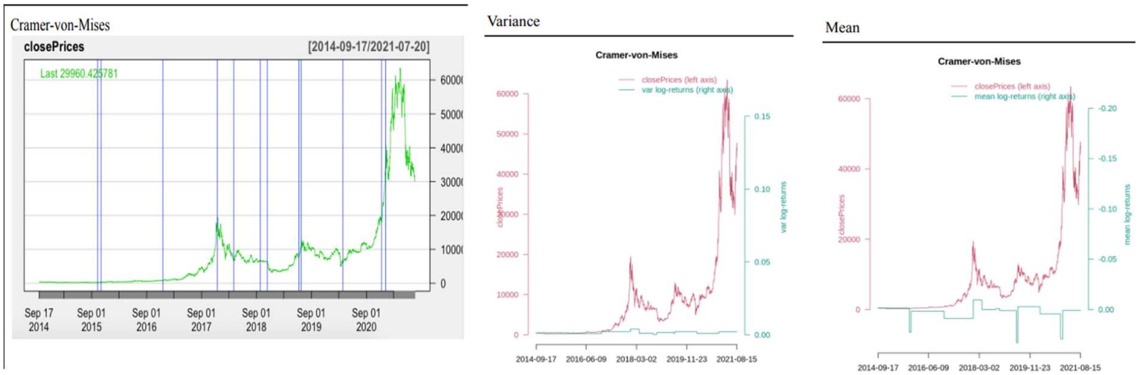


Fig 1,2,3: - Plot 1 shows change points from log returns. Plot(2) and plot three show log mean and variance graphs. In plot(1), green lines indicate USD returns of Bitcoin, and Blue lines indicate change points. In plot(2) green line indicates Log mean returns, and in plot(3) green line indicates the log variance returns.

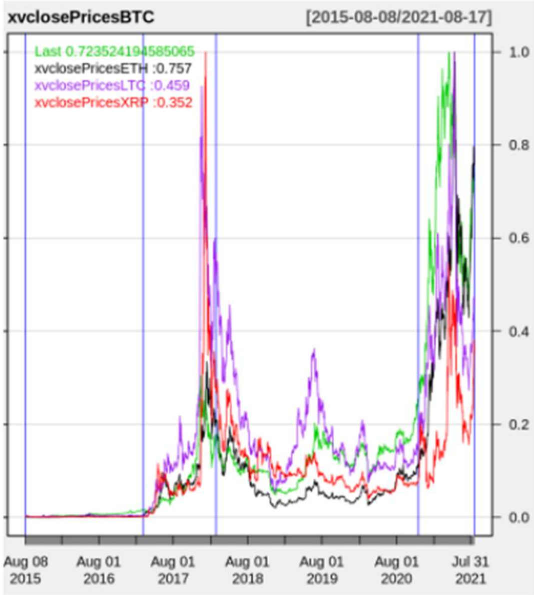
(2) Results of the Univariate analysis of Bitcoin: -

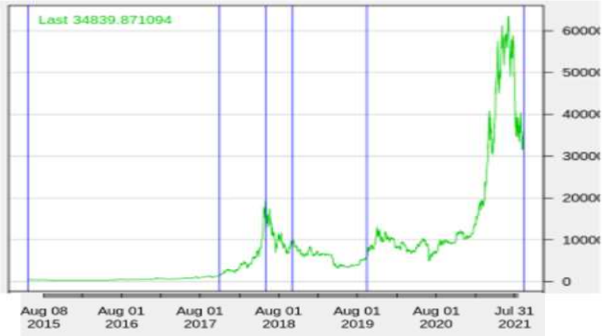
In the Log mean returns plot, the spikes indicate the change in the Log mean and detect growth points in that period. The Cramer-von-mises statistic indicates small growth periods in

2015,2017, and 2019, followed by periods where there was growth or decline at a slower pace. The analysis also captures the financial boom and crash from 2017 to 2018. The effects of the boom in 2017 balanced the crash in 2018.

(B)Multivariate Change point detection of Cryptocurrencies: -

This analysis takes the combined mixed log returns for the four cryptocurrencies. For the simplicity of analysis, the log return time series is converted to a normalized log return time series by normalization using maximum and minimum scaling. The e.divisive methodology is applied to the ECP R package to detect combined change points. The e.divisive methodology detects change points by comparing the distributions before and after a point. The change point is detected when there is a difference between the sides of the time series. The e.divisive method is also applied to the bitcoin returns. The results obtained are given in the below figure.





cryptocurrencies

(1) Results from Multivariate time series of Cryptocurrencies: -

In the combined log return plot for cryptocurrencies, change points are detected in periods 2016-2017, 2017-2018, and 2020 2021. The change point detected at the point between 2018 to 2019 on a univariate plot of bitcoin has not been detected on a multivariate as Litecoin sustained growth in that period. Also, the change point detected on multivariate in 2020 was not detected by univariate. The change point in 2020 indicated growth in Ethereum and Ripple.

(C) Socioeconomics behind changes in the crypto market: -

Socioeconomics drives major changes in the crypto market as they consider events like regulations in the market, robberies, investments, and rumours, which can lead to market turmoil. Many such events have been key contributors to the crypto market crash in 2018. In 2016 a Swiss railway operator upgraded their ticket scanning systems to steal Bitcoin addresses from phone apps. In 2018 Coin check was hacked, and NEM worth 530 million USD was robbed. This led to the immediate suspension of all trading on Coin checks. In 2018 token sales were banned on initial coin offerings. All these events led to the crash of the crypto market in 2018.

So, it is important to consider such factors for change point detection. The fear and greed index proposed by the website alternative.me can be used to check the emotional context of shareholders and investors in a particular instant. Fear indicates factors like volatility in the market and increased negative rumours about a stock. Greed indicates positive investments, an increase in buy volumes, and good news in the context of the stock. Change points are detected for log Bitcoin returns using Cramer-von-mises in considering the fear and greed index at ARL0=1000,4000. The change points are plotted on the fear-greed index graph for analysis. Change points are detected during 2017 when there was an increase in greed with the boom in the market, and a change point at the 2018 crash is detected where an increase of fear can be seen in fig 6.

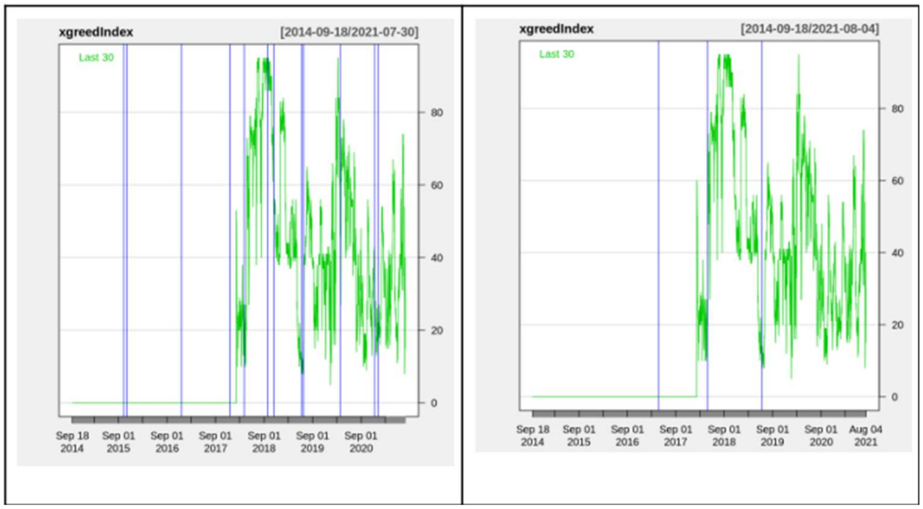


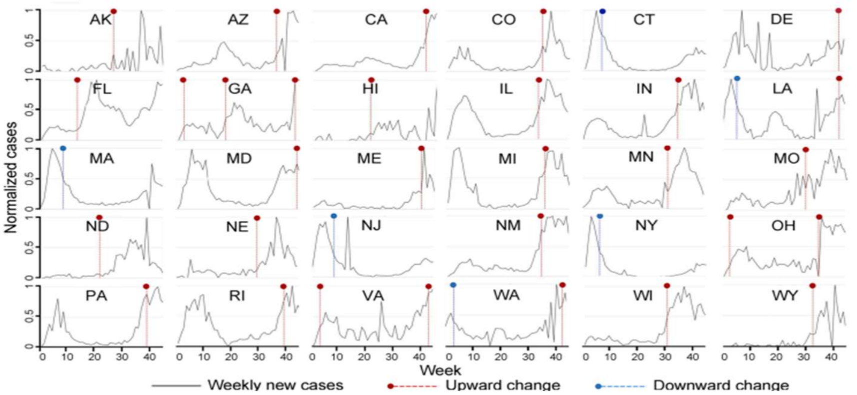
Fig 6: - Plots for change points on fear greed index at ARL0=1000 and 4000, respectively. Blue lines indicate change points.

4 APPLICATION 2: MONITORING COVID 19 DISEASE

As we all know, the COVID-19 pandemic has disturbed human activities. Due to this governments put some restrictions such as no face-to-face interaction, closed non-essential business, low-down, and mask is mandatory. However, different states have different timeline policies and strengths. Here we took weekly COVID-19 Cases from 50 States of the United States. In this application, we took 45 weeks of data on COVID-19 cases in all 50 states. Brief example of implementation of sequential MKS test on this same COVID case dataset is already discussed in Section 2.4 of this article.

RESULTS

After applying the MKS test in all 50 states of united states, we found that 30 states have at least one change point detection within 95% confidence intervals. Among this, 25 states have one change point detection within 95% confidence intervals, 4 states (LA, OH, VA, WA) have two change point detection within 95% confidence intervals, and one state(GA) have three change point detection within 95% confidence intervals. Only one state Vermont has on change point detection within 95% confidence intervals means in this state there was on sudden increment and decrement of COVID-19 cases in entire 45 weeks of study time. We can see in figure 2, 30 states which have at least one change point in within 95% CIs. In the figure we can see upward change (Red line), and downward change (blue line) which indicate sudden increment and decrement in COVID cases during that period.



Based on three clusters of chronological change point detection based on disease Development Stage is divided into three stages. So, weeks 1-10 (March 23 - May 31, 2020) are considered as the 1st stage, weeks 11-31 (June 1 – November 19, 2020) are considered as a 2nd stage, and weeks 31-45 (November 19, 2020 – January 31, 2021, considered 3rd stage of this disease development.

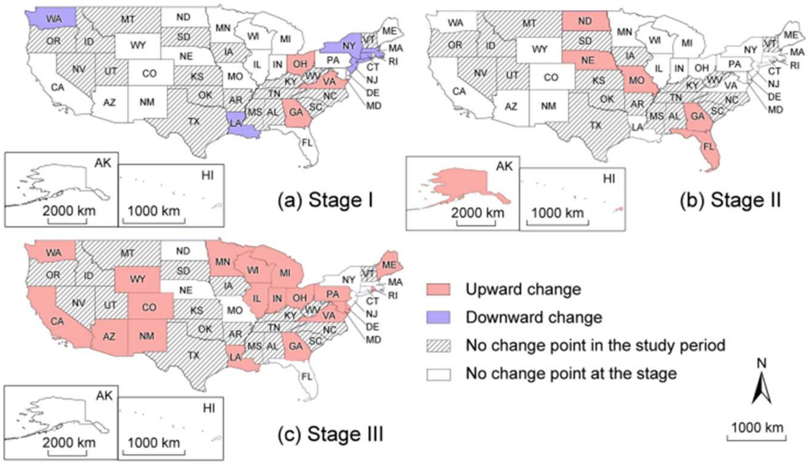


Figure 3:- The emergence change point of each state in stages wise.

From figure 3, in the stage, one can see a downward change in the Northeastern state (MA, NY, NJ, CT) this pattern explains that during that time their state government put some restrictions on face-to-face interaction, and also closed non-essential businesses and schools due to this in that time COVID cases were low as compared to stage 3, whereas in stage 3 we can see upward change in western and in mid-western states, and cases rise in the last summer and fall. The main reason behind the increase in covid cases is the government removing or reducing restrictions on face-to-face interaction, and reopening public places without a mandatory mask.

Comparison results with other methods

We compared this MKS method with other two change point detection methods which are Pruned exact linear time(PELT) Method and the regression-based method. Both methods are used for change point detection in time series data. Mainly, this method minimizes the cost of function over the possible numbers and locations. If the MKS-identified change point can be confirmed by the two other methods, the validated test. A confirmation is accepted if MKS identified change point is validated within two weeks. So based on the 36 MKS- identified change points. The MKS test reaches 41.7% with the PELT method and 47.2% with the regression method. The other two methods identified at least one change point in all the states, even if there is no change of direction. From the comparison, we found that the MKS test only detects sudden changes and avoids false-positive results. This test is a relatively conservative method.

SUMMARY

As the data sequence (online data) does not have a fixed length, the observations are received and processed sequentially over time. When each observation is received, a decision is made using hypothesis testing on whether the change has occurred or not based on the data which has been received so far. If no change is found, the next observation in the data sequence is processed. The change detector is restarted from the following observation after the change point is detected.

The study of Covid 19 cases using Mann-Kendall-Sneyer (MKS) is among the first to implement the MKS model to change point detection. The MKS model is characterized by several advantages. Firstly, it has high computational efficiency and is easy to implement, user can easily implement this model with statistical knowledge using Microsoft excel. Secondly, this model indicates the direction of change whereas a model such as PELT can only detect the change. However, due to conservative and agreement with another slower, sensitive method the MKS test is preferred for initial pattern identification and data pruning Using different software packages, the change points of the cryptocurrency prices were detected. Using social-economic events they were able to analyze why there was a crash or boom in the prices. Also, similar results were found when compared to change points detected from the greed and fear index

FUTURE WORKS

Conservative nature and moderate agreement with other slower and sensitive methods the MKS method is recommended only for initial pattern identification. So further research is to overcome this drawback. Also, with further modification and validation, the MKS method can be applied to other health data such as injuries, mortality etc.

The introduction of a topological data analysis tool can enhance change point detection of the process. softwares need to be developed to include this tool thereby increase the accuracy of the detection

REFERECES

[[1]Ku, LL., Huang, TC. Sequential monitoring of manufacturing processes: an application of grey forecasting models. *Int J Adv Manuf Technol* **27**, 543–546 (2006). https://doi.org/10.1007/s00170-004-2198-0

[2] ong Liu, Makoto Yamada, Nigel Collier, Masashi Sugiyama,Change-point detection in time-series data by relative density-ratio estimation,Neural Networks,

[3]Detection and Characterization of Anomalies in Multivariate Time Series, Haibin Cheng and Pang-Ning Tan and Christopher Potter and Steven A. Klooster

[4] Chen, X., Wang, H., Lyu, W. *et al.* The Mann-Kendall-Sneyers test to identify the change points of COVID-19 time series in the United States. *BMC Med Res Methodol* **22**, 233 (2022). https://doi.org/10.1186/s12874-022-01714-6

[5] Zhang, Han, Qian Zhou, and Dr Pablo Roldan. "Change Point Detection methods applied to Financial Time Series Research Document."

[6] Ross, Gordon J., Dimitris K. Tasoulis, and Niall M. Adams. "Sequential monitoring of a Bernoulli sequence when the pre-change parameter is unknown." *Computational Statistics* 28.2 (2013): 463-479.

[7] Polunchenko, Aleksey S., and Alexander G. Tartakovsky. "State-of-the-art in sequential change-point detection." *Methodology and computing in applied probability* 14.3 (2012): 649-684.

[8] Ross, . G. J. (2015). Parametric and Nonparametric Sequential Change Detection in R: The cpm Package. *Journal of Statistical Software*, *66*(3), 1–20. https://doi.org/10.18637/jss.v066.i03

[9] Josua G¨osmann, Christina Stoehr, Johannes Heiny and Holger Dette. “Sequential change point detection in high dimensional time series”

[10] By Louis Gordon and Moshe Pollak “An Efficient Sequential Nonparametric Scheme For Detecting A Change Of Distribution”

[11] Selçuk Ta¸scıo˘glu ,* , Memduh Köse and Gökhan Soysal “Sequential Transient Detection for RF Fingerprinting”

[12] Yoshinobu Kawahara; Masashi Sugiyama “Change-Point Detection in Time-Series Data by Direct Density-Ratio Estimation”

[13] Aleksey S. Polunchenko; Alexander G. Tartakovsky “State-of-the-Art in Sequential Change-Point Detection”

[14] Odalric-Ambrym Maillard. Sequential change-point detection: Laplace concentration of scan statistics and non-asymptotic delay bounds. Algorithmic Learning Theory, 2019, Chicago, United States. pp.1 - 23. ffhal-02351665f

[15] Michalis K. Titsias; Jakub Sygnowski; Yutian Chen; Sequential Changepoint Detection in Neural Networks with Checkpoints