An Analysis of a Circular Changepoint Model – A Covid-19 Case Study –

Matthew Scholes with Owen Li

University of Lancaster - STOR-i

September 17, 2022

Contents

1. Introducing Changepoint Models
   - Linear Models
   - Circular Model
2. Circular Model Performance
3. Covid-19 Case Study
   - Data Suitability
   - Segment Neighbourhood vs Circular
   - Potential Error

Contents

1. Introducing Changepoint Models
   - Linear Models
   - Circular Model
2. Circular Model Performance
3. Covid-19 Case Study
   - Data Suitability
   - Segment Neighbourhood vs Circular
   - Potential Error

Introducing Changepoint Models

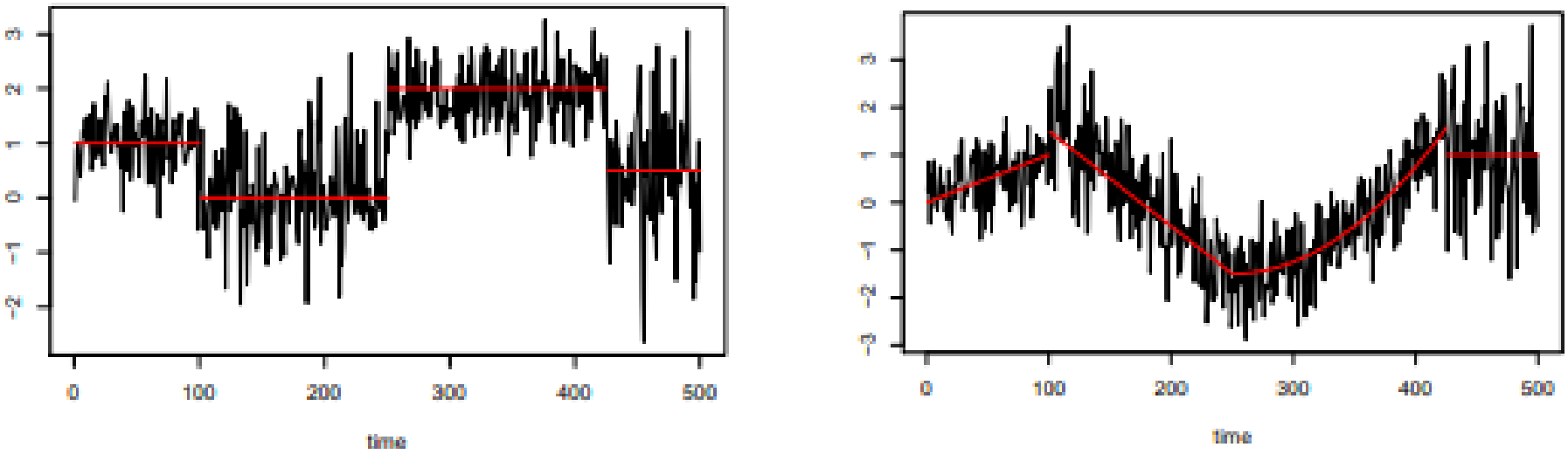


Figure: Graphs show examples of how parameters change given at changepoint locations: LHS - mean; RHS - trend (1).

- The aim of changepoints regression is to reduce the cost function, a common example is using the negative log likelihood (for normalised data).
- $\mathrm{L}(\mathcal{M}_\kappa) = -2\mathrm{log}\mathrm{max}\mathrm{L}(\theta_\kappa) + \mathrm{p}_\kappa\phi(\mathrm{n})$ (2).

Linear Models

- Binary Segmentation (BinSeg):
  - Computationally fast $\mathcal{O}(n)$ (3).
  - Only able to find local minimum.

Linear Models

- Binary Segmentation (BinSeg):
  - Computationally fast $\mathcal{O}(n)$ (3).
  - Only able to find local minimum.
- Segment Neighbourhood (SegN):
  - For max m changepoints at time τ.
  - Minimize $\sum_{i=0}^{m} \mathcal{C}(y_{(\tau_i+1):\tau_{i+1}})$.
  - Computationally slow $\mathcal{O}(n^2)(4)$.
  - Exact, will find global minimum.

Linear Models

- Binary Segmentation (BinSeg):
  - Computationally fast $\mathcal{O}(n)$ (3).
  - Only able to find local minimum.
- Segment Neighbourhood (SegN):
  - For max m changepoints at time τ.
  - Minimize $\sum_{i=0}^{m} \mathcal{C}(y_{(\tau_i+1):\tau_{i+1}})$.
  - Computationally slow $\mathcal{O}(n^2)(4)$.
  - Exact, will find global minimum.
- Pruned Exact Linear Time (PELT):
  - Removes values of τ that can never be considered a changepoint.
  - Computationally fast $\mathcal{O}(nlog(n))(5)$.
  - Exact, will find global minimum.

Circular Model

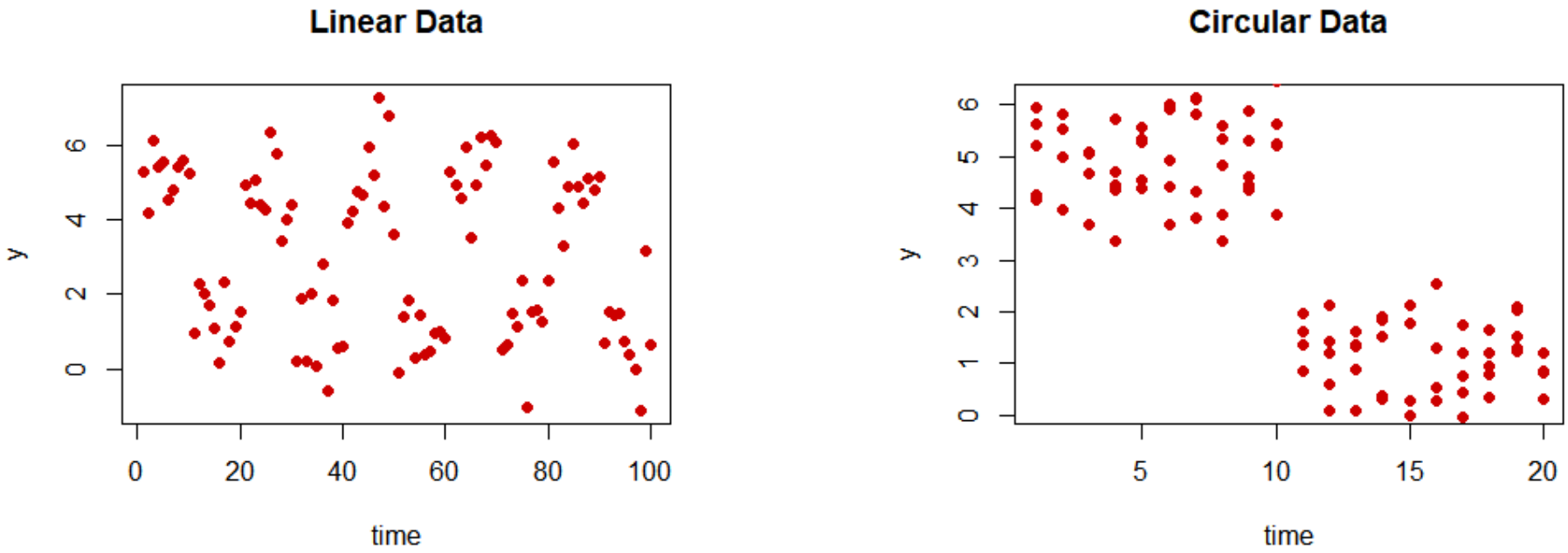


Figure: Normally distributed data; LHS initial linear form, RHS data wrapped around a period of 20.

- Instead of representing the data linearly, wrap the time axis on itself so every time point has multiple data entries.
- Then minimize $\sum_{i=0}^{m} \mathcal{C}(\mathbf{x}_{(\tau_i+1):\tau_{i+1}})$ where $\mathbf{x}$ is a vector of points.

Contents

1. Introducing Changepoint Models
   - Linear Models
   - Circular Model
2. Circular Model Performance
3. Covid-19 Case Study
   - Data Suitability
   - Segment Neighbourhood vs Circular
   - Potential Error

Circular Model Performance

- generated normalised, periodic, data with a constant variance of $\sigma = 1$ and a varying mean (μ).
- Success is defined only if both changepoints are located at the correct place.
- Circular method can detect changepoints that are more 'subtle'.
- Finds both changepoints with a success of 99% the up to and including $\Delta\mu = 2$.

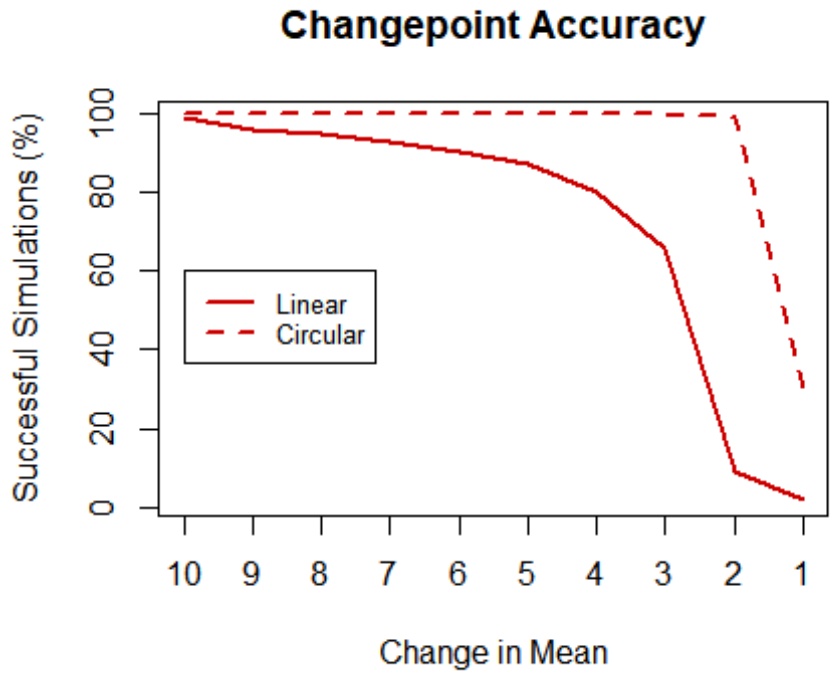


Figure: A comparison of linear and circular methods for detecting a change in mean for a time series.

Contents

1. Introducing Changepoint Models
   - Linear Models
   - Circular Model
2. Circular Model Performance
3. Covid-19 Case Study
   - Data Suitability
   - Segment Neighbourhood vs Circular
   - Potential Error

Background

- Data of positive Covid-19 cases recorded in England from $30/01/20$.
- Due to multiple social factors, less tests are processed at weekends.
- This causes a dip in positive cases over the weekend which leads to periodicity (6).
- Aim: Identify this periodicity and see if circular method can identify changepoints hidden to traditional Segment Neighbourhood search.

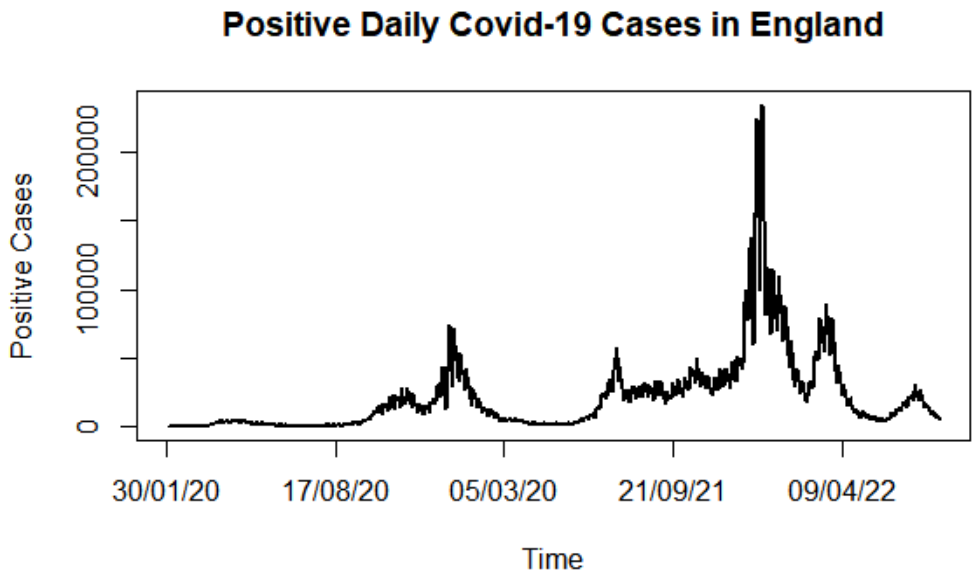
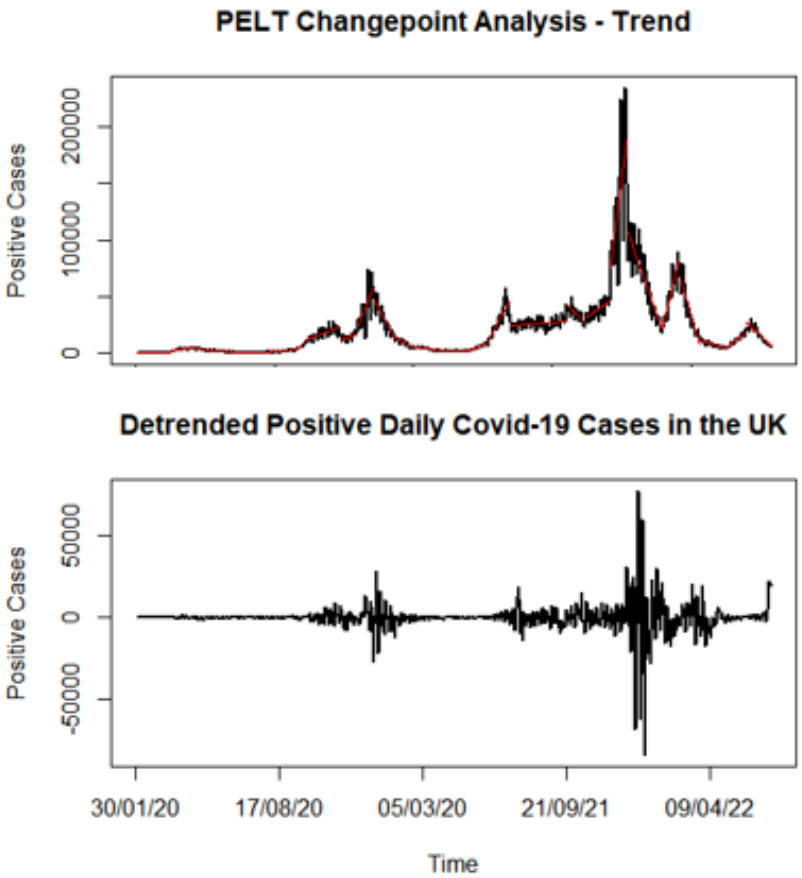
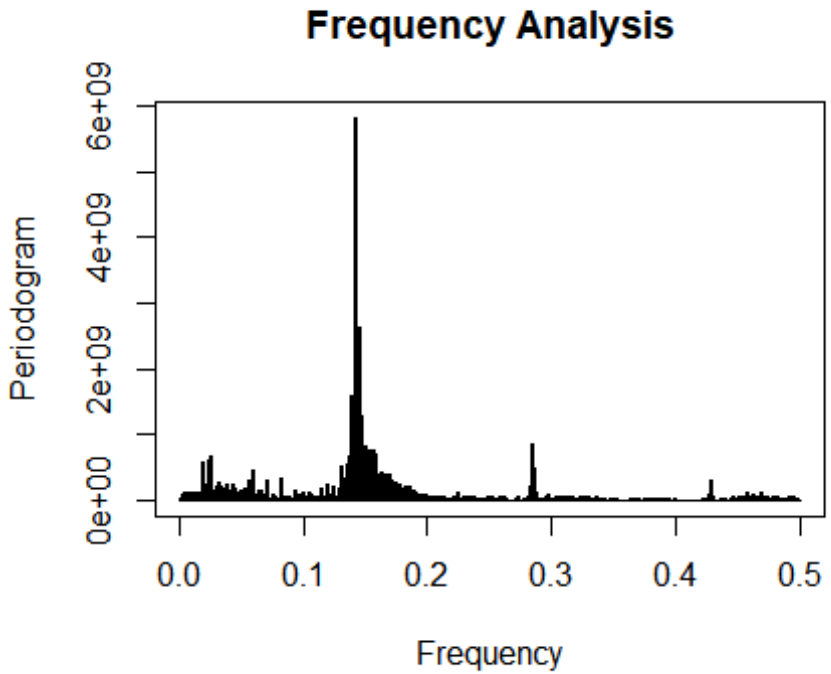


Figure: Number of positive cases in England from $30/01/20$.

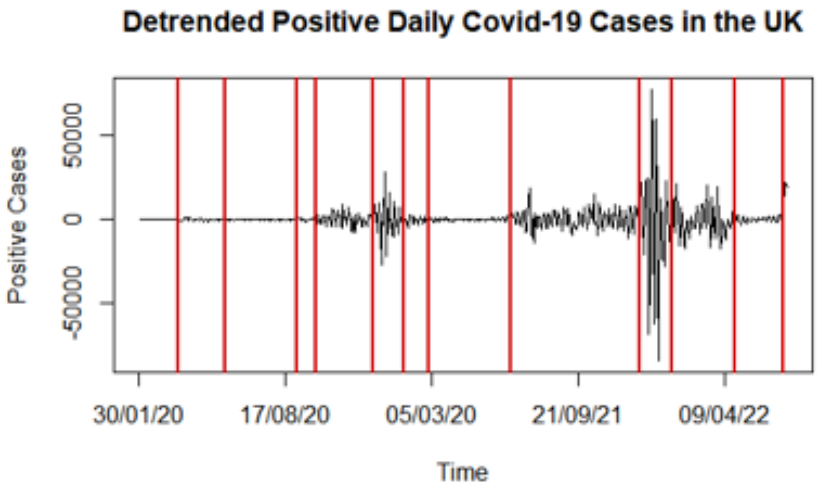
Checking Periodicity

- Circular Method requires period of data as an input.
- Trends are determined using PELT.
- Conducting a FFT on the cleaned data a peak frequency of 0.143 was found.
- This corresponds to a period of 6.98 days.





Segment Neighbourhood vs Circular



Mon	Tue	Wed	Thu	Fri	Sat	Sun
2	2	1	3	5	3	7

Figure: LHS, Results of SegN and RHS Results of Circular Method.

- Segment Neighbourhood picks up similar changepoints to PELT, failing to register the periodicity.
- 43% of cpts are found on Friday and Sunday, but why is this inconsistent?

Areas of Error

- The times where the method does not locate Friday or Sunday as a changepoint are boxed in red.
- The majority of these are areas of high variance, which correspond to times of numerous positive cases in raw data.
- Could the increased variance account for erroneous changepoint location?
- What about the first box? This is low variance.

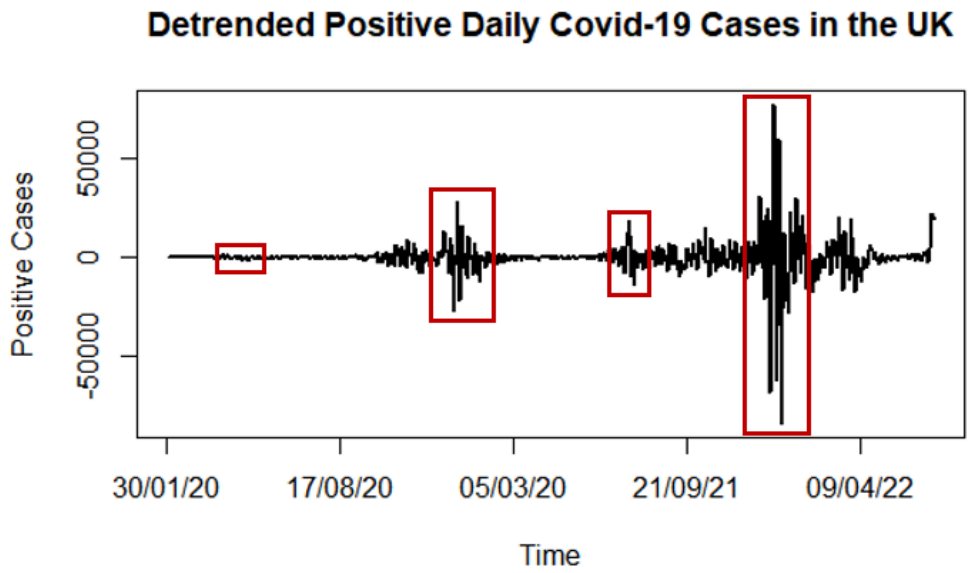
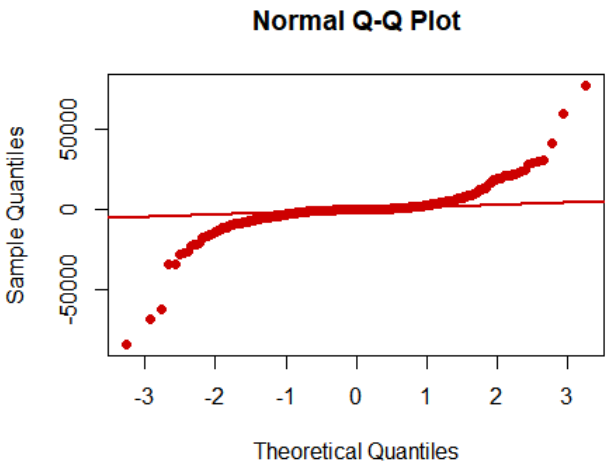


Figure: Boxes areas are locations where neither Friday or Saturday are located as changepoints.

Areas of Error

- Requirement of the method is that the data is normalised.
- The QQ plot shows that the data is too heavy at the tails to be completely normalised.
  - Areas of low variance are more normalised than the boxed areas.
  - This does not mean that the high variance areas cannot be analysed, but a different cost function would be needed.
- Could political reasoning explain the error from low variance?



Coronavirus: Matt Hancock sets aim of 100,000 tests a day by end of April

2 April 2020 · Comments

Coronavirus: Target reached as UK tests pass 100,000 a day

1 May 2020 · Comments

Coronavirus: Testing falls short of 100,000 daily target

3 May 2020

Figure: LHS: QQ plot of cleaned data, RHS headlines from BBC regarding 100k testing target.

Conclusion

- Circular method was developed to locate changepoint locations, normally hidden from linear methods.
- It achieves this as can detect smaller changes.
- When used with a Covid-19 case study, weekly points of change were located as expected.
- Revealed further information about the time series, prompting areas for further research.
- A very powerful tool, when used in isolation and in conjunction with other linear changepoint models.

Thank you for listening

Are there any questions?

References

[1] I. A. Eckley *et al.*, "Analysis of changepoint models," *Bayesian time series models*, vol. 205, p. 224, 2011.

[2] Hinkley *et al.*, "Inference about the change-point in a sequence of random variables," 1970.

[3] A. J. Scott and M. Knott, "A cluster analysis method for grouping means in the analysis of variance," *Biometrics*, pp. 507–512, 1974.

[4] J. V. Braun and H.-G. Muller, "Statistical methods for dna sequence segmentation," *Statistical Science*, pp. 142–162, 1998.

[5] R. Killick, P. Fearnhead, and I. A. Eckley, "Optimal detection of changepoints with a linear computational cost," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1590–1598, 2012.

[6] Q. Bukhari *et al.*, "Periodic oscillations in daily reported infections and deaths for coronavirus disease 2019," *JAMA network open*, vol. 3, no. 8, pp. e2 017 521–e2 017 521, 2020.