An Evaluation of Change Point Detection Algorithms

Gerrit J. J. van den Burg[1] Christopher K. I. Williams[2,1]
gvandenburg@turing.ac.uk ckiw@inf.ed.ac.uk

[1] The Alan Turing Institute, London, UK
[2] The University of Edinburgh, Edinburgh, UK

May 26, 2020

Abstract

Change point detection is an important part of time series analysis, as the presence of a change point indicates an abrupt and significant change in the data generating process. While many algorithms for change point detection exist, 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. Therefore, instead of developing yet another change point detection method, we consider it vastly more important to properly evaluate existing algorithms on real-world data. To achieve this, we present the first data set specifically designed for the evaluation of change point detection algorithms, consisting of 37 time series from various domains. Each time series was annotated by five expert human annotators to provide ground truth on the presence and location of change points. We analyze the consistency of the human annotators, and describe evaluation metrics that can be used to measure algorithm performance in the presence of multiple ground truth annotations. Subsequently, we present a benchmark study where 14 existing algorithms are evaluated on each of the time series in the data set. This study shows that binary segmentation (Scott and Knott, 1974) and Bayesian online change point detection (Adams and MacKay, 2007) are among the best performing methods. Our aim is that this data set will serve as a proving ground in the development of novel change point detection algorithms.

Keywords: change point detection, time series analysis, benchmark evaluation

1 Introduction

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. The literature on change point detection (CPD) is concerned with accurately detecting these events in time series data, as well as with modeling time series in which such events occur. Applications of CPD are found in finance (Lavielle and Teyssiere, 2007), business (Taylor and Letham, 2018), quality control (Page, 1954), and network traffic analysis (Kurt et al., 2018), among many other domains.

There is a large body of work on both the theoretical and practical aspects of change point detection. However, little attention has been paid to the evaluation of CPD algorithms on real-world time series. Existing work often follows a predictable strategy when presenting a novel detection algorithm: the method is first evaluated on a set of *simulated* time series with known change points where both the model fit and detection accuracy are evaluated. An obvious downside of such experiments is that the dynamics of the simulated data are often particular to the paper, and any model that corresponds to these dynamics has an unfair advantage.

Subsequently the proposed method is typically evaluated on only a small number of real-world time series. These series are often reused (e.g. the univariate well-log data of Ó Ruanaidh and Fitzgerald, 1996), may have been preprocessed (e.g. by removing outliers or seasonality), and may not have unambiguous ground truth available. When evaluating a proposed method on real-world data, post-hoc analysis is frequently applied to argue why the locations detected by the algorithm correspond to known events. Clearly, this is not a fair or accurate approach to evaluating novel change point detection algorithms.

Rather than proposing yet another detection method and evaluating it in the manner just described, we argue that it is vastly more important to create a realistic benchmark data set for CPD algorithms. Such a data set will allow researchers to evaluate their novel algorithms in a systematic and realistic way against existing alternatives. Furthermore, since change point detection is an unsupervised learning problem, an evaluation on real-world data is the most reliable approach to properly comparing detection accuracy. A benchmark data set can also uncover important failure modes of existing methods that may not occur in simulated data, and suggest avenues for future research based on a quantitative analysis of algorithm performance.

Change point detection can be viewed as partitioning an input time series into a number of segments. It is thus highly similar to the problem of image segmentation, where an image is partitioned into distinct segments based on different textures or objects. Historically, image segmentation suffered from the same problem as change point detection does today, due to the absence of a large, high-quality data set with objective ground truth (as compared to, for instance, image classification). It was not until the Berkeley Segmentation Data Set (BSDS; Martin et al., 2001) that quantitative comparison of segmentation algorithms became feasible. We believe that it is essential to develop a similar data set for change point detection, to enable both the research community as well as data analysts who use these methods to quantitatively compare CPD algorithms.

To achieve this goal, we present the first data set specifically designed to evaluate change point detection algorithms. We developed an annotation tool and collected annotations from data scientists for 37 real-world time series. Using this data set, we performed the first large-scale quantitative evaluation of existing algorithms. We analyze the results of this comparison to uncover failure cases of present methods and discuss various potential improvements. To summarize, our main contributions are as follows.

1. We describe the design and collection of a data set for change point detection algorithms and analyze the data set for annotation quality and consistency. The data set is made freely available to accelerate research on change point detection.[1]

2. We present metrics for the evaluation of change point detection algorithms that take multiple annotations per data set into account.

3. We evaluate a significant number of existing change point detection algorithms on this data set and provide the first benchmark study on a large number of real-world data sets using two distinct experimental setups.[2]

The remainder of this paper is structured as follows. In Section 2 we give a high-level overview and categorization of existing work on CPD. Section 3 presents a framework for evaluating change point methods when ground truth from multiple human annotators is available. The annotation tool and data set collection process are described in Section 4, along with an analysis of data quality and annotator consistency. The design of the benchmark study and an analysis of the results is presented in Section 5. Section 6 concludes.

[1] See `https://github.com/alan-turing-institute/TCPD`.

[2] All code needed to reproduce our results is made available through an online repository: `https://github.com/alan-turing-institute/TCPDBench`.

2 Related Work

Methods for CPD can be roughly categorized as follows: (1) online vs. offline, (2) univariate vs. multivariate, and (3) model-based vs. nonparametric. We can further differentiate between Bayesian and frequentist approaches among the model-based methods, and between divergence estimation and heuristics in the nonparametric methods. Due to the large amount of work on CPD we only provide a high-level overview of some of the most relevant methods here, which also introduces many of the methods that we subsequently consider in our evaluation study. For a more expansive review of change point detection algorithms, see e.g. Aminikhanghahi and Cook (2017) or Truong et al. (2020).

Let $\mathbf{y}_t \in \mathcal{Y}$ denote the observations for time steps $t = 1, 2, \ldots$, where the domain $\mathcal{Y}$ is d-dimensional and typically assumed to be a subset of $\mathbb{R}^d$. A segment of the series from $t = a, a+1, \ldots, b$ will be written as $\mathbf{y}_{a:b}$. The ordered set of change points is denoted by $\mathcal{T} = \{\tau_0, \tau_1, \ldots, \tau_n\}$, with $\tau_0 = 1$ for notational convenience. In *offline* change point detection we further use T to denote the length of the series and define $\tau_{n+1} = T + 1$. Note that we assume that a change point marks the first observation of a new segment.

Early work on CPD originates from the quality control literature. Page (1954) introduced the CUSUM method that detects where the corrected cumulative sum of observations exceeds a threshold value. Theoretical analysis of this method was subsequently provided by Lorden (1971). Hinkley (1970) generalized this approach to testing for differences in the maximum likelihood, i.e., testing whether $\log p(\mathbf{y}_{1:T} \mid \theta)$ and $\max_\tau[\log p(\mathbf{y}_{1:\tau-1} \mid \theta_1) + \log p(\mathbf{y}_{\tau:T} \mid \theta_2)]$ differ significantly, for model parameters θ, θ_1, and θ_2. Alternatives from the literature on structural breaks include the use of a Chow test (Chow, 1960) or multiple linear regression (Bai and Perron, 2003).

In offline *multiple* change point detection the optimization problem is given by

$$\min_{\mathcal{T}} \sum_{i=1}^{n+1} \ell(\mathbf{y}_{\tau_{i-1}:\tau_i - 1}) + \lambda P(n), \tag{1}$$

with $\ell(\cdot)$ a cost function for a segment (e.g., the negative maximum log-likelihood), $\lambda \geq 0$ a hyperparameter, and $P(n)$ a penalty on the number of change points. An early method by Scott and Knott (1974) introduced binary segmentation as an approximate CPD algorithm that greedily splits the series into disjoint segments based on the above cost function. This greedy method thus obtains a solution in $\mathcal{O}(T \log T)$ time. An efficient dynamic programming approach to compute an exact solution for the multiple change point problem was presented in Auger and Lawrence (1989), with $\mathcal{O}(mT^2)$ time complexity for a given upper bound m on the number of change points.

Improvements to these early algorithms were made by requiring the cost function to be additive, i.e., $\ell(\mathbf{y}_{a:b}) = \ell(\mathbf{y}_{a:\tau-1}) + \ell(\mathbf{y}_{\tau:b})$. Using this assumption, Jackson et al. (2005) present a dynamic programming algorithm with $\mathcal{O}(T^2)$ complexity. Killick et al. (2012) subsequently extended this method with a pruning step on potential change point locations, which reduces the time complexity to be approximately linear under certain conditions on the data generating process and the distribution of change points. A different pruning approach was proposed by Rigaill (2015) using further assumptions on the cost function, which was extended by Maidstone et al. (2017) and made robust against outliers by Fearnhead and Rigaill (2019). Recent work in offline CPD also includes the Wild Binary Segmentation method (Fryzlewicz, 2014) that extends binary segmentation by applying the CUSUM statistic to randomly drawn subsets of the time series. Furthermore, the Prophet method for Bayesian time series forecasting (Taylor and Letham, 2018) supports fitting time series with change points, even though it is not a dedicated CPD method. These offline detection methods are generally restricted to univariate time series.

Bayesian *online* change point detection was simultaneously proposed by Fearnhead and Liu (2007) and Adams and MacKay (2007). These models learn a probability distribution over the "run length", which is the time since the most recent change point. This gives rise to a recursive message-passing algorithm for the joint distribution of the observations and the run lengths. The formulation of Adams and MacKay (2007) has been extended in several works to include online hyperparameter optimization (Turner et al., 2009) and Gaussian Process segment models (Garnett et al., 2009; Saatçi et al., 2010). Recent work by Knoblauch and Damoulas (2018) has added support for model selection and spatio-temporal models, as well as robust detection using β-divergences (Knoblauch et al., 2018).

Nonparametric methods for CPD are typically based on explicit hypothesis testing, where a change point is declared when a test statistic exceeds a certain threshold. This includes kernel change point analysis (Harchaoui et al., 2009), the use of an empirical divergence measure (Matteson and James, 2014), as well as the use of histograms to construct a test statistic (Boracchi et al., 2018). Both the Bayesian online change point detection methods and the nonparametric methods have support for multidimensional data.

Finally, we observe that Hocking et al. (2013, 2014) propose a database of DNA copy number profiles that can be used to evaluate change point algorithms (Maidstone et al., 2017). While this is an important application domain for CPD algorithms, we are interested in evaluating the performance of change point algorithms on general real-world time series.

3 Evaluation Metrics

In this section we review common metrics for evaluating CPD algorithms and provide those that can be used to compare algorithms against multiple ground truth annotations. These metrics will also be used to quantify the consistency of the annotations of the time series, which is why they are presented here. Our discussion bears many similarities to that provided by Martin et al. (2004) and Arbelaez et al. (2010) for the BSDS, due to the parallels between change point detection and image segmentation.

Existing metrics for change point detection can be roughly divided between clustering metrics and classification metrics. These categories represent distinct views of the change point detection problem, and we discuss them in turn. The locations of change points provided by annotator $k \in \{1, \dots, K\}$ are denoted by the ordered set $\mathcal{T}_k = \{\tau_1, \dots, \tau_{n_k}\}$ with $\tau_i \in [1, T]$ for $i = 1, \dots, n_k$ and $\tau_i < \tau_j$ for $i < j$. Then $\mathcal{T}_k$ implies a *partition*, $\mathcal{G}_k$, of the interval $[1, T]$ into disjoint sets $\mathcal{A}_j$, where $\mathcal{A}_j$ is the segment from τ_{j-1} to $\tau_j - 1$ for $j = 1, \dots, n_k + 1$. Recall that we use $\tau_0 = 1$ and $\tau_{n_k+1} = T + 1$ for notational convenience.

Evaluating change point algorithms using clustering metrics corresponds to the view that change point detection inherently aims to divide the time series into distinct regions with a constant data generating process. Clustering metrics such as the variation of information (VI), adjusted Rand index (ARI; Hubert and Arabie, 1985), Hausdorff distance (Hausdorff, 1927), and segmentation covering metric (Everingham et al., 2010; Arbelaez et al., 2010) can be readily applied. We disregard the Hausdorff metric because it is a maximum discrepancy metric and can therefore obscure the true performance of methods that report many false positives. Following the discussion in Arbelaez et al. (2010) regarding the difficulty of using VI for multiple ground truth partitions and the small dynamic range of the Rand index, we choose to use the covering metric as the clustering metric in the experiments.

For two sets $\mathcal{A}, \mathcal{A}' \subseteq [1, T]$ the Jaccard index, also known as Intersection over Union, is given by

$$J(\mathcal{A}, \mathcal{A}') = \frac{|\mathcal{A} \cap \mathcal{A}'|}{|\mathcal{A} \cup \mathcal{A}'|}. \tag{2}$$

Following Arbelaez et al. (2010), we define the covering metric of partition $\mathcal{G}$ by partition $\mathcal{G}'$ as

$$C(\mathcal{G}', \mathcal{G}) = \frac{1}{T} \sum_{\mathcal{A} \in \mathcal{G}} |\mathcal{A}| \cdot \max_{\mathcal{A}' \in \mathcal{G}'} J(\mathcal{A}, \mathcal{A}'). \tag{3}$$

For a collection $\{\mathcal{G}_k\}_{k=1}^K$ of ground truth partitions provided by the human annotators and a partition $\mathcal{S}$ given by an algorithm, we compute the average of $C(\mathcal{S}, \mathcal{G}_k)$ for all annotators as a single measure of performance.

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). The F_β-measure provides a single quantity that incorporates both precision, P, and recall, R, through

$$F_\beta = \frac{(1+\beta^2)PR}{\beta^2 P + R}, \tag{4}$$

with $\beta = 1$ corresponding to the well-known F1-score (Van Rijsbergen, 1979).

When evaluating CPD algorithms using classification metrics it is common to define a margin of error around the true change point location to allow for minor discrepancies (e.g., Killick et al., 2012; Truong et al., 2020). However with this margin of error we must be careful to avoid double counting, so that for multiple detections within the margin around a true change point only one is recorded as a true positive (Killick et al., 2012). To handle the multiple sets of ground truth change points provided by the annotator, we follow the precision-recall framework from the BSDS proposed in Martin et al. (2004).

Let $\mathcal{X}$ denote the set of change point locations provided by a detection algorithm and let $\mathcal{T}^* = \bigcup_k \mathcal{T}_k$ be the combined set of all human annotations. For a set of ground truth locations $\mathcal{T}$ we define the set of true positives $\mathrm{TP}(\mathcal{T}, \mathcal{X})$ of $\mathcal{X}$ to be those $\tau \in \mathcal{T}$ for which $\exists x \in \mathcal{X}$ such that $|\tau - x| \leq M$, while ensuring that only one $x \in \mathcal{X}$ can be used for a single $\tau \in \mathcal{T}$. The latter requirement is needed to avoid the double counting mentioned previously, and $M \geq 0$ is the allowed margin of error. The precision and recall are then defined as

$$P = \frac{|\mathrm{TP}(\mathcal{T}^*, \mathcal{X})|}{|\mathcal{X}|}, \tag{5}$$

$$R = \frac{1}{K} \sum_{k=1}^{K} \frac{|\mathrm{TP}(\mathcal{T}_k, \mathcal{X})|}{|\mathcal{T}_k|}. \tag{6}$$

With this definition of precision, we consider as false positives only those detections that correspond to *no* human annotation. Further, by defining the recall as the average of the recall for each individual annotator we encourage algorithms to explain *all* human annotations without favoring any annotator in particular.[3] Note that precision and recall are undefined if the annotators or the algorithm declare that a series has no change points. To avoid this, we include $t = 1$ as a trivial change point in $\mathcal{T}_k$ and $\mathcal{X}$. Since change points are interpreted as the start of a new segment this choice does not affect the meaning of the detection results.

Using both the segmentation covering metric and the F1-score in the benchmark study allows us to explore both the clustering and classification view of change point detection. In the experiments we use a margin of error of $M = 5$.

[3] This definition of recall is the *macro*-average of the recall for each annotator, as opposed to the *micro*-average, $R' = (\sum_k |\mathrm{TP}(\mathcal{T}_k, \mathcal{X})|)/(\sum_k |\mathcal{T}_k|)$. When using R' situations can arise where it is favorable for the detector to agree with the annotator with the *most* annotations, which is not a desirable property.

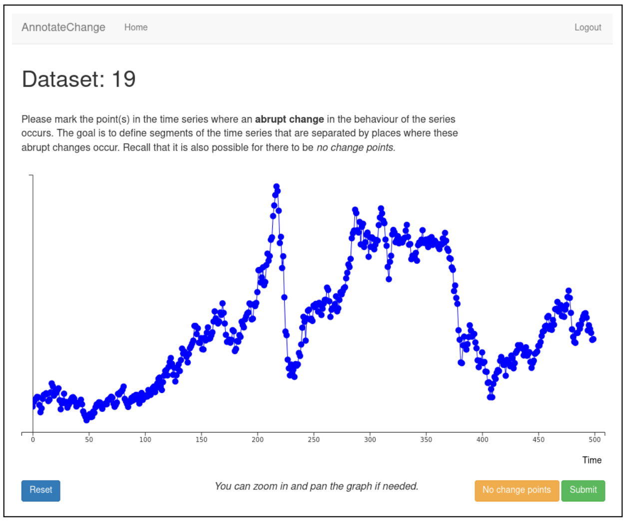


Figure 1: The annotation tool created to collect time series annotations. Note that no contextual information of the time series is provided to ensure annotators do not use external knowledge during the process. The *rubric* presented to the annotators is also visible.

4 Change Point Dataset

This section describes the annotation tool used for collecting the change point data set, as well as the data sources, the experimental protocol, and an analysis of the obtained annotations.

4.1 Annotation Tool

To facilitate the collection of the annotated data set, we created a web application that enables annotators to mark change points on a plot of a time series (see Figure 1).[4] Annotators can select multiple change points in the data or can signal that they believe the time series contains no change points. Before annotators are provided any real-world time series, they are required to go through a set of introductory pages that describe the application and show various types of change points (including changes in mean, variance, trend, and seasonality). These introductory pages also illustrate the difference between a change point and an outlier. During this introduction the annotator is presented with several synthetic time series with known change points and provided with feedback on the accuracy of their annotations. If the annotator's average F1-score is too low on these introductory data sets, they are kindly asked to repeat the introduction. This ensures that all annotators have a certain baseline level of familiarity with annotating change points.

To avoid that annotators use their knowledge of historical events or are affected by other biases, no dates are shown on the time axis and no values are present on the vertical axis. The names of the series are also not shown to the annotators. Furthermore, the annotation page always displays the following task description to the annotators:

> Please mark the point(s) in the time series where an **abrupt change** in the behavior of the series occurs. The goal is to define segments of the time series that are separated by places where these abrupt changes occur. Recall that it is also possible for there to be *no change points*.

[4] The web application is made available as open-source software, see: `https://github.com/alan-turing-institute/AnnotateChange`.

This description is purposefully left somewhat vague to avoid biasing the annotators in a specific direction, and is inspired by the rubric in the BSDS (Martin et al., 2001).

4.2 Data Sources

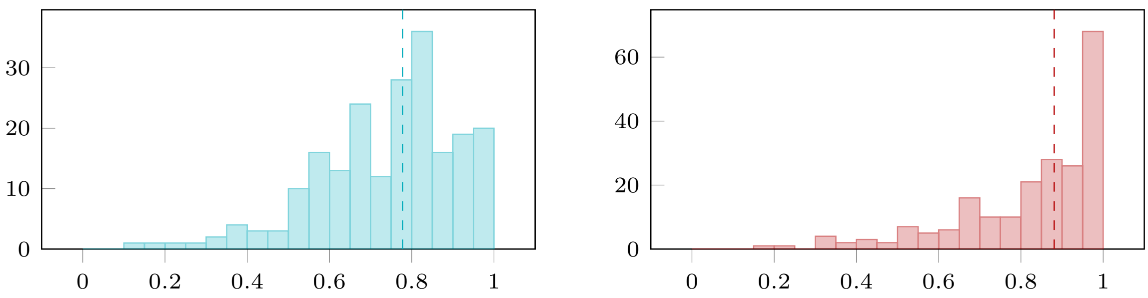
Time series were collected from various online sources including the WorldBank, EuroStat, U.S. Census Bureau, GapMinder, and Wikipedia. A number of time series show potential change point behavior around the financial crisis of 2007-2008, such as the price for Brent crude oil (Figure 1), U.S. business inventories, and U.S. construction spending, as well as the GDP series for various countries. Other time series are centered around introduced legislation, such as the mandatory wearing of seat belts in the U.K., the Montreal Protocol regulating CFC emissions, or the regulation of automated phone calls in the U.S. Various data sets are also collected from existing work on change point detection and structural breaks, such as the well-log data set (Ó Ruanaidh and Fitzgerald, 1996), a Nile river data set (Durbin and Koopman, 2012), and a sequence from the bee-waggle data set (Oh et al., 2008). The main criterion for including a time series was whether it displayed interesting behavior that might include an abrupt change. Several time series were included that may not in fact contain a change point but are nevertheless considered interesting for evaluating CPD algorithms due to features such as seasonality or outliers.

A total of 37 real time series were collected for the change point data set. This includes 33 univariate and 4 multivariate series (with either 2 or 4 dimensions). One of the univariate time series (`uk_coal_employ`, Figure 37 in Appendix C) has two missing observations, and nine other series show seasonal patterns. Unbeknownst to the annotators we added 5 simulated "quality control" series with known change points in the data set, which allows us to evaluate the quality of the annotators *in situ* (see below). Four of these series have a change point with a shift in the mean and the fifth has no change points. Of the quality control series with a change point two have a change in noise distribution, one contains an outlier, and another has multiple periodic components. The change point data set thus consists of 42 time series in total. The average length of the time series in the dataset is 327.7, with a minimum of 15 and a maximum of 991. See Appendix C for a complete overview of all series in the data set.

4.3 Annotation Collection

Annotations were collected in several sessions by inviting data scientists and machine learning researchers to annotate the series. We deliberately chose not to ask individuals who are experts in the various domains of the time series, to avoid the use of outside knowledge in the annotation process. The annotation tool randomly assigned time series to annotators whenever they requested a new series to annotate, with a bias towards series that already had annotations. The same series was never assigned to an annotator more than once. In total, eight annotators provided annotations for the 42 time series, with five annotators assigned to each time series. On average the annotators marked 7.4 unique change points per series, with a standard deviation of 7.0, a minimum of 0 and a maximum of 26. The provided annotations for each series are shown in Appendix C.

Although all annotators successfully completed the introduction before annotating any real time series, some nonetheless commented on the difficulty of deciding whether a series had a change point or not. In some cases this was due to a significant change spanning multiple time steps, which led to ambiguity about whether the transition period should be marked as a separate segment or not (see e.g. Figure 1). Another source of ambiguity were periodic time series with abrupt changes (e.g. `bank`, Figure 6), which could be regarded as a stationary switching distribution without abrupt changes or a series with frequent change points. In these cases annotators were advised to use their intuition and experience as data scientists to evaluate whether they considered a potential change to be an *abrupt* change.



(a) Covering Metric

(b) F1-score

Figure 2: Histograms of one vs. rest annotator performance for both metrics. Median scores over all data sets are shown with dashed lines.

4.4 Consistency

As mentioned above we included five "quality control" series to evaluate the accuracy and consistency of the human annotators in practice, which are shown in Appendix C.2. On two of the four series with known change points (Figure 42 and Figure 44) all annotators identified the true change point within three time steps of the correct location. On another time series, `quality_control_2`, four of the five annotators identified the change point within two time steps, with a fifth annotator believing there to be no change points (Figure 43). The most difficult series to annotate was the one with periodic features (`quality_control_4`, Figure 45). Here, two of the annotators voted there to be no change points while the three other annotators identified the change correctly within three time steps, with one also incorrectly marking several other points as change points. The series without change points, `quality_control_5`, was correctly identified as such by all annotators. This suggests that human annotators can identify the presence and location of change points with high accuracy on most kinds of time series, with seasonal effects posing the biggest challenge.

To quantitatively assess the consistency of the annotators we evaluate each individual annotator against the four others on every time series, using the metrics presented previously. Figure 2 shows the histograms of these scores for both metrics. The figures show that on average there is a high degree of agreement between annotators. It can be seen that the values are generally higher on the F1-score, which is likely due to the margin of error in this metric. The cases with low agreement are often due to one or two annotators disagreeing with the others. An example of this is the `lga_passengers` series, where there is a high level of agreement between four of the annotators about the number and approximate location of the change points, but one annotator believes there are no change points (Figure 24). The histograms show that the median annotator score is approximately 0.8 for the covering metric and 0.9 for the F1-score, which we consider to be a sufficient level of consistency for our purposes.

5 Benchmark Study

This section presents a benchmark study of existing change point algorithms on the data set of human-annotated time series. We describe the methods included in the study, the setup of the experiments, and the results of the evaluation. The code and data used in this benchmark study are available through an online repository.[5]

[5]See: `https://github.com/alan-turing-institute/TCPDBench`.

Table 1: The change point detection algorithms considered in the benchmark study.

Name	Method	Reference
AMOC	At Most One Change	Hinkley (1970)
BINSEG	Binary Segmentation	Scott and Knott (1974)
BOCPD	Bayesian Online Change Point Detection	Adams and MacKay (2007)
BOCPDMS	BOCPD with Model Selection	Knoblauch and Damoulas (2018)
CPNP	Nonparametric Change Point Detection	Haynes et al. (2017)
ECP	Energy Change Point	Matteson and James (2014)
KCPA	Kernel Change-Point Analysis	Harchaoui et al. (2009)
PELT	Pruned Exact Linear Time	Killick et al. (2012)
PROPHET	Prophet	Taylor and Letham (2018)
RBOCPDMS	Robust BOCPDMS	Knoblauch et al. (2018)
RFPOP	Robust Functional Pruning Optimal Partitioning	Fearnhead and Rigaill (2019)
SEGNEIGH	Segment Neighborhoods	Auger and Lawrence (1989)
WBS	Wild Binary Segmentation	Fryzlewicz (2014)
ZERO	No Change Points	

5.1 Experimental Setup

The aim of the study is to obtain an understanding of the performance of existing change point methods on real-world data. With that in mind we evaluate a large selection of methods that are either frequently used or that have been recently developed. Table 1 lists the methods included in the study. To avoid issues with implementations we use existing software packages for all methods, which are almost always made available by the authors themselves. We also include the baseline ZERO method, which always returns that a series contains no change points.

Many of the methods have (hyper)parameters that affect the predicted change point locations. To obtain a realistic understanding of how accurate the methods are in practice, we create two separate evaluations: one showing the performance using the *default* settings of the method as defined by the package documentation, and one reporting the *maximum* score over a grid search of parameter configurations. We refer to these settings as the “Default” and “Best” experiments, respectively. The Default experiment aims to replicate the common practical setting of a data analyst trying to detect potential change points in a time series without any prior knowledge of reasonable parameter choices. The Best experiment aims to identify the highest possible performance of each algorithm by running a full grid search over its hyperparameters. It should be emphasized that while the Best experiment is important from a theoretical point of view, it is not necessarily a realistic measure of algorithm performance. Due to the unsupervised nature of the change point detection problem we can not in general expect those who use these methods to be able to tune the parameters such that the results match the *unknown* ground truth. For a detailed overview of how each method was applied, see Appendix A.

Since not all methods support multidimensional time series or series with missing values, we separate the analysis further into univariate and multivariate time series. While it is possible to use the univariate methods on multidimensional time series by, for instance, combining the detected change points of each dimension, we chose not to do this as it could skew the reported performance with results on series that the methods are not explicitly designed to handle. Furthermore, because the programming language differs between implementations we do not report on the computation time of the methods. 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.

The Bayesian online change point detection algorithms that follow the framework of Adams and MacKay (2007) compute a full probability distribution over the location of the change points and therefore do not return a single segmentation of the series. To facilitate comparison to the other methods we use the maximum a posteriori (MAP) segmentation of the series (see, e.g., Fearnhead and Liu, 2007).

Table 2: Scores for both experiments on each metric, separated among univariate and multivariate series. For the Default experiment the value corresponds to the average of a single run of the algorithm, whereas for the Best experiment it is the average of the highest score obtained for each series. Averages are computed on series where we have results for all methods. The highest value in each column is highlighted in bold.

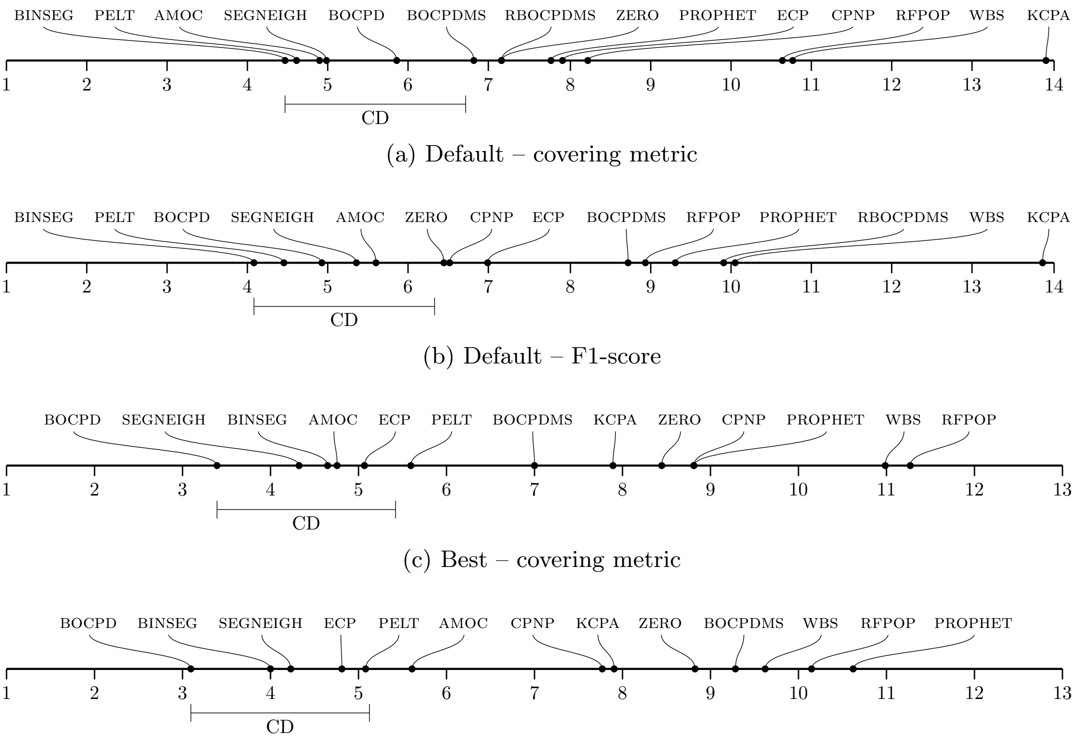
	Univariate				Multivariate			
	Default		Best		Default		Best	
	Cover	F1	Cover	F1	Cover	F1	Cover	F1
AMOC	0.702	0.704	0.746	0.799				
BINSEG	**0.706**	**0.744**	0.780	0.856				
BOCPD	0.636	0.690	**0.789**	**0.880**	0.455	**0.610**	**0.801**	**0.941**
BOCPDMS	0.633	0.507	0.744	0.620	**0.486**	0.429	0.623	0.533
CPNP	0.535	0.607	0.552	0.666				
ECP	0.523	0.598	0.720	0.797	0.402	0.545	0.590	0.725
KCPA	0.062	0.111	0.626	0.683	0.047	0.071	0.626	0.747
PELT	0.689	0.710	0.725	0.787				
PROPHET	0.540	0.488	0.576	0.534				
RBOCPDMS	0.629	0.447			0.402	0.356		
RFPOP	0.392	0.499	0.414	0.531				
SEGNEIGH	0.676	0.676	0.784	0.855				
WBS	0.330	0.412	0.428	0.533				
ZERO	0.583	0.669	0.579	0.662	0.464	0.577	0.464	0.577

5.2 Results

Table 2 shows the average scores on each of the considered metrics for both experiments (complete results are available in Appendix B.1).[6] We see that for the Default experiment the BINSEG method of Scott and Knott (1974) achieves the highest average performance on the univariate time series, closely followed by AMOC and PELT. For this experiment the BOCPDMS method (Knoblauch and Damoulas, 2018) does best for the multivariate series when using the covering metric, but BOCPD (Adams and MacKay, 2007) does best on the F1-score. In the Best experiment – where hyperparameters of the methods are varied to achieve better performance – we see that BOCPD does best for both the univariate and multivariate time series. Note that for the univariate series the maximum achieved performance of BOCPD is only slightly higher than that of BINSEG and SEGNEIGH, indicating that parameter tuning makes these three methods competitive with each other. These methods assume constant Gaussian distributions per segment by default, with BINSEG and SEGNEIGH using the modified Bayesian Information Criterion of Zhang and Siegmund (2007) as a means of regularization (as in Equation 1).

We note that the ZERO method outperforms many of the other methods, especially in the Default experiment. This is in part due to the relatively small number of change points per series, but also suggests that many methods return a large number of false positives. Furthermore, we see that for CPNP, PROPHET, and RFPOP the hyperparameter tuning has a relatively small effect on performance. This is unsurprising however, since these methods also have a small number of hyperparameters that can be varied. For instance, for CPNP only the choice of penalty function and the number of quantiles used in the nonparametric estimation of the distribution can be

[6] The RBOCPDMS method failed in two of the univariate data sets (`bitcoin` and `scanline_126007`), which are therefore not included in the averages for the Default experiment. Results on `uk_coal_employ` are excluded for both experiments because not all methods can handle missing values. This explains the slight differences in the performance of the ZERO method between experiments. RBOCPDMS is absent from the results for the Best experiment due to the restriction on computation time that was put in place (see Appendix A for details).



(a) Default – covering metric

(b) Default – F1-score

(c) Best – covering metric

(d) Best – F1-score

Figure 3: Average ranks of the algorithms for both experiments and both metrics for univariate time series. Only time series for which we have results for every method are included in the rank calculation. CD denotes the critical difference interval, outside of which methods perform significantly worse than the best method. Lower ranks are better.

tuned, and the results show that this is insufficient to achieve competitive performance. By contrast, parameter tuning has a significant effect on the performance of KCPA, and it makes ECP competitive with AMOC and PELT when considering the F1-score.

We can analyze how frequently a method scores as one of the best methods by ranking their performance on each individual time series and taking averages, following Demšar (2006). Figure 3 shows these average ranks of the methods for both experiments on the univariate time series (rank plots for multivariate series can be found in Appendix B.2). The plots in Figure 3 illustrate that in the Default experiment the offline segmentation methods (BINSEG, PELT, AMOC, and SEGNEIGH) and BOCPD frequently perform well on both metrics. The offline segmentation algorithms are also some of the best methods in the experiment where hyperparameters can be varied, but are outperformed by the BOCPD method. This is in line with the average scores reported in Table 2, which again shows that parameter tuning can improve the performance of BOCPD to the point where it generally outperforms all other methods. It is noteworthy that even though BOCPD assumes a simple Gaussian model with constant mean for each segment, it outperforms methods that allow for more complex time series behavior such as trend and seasonality (PROPHET), or autoregressive structures (BOCPDMS). The figures further show that WBS and RFPOP perform poorly regardless of whether hyperparameter tuning is applied.

As described in Demšar (2006), the average ranks of the algorithms allow us to apply statistical significance tests on the difference in performance between the methods. We use Holm's procedure (Holm, 1979) to test for significant differences between the best method and the other methods, while controlling for the family-wise error rate (using $\alpha = 0.05$). Then, the performance of two methods is significantly different if their ranks differ by at least the *critical difference* (CD). For each experiment and metric the critical difference is shown in Figure 3, using the best performing method as reference point. For both experiments we see that there is no statistically significant difference between BINSEG, AMOC, SEGNEIGH, and BOCPD. However, in the Best experiment we observe that BOCPD significantly outperforms PELT on the covering metric, whereas this is not the case with the F1-score or in the Default experiment. Moreover, ECP is significantly outperformed by BOCPD in the Default experiment, but not in the Best experiment. The methods CPNP, KCPA, PROPHET, RFPOP, and WBS are significantly outperformed by the best method regardless of the experiment or evaluation metric.

Finally, we observe that when comparing the results for different evaluation metrics the choice of metric affects the ranking of the algorithms (see Figure 3). Because the annotations are combined in a single set of ground truth change points when computing the precision, it is easier to achieve a high F1-score than it is to obtain a high score on the covering metric. Indeed, a high score on the covering metric indicates a high degree of agreement with each *individual* annotator, whereas a high value on the F1-score indicates a strong agreement with the *combined* set of annotations. Nevertheless, the best performing method is the same regardless of the metric, which reflects a degree of stability in the relative performance of the methods.

6 Discussion

We have introduced the first dedicated data set of human-annotated real-world time series for the evaluation and development of change point detection algorithms. Moreover, we have presented a framework for evaluating change point detection algorithms against ground truth provided by multiple human annotators, with two change point metrics that take distinct but complementary approaches to evaluating detection performance. With this data set we have conducted the first thorough evaluation of a large selection of existing change point algorithms in an effort to discover those that perform best in practice. This benchmark study has shown that the binary segmentation algorithm of Scott and Knott (1974) performs best on univariate time series when using the default algorithm parameters from Killick and Eckley (2014), but its performance is not statistically significantly different from PELT, AMOC, SEGNEIGH, and BOCPD. When hyperparameter tuning is performed, the BOCPD method of Adams and MacKay (2007) outperforms all other methods on both univariate and multivariate time series. However, the differences between it and several other algorithms (including SEGNEIGH, BINSEG, and ECP) have been shown to not be statistically significantly different.

Both the introduced data set and the benchmark study can be extended and improved in the future. However, the present paper has already provided novel insights in the practical performance of change point detection algorithms and indicated potential topics for further research on algorithm development. For instance, future work could focus on incorporating automated hyperparameter tuning for BOCPD to improve the default performance of this method (e.g., following Turner et al., 2009). Alternatively, it may be possible to develop ways in which the performance of the binary segmentation or segment neighborhood methods can be improved. The presented change point data set and benchmark study motivate these research directions based on a quantitative analysis of algorithm performance, which was not previously possible.

Because of the unsupervised nature of the change point detection problem, it is difficult to compare algorithms in settings other than through simulation studies. But since the goal is ultimately to apply these algorithms to real-world time series, we should strive to ensure that they work well in practice. The change point data set presented here was inspired by the Berkeley Segmentation Data Set introduced in Martin et al. (2001) as well as the PASCAL VOC challenge (Everingham et al., 2010). In the past, these data sets have been crucial in accelerating the state of the art in image segmentation and object recognition, respectively. Our hope is that the data set and the benchmark study presented in this work can do the same for change point detection.

Acknowledgments

The authors would like to thank James Geddes as well as the other annotators who preferred to remain anonymous. We furthermore thank Siamak Zamani Dadaneh for helpful comments on an earlier version of this manuscript, as well as for suggesting the ZERO baseline method.

Changelog

Below is a list of changes to the arXiv version of this paper.

- v2
  - Added the ZERO method to the comparison.
  - Added summary statistics in the text regarding the length of the time series and the number of annotated change points.
  - Added rank plots for multivariate time series to the appendix.
  - Corrected an error in the computation of the F1 score and updated the results and the online code repository. This correction had no major effect on our conclusions.
  - Made minor revisions in Section 5.2 and Section 6 to reflect minor changes in the relative performance of the algorithms.
  - Updated acknowledgements.
- v1
  - Initial version.

A Simulation Details

This section presents additional details of the implementations and configurations used for each of the algorithms. Where possible, we used packages made available by the authors of the various methods. During the grid search parameter configurations that led to computational errors (e.g. overflow) were skipped.

- AMOC, BINSEG, PELT, and SEGNEIGH were implemented using the R `changepoint` package, v2.2.2 (Killick and Eckley, 2014).
  - As default parameters we used the `cpt.mean` function, following Fryzlewicz (2014), with the Modified Bayes Information Criterion (MBIC) penalty, the Normal test statistic, and minimum segment length of 1. For BINSEG and SEGNEIGH we additionally used the default value of 5 for the maximum number of change points/segments.
  - During the grid search, we varied the parameters as follows:

    Function: `cpt.mean`, `cpt.var`, `cpt.meanvar`

    Penalty: None, SIC, BIC, MBIC, AIC, Hannan-Quinn, and Asymptotic (using p-value 0.05)

    Test Statistic: Normal, CUSUM, CSS, Gamma, Exponential, and Poisson

    Max. CP: 5 (default) or $T/2+1$ (max). Only for BINSEG or SEGNEIGH.

    Combinations of parameters that were invalid for the method were ignored.

- CPNP was implemented with the R `changepoint.np` package (v0.0.2). The penalty was varied in the same way as for PELT (see above) with the addition that in the grid search the number of quantiles was varied on the grid $\{10, 20, 30, 40\}$.

- RFPOP was called using the `robseg` package available on GitHub.[7] As default setting the biweight loss function was used. During the grid search we additionally considered the L1, L2, and Huber loss. Other parameters (`lambda` and `lthreshold`) were set as described in the experimental section of Fearnhead and Rigaill (2019).

- ECP and KCPA were accessed through the R `ecp` package (v3.1.1) available on CRAN (James and Matteson, 2013). For ECP the parameter settings were:
  - As default parameters we used the `e.divisive` function with $\alpha = 1$, minimum segment size of 30, 199 runs for the permutation test and a significance level of 0.05.
  - The grid search consisted of:

    Algorithm: `e.divisive` or `e.agglo`

    Significance level: 0.05 or 0.01.

    Minimum segment size: 2 (minimal) or 30 (default)

    α : $(0.5, 1.0, 1.5)$

    For the `e.agglo` algorithm only the α parameter has an effect.

  The KCPA method takes two parameters: a cost and a maximum number of change points. In the default experiment the cost parameter was set to 1 and we allowed the maximum number of change points possible. In the grid search the cost parameter was varied on the set $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$ and the maximum number of change points was varied between the maximum possible and 5, which corresponds to the methods in the `changepoint` package.

[7] `https://github.com/guillemr/robust-fpop`, version 2019-07-02.

- WBS was evaluated using the R package of the same name available on CRAN (v1.4).
    - As defaults we used the SSIC penalty, the integrated version, and the default of 50 for the maximum number of change points.
    - During the grid search we used:

      Penalty: SSIC, BIC, MBIC
      Integrated: true, false
      Max CP: 50 (default) or T (max)

- BOCPD was called using the R package `ocp` (v0.1.1), available on CRAN (Pagotto, 2019). For the experiments we used the Gaussian distribution with the negative inverse gamma prior. The prior mean was set to 0, since the data is standardized. No truncation of the run lengths was applied.
    - For the default experiment the intensity (λ_{gap}) was set to 100, the prior variables were set to $\alpha_0 = 1$, $\beta_0 = 1$, and $\kappa_0 = 1$.
    - For the grid search, we varied the parameters as follows:

      Intensity: $(50, 100, 200)$
      α_0**:** $(0.01, 1, 100)$
      β_0**:** $(0.01, 1, 100)$
      κ_0**:** $(0.01, 1, 100)$

- BOCPDMS and RBOCPDMS were run with the code available on GitHub.[8] The settings for BOCPDMS were the same as that for BOCPD above, with the exception that during the grid search run length pruning was applied by keeping the best 100 run lengths. This was done for speed considerations. No run length pruning was needed for BOCPDMS with the default configuration, but this was applied for RBOCPDMS. For RBOCPDMS we used the same settings, but supplied additional configurations for the α_0 and α_{rld} parameters. As default values and during the grid search we used $\alpha_0 = \alpha_{\text{rld}} = 0.5$. For both methods we applied a timeout of 30 minutes during the grid search and we used a timeout of 4 hours for RBOCPDMS in the default experiment.

- PROPHET was run in R using the `prophet` package (v0.4). Since Prophet requires the input time vector to be a vector of dates or datetimes, we supplied this where available and considered the time series that did not have date information to be a daily series. By supplying the dates of the observations Prophet automatically chooses whether to include various seasonality terms. For both experiments we allowed detecting of change points on the entire input range and used a threshold of 0.01 for selecting the change points (in line with the change point plot function in the Prophet package). During the grid search we only varied the maximum number of change points between the default (25) and the theoretical maximum (T).

[8] `https://github.com/alan-turing-institute/bocpdms/` at Git commit hash f2dbdb5, and `https://github.com/alan-turing-institute/rbocpdms/` at Git commit hash 643414b, respectively.

B Additional Results

B.1 Tables

Table 3: **Default-Covering.** Results for each method on each of the time series for the experiment with default settings using the covering metric. Values are unavailable either due to a failure (F), because of missing values in the series unsupported by the method (M), or because the method is not suited for multidimensional series (blank). Highest values for each time series are shown in bold.

Dataset	AMOC	BINSEG	BOCPD	BOCPDMS	CPNP	ECP	KCPA	PELT	PROPHET	RBOCPDMS	RFPOP	SEGNEIGH	WBS	ZERO
bank	0.967	0.509	0.048	0.188	0.053	0.127	0.036	0.509	0.361	0.644	0.036	0.509	0.048	**1.000**
bitcoin	**0.764**	0.754	0.717	0.748	0.364	0.209	0.046	0.758	0.723	F	0.168	0.758	0.304	0.516
brent_spot	0.423	0.627	0.586	0.265	0.411	0.387	0.022	0.627	0.527	0.504	0.225	**0.630**	0.242	0.266
businv	0.574	0.562	0.463	0.459	0.402	0.311	0.013	**0.603**	0.478	0.559	0.123	0.494	0.108	0.461
centralia	**0.675**	0.564	0.612	0.650	0.564	**0.675**	0.440	0.564	**0.675**	0.624	0.528	0.564	0.253	**0.675**
children_per_woman	**0.798**	**0.798**	0.758	0.427	0.486	0.397	0.048	**0.798**	0.521	0.745	0.154	0.771	0.186	0.429
co2_canada	0.527	0.608	**0.716**	0.276	0.514	0.639	0.263	0.611	0.540	0.432	0.497	0.612	0.480	0.278
construction	0.525	0.466	0.395	0.571	0.334	0.352	0.016	0.423	0.502	**0.581**	0.092	0.423	0.198	0.575
debt_ireland	0.321	**0.635**	0.584	0.312	**0.635**	0.321	0.210	**0.635**	0.321	0.306	0.489	**0.635**	0.248	0.321
gdp_argentina	0.667	0.667	0.630	0.711	0.451	**0.737**	0.061	0.667	0.534	0.711	0.332	0.631	0.068	**0.737**
gdp_croatia	0.605	0.605	0.642	0.655	0.642	**0.708**	0.108	0.605	**0.708**	0.655	0.353	0.605	0.108	**0.708**
gdp_iran	0.484	0.477	0.520	0.580	0.448	**0.583**	0.062	0.477	**0.583**	0.580	0.248	0.503	0.066	**0.583**
gdp_japan	0.654	0.654	0.578	0.644	0.522	**0.802**	0.041	0.654	**0.802**	0.645	0.269	0.654	0.048	**0.802**
global_co2	0.743	0.748	0.640	0.738	0.313	0.535	0.173	0.638	0.284	0.745	0.368	0.634	0.187	**0.758**
homeruns	**0.683**	**0.683**	0.526	0.677	0.516	0.611	0.047	**0.683**	0.574	0.554	0.354	0.590	0.312	0.511
iceland_tourism	0.764	0.764	0.595	0.616	0.383	0.382	0.017	0.764	0.453	0.855	0.293	0.650	0.512	**0.946**
jfk_passengers	**0.839**	**0.839**	0.541	0.831	0.561	0.438	0.008	**0.839**	0.373	0.628	0.264	0.791	0.409	0.630
lga_passengers	0.427	0.458	0.496	0.444	0.473	0.386	0.013	0.474	0.434	0.382	0.412	**0.543**	0.477	0.383
measles	**0.951**	**0.951**	0.081	0.255	0.098	0.060	0.005	0.213	0.603	0.296	0.046	0.367	0.081	**0.951**
nile	0.880	0.880	**0.888**	0.870	0.880	0.857	0.032	0.880	0.758	0.753	0.880	0.880	0.880	0.758
ozone	**0.701**	0.600	0.602	0.560	0.441	0.574	0.070	0.635	0.574	0.577	0.309	0.627	0.070	0.574
quality_control_1	**0.992**	**0.992**	0.887	0.982	0.683	0.620	0.010	**0.992**	0.693	0.990	0.655	**0.992**	0.687	0.503
quality_control_2	0.922	0.922	**0.927**	0.921	0.922	0.927	0.010	0.922	0.723	0.912	0.922	0.922	0.922	0.638
quality_control_3	0.996	0.996	**0.997**	0.977	0.831	**0.997**	0.008	0.996	0.500	0.743	0.658	0.996	0.996	0.500
quality_control_4	0.563	0.518	0.511	0.670	0.481	0.257	0.009	0.538	0.508	0.661	0.059	0.538	0.080	**0.673**
quality_control_5	**1.000**	**1.000**	**1.000**	0.994	**1.000**	**1.000**	0.006	**1.000**	**1.000**	0.994	**1.000**	**1.000**	**1.000**	**1.000**
rail_lines	**0.786**	**0.786**	0.767	0.416	0.738	0.428	0.103	**0.786**	0.534	0.416	0.439	**0.786**	0.103	0.428
ratner_stock	0.873	0.913	0.796	0.866	0.380	0.185	0.021	0.908	0.444	0.862	0.162	**0.914**	0.176	0.450
robocalls	0.641	0.641	**0.808**	0.622	0.677	0.601	0.069	0.641	0.601	0.607	0.447	0.760	0.069	0.601
scanline_126007	0.519	0.464	0.346	0.433	0.433	0.327	0.025	0.444	0.503	F	0.316	**0.567**	0.263	0.503
scanline_42049	0.424	0.631	**0.892**	0.832	0.529	0.490	0.121	0.745	0.441	0.421	0.257	0.730	0.432	0.211
seatbelts	0.683	**0.797**	0.757	0.526	0.750	0.615	0.020	**0.797**	0.628	0.526	0.484	0.765	0.727	0.528
shanghai_license	**0.920**	**0.920**	0.856	0.616	0.474	0.518	0.020	**0.920**	0.768	0.754	0.381	0.826	0.209	0.547
uk_coal_employ	M	M	M	0.429	M	0.356	0.356	M	**0.481**	M	M	M	M	0.356
unemployment_nl	0.507	**0.669**	0.495	0.570	0.503	0.470	0.039	0.648	0.507	0.539	0.243	0.648	0.222	0.507
us_population	0.736	0.391	0.219	0.801	0.391	0.089	0.003	0.506	0.096	0.555	**0.803**	0.307	0.043	**0.803**
usd_isk	0.853	0.735	0.672	**0.866**	0.480	0.616	0.023	0.730	0.436	0.582	0.163	0.730	0.194	0.436
well_log	0.453	0.695	0.776	0.778	0.772	0.601	0.020	0.679	0.411	0.665	**0.787**	0.647	0.719	0.225
apple			0.365	0.298		0.305	0.007			0.424				**0.425**
bee_waggle_6			0.089	0.887		0.116	0.004			0.277				**0.891**
occupancy			**0.549**	0.459		0.530	0.050			0.467				0.236
run_log			**0.815**	0.302		0.657	0.128			0.439				0.304

Table 4: **Default-F1.** Results for each method on each of the time series for the experiment with default settings using the F1-score. Values are unavailable either due to a failure (F), because of missing values in the series unsupported by the method (M), or because the method is not suited for multidimensional series (blank). Highest values for each time series are shown in bold.

Dataset	AMOC	BINSEG	BOCPD	BOCPDMS	CPNP	ECP	KCPA	PELT	PROPHET	RBOCPDMS	RFPOP	SEGNEIGH	WBS	ZERO
bank	0.667	0.400	0.047	0.118	0.044	0.154	0.008	0.400	0.154	0.333	0.015	0.333	0.039	**1.000**
bitcoin	0.367	0.426	**0.692**	0.269	0.463	0.338	0.092	0.672	0.446	F	0.284	0.672	0.464	0.450
brent_spot	0.272	0.483	0.521	0.239	**0.607**	0.478	0.104	0.465	0.249	0.321	0.490	0.431	0.516	0.315
businv	0.455	0.370	0.270	0.370	0.304	0.301	0.047	0.370	0.275	0.312	0.245	0.312	0.230	**0.588**
centralia	0.763	0.909	0.909	0.846	0.909	0.763	0.714	0.909	0.763	0.846	**1.000**	0.909	0.556	0.763
children_per_woman	0.618	0.618	**0.637**	0.337	0.326	0.349	0.068	0.618	0.310	0.288	0.246	0.337	0.271	0.507
co2_canada	0.544	0.691	0.619	0.265	0.578	**0.817**	0.169	0.661	0.482	0.381	0.569	0.661	0.520	0.361
construction	0.516	**0.709**	0.634	0.410	0.602	0.574	0.038	**0.709**	0.324	0.480	0.185	**0.709**	0.316	0.696
debt_ireland	0.469	0.760	0.760	0.611	0.760	0.469	0.519	0.760	0.469	0.530	**0.824**	0.760	0.538	0.469
gdp_argentina	0.889	0.889	**0.947**	0.583	0.818	0.824	0.131	0.889	0.615	0.452	0.571	**0.947**	0.148	0.824
gdp_croatia	0.583	0.583	**1.000**	0.583	**1.000**	0.824	0.160	0.583	0.824	0.452	0.400	0.583	0.167	0.824
gdp_iran	0.492	0.492	0.622	0.492	0.330	**0.652**	0.219	0.492	**0.652**	0.395	0.609	0.395	0.246	**0.652**
gdp_japan	0.615	0.615	0.800	0.471	0.667	**0.889**	0.068	0.615	**0.889**	0.667	0.190	0.615	0.077	**0.889**
global_co2	**0.929**	**0.929**	0.754	0.458	0.421	0.754	0.167	0.754	0.463	0.458	0.293	0.754	0.179	0.846
homeruns	**0.812**	**0.812**	0.729	0.552	0.604	0.675	0.133	**0.812**	0.723	0.508	0.661	0.397	0.593	0.659
iceland_tourism	0.643	0.643	0.486	0.391	0.391	0.391	0.021	0.643	0.220	0.667	0.200	0.486	0.200	**0.947**
jfk_passengers	**0.776**	**0.776**	0.437	0.559	0.574	0.437	0.026	**0.776**	0.354	0.347	0.344	0.650	0.437	0.723
lga_passengers	0.422	0.438	**0.630**	0.297	0.606	0.392	0.054	0.438	0.366	0.348	0.498	0.395	0.524	0.535
measles	**0.947**	**0.947**	0.069	0.124	0.118	0.069	0.004	0.124	0.391	0.117	0.030	0.327	0.039	**0.947**
nile	**1.000**	**1.000**	**1.000**	0.800	**1.000**	**1.000**	0.040	**1.000**	0.824	0.452	**1.000**	**1.000**	**1.000**	0.824
ozone	0.531	0.650	0.650	0.531	0.750	0.723	0.109	**1.000**	0.723	0.559	0.375	**1.000**	0.113	0.723
quality_control_1	**1.000**	**1.000**	0.800	0.667	0.667	0.667	0.025	**1.000**	0.500	0.667	0.667	**1.000**	0.571	0.667
quality_control_2	**1.000**	**1.000**	**1.000**	0.667	**1.000**	**1.000**	0.028	**1.000**	0.545	0.667	**1.000**	**1.000**	**1.000**	0.750
quality_control_3	**1.000**	**1.000**	**1.000**	0.667	0.571	**1.000**	0.022	**1.000**	0.667	0.333	0.286	**1.000**	**1.000**	0.667
quality_control_4	0.810	**0.873**	0.658	0.438	0.608	0.393	0.028	0.726	0.335	0.360	0.233	0.726	0.246	0.780
quality_control_5	**1.000**	**1.000**	**1.000**	0.500	**1.000**	**1.000**	0.006	**1.000**	**1.000**	0.500	**1.000**	**1.000**	**1.000**	**1.000**
rail_lines	0.846	0.846	0.846	0.349	**0.966**	0.537	0.200	0.846	0.423	0.349	0.571	0.846	0.205	0.537
ratner_stock	0.776	**0.824**	0.783	0.500	0.306	0.282	0.034	0.650	0.280	0.559	0.203	**0.824**	0.250	0.571
robocalls	0.800	0.800	**0.966**	0.696	0.832	0.636	0.179	0.800	0.636	0.593	0.714	**0.966**	0.182	0.636
scanline_126007	0.710	0.739	0.820	0.767	**0.889**	0.732	0.101	0.759	0.644	F	0.649	0.732	0.616	0.644
scanline_42049	0.485	0.713	**0.962**	0.844	0.705	0.580	0.164	0.804	0.269	0.390	0.460	0.748	0.571	0.276
seatbelts	0.474	**0.683**	0.583	0.383	0.509	0.321	0.051	**0.683**	0.452	0.383	0.494	0.583	0.583	0.621
shanghai_license	**0.868**	**0.868**	0.713	0.491	0.437	0.698	0.048	**0.868**	0.532	0.326	0.357	0.713	0.208	0.636
uk_coal_employ	M	M	M	0.495	M	0.513	0.513	M	**0.639**	M	M	M	M	0.513
unemployment_nl	0.566	**0.876**	0.683	0.592	0.683	0.620	0.145	0.773	0.566	0.495	0.549	0.773	0.571	0.566
us_population	**1.000**	0.667	0.216	0.471	0.216	0.174	0.007	0.471	0.159	0.242	0.889	0.320	0.077	0.889
usd_isk	**0.785**	0.657	0.609	0.678	0.504	0.661	0.079	0.657	0.489	0.282	0.390	0.657	0.513	0.489
well_log	0.279	0.534	0.796	0.769	0.822	0.818	0.069	0.555	0.149	0.578	**0.923**	0.485	0.724	0.237
apple			0.513	0.318		0.513	0.029			0.373				**0.594**
bee_waggle_6			0.121	0.481		0.124	0.010			0.218				**0.929**
occupancy			**0.807**	0.496		0.750	0.107			0.290				0.341
run_log			**1.000**	0.420		0.792	0.139			0.543				0.446

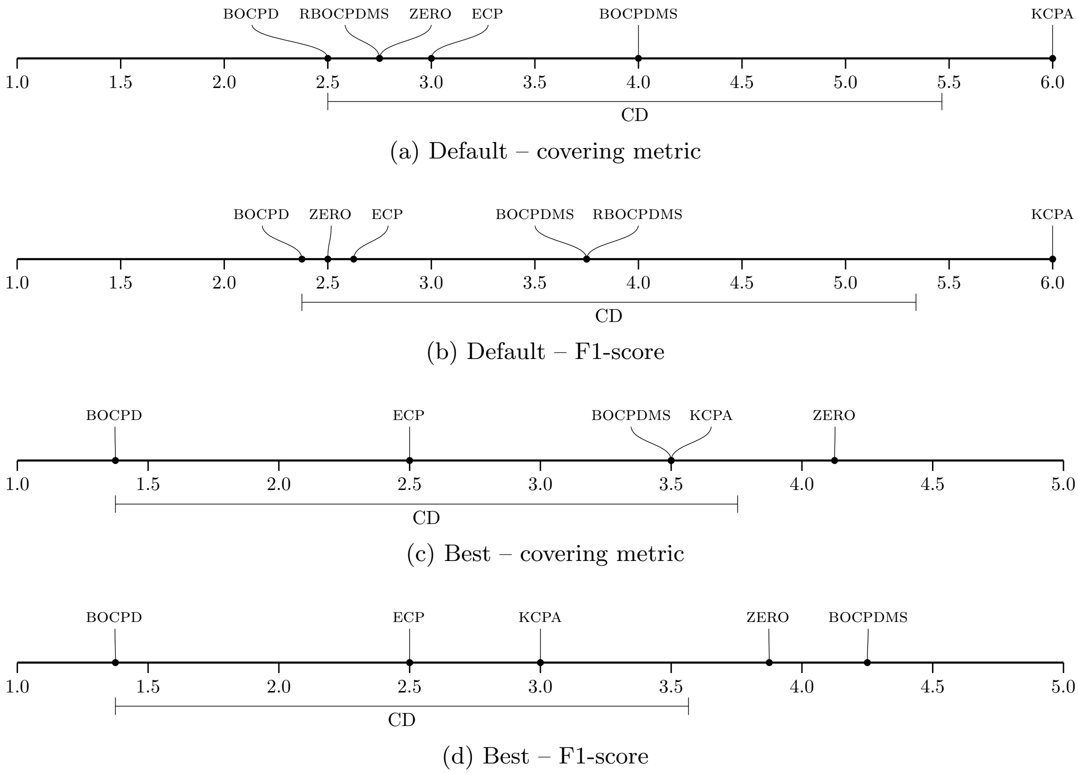
Table 5: **Best-Covering.** Maximum segmentation covering score for each method on each of the time series after running the grid search. Values are unavailable either due to a failure (F), because of missing values in the series unsupported by the method (M), the method timing out (T), or because the method is not suited for multidimensional series (blank). Highest values for each time series are shown in bold.

Dataset	AMOC	BINSEG	BOCPD	BOCPDMS	CPNP	ECP	KCPA	PELT	PROPHET	RBOCPDMS	RFPOP	SEGNEIGH	WBS	ZERO
bank	**1.000**	**1.000**	**1.000**	0.997	0.103	0.238	0.509	0.509	**1.000**	T	0.036	**1.000**	0.053	**1.000**
bitcoin	0.771	0.771	**0.822**	0.773	0.364	0.772	0.778	0.794	0.723	T	0.168	0.771	0.409	0.516
brent_spot	0.503	0.650	**0.667**	0.265	0.437	0.653	0.571	0.659	0.527	T	0.235	0.659	0.310	0.266
businv	0.574	0.562	**0.693**	0.459	0.402	0.690	0.405	0.647	0.539	0.459	0.123	0.647	0.154	0.461
centralia	0.675	0.675	**0.753**	0.650	0.675	**0.753**	0.675	0.675	0.675	0.624	0.568	0.675	0.253	0.675
children_per_woman	**0.804**	0.798	0.801	0.427	0.486	0.718	0.613	0.798	0.521	0.715	0.154	0.798	0.284	0.429
co2_canada	0.527	0.747	**0.773**	0.584	0.528	0.751	0.739	0.705	0.605	0.617	0.497	0.752	0.552	0.278
construction	**0.629**	0.575	0.585	0.571	0.375	0.524	0.395	0.423	0.561	0.575	0.154	0.575	0.300	0.575
debt_ireland	0.635	0.717	0.688	0.729	0.777	**0.798**	0.747	0.777	0.321	0.396	0.489	0.777	0.248	0.321
gdp_argentina	**0.737**	**0.737**	**0.737**	0.711	0.461	**0.737**	0.367	0.667	0.592	0.711	0.346	**0.737**	0.326	**0.737**
gdp_croatia	**0.708**	**0.708**	**0.708**	0.675	**0.708**	**0.708**	0.623	**0.708**	**0.708**	**0.708**	0.353	**0.708**	0.108	**0.708**
gdp_iran	0.583	0.583	0.583	0.611	0.448	0.583	0.505	0.505	0.583	**0.692**	0.248	0.583	0.295	0.583
gdp_japan	**0.802**	**0.802**	**0.802**	0.777	0.522	**0.802**	0.525	**0.802**	**0.802**	**0.802**	0.283	**0.802**	0.283	**0.802**
global_co2	**0.758**	**0.758**	**0.758**	0.745	0.381	0.665	0.602	0.743	0.284	0.745	0.368	**0.758**	0.338	**0.758**
homeruns	**0.694**	0.683	**0.694**	0.681	0.537	**0.694**	0.501	0.683	0.575	0.506	0.392	0.683	0.407	0.511
iceland_tourism	**0.946**	**0.946**	**0.946**	0.936	0.393	0.842	0.655	0.830	0.498	0.936	0.293	**0.946**	0.512	**0.946**
jfk_passengers	0.844	0.839	0.837	**0.855**	0.583	0.807	0.563	0.839	0.373	T	0.393	0.839	0.514	0.630
lga_passengers	0.434	0.541	0.547	0.475	0.478	**0.653**	0.536	0.534	0.446	T	0.426	0.543	0.501	0.383
measles	**0.951**	**0.951**	**0.951**	0.950	0.098	0.105	0.400	0.232	0.616	F/T	0.046	**0.951**	0.084	**0.951**
nile	**0.888**	0.880	**0.888**	0.870	0.880	**0.888**	0.758	0.880	0.758	0.876	0.880	0.880	0.880	0.758
ozone	0.721	0.600	0.602	0.735	0.458	0.676	0.451	0.635	0.574	**0.755**	0.358	0.627	0.309	0.574
quality_control_1	**0.996**	0.992	**0.996**	0.990	0.721	0.989	0.620	0.992	0.693	0.780	0.706	0.992	0.687	0.503
quality_control_2	0.927	0.922	**0.927**	0.921	0.922	0.927	**0.927**	0.922	0.723	0.637	0.922	0.922	0.922	0.638
quality_control_3	**0.997**	0.996	**0.997**	0.987	0.831	**0.997**	**0.997**	0.996	0.500	T	0.978	0.996	0.996	0.500
quality_control_4	**0.742**	0.673	0.673	0.670	0.508	0.540	0.535	0.673	0.673	T	0.077	0.673	0.506	0.673
quality_control_5	**1.000**	**1.000**	**1.000**	0.994	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.994	**1.000**	**1.000**	**1.000**	**1.000**
rail_lines	0.786	0.786	0.768	0.769	**0.786**	0.768	0.773	0.786	0.534	0.769	0.482	0.786	0.103	0.428
ratner_stock	0.874	0.914	0.906	0.872	0.382	0.874	0.771	**0.914**	0.444	T	0.162	**0.914**	0.182	0.450
robocalls	0.666	0.788	**0.808**	0.741	0.677	**0.808**	**0.808**	0.760	0.601	0.682	0.569	0.760	0.559	0.601
scanline_126007	0.634	0.633	0.631	**0.677**	0.433	0.390	0.494	0.444	0.503	T	0.316	0.633	0.329	0.503
scanline_42049	0.425	0.861	**0.892**	0.875	0.577	0.862	0.860	0.862	0.441	T	0.257	0.862	0.690	0.211
seatbelts	0.683	0.797	0.800	0.630	**0.809**	0.800	0.702	0.797	0.635	0.533	0.660	0.797	0.727	0.528
shanghai_license	**0.930**	0.920	0.920	0.911	0.474	0.920	0.497	0.920	0.804	0.763	0.381	0.920	0.351	0.547
uk_coal_employ	M	M	M	**0.504**	M	0.356	0.356	M	0.481	M	M	M	M	0.356
unemployment_nl	0.627	0.669	**0.669**	0.572	0.503	0.501	0.476	0.650	0.507	F/T	0.246	0.650	0.447	0.507
us_population	**0.803**	**0.803**	0.737	0.801	0.391	0.508	0.271	0.506	0.135	T	**0.803**	**0.803**	0.050	**0.803**
usd_isk	0.865	0.764	0.853	**0.866**	0.540	0.858	0.714	0.791	0.436	0.770	0.166	0.791	0.401	0.436
well_log	0.463	0.828	0.793	0.769	0.779	**0.846**	0.798	0.782	0.411	T	0.787	0.782	0.768	0.225
apple			**0.846**	0.720		0.758	0.462			F/T				0.425
bee_waggle_6			**0.891**	0.889		0.116	0.730			0.205				**0.891**
occupancy			0.645	0.556		**0.666**	0.581			F/T				0.236
run_log			**0.824**	0.327		0.819	0.729			0.329				0.304

Table 6: **Best-F1.** Maximum F1-score for each method on each of the time series after running the grid search. Values are unavailable either due to a failure (F), because of missing values in the series unsupported by the method (M), the method timing out (T), or because the method is not suited for multidimensional series (blank). Highest values for each time series are shown in bold.

Dataset	AMOC	BINSEG	BOCPD	BOCPDMS	CPNP	ECP	KCPA	PELT	PROPHET	RBOCPDMS	RFPOP	SEGNEIGH	WBS	ZERO
bank	**1.000**	**1.000**	**1.000**	0.500	0.054	0.200	0.333	0.400	**1.000**	T	0.015	**1.000**	0.043	**1.000**
bitcoin	0.507	0.690	0.733	0.533	0.611	0.625	0.665	**0.735**	0.446	T	0.284	**0.735**	0.690	0.450
brent_spot	0.465	**0.670**	0.609	0.239	0.607	0.636	0.553	0.586	0.249	T	0.521	0.586	0.564	0.315
businv	**0.588**	**0.588**	**0.588**	0.455	0.386	0.370	0.294	0.490	0.275	0.370	0.261	**0.588**	0.289	**0.588**
centralia	0.909	**1.000**	**1.000**	**1.000**	**1.000**	0.909	**1.000**	**1.000**	0.763	0.846	**1.000**	**1.000**	0.556	0.763
children_per_woman	0.678	0.663	**0.712**	0.405	0.344	0.551	0.525	0.637	0.310	0.504	0.246	0.637	0.500	0.507
co2_canada	0.544	0.856	**0.924**	0.479	0.642	0.875	0.867	0.670	0.482	0.542	0.569	0.872	0.681	0.361
construction	0.696	**0.709**	**0.709**	0.410	0.602	**0.709**	0.634	**0.709**	0.324	0.340	0.185	**0.709**	0.523	0.696
debt_ireland	0.760	**1.000**	**1.000**	0.892	0.958	0.980	**1.000**	**1.000**	0.469	0.748	0.824	**1.000**	0.538	0.469
gdp_argentina	0.889	**0.947**	**0.947**	0.583	0.818	0.889	0.800	**0.947**	0.615	0.452	0.615	**0.947**	0.421	0.824
gdp_croatia	**1.000**	0.824	**1.000**	0.583	**1.000**	0.824	0.583	0.824	0.824	0.824	0.400	0.824	0.167	0.824
gdp_iran	0.696	0.652	**0.862**	0.492	0.620	0.824	0.734	0.808	0.652	0.737	0.636	0.808	0.576	0.652
gdp_japan	**1.000**	0.889	**1.000**	0.615	0.667	**1.000**	0.500	0.889	0.889	0.889	0.222	0.889	0.222	0.889
global_co2	0.929	0.929	0.889	0.458	0.667	**0.929**	0.667	0.929	0.463	0.547	0.293	0.929	0.250	0.846
homeruns	0.812	**0.829**	**0.829**	0.650	0.650	0.829	0.829	0.812	0.723	0.397	0.661	0.812	0.664	0.659
iceland_tourism	0.947	0.947	0.947	0.486	0.391	**1.000**	0.486	0.643	0.220	0.667	0.200	0.947	0.200	0.947
jfk_passengers	**0.776**	**0.776**	**0.776**	0.650	0.602	0.651	0.437	**0.776**	0.354	T	0.491	**0.776**	0.437	0.723
lga_passengers	0.561	0.620	0.704	0.563	0.606	**0.892**	0.526	0.537	0.366	T	0.592	0.537	0.674	0.535
measles	**0.947**	**0.947**	**0.947**	0.486	0.118	0.080	0.281	0.153	0.391	F/T	0.030	**0.947**	0.041	**0.947**
nile	**1.000**	**1.000**	**1.000**	0.800	**1.000**	**1.000**	0.824	**1.000**	0.824	0.667	**1.000**	**1.000**	**1.000**	0.824
ozone	0.776	0.723	0.857	0.778	0.750	**1.000**	0.667	**1.000**	0.723	0.651	0.429	**1.000**	0.286	0.723
quality_control_1	**1.000**	**1.000**	**1.000**	0.667	0.667	**1.000**	0.667	**1.000**	0.500	0.286	0.667	**1.000**	0.667	0.667
quality_control_2	**1.000**	**1.000**	**1.000**	0.667	**1.000**	**1.000**	**1.000**	**1.000**	0.750	0.429	**1.000**	**1.000**	**1.000**	0.750
quality_control_3	**1.000**	**1.000**	**1.000**	0.766	0.571	**1.000**	**1.000**	**1.000**	0.667	T	0.800	**1.000**	**1.000**	0.667
quality_control_4	0.810	**0.873**	0.787	0.561	0.658	0.726	0.658	0.780	0.780	T	0.241	0.780	0.608	0.780
quality_control_5	**1.000**	**1.000**	**1.000**	0.500	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.500	**1.000**	**1.000**	**1.000**	**1.000**
rail_lines	0.846	0.846	**0.966**	0.889	**0.966**	**0.966**	0.800	0.846	0.537	0.730	0.615	0.889	0.205	0.537
ratner_stock	0.776	0.824	**0.868**	0.559	0.396	0.776	0.754	0.824	0.280	T	0.203	0.824	0.378	0.571
robocalls	0.800	**0.966**	**0.966**	0.750	0.862	**0.966**	**0.966**	**0.966**	0.636	0.846	0.714	**0.966**	0.714	0.636
scanline_126007	0.710	0.920	**0.921**	0.829	0.906	0.870	0.838	0.889	0.644	T	0.649	0.889	0.818	0.644
scanline_42049	0.485	0.879	**0.962**	0.889	0.713	0.910	0.908	0.910	0.269	T	0.460	0.910	0.650	0.276
seatbelts	0.824	**0.838**	0.683	0.583	0.735	0.683	0.621	0.683	0.452	0.383	0.563	0.735	0.583	0.621
shanghai_license	**0.966**	0.868	0.868	0.605	0.600	0.868	0.465	0.868	0.532	0.389	0.357	0.868	0.385	0.636
uk_coal_employ	M	M	M	0.617	M	0.513	0.513	M	**0.639**	M	M	M	M	0.513
unemployment_nl	0.742	**0.889**	0.876	0.592	0.747	0.755	0.744	0.788	0.566	F/T	0.628	0.788	0.801	0.566
us_population	**1.000**	0.889	**1.000**	0.615	0.232	0.471	0.276	0.500	0.159	T	0.889	0.889	0.113	0.889
usd_isk	**0.785**	0.704	**0.785**	0.678	0.674	**0.785**	0.601	0.657	0.489	0.510	0.462	0.678	0.636	0.489
well_log	0.336	0.914	0.832	0.743	0.822	**0.928**	0.776	0.873	0.149	T	0.923	0.873	0.832	0.237
apple			**0.916**	0.445		0.745	0.634			F/T				0.594
bee_waggle_6			**0.929**	0.481		0.233	0.634			0.245				**0.929**
occupancy			0.919	0.735		**0.932**	0.812			F/T				0.341
run_log			**1.000**	0.469		0.990	0.909			0.380				0.446

B.2 Multivariate rank plots



(a) Default – covering metric

(b) Default – F1-score

(c) Best – covering metric

(d) Best – F1-score

Figure 4: Average ranks of the algorithms for both experiments and both metrics for the multivariate time series. RBOCPDMS is excluded in the Best experiment due to the limit on computation time (see Appendix A). CD denotes the critical difference interval, outside of which methods perform significantly worse than the best method. Lower ranks are better.

C Dataset Overview

In the following pages we show the time series included in the change point data set, as well as the provided annotations. Annotations are marked on the graphs by a dashed vertical line with one or more triangles on the horizontal axis that correspond to the number of annotators that marked that point as a change point. The colors of the triangles correspond to each of the five annotators for the time series. The number in the box on the left side of the figure is the number of annotators who believed there were no change points in the time series. A brief description including the source of the series is provided.[9]

C.1 Real Time Series

This section lists the real time series from various sources and application domains.

[9]See `https://github.com/alan-turing-institute/TCPD` for the complete data set and further information on each individual time series.

Figure 5: 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. Data retrieved from Yahoo Finance.

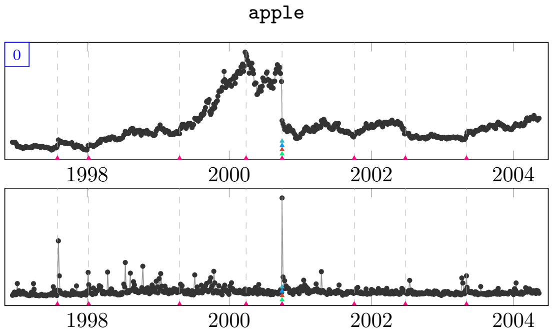


Figure 6: The amount of money in someone's current account. Significant changes occur, but the series is also periodic.

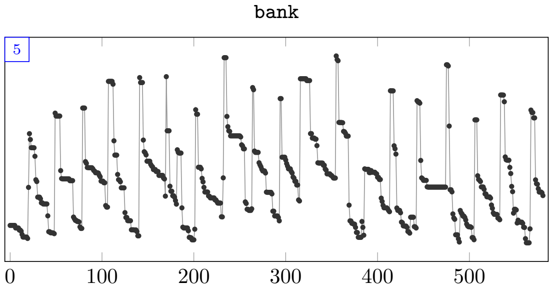


Figure 7: Honey bee movement switches between three states: left turn, right turn, and waggle. This series contains the x position, y position, and the sine and cosine of the head angle of a single bee. Data obtained from Oh et al. (2008).

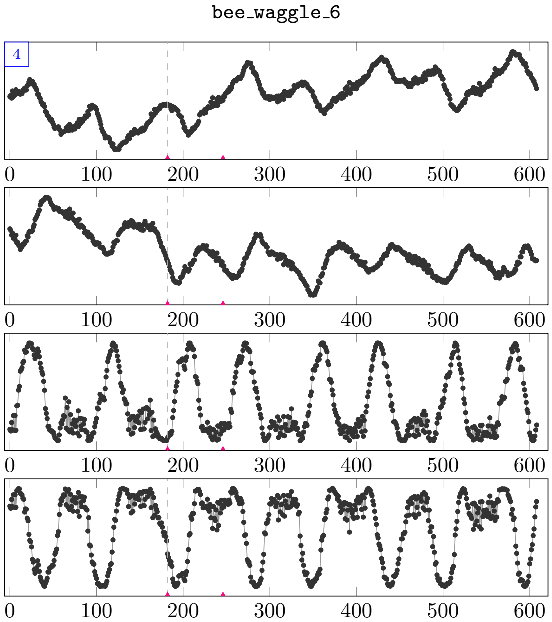


Figure 8: The price of Bitcoin in USD. Data obtained from Blockchain.com.

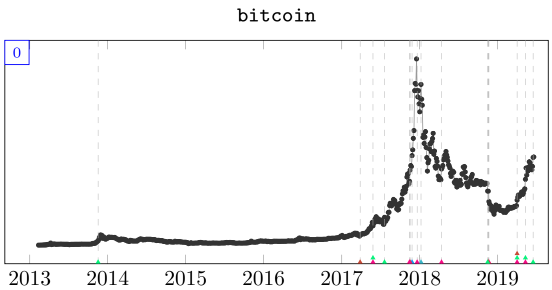
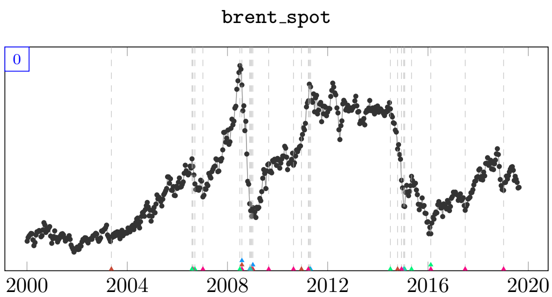


Figure 9: The price of Brent Crude oil in USD per barrel. Significant changes appear around 2008 and 2014. The data is sampled at every 10 original observations to reduce the length of the series. Obtained from the U.S. Energy Information Administration.



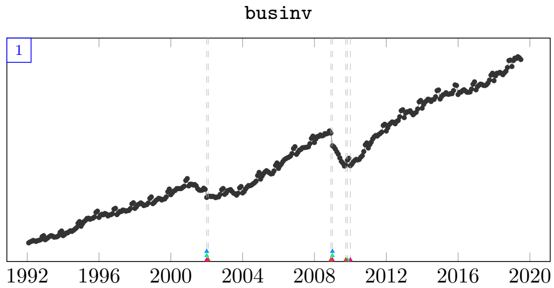


Figure 10: Monthly total business inventories (in USD), obtained from the U.S. Census Bureau. Effects of the financial crisis are visible. The non-seasonally-adjusted series is used to maintain the periodic component.

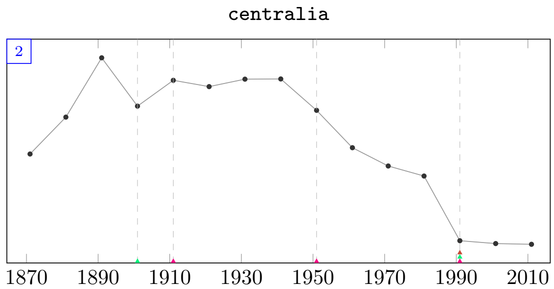


Figure 11: The population of the mining town of Centralia, Pennsylvania, U.S., where a mine fire has been burning since 1962. Eminent domain was invoked in 1992, condemning all buildings in the town. Data obtained from Wikipedia.

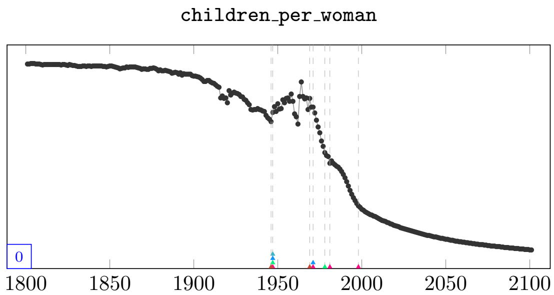


Figure 12: The global average number of children born per woman. Data obtained from GapMinder.

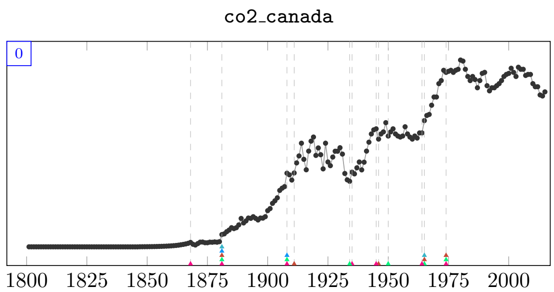


Figure 13: CO_2 emissions per person in Canada. Data obtained from GapMinder.

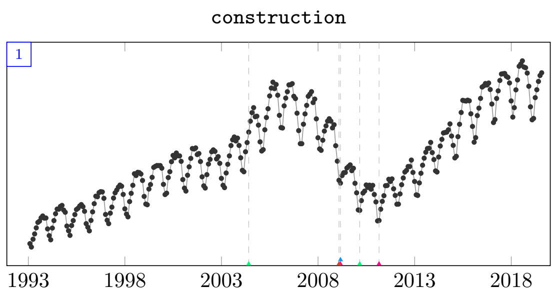


Figure 14: Total private construction spending in the U.S. Data obtained from the U.S. Census Bureau.

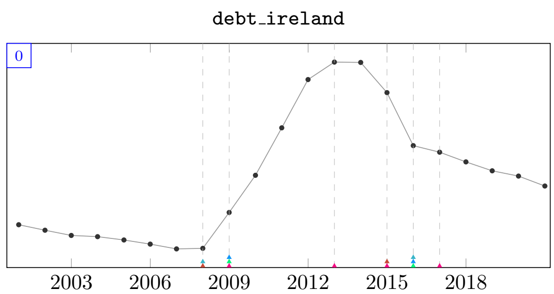


Figure 15: The government debt ratio of Ireland. Effects of the financial crisis of 2007-2008 are visible. Data obtained from EuroStat.

Figure 16: The GDP of Argentina in constant local currency. Data obtained from the World Bank.

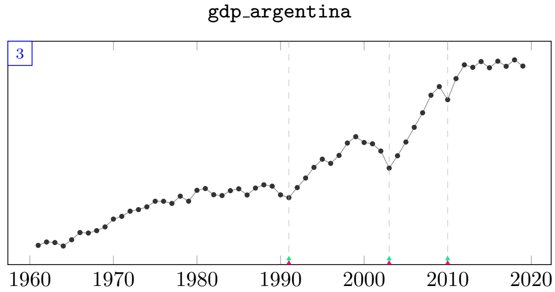


Figure 17: The GDP of Croatia in constant local currency. Data obtained from the World Bank.

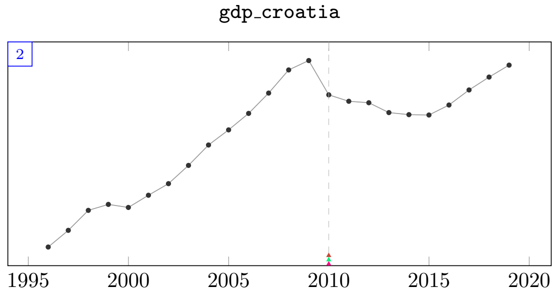


Figure 18: The GDP of Iran in constant local currency. Data obtained from the World Bank.

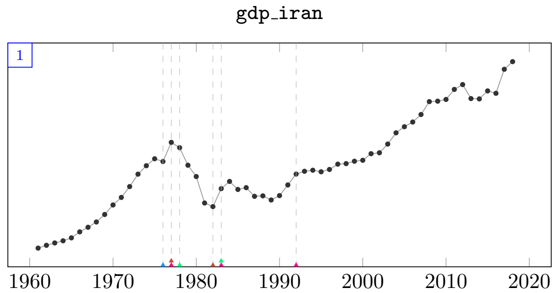


Figure 19: The GDP of Japan in constant local currency. Data obtained from the World Bank.

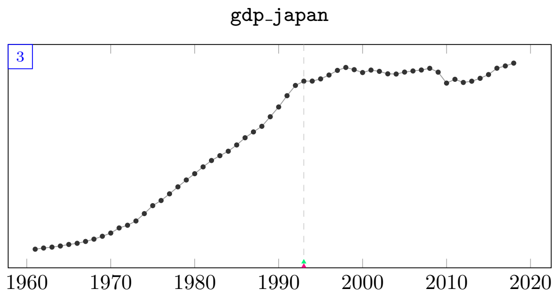


Figure 20: Monthly Global CO_2 levels. The series has been sampled every 48 months and cropped to recent history to reduce the length of the series. Data obtained from Meinshausen et al. (2017).

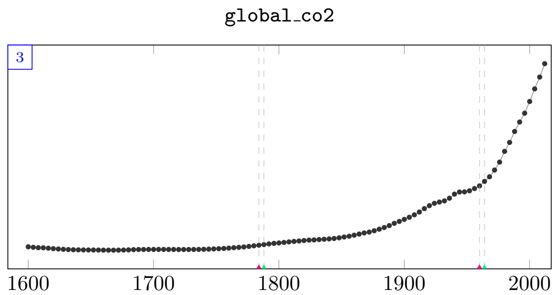


Figure 21: The number of home runs in the American League of baseball by year. Potentially significant events are the second world war and the Major League Baseball expansions. Data retrieved from the Baseball Databank.

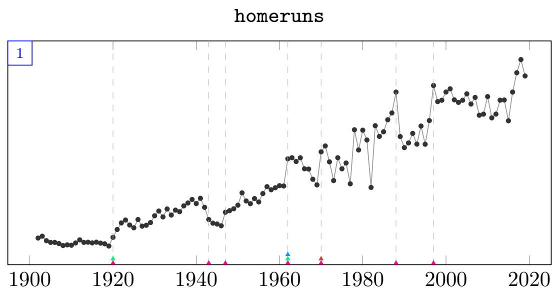


Figure 22: Monthly visitors to Iceland through Keflavik airport. Data obtained from the Icelandic Tourist Board.

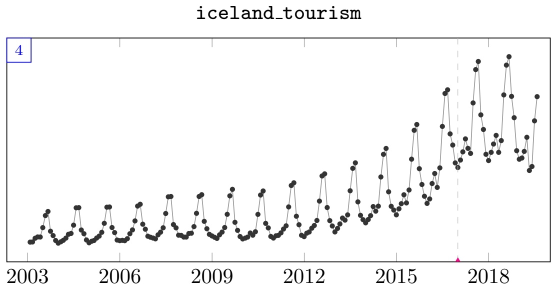


Figure 23: The number of passengers arriving and departing at John F. Kennedy airport in New York City. Data obtained from the Port Authority of New York and New Jersey.

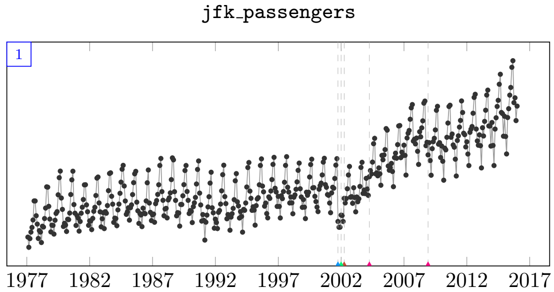


Figure 24: The number of passengers arriving and departing at LaGuardia airport in New York City. Data obtained from the Port Authority of New York and New Jersey.

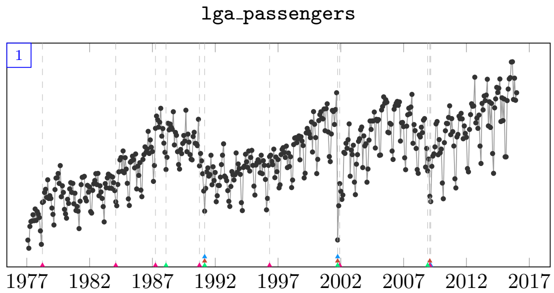


Figure 25: Number of measles cases in England and Wales over time. Data obtained from `https://ms.mcmaster.ca/~bolker`.

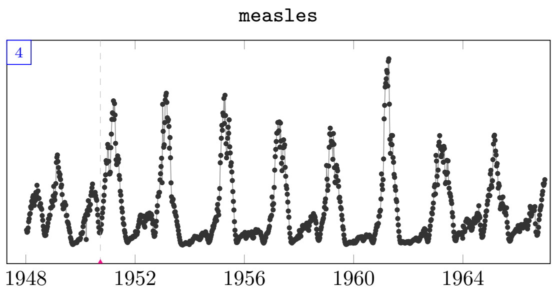
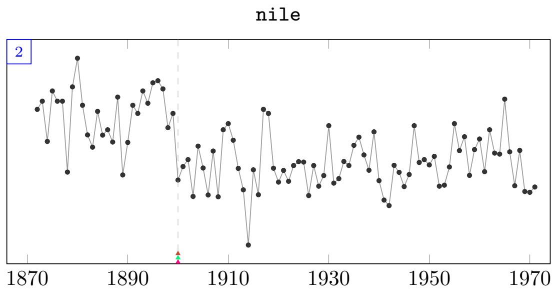


Figure 26: Yearly volume of the Nile river at Aswan. A dam was built in 1898. Data obtained from the website for the book by Durbin and Koopman (2012). This data set differs from the Nile data set for the period 622 to 1284 AD, used in other change point work.



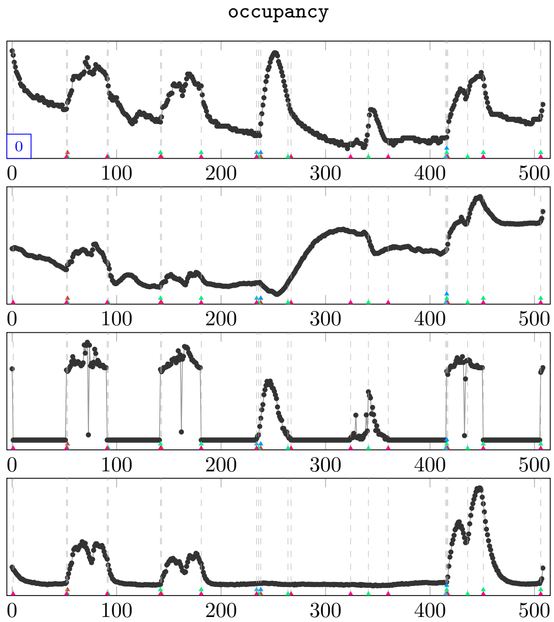


Figure 27: Room occupancy time series by Candanedo and Feldheim (2016) obtained from the UCI repository (Bache and Lichman, 2013). The four dimensions correspond to measurements of temperature, relative humidity, light, and CO_2. The original data has been sampled at every 16 observations to reduce the length of the series.

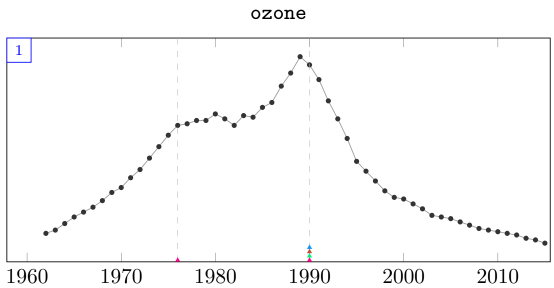


Figure 28: Levels of ozone-depleting substances in the atmosphere. The Montreal Protocol came into force in September 1989. Data from `www.ourworldindata.org` and originally due to Hegglin et al. (2015).

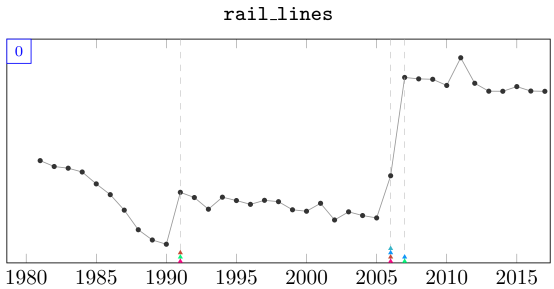


Figure 29: Total kilometers of rail lines in the world. Data obtained from the World Bank.

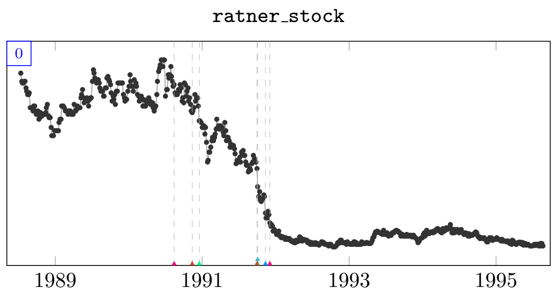


Figure 30: Stock price of the Ratner Group for a period in the early 1990s. Data obtained from Yahoo Finance and sampled every 3 observations to reduce the length of the series.

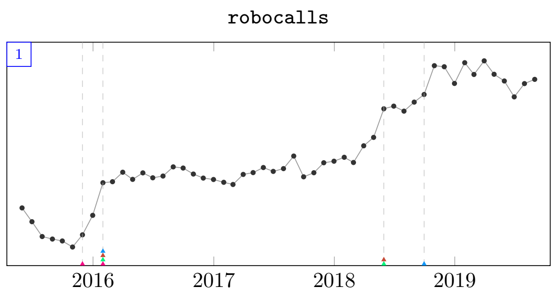


Figure 31: The number of automated phone calls in the U.S., which have been subject to varying degrees of regulation. Obtained from `www.robocallindex.com`.

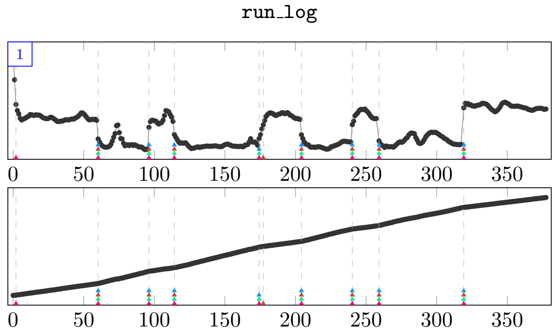


Figure 32: The pace and total distance traveled by a runner following an interval training program (alternating between running and walking). Data provided by the author.

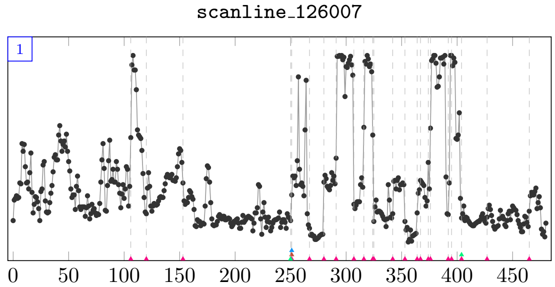


Figure 33: A horizontal scan line of image no. 126007 from the Berkeley Segmentation Dataset (Martin et al., 2001).

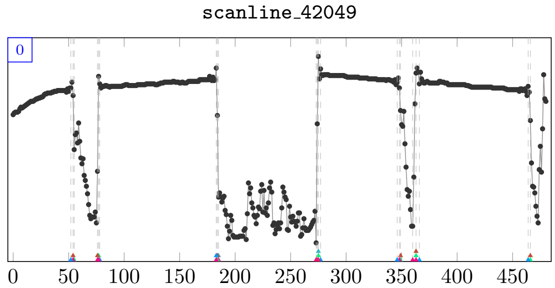


Figure 34: A horizontal scan line of image no. 42049 from the Berkeley Segmentation Dataset (Martin et al., 2001).

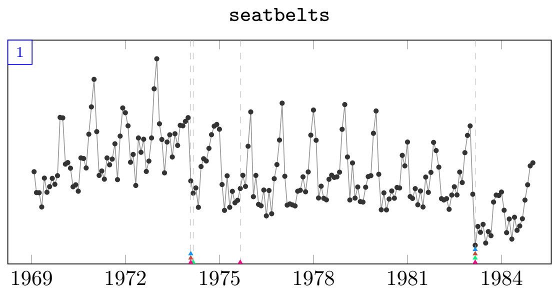


Figure 35: Number of drivers killed or seriously injured in the U.K. around the period of the introduction of seatbelts. Seatbelts were compulsory in new cars starting in 1972 and were mandatory to be worn from 1983 onwards. Obtained from the `datasets` package in R.

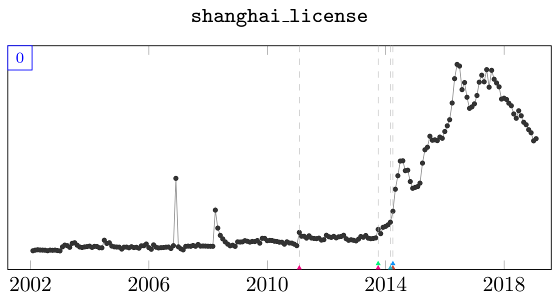


Figure 36: The number of applicants for a license plate in Shanghai. Data obtained from Kaggle.

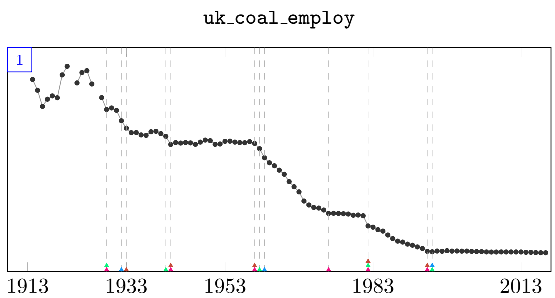


Figure 37: Number of workers employed in British coal mines. Data obtained from the U.K. government.

Figure 38: Unemployment rates in The Netherlands. Data obtained from Statistics Netherlands.

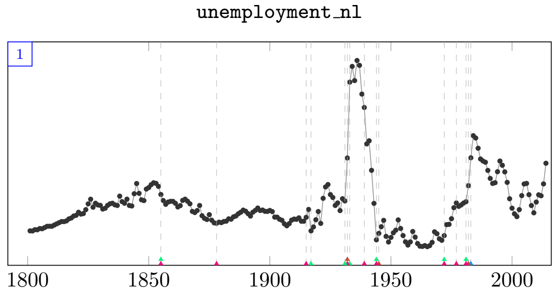


Figure 39: Exchange rate between the US Dollar and the Icelandic Króna in the years around the financial crisis. Data obtained from EuroStat.

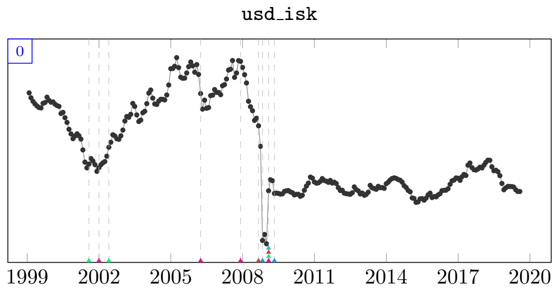


Figure 40: Population of the U.S. over time. Data obtained from the U.S. Census Bureau.

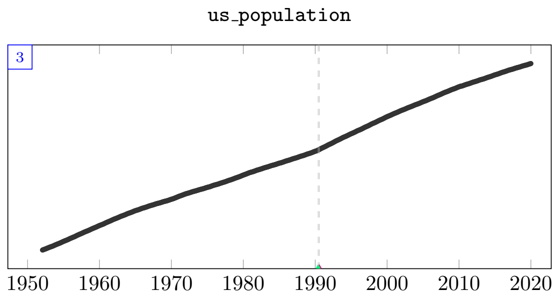
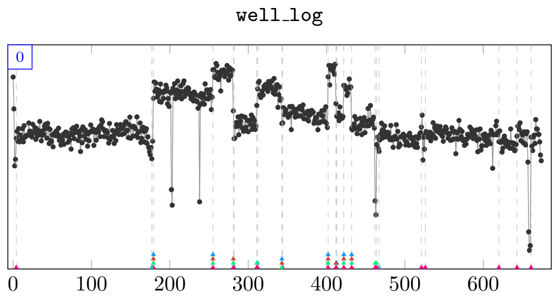


Figure 41: Well-log data from Ó Ruanaidh and Fitzgerald (1996). The series has been sampled every 6 iterations to reduce the length of the series.



C.2 Quality Control

This section lists the quality control data sets with known change points.

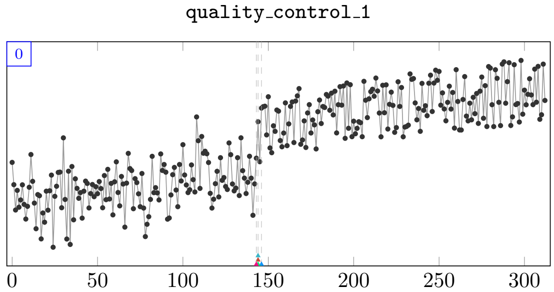


Figure 42: A series with a known change point at index 146. The series has Gaussian noise and a small trend before the change point and an offset and uniform noise after the change point.

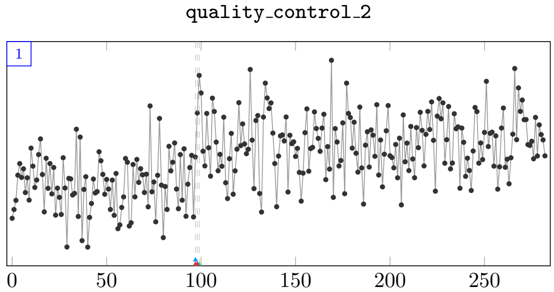


Figure 43: A series with a known change point at index 97. The series has constant noise and a mean shift change point.

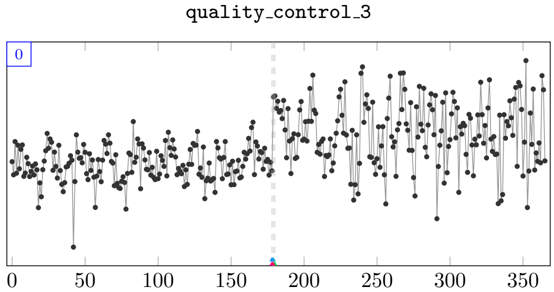


Figure 44: A series with a known change point at index 179. The series has noise $\mathcal{N}(0,1)$ before the change point and $\mathcal{N}(2,2)$ after the change point, as well as an outlier at index 42.

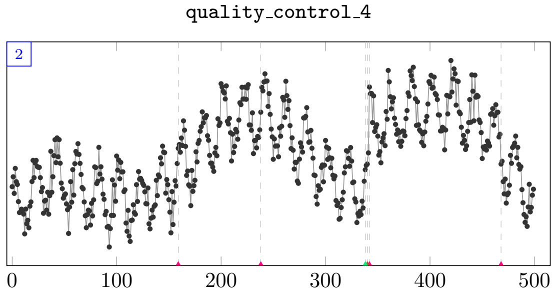


Figure 45: A series with a known change point at index 341. The series has multiple periodic components of different amplitude and a mean shift change point.

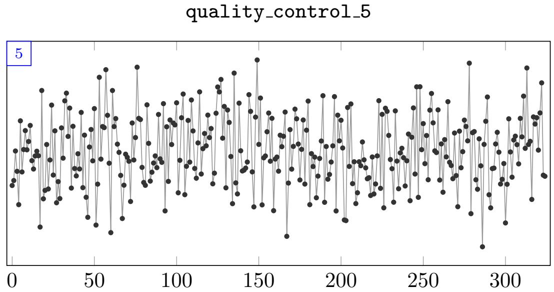


Figure 46: A series of $\mathcal{N}(0,1)$ noise without a change point.

C.3 Introductory Series

This section lists the synthetic time series shown to the annotators during the introduction. Since all annotators had to successfully complete the introduction before they could continue with real time series, the annotations are not shown for these series.

Figure 47: Series illustrating mean shift change point (index 50).

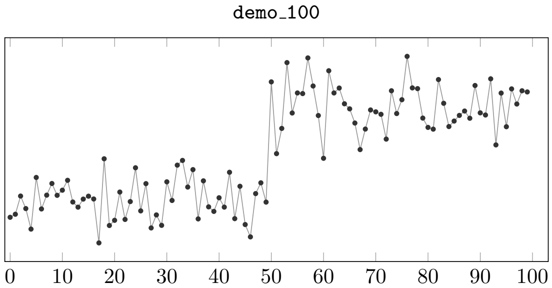


Figure 48: Series illustrating multiple change points (index 33 and 79).

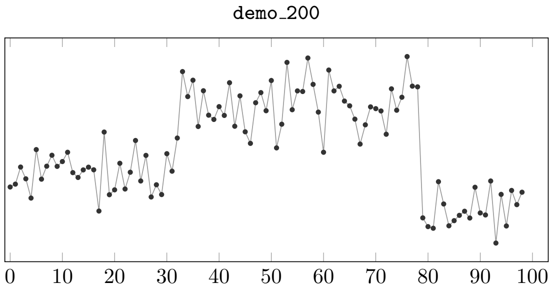


Figure 49: Series illustrating variance change (index 43).

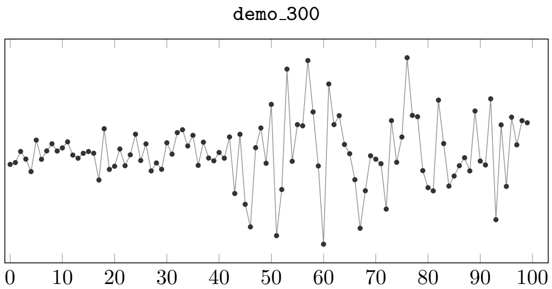


Figure 50: Series illustrating no change points.

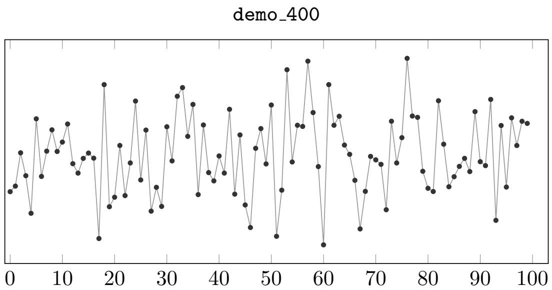


Figure 51: Series illustrating outliers and a mean shift change point (index 80).

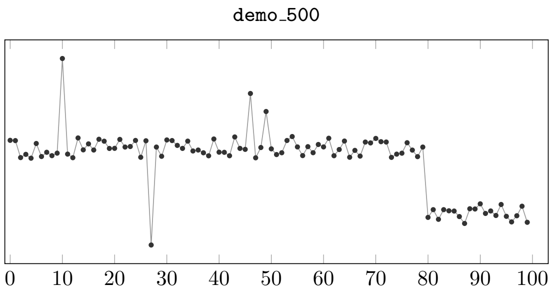


Figure 52: Series illustrating a trend change (index 65).

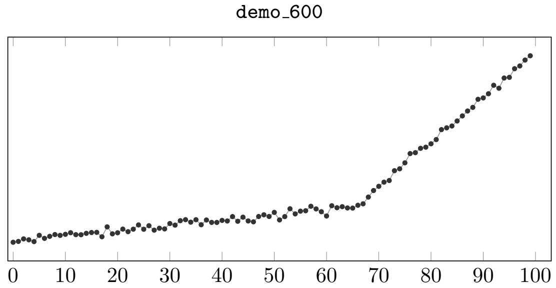


Figure 53: Series illustrating a random walk (no change point).

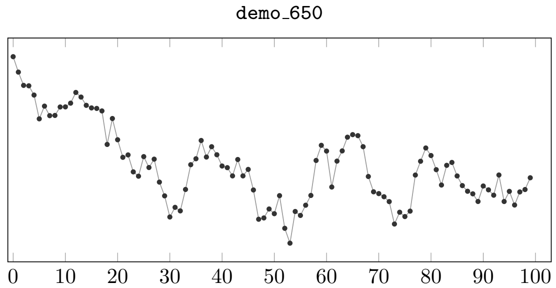


Figure 54: Series illustrating a change in periodicity (index 57).

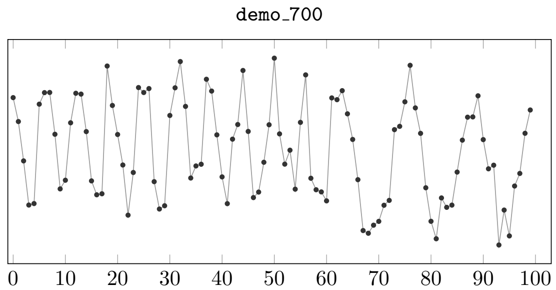
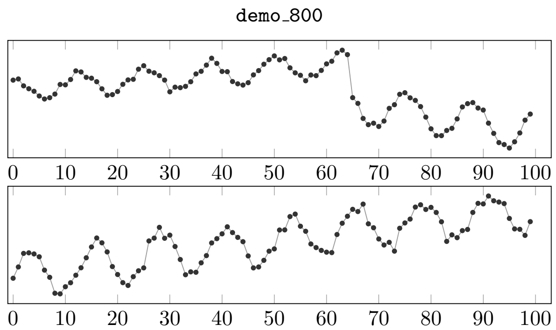


Figure 55: Series illustrating a multidimensional series with a change point in one dimension (index 65).



References

R. P. Adams and D. J. C. MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

S. Aminikhanghahi and D. J. Cook. A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51(2):339–367, 2017.

P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 33(5):898–916, 2010.

I. E. Auger and C. E. Lawrence. Algorithms for the optimal identification of segment neighborhoods. *Bulletin of Mathematical Biology*, 51(1):39–54, 1989.

K. Bache and M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

J. Bai and P. Perron. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1):1–22, 2003.

G. Boracchi, D. Carrera, C. Cervellera, and D. Macciò. QuantTree: Histograms for change detection in multivariate data streams. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 639–648, 2018.

L. M. Candanedo and V. Feldheim. Accurate occupancy detection of an office room from light, temperature, humidity and CO_2 measurements using statistical learning models. *Energy and Buildings*, 112:28–39, 2016.

G. C. Chow. Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, 28:591–605, 1960.

J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

J. Durbin and S. J. Koopman. *Time series analysis by state space methods*. Oxford University Press, 2012.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2): 303–338, 2010.

P. Fearnhead and Z. Liu. On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):589–605, 2007.

P. Fearnhead and G. Rigaill. Changepoint detection in the presence of outliers. *Journal of the American Statistical Association*, 114(525):169–183, 2019.

P. Fryzlewicz. Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281, 2014.

R. Garnett, M. A. Osborne, and S. J. Roberts. Sequential Bayesian prediction in the presence of changepoints. In *Proceedings of the 26th International Conference on Machine Learning*, pages 345–352. ACM, 2009.

Z. Harchaoui, E. Moulines, and F. R. Bach. Kernel change-point analysis. In *Advances in Neural Information Processing Systems*, pages 609–616, 2009.

F. Hausdorff. *Mengenlehre*. Walter de Gruyter Berlin, 1927.

K. Haynes, P. Fearnhead, and I. A. Eckley. A computationally efficient nonparametric approach for changepoint detection. *Statistics and Computing*, 27(5):1293–1305, 2017.

M. I. Hegglin, D. W. Fahey, M. McFarland, S. A. Montzka, and E. R. Nash. *Twenty Questions and Answers About the Ozone Layer: 2014 Update, Scientific Assessment of Ozone Depletion: 2014*. World Meteorological Organization, Geneva, Switzerland, 2015.

D. V. Hinkley. Inference about the change-point in a sequence of random variables. *Biometrika*, 57(1):1–17, 4 1970.

T. D. Hocking, G. Schleiermacher, I. Janoueix-Lerosey, V. Boeva, J. Cappo, O. Delattre, F. Bach, and J.-P. Vert. Learning smoothing models of copy number profiles using breakpoint annotations. *BMC Bioinformatics*, 14(1):164, 2013.

T. D. Hocking, V. Boeva, G. Rigaill, G. Schleiermacher, I. Janoueix-Lerosey, O. Delattre, W. Richer, F. Bourdeaut, M. Suguro, M. Seto, F. Bach, and J.-P. Vert. SegAnnDB: interactive web-based genomic segmentation. *Bioinformatics*, 30(11):1539–1546, 2014.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

B. Jackson, J. D. Scargle, D. Barnes, S. Arabhi, A. Alt, P. Gioumousis, E. Gwin, P. Sangtrakulcharoen, L. Tan, and T. T. Tsai. An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters*, 12(2):105–108, 2005.

N. A. James and D. S. Matteson. `ecp`: An R package for nonparametric multiple change point analysis of multivariate data. *arXiv preprint arXiv:1309.3295*, 2013.

R. Killick and I. Eckley. `changepoint`: An R package for changepoint analysis. *Journal of Statistical Software*, 58(3):1–19, 2014.

R. Killick, P. Fearnhead, and I. A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

J. Knoblauch and T. Damoulas. Spatio-temporal Bayesian on-line changepoint detection with model selection. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 2718–2727, 2018.

J. Knoblauch, J. E. Jewson, and T. Damoulas. Doubly robust Bayesian inference for non-stationary streaming data with β-divergences. In *Advances in Neural Information Processing Systems*, pages 64–75, 2018.

B. Kurt, Ç Yıldız, T. Y. Ceritli, B. Sankur, and A. T. Cemgil. A Bayesian change point model for detecting SIP-based DDoS attacks. *Digital Signal Processing*, 77:48–62, 2018.

M. Lavielle and G. Teyssiere. Adaptive detection of multiple change-points in asset price volatility. In *Long Memory in Economics*, pages 129–156. Springer, 2007.

G. Lorden. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42(6):1897–1908, 1971.

R. Maidstone, T. Hocking, G. Rigaill, and P. Fearnhead. On optimal multiple changepoint algorithms for large data. *Statistics and Computing*, 27(2):519–533, 2017.

D. R. Martin, C. C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the 8th IEEE International Conference on Computer Vision*, volume 2, pages 416–423, 2001.

D. R. Martin, C. C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 26(5):530–549, 2004.

D. S. Matteson and N. A. James. A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109(505):334–345, 2014.

M. Meinshausen, E. Vogel, A. Nauels, K. Lorbacher, N. Meinshausen, D. M. Etheridge, P. J. Fraser, S. A. Montzka, P. J. Rayner, C. M. Trudinger, P. B. Krummel, U. Beyerle, J. G. Canadell, J. S. Daniel, I. G. Enting, R. M. Law, C. R. Lunder, S. O'Doherty, R. G. Prinn, S. Reimann, M. Rubino, G. J. M. Velders, M. K. Vollmer, R. H. J. Wang, and R. Weiss. Historical greenhouse gas concentrations for climate modelling (CMIP6). *Geoscientific Model Development*, 10:2057–2116, 2017.

S. M. Oh, J. M. Rehg, T. Balch, and F. Dellaert. Learning and inferring motion patterns using parametric segmental switching linear dynamic systems. *International Journal of Computer Vision*, 77(1-3):103–124, 2008.

J. J. K. Ó Ruanaidh and W. J. Fitzgerald. *Numerical Bayesian Methods Applied to Signal Processing*. Springer, 1996.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

A. Pagotto. *ocp: Bayesian Online Change Point Detection*, 2019. R package version 0.1.1.

G. Rigaill. A pruned dynamic programming algorithm to recover the best segmentations with 1 to K_{max} change-points. *Journal de la Société Française de Statistique*, 156(4):180–205, 2015.

Y. Saatçi, R. Turner, and C. E. Rasmussen. Gaussian process change point models. In *Proceedings of the 27th International Conference on Machine Learning*, pages 927–934, 2010.

A. J. Scott and M. Knott. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, 30(3):507–512, 1974.

S. J. Taylor and B. Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018.

C. Truong, L. Oudre, and N. Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167, 2020.

R. Turner, Y. Saatçi, and C. E. Rasmussen. Adaptive sequential Bayesian change point detection. In Zaid Harchaoui, editor, *Temporal Segmentation Workshop at NIPS 2009*, Whistler, BC, Canada, 2009.

C. J. Van Rijsbergen. *Information Retrieval*. Butterworths, 1979.

N. R. Zhang and D. O. Siegmund. A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32, 2007.