**National Research University Higher School of Economics**

**Faculty of Computer Science**

**HSE and University of London Double Degree Programme in Data Science and Business Analytics**

**BACHELOR'S THESIS**

**Software Project**

**Detecting Structural Breaks Using Time-Series Embeddings**

**Prepared by the student of Group 181, Year 4,**

**Mikhail Ovyan**

**Thesis Supervisor:**

**< Academic Degree>, <Academic Title >, <Full Name ()>**

**Advisor:**

**< Academic Degree >, < Academic Title >, <Full Name ()>**

**Moscow**

**2022**

**Table of contents**

**Introduction**

Traditionally, in the study of time series there is an assumption that the statistical properties of the observed are constant in time or change slowly. However, working with the real-life datasets this property is not always satisfied. Therefore, for many practical purposes, the detection of an abrupt change in the properties of the observed series occurring at an unknown point in time in advance is an important task.

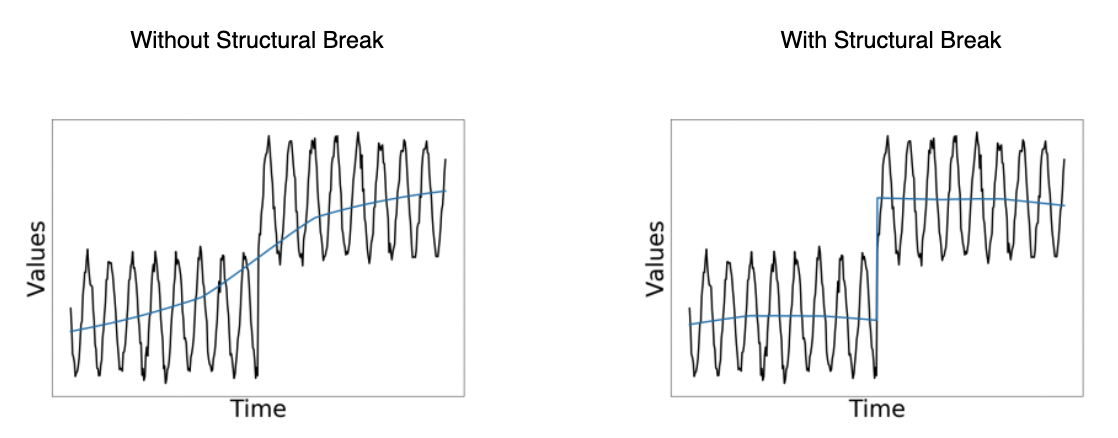


Figure 1. Analyzing a trend line in the time-series with a break

Moments of abrupt change in the behavior of a time series are often cause for alarm as they may signal a significant alteration to the data generating process. Detecting such change points is therefore of high importance for data analysts and engineers. Identifying structural change is a crucial step in analysis of time series. The longer the time span, the higher the likelihood that the model parameters have changed because of major disruptive events, such as the 2007–2008 financial crisis and the 2020 COVID–19 outbreak [1]. Models that previously worked well become imprecise and misleading in these cases. The identification of periods with different dynamics is a very important task. This task is called segmentation.

There are generally two types of such segmentation: a priori and a posteriori. A priori segmentation is the task of predicting the exact moment of discontinuity. A posteriori segmentation means the task of dividing the available time series into several segments with different dynamics. In this work I will focus on the posteriori segmentation. Detecting the existence of breaks and dating them is therefore necessary not only for estimation purposes but also for understanding drivers of change and their effect on relationships.

The times in which the parameters change is called “change points” in the statistics literature and “structural breaks” in economics. As both terms are synonymous to each other, in the paper I will use the latter term or just “breaks”.

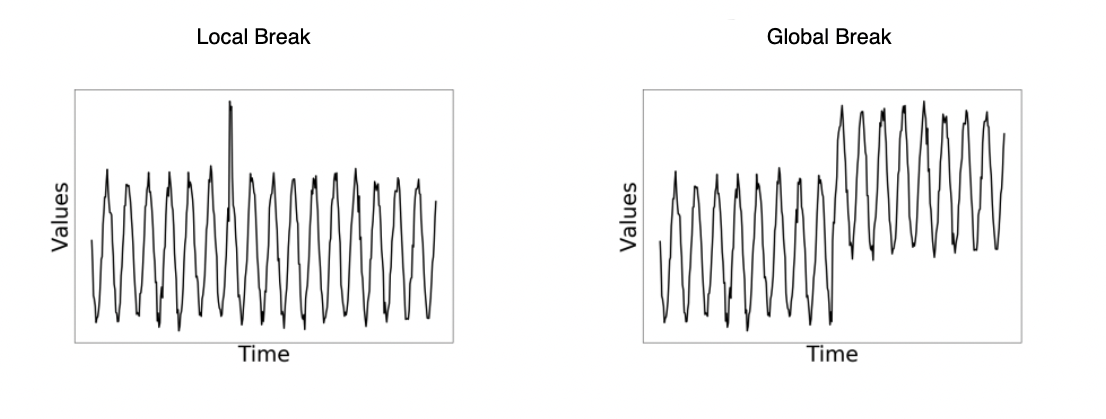


Figure 2. Local and Global structural breaks

There are two general types of structural breaks: a local break and a global break, which are demonstrated on the Figure 2. In this paper, I will focus on the global type of the structural breaks because of the features of the method I will propose later in this work.

The present paper dived into five sections. The first section outlines the overview

of a structural break, the second section deals with the different approaches which are currently used to analyze time-series and to segment disruptive intervals, including tests for known breakpoints and unknown breakpoints. The third section focuses on the proposed method of using vector embeddings to solve the problem of identifying structural breaks. The final sections describe the conducted experiments on the different datasets, including both real-world and synthetic data, and compare different methods using the appropriate metrics.

**Literature review**

In statistics and econometrics literature there is an extensive amount of work on testing for structural breaks. Earlier works such as Chow (1960) tests for parameter stability under a known break date using a F-statistic, and Quandt (1958, 1960) suggests using a maximum F-statistic over all values of the potential break date when the break date is unknown. Andrews (1993) studies the properties of such tests and derives the asymptotic distribution of the test statistic. Brown, Durbin, and Evans (1978) provide a test on the stability of parameters by considering partial sums of the standardized forecast errors of rolling regressions, referred as the CUSUM test.

### **The Chow Test**

The Chow (1960) test was one of the first tests which set the foundation for structural break testing. It is built on the theory that if parameters are constant then out-of-sample forecasts should be unbiased. It tests the null hypothesis that there is no structural break against the alternative that there is a known structural break at time Tb. The test considers a linear model split into samples at a predetermined break point such that:

**Yt = x′tβ1 + ut, for t ≤ Tb**

and

**Yt = x′tβ2 + ut, for t > Tb**

The test estimates coefficients for each period and uses the out-of-sample forecast errors to compute an F-test comparing the stability of the estimated coefficients across the two periods. One key issue with the Chow test is that the break point must be predetermined prior to implementing the test. Furthermore, the break point must be exogenous, or the standard distribution of the statistic is not valid.

### **The Quandt Likelihood Ratio Test**

The Quandt Likelihood Ratio (QLR) (1960) test builds on the Chow test and attempts to eliminate the need for picking a break point by computing the Chow test at all possible break points. The largest Chow test statistic across the grid of all potential break points is chosen as the Quandt statistic as it indicates the most likely break point.

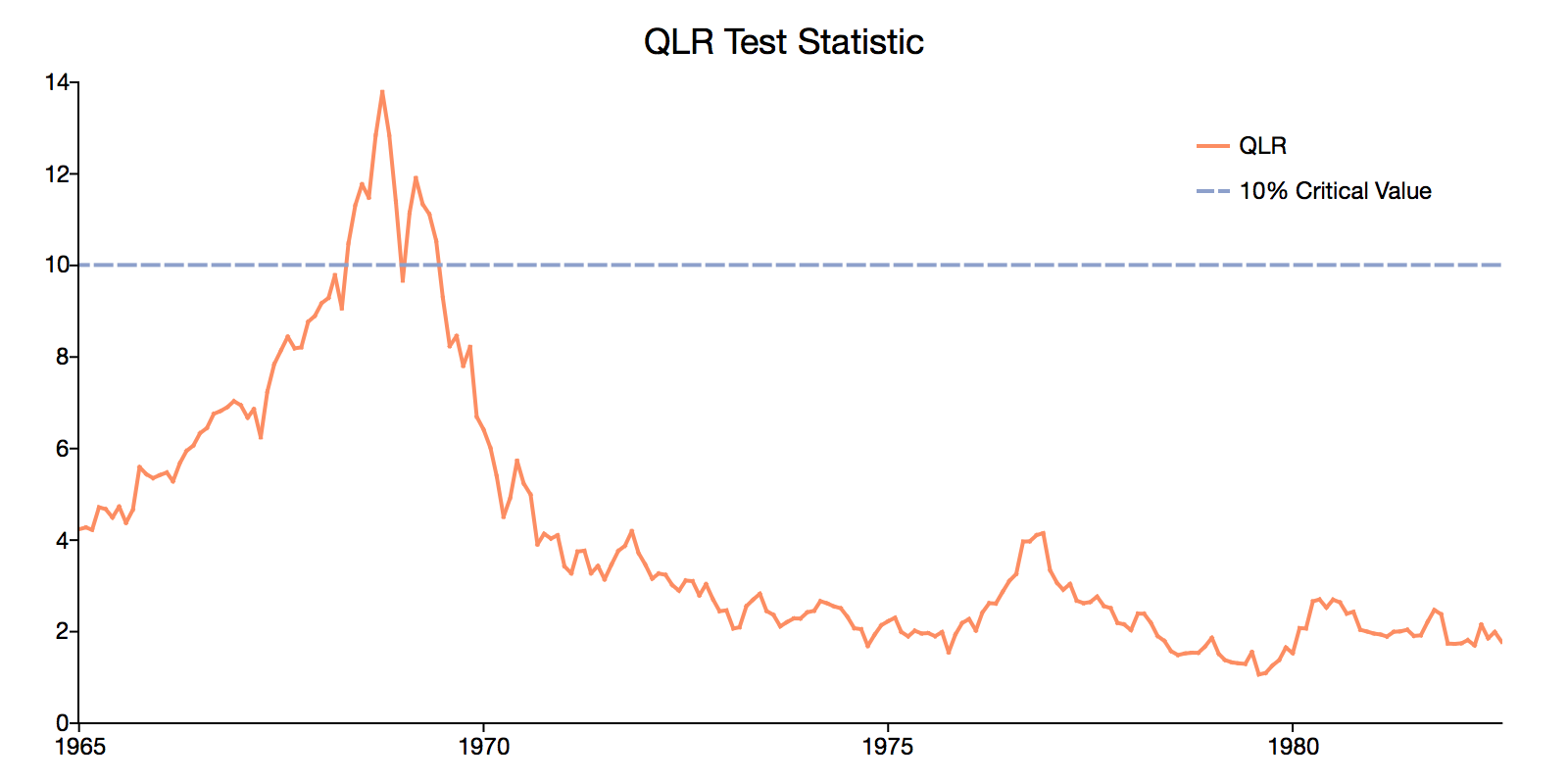


Figure 8. QLR test

This test was widely unusable because the limiting distribution of the test statistic under the assumption of an unknown break point was not known. However, the test became statistically relevant when Andrews and Ploberg (1994), developed an applicable distribution for the test-statistic for cases such as the Quandt test.

### **The CUSUM Test**

In their 1975 paper Brown, Durban, and Evans propose the CUSUM test of the null hypothesis of parameter stability. The CUSUM test for instability is appropriate for testing for parameter instability in the intercept term. It is best described as a test for instability of the variance of post-regression residuals.

The CUSUM test is based on the recursive least square estimation of the model:

**Yt = x′t****βt + ut, for all k+1** **≤ t ≤ T**

This yields a set of estimates for β k, β k+1, …, β T. The CUSUM test statistic is computed from the one-step-ahead residuals of the recursive least squares model. It is based on the intuition that if β changes from one period to the next then the one-step-ahead forecast will not be accurate, and the forecast error will be greater than zero.

This means the greater the CUSUM test statistic, the greater the forecast error, and the greater the statistical evidence in favor of parameter instability.

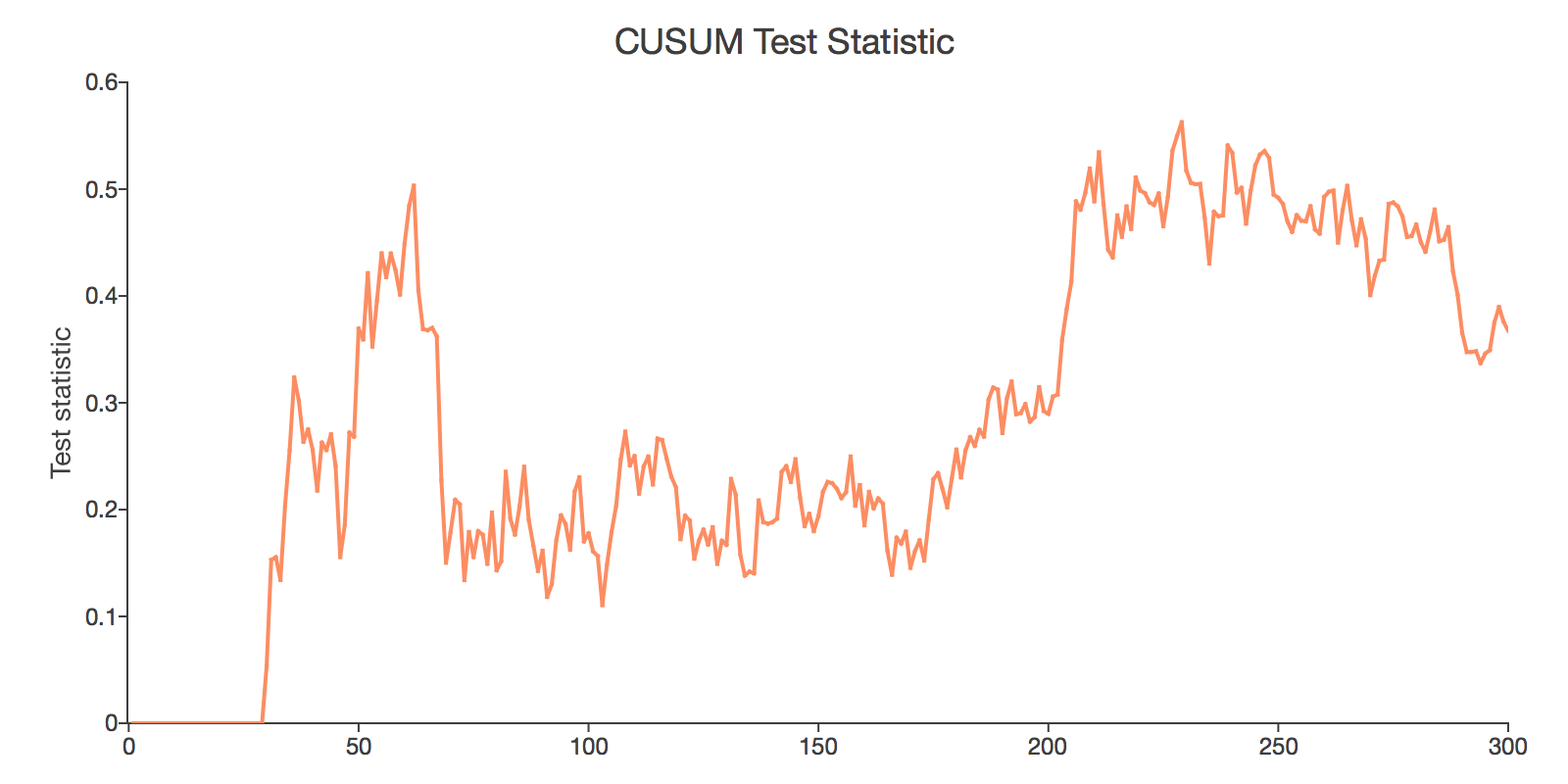


Figure 9. CUSUM Test Statistic value

### **Xtbreak**

In article “Testing and Estimating Structural Breaks in Time Series and Panel Data in Stata” (2021) [6] authors introduced a new community contributed command called xtbreak, which provides researchers with a complete toolbox for analyzing multiple structural breaks in time series and panel data. Xtbreak model has the following structure:

Text, letter

Description automatically generated

xtbreak can detect the existence of breaks, determine their number and location, and provide break date confidence intervals. The toolbox is based on asymptotically valid tests for the presence of breaks, a consistent break date estimator, and a break date confidence interval with correct asymptotic coverage. In fact, xtbreak includes three tests:

1. A test of no structural change against the alternative of a specific number of changes
2. A test the null hypothesis of no structural change against the alternative of an unknown number of structural changes, and
3. A test of the null of s changes against the alternative of s + 1 changes.

The package also includes an algorithm that employs the last test consecutively to estimate the true number of breaks. The tested break dates can be unknown or user-defined, as when researchers have additional information and wish to examine whether there was a break in a specific point in time. Once the presence of breaks has been tested and confirmed, xtbreak estimates the locations of the breaks and provides the associated confidence intervals.

However, we are interested in the way this model identifies all breakpoints and segments the intervals of time-series. This is done using the idea that if the model has the correct number of breaks and the correct points in time, then the SSR should be smaller than for a model with a larger or smaller number of breaks.

Diagram

Description automatically generated

Figure 10. Dynamic Programming algorithm, Bai and Perron (2003)

Thus, the only task is to minimize SSR, which is implemented in xtbreak using the dynamic programming algorithm from Bai and Perron (2003), Figure 10.

### **Matrix Profile**

<https://www.cs.ucr.edu/~eamonn/PID4481997_extend_Matrix%20Profile_I.pdf>

<https://www.cs.ucr.edu/~eamonn/Matrix_Profile_Tutorial_Part1.pdf>

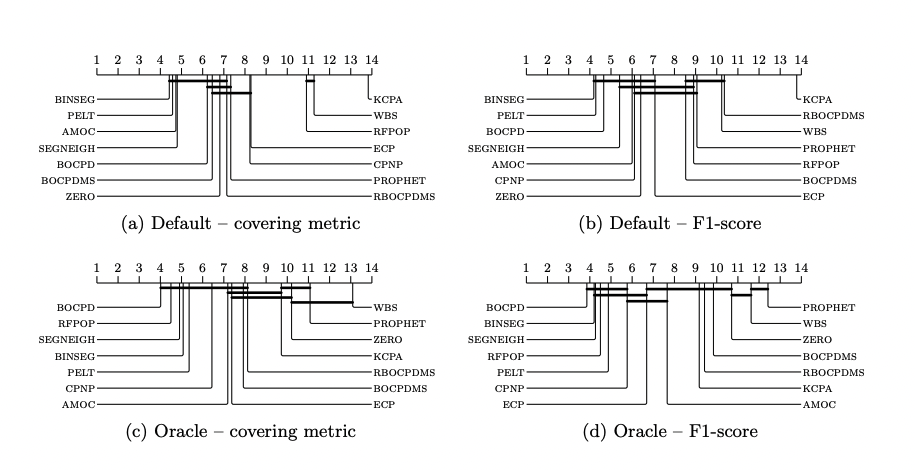
<https://www.cs.ucr.edu/~eamonn/Matrix_Profile_Tutorial_Part2.pdf>

<https://www.cs.ucr.edu/~eamonn/MatrixProfile.html>

<https://www.cs.ucr.edu/~eamonn/100_Time_Series_Data_Mining_Questions__with_Answers.pdf>

**Main Part**

Метрики:



Speed + memory



https://www.statmod.ru/\_diploma/2019m/13m\_02\_merzlyakov.pdf

An alternative view of evaluating CPD algorithms considers change point detection

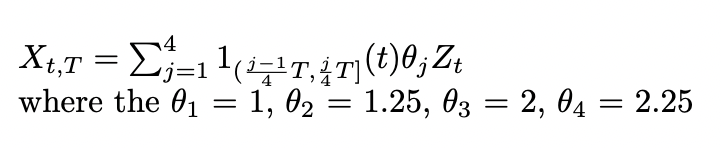
as a classification problem between the “change point” and “non-change point” classes (Killick et al., 2012; Aminikhanghahi and Cook, 2017). Because the number of change points is generally a small proportion of the number of observations in the series, common classification metrics such as the accuracy score will be highly skewed. It is more useful to express the effectiveness of an algorithm in terms of precision (the ratio of correctly detected change points over the number of detected change points) and recall (the ratio of correctly detected change points over the number of true change points). [[x]](https://arxiv.org/pdf/2003.06222.pdf)

**Data**

**Synthetic Datasets**

Testing different solutions to the problem of the structural breaks’ segmentation has been always a cornerstone in the papers. Given that there is no strict mathematical definition of the structural break, it is very difficult to find a labeled time-series featuring breaks and different shapes.

For that reason, it is usual to generate several various synthetic dataset, which would include breaks at the known points and would feature different shapes, trends and etc. For example, authors of the “Detection of multiple structural breaks in multivariate time series” [7] paper propose the following model:



and Z ​​is a Gaussian white noise process. It features 4 segments which differ by the MA parameters. These breaks are also can be modified with the different variance of Z between them.

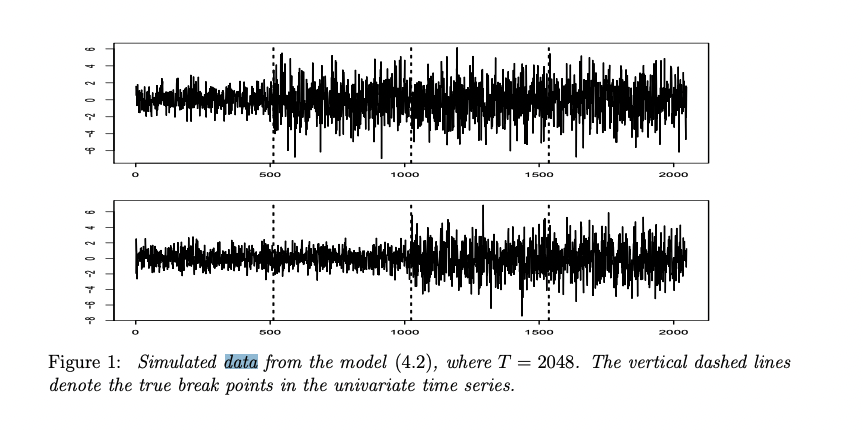
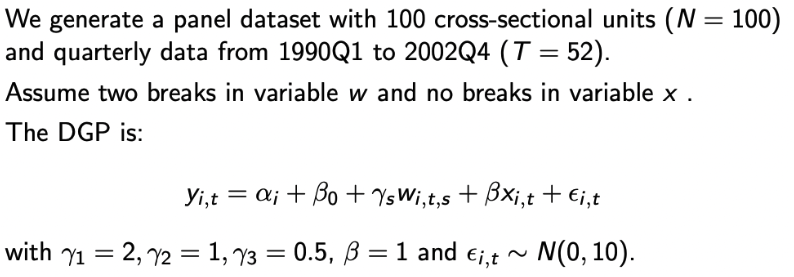


Figure 7. Synthetic Dataset with different theta parameters

Another synthetic dataset was proposed in the original “xtbreak” paper [6]. It has the following structure:



Thus, featuring 3 breakpoints or 4 different segments with the known location of the bounds.

**Turing Change Point Dataset**

In the paper “An Evaluation of Change Point Detection Algorithms” [5], the authors state, that while many algorithms for change point detection have been proposed, comparatively little attention has been paid to evaluating their performance on real-world time series. Algorithms are typically evaluated on simulated data and a small number of commonly used series with unreliable ground truth.

Clearly this does not provide sufficient insight into the comparative performance of these algorithms in the field. Therefore, instead of developing yet another change point detection method, they considered it vastly more important to properly evaluate existing algorithms on real-world data. To achieve this, they presented a dataset specifically designed for the evaluation of change point detection algorithms that consists of 37 time-series from various application domains and was manually annotated by volunteers.

All series are standardized to zero mean and unit variance to avoid issues with numerical precision arising for some methods on some of the time series.

Out of the 37 time-series I have decided to choose 4 the most distinctive. Those include the financial data, stock price, and even a scanline of an image.

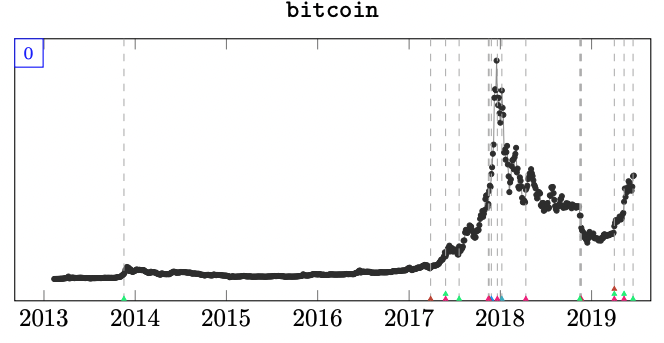
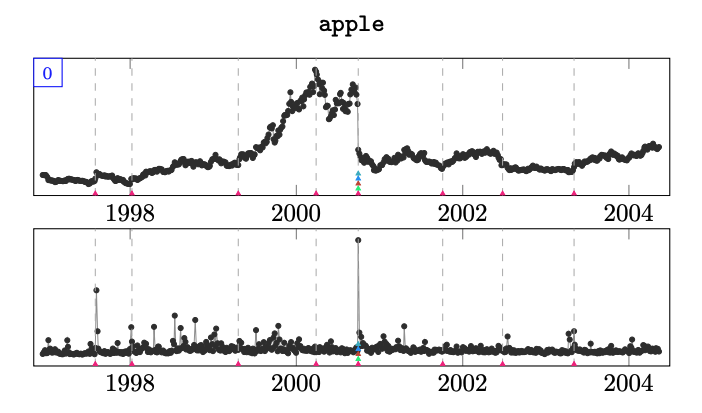


Figure 3. Apple stocks and Bitcoin price

Figure 3, on the left, features the daily closing price and volume of Apple, Inc. stock for a period around the year 2000. The series was sampled at every three-time steps to reduce the length of the series. A significant drop in stock price occurs on 2000.09.29. On the right, the price of Bitcoin in USD from 2013 to 2020.

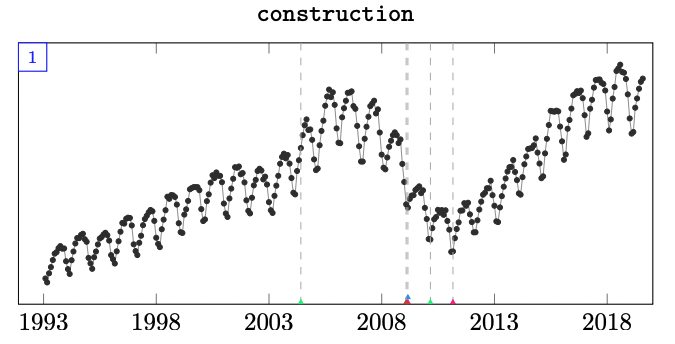
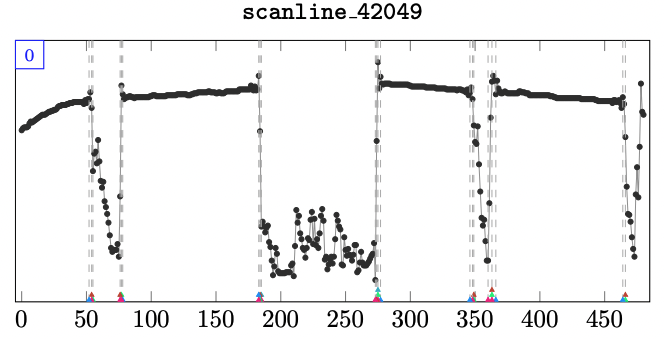


Figure 4. Scanline and Construction data plots

Figure 4, on the left, features a horizontal scan line of image no. 42049 from the Berkeley Segmentation Data Set. On the right, Total private construction spending in the U.S.

**COVID-19 Dataset**

This dataset was featured in the xtbreak paper [6] to find the evidence of multiple breaks. The data consists of both aggregate country and disaggregated state level US data about the covid cases and deaths.

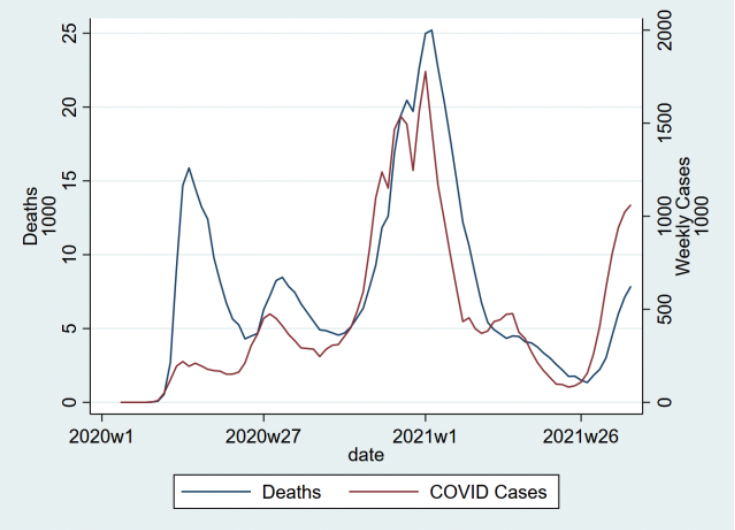


Figure 5. Covid-19 cases and deaths from 2020 to 2021

The data is made up by the weekly data on the number of deaths and new cases from the CDC. Starting in mid-December 2020, vaccines were introduced, and the authors of the paper assume the relationship between the number of cases and deaths to be changed. Therefore, it seems reasonable to expect at least two breaks. Figure 6 features the two structural breaks proposed by the original authors.

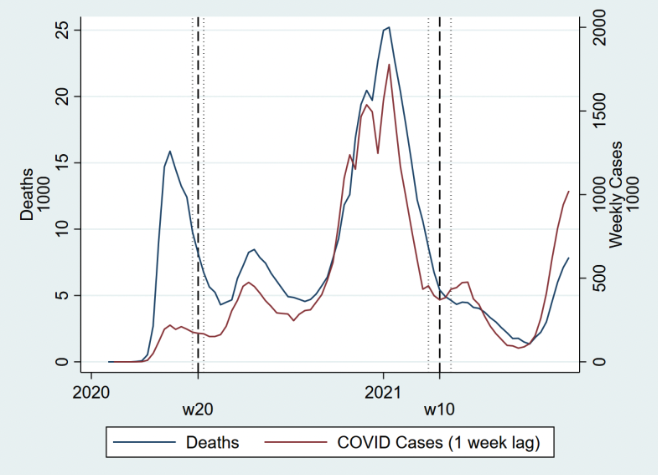
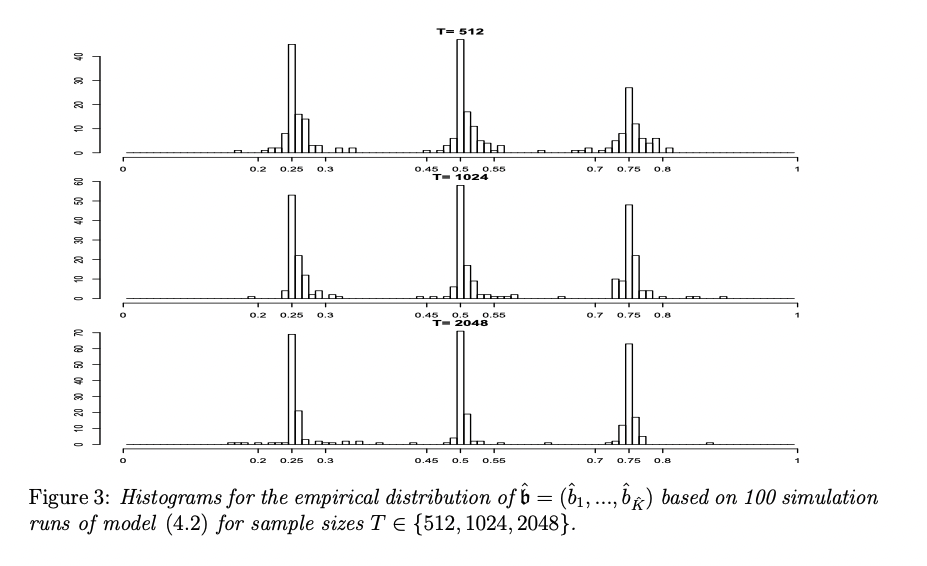


Figure 6. Structural breaks in the Covid dataset

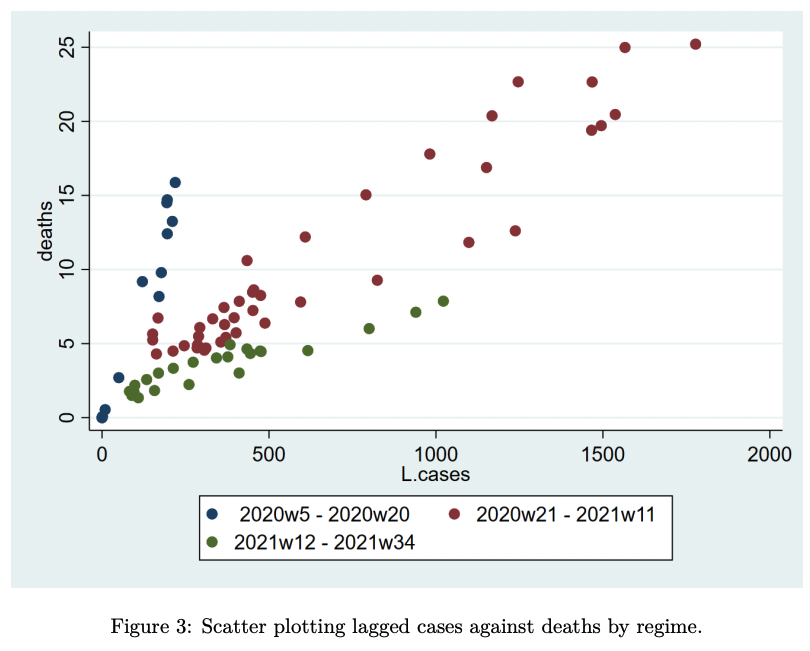
**Experiments**

****

https://www.statmod.ru/\_diploma/2019m/13m\_02\_merzlyakov.pdf

****

Прикольная идея делать бутстреп и показывать куда точки разрыва попадали.



**Interpretation of the Results**

**Conclusion**