SEQUENTIAL CHANGE POINT DETECTION

Kushal Bhyregowda
Shubhvaratha Dutta
Parth Mahendrabhai patel
Venkata Surya Teja Sistla

Arizona State University

Agenda

- Introduction
- Methodology
- Application 1
- Application 2
- Summary and conclusions
- Future works

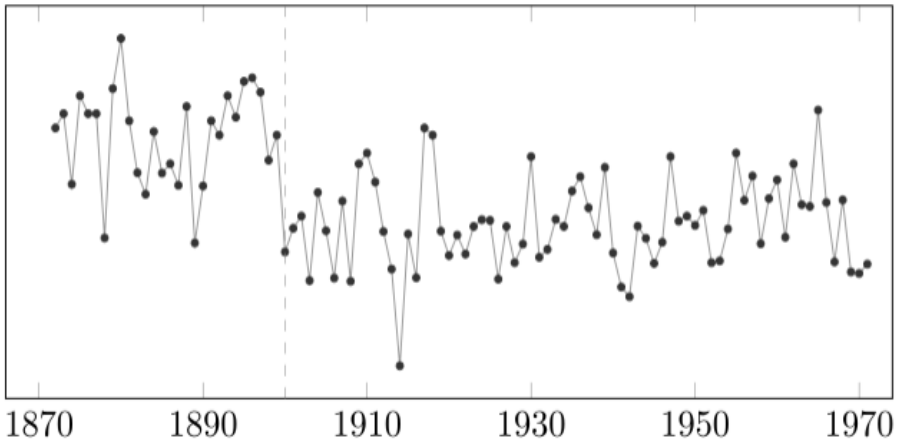
INTRODUCTION

Change Point Detection

Identify the times or points when the distribution of the process or a time series changes abruptly

Sequential change point Detection

Identifying the point or time of change for sequential samples or data



Yearly volume of nile river at Aswan

What is the goal?

Has the change occurred?

If yes,where it occurred?

What is the difference between the pre and post change data?

How certain are we of the change point location?

How many changes have occurred?

Why has there been a change?

MOTIVATION

- In real world applications most of the datasets are sequential, in order to detect changes as early as possible, we need have efficient model.

Methodology - Sequential CPD

➢ **Batch change detection (Phase I)** :-

Consider fixed length sequence of 'n' observations x1, . . . , xn which may or may not contain a change point. If a change point exists at some time $\boldsymbol{\tau}$, then the observations have a distribution F_0 prior to this point, and a distribution F_1 afterwards, where F0 ≠ F1.

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

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

Determine two-sample test statistic Dn. Since x_k = observation after which sequence changed , not known we can find -

$$D_n = \max_{k=2,\ldots,n-1} D_{k,n} = \max_{k=2,\ldots,n-1} \left| \frac{\widetilde{D_{k,n}} - \mu_{\widetilde{Dk,n}}}{\sigma_{\widetilde{Dk,n}}} \right|$$

- if $D_n > h_n$ ➔ Reject Null Hypothesis (means change in sequence detected)
- If $D_n < h_n$ ➔ Fail to Reject Null Hypothesis (means no change in sequence)
- h_n= threshold value chosen to bound the Type 1 error rate.

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

$$\hat{\boldsymbol{\tau}} = \arg\max_k D_{k,n}.$$

➢ Sequential change detection (Phase II)

Let x_t denote the t^{th} observation that has been received, where $t \in \{1, 2, \ldots\}$.

Whenever a new observation x_t is received, the CPM approach treats x1, . . . , xt as being a fixed length sequence and computes Dt using the previous batch methodology, where we are using the notation Dt rather than D_n to highlight the sequential nature of the procedure. A change is then flagged if $D_t > h_t$ for some appropriately chosen threshold. If no change is detected, the next observation xt+1 is received, then D_{t+1} is computed and compared to h_{t+1}, and so on.

The procedure therefore consists of a repeated sequence of hypothesis tests. Whenever a change point is detected, the change detector is simply restarted from the following observation in the sequence.

In the sequential setting, h_t is chosen so that the probability of incurring a Type 1 error is constant over time, so that under the null hypothesis of no change:

$$P(D_1 > h_1) = \alpha$$
$$P(D_t > ht \,|D_{t-1} \leq h_{t-1}, \ldots, D1 \leq h1) = \alpha,\ t > 1$$

Goals –

- Low False Alarm Rate
- Quick detection of changepoint with lowest ARL1 values
- Highly computationally efficient

- **Different CPM Packages available –**

  - *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.

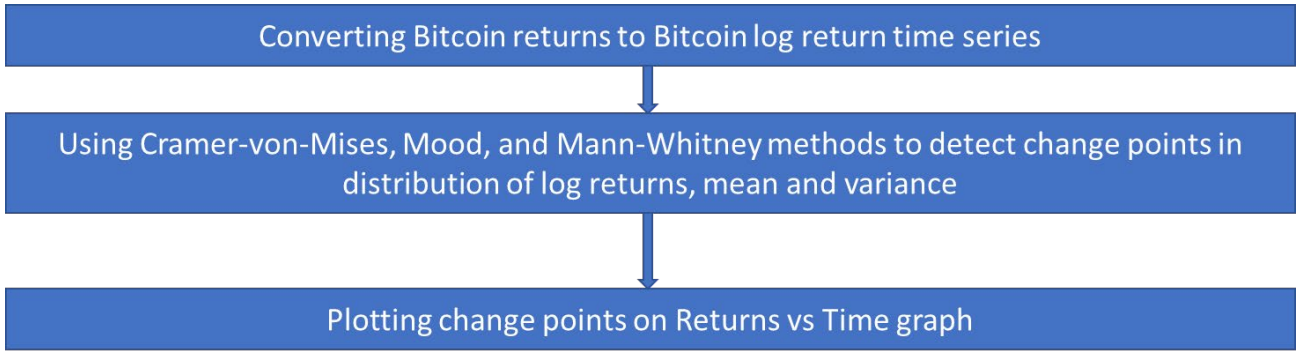
  - *The Lepage, Kolmogorov-Smirnov, and Cramer-von-Mises statistics* for detecting more general distributional changes where again no assumptions are made about the sequence distribution.

Application 1: Sequential CPD methods applied to Financial Time Series

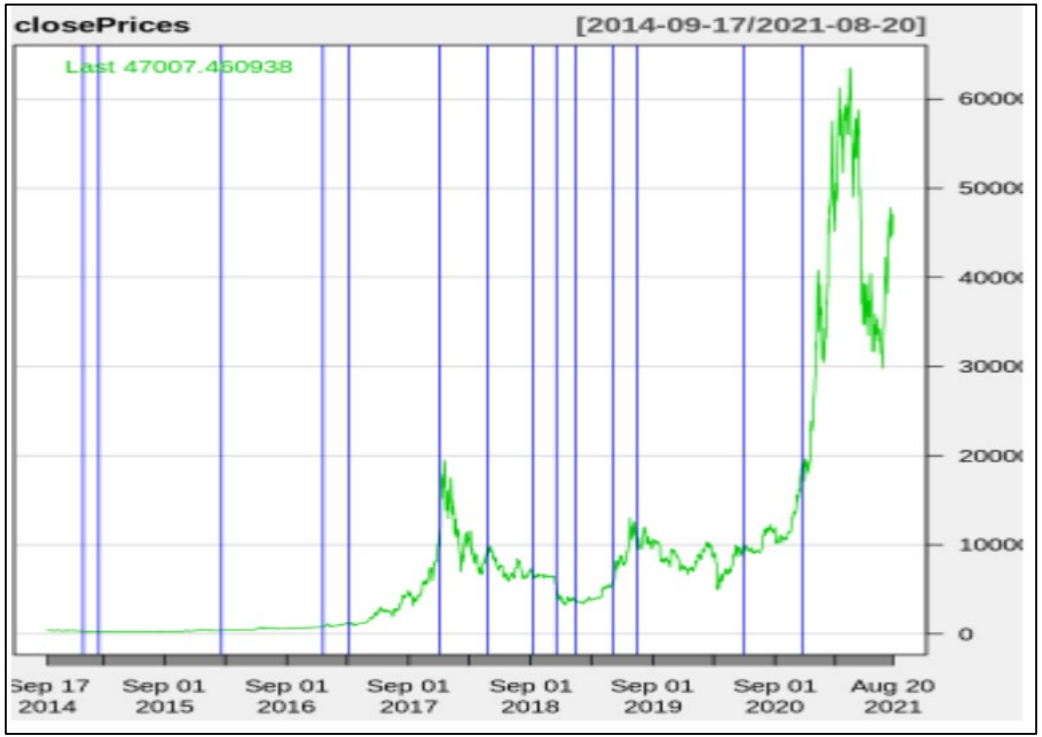
• Can be used to analyze the distributional changes and detects the change points of univariate times series of bitcoin prices in real-time and multivariate time series considering other crypto currencies from 2014 to 2021.

•Univariate time series of bitcoin

1. Batch Detection-fixed length
2. Sequential Detection-no fixed length can take new values

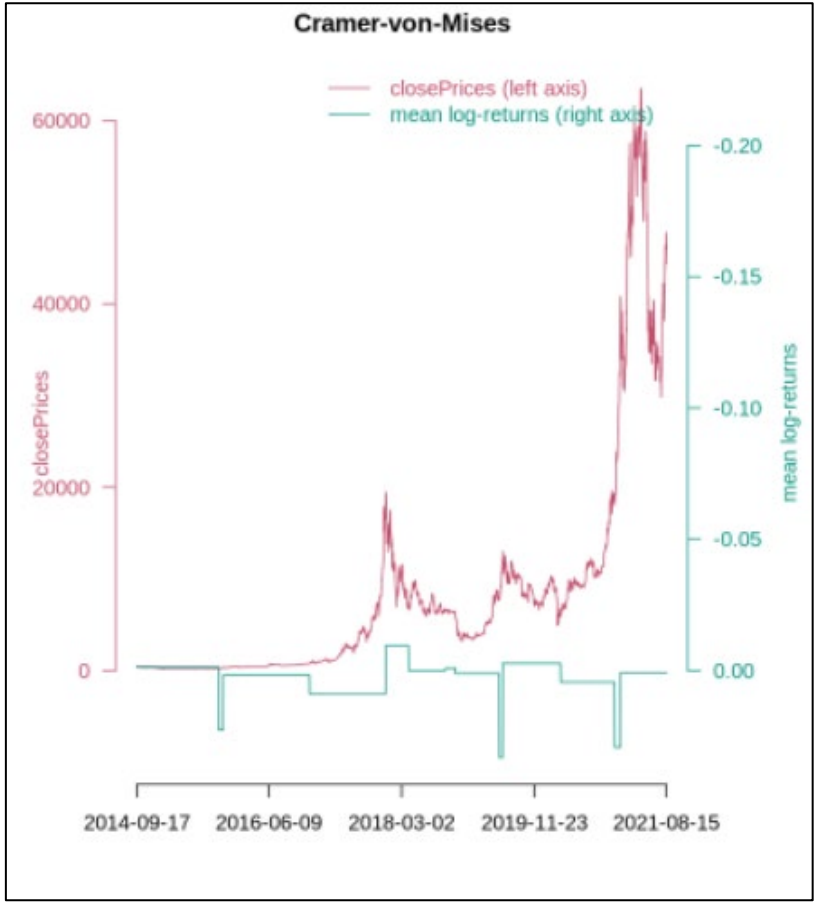


Sequential detection of change points in the log return distribution, log mean and log variance using Cramer-von-Mises statistic

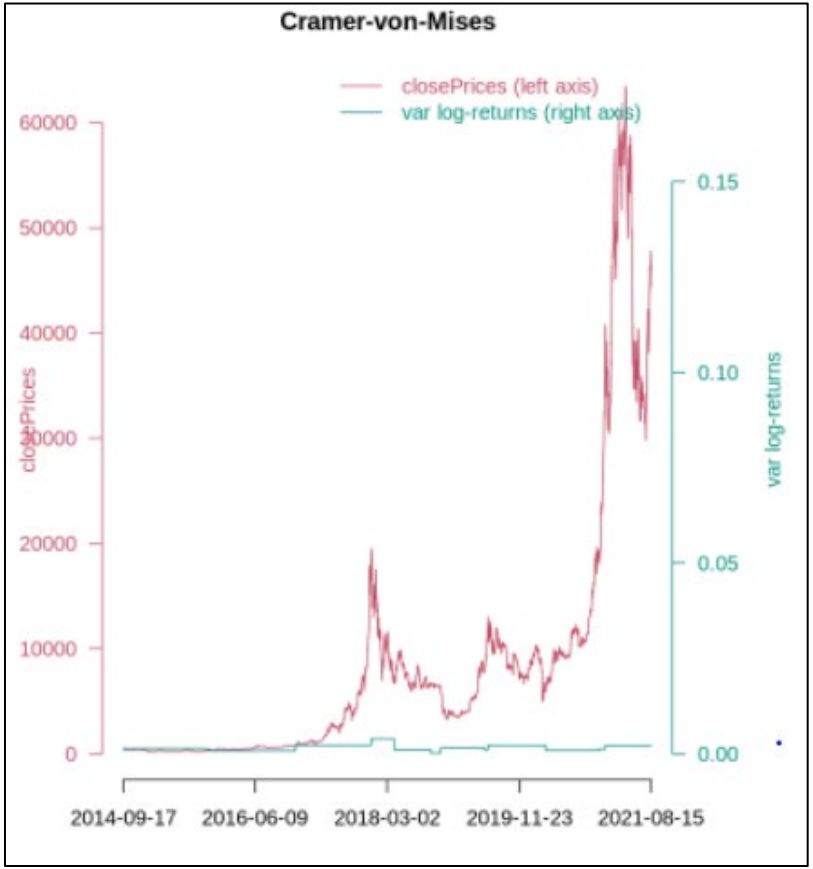


ARL0 = 1000 is kept constant through the times series

Mean change point detection



Variance change point detection



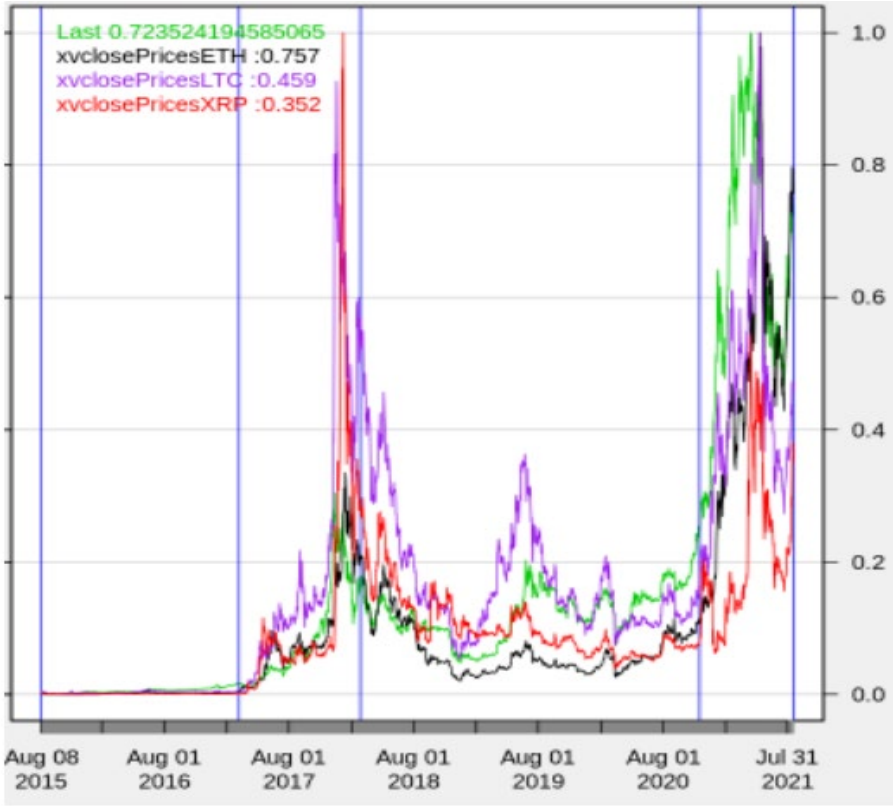
Multivariate Time Series of Cryptocurrencies

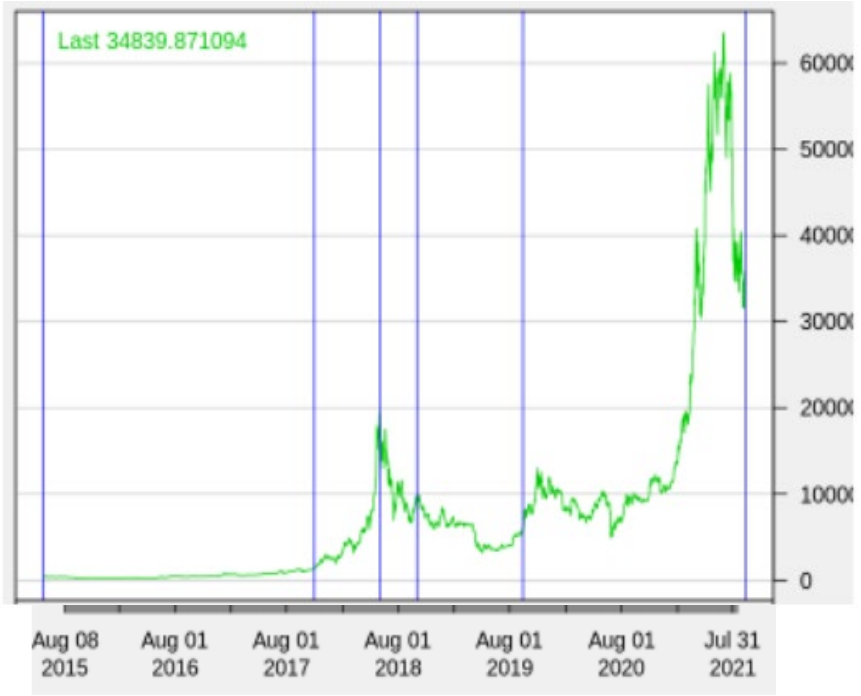
•The entire crypto currency market can be analyzed by building multivariate time series for multiple cryptocurrencies. The paper considered Bitcoin, Ethereum, Litecoin, and Ripple to detect change points in crypto market.

•The method converts the returns of the crypto currencies to log returns time series then these are analyzed together to detect change points in the crypto market.

•The paper utilizes e.divisive method on ecp R package to detect shift by comparing the permutations of time series before and after a point. Shift is detected when great difference is observed between these points.

Comparison of change points in multivariate and univariate time series plots generated by ECP.





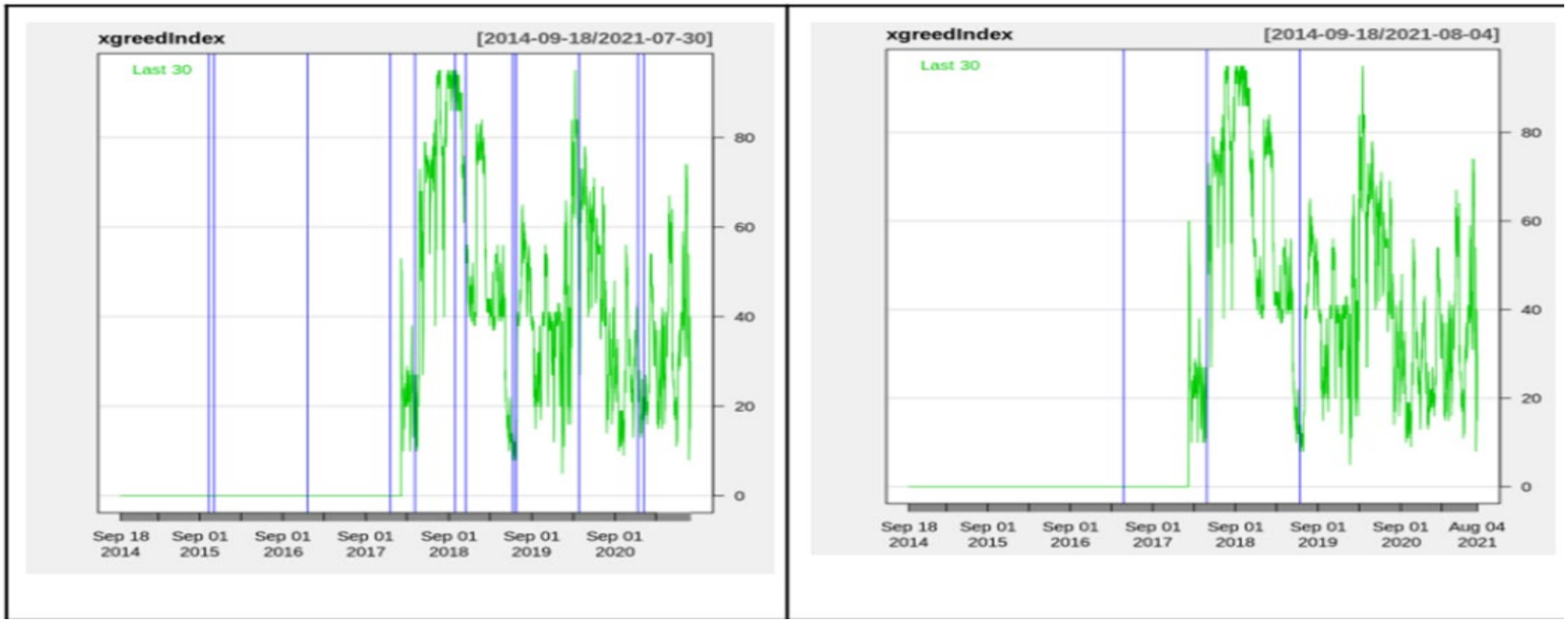
Socioeconomic events behind bitcoin change points:-

1. Jan 2018: Coincheck was hacked and 518 million USD of NEM stolen lead to suspension of coincheck trading.
2. March 2018: Compromised Finance API keys were used to execute irregular trades
3. Facebook, Google, and Twitter banned advertisements for initial coin offerings (ICO) and token sales.

Fear and Greed index:-

- It is a number based on analysis of emotions and sentiments from different sources on regular basis.
- Low value represents “Fear” and high value represents “Greed”.
- Fear - Sudden rise in volatility (Can be a chance to buy)
- Greed-High buying volumes in a positive market (Market is due for correction)

Change points in log-returns of the closing prices of Bitcoin in the context of the Fear and Greed Index for ARL0=1000 and ARL0=4000 respectively.



Application 2 – Sequential CPD on COVID-19 time series in the United States

- ***MKST (Mann-Kendall-Sneyers) Test :-*** a sequential extension of the MK test which is useful in detecting monotonic changes in the trends and the corresponding change points, making it useful for disease tracking and monitoring in the mid to long term.

- **Dataset**:-

  Weekly New COVID Case Data (For Each State) : $X = \{x1, x2, x3 \ldots xN\}$,
  where N = total number of weeks under observation = 45

- ***Method :-***

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

$$S_k = \sum_{i=1}^{k} m_i \ , \quad where \ (k = 1, 2, 3, \ldots, N)$$

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

  Backward Sequence, U_b *= Obtained by reversing time series data X and deriving Xr. We derived the intermediate sequence* U_{fr} *by applying Eq. (4) 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.*

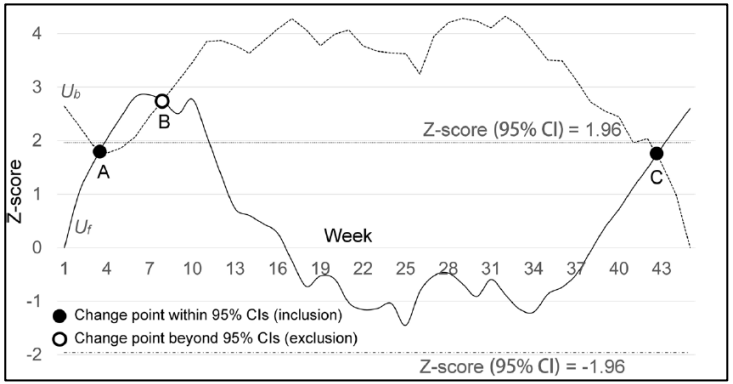


Fig. 1 MKS test of new weekly cases in Virginia with the forward sequence (solid line) and the backward sequence (dashed line). The black dot is the identified change point, and the white dot is the excluded change point.

Application 2 – Test Observations

By applying the MKS test to weekly new COVID-19 cases in 50 states, it was identified that

- 30 states found having at least one change point within the 95% CIs.
- Among these, 25 states = single change point,
- 4 states(i.e., LA, OH, VA, and WA) = two change points and
- 1 state (i.e., GA) = 3 change point

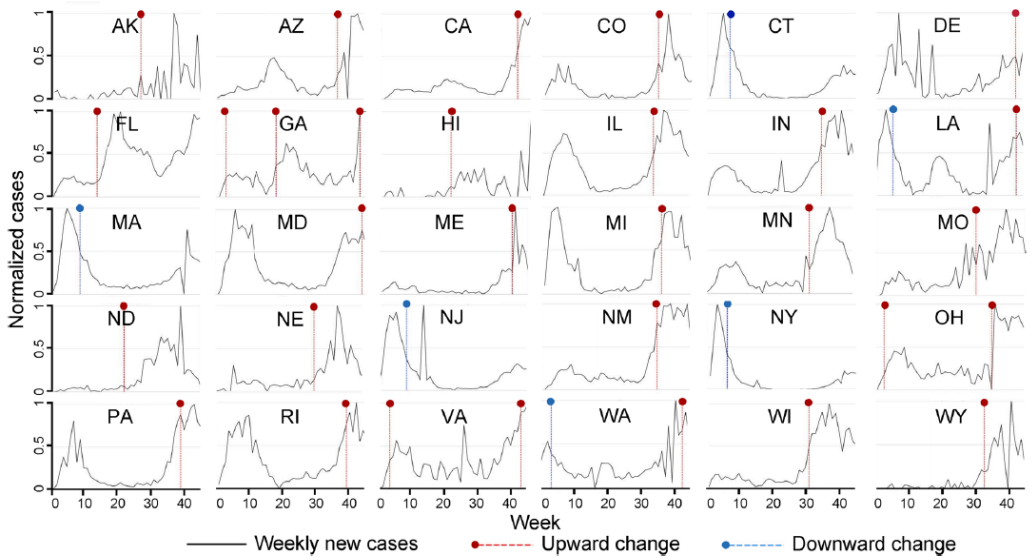


Fig. 4 States with at least one change point identified. The horizontal axis is the week; the vertical axis is the weekly new cases normalized to 0–100% with respect to the maximum weekly new cases in each state

Application 2 - Disease Development Stage

Based on the 3 clusters of chronologically ordered CP following 3 disease development stage observed. Based on the three development stages, we then mapped out the emergence of the change point for each state, as shown in Fig. 3. –

1. Weeks 1–10 (March 23 - May 31, 2020),
2. Weeks 11–30 (June 1 - November 19, 2020),
3. Weeks 31–45 (November 19, 2020 - January 31, 2021).

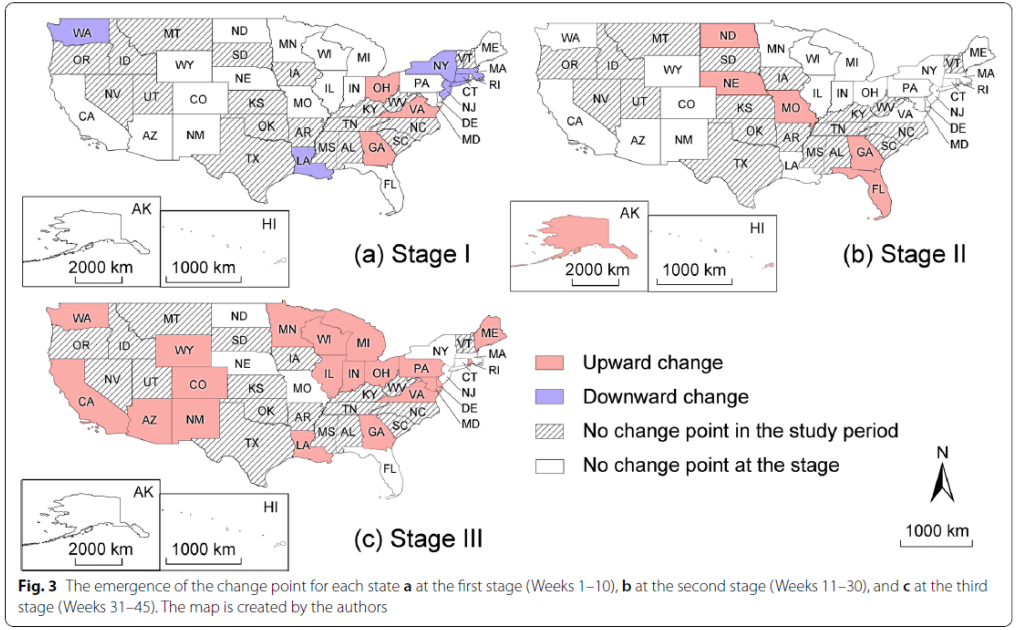


Fig. 3 The emergence of the change point for each state **a** at the first stage (Weeks 1–10), **b** at the second stage (Weeks 11–30), and **c** at the third stage (Weeks 31–45). The map is created by the authors

SUMMARY AND CONCLUSION

- The sequential change detection model uses hypothesis testing for determining whether the change as occurred or not.
- CPM performs good for small degrees of parameter misspecification
- The MKS test is characterised by high efficiency and easy implementation
- The method can detect change of direction
- The MKS is non parametric models so the model can be applied to time series data where the data is not normal distributed or has extreme variability

FUTUREWORK

- The change point model has good performance when change occurs relatively late but performs poorly when change occurs early in the stream.
- With further modification and validation the MKS test model can be applied to other health data like injuries, disabilities and mortalities.

REFERENCES

[1]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.

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

[3]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.

[4]Chen, Xiang, et al. "The Mann-Kendall-Sneyers test to identify the change points of COVID-19 time series in the United States." *BMC Medical Research Methodology* 22.1 (2022): 1-9.

[5] 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