Karlsruher Institut für Technologie

Karlsruhe Reports in Informatics 2015,9
Edited by Karlsruhe Institute of Technology,
Faculty of Informatics
ISSN 2190-4782

An Evaluation of Combinations of Lossy Compression and Change-Detection Approaches (Technical Report '15)

Gregor Hollmig, Matthias Horne, Simon Leimkühler, Frederik Schöll,
Carsten Strunk, Pavel Efros, Erik Buchmann, Klemens Böhm

2015

Karlsruher Institut für Technologie

KIT – University of the State of Baden-Wuerttemberg and National Research Center of the Helmholtz Association

An Evaluation of Combinations of Lossy Compression and Change-Detection Approaches (Technical Report '15)

Gregor Hollmig Matthias Horne Simon Leimkühler

Frederik Schöll Carsten Strunk Pavel Efros

Erik Buchmann Klemens Böhm

Karlsruhe Institute of Technology (KIT), Germany
{gregor.hollmig, matthias.horne, simon.leimkuehler, carsten.strunk}@student.kit.edu
schoell@ira.uka.de
{pavel.efros, erik.buchmann, klemens.boehm}@kit.edu

ABSTRACT

Today, time series of numerical data are ubiquitous, for instance in the Internet of Things. In such scenarios, it is often necessary to compress the data and to detect changes on it. More specifically, both methods are used *in combination*, i.e., data is lossily compressed and later decompressed, and then change detection takes place. There exists a broad variety of compression as well as of change-detection techniques. This calls for a systematic comparison of different combinations of compression and change-detection techniques, for different data sets, together with recommendations on how the values of the various (typically non-linear) parameters should be chosen. This article is such an evaluation. Its design is not trivial, necessitating a number of decisions. We work out the details and the rationale behind our design choices. Next to other results, our study shows that the choice of combinations of change detection and compression algorithm and their parameterization does affect result quality significantly. Our evaluation also indicates that results are highly contingent on the nature of the data.

1. INTRODUCTION

Nowadays, time-series data is ubiquitous. More and more applications like the Smart Grid or the Internet of Things that produce and/or process time-series data are proliferating. Such data is often used to detect certain events and to react to them as soon as possible, i.e., change-detection methods are indispensable. On the other hand, because of the many devices generating data, the huge amount of data and the high data-transfer rates, an efficient compression is essential. Lossy compression yields compression rates that applications can cope with in many situations. In this study, we focus on this kind of compression. Putting things together, it often is necessary to combine lossy compression[1] and change-detection techniques.

Example 1. *Smart meters may deliver data to a central analysis system via a wireless network. To save bandwidth and to reduce costs, the data is compressed directly on the device. The central data-analysis system can then do change detection to react to events such as a sudden increase in overall power consumption.*

When combining lossy compression and change detection, several issues arise. First, lossy compression introduces errors. In particular, changes can be lost, or new false changes can occur. Therefore a lossy compression method must be chosen which preserves the change information as much as possible. Furthermore, different use cases generate different kinds of time-series data, as we will explain. Thus it is necessary to choose a good combination of compression and change-detection technique *per use case*. This is difficult due to the large number of possible combinations. Next, compression as well as change-detection algorithms usually have several parameters, which often have non-obvious effects on the outcome. The expectation typically is that domain experts select the parameter values. This means that these experts must have a deep insight into the algorithms used. But even if they have selected the values, it is hard to determine whether their selection is a good one. To investigate how combinations of compression and change-detection algorithms perform on different datasets, a systematic comparison is necessary. This article is such a study.

Designing our study has been challenging, partly due to the issues just mentioned. To illustrate, one of the various design decisions is as follows: It is difficult to choose the parameterization of the compression and the change-detection algorithms such that the comparison is fair. Reusing the parameter values suggested in the original publications may not be the best option. This is because proper choices of

[1]For improved readability we usually refer to compression and later decompression simply as compression.

parameter values depend on the data the algorithms are applied to. Thus, we have decided to perform an optimization on each dataset that yields the parameter values that give way to change-detection results after compression that are closest to some carefully chosen reference point. This article lists the design questions encountered in the context of our comparison, together with explanations behind our choices.

In line with these design decisions, we have implemented a framework that can be used for the evaluation of virtually any combination of compression and change-detection methods. In our specific study, we examine five compression algorithms like APCA [16] and five change-detection algorithms like Online-Kernel Change Detection [9] on five datasets, resulting in 125 possible combinations. We focus on result quality and leave aside criteria such as runtime performance or total cost of ownership, which highly depend on specifics of the implementations and the runtime environment as well as on characteristics of the underlying optimization framework.

The study shows that, while the choice of the dataset does have a huge impact on which combination of compression and change-detection technique performs best, some algorithms like Chebyshev Approximation [6, 7, 4] and Bayesian Online Change Detection [2] yield good results in many settings. We also observe that a good change detection is possible even on strongly compressed data. Next, our results are particularly interesting because studying the algorithms in isolation (e.g., compression without subsequent change detection) may yield a different picture. In [14] for instance, competing algorithms have outperformed Chebyshev Approximation with regard to the compression ratio. In our context in turn, this algorithm has proven to be suitable in combination with many change-detection algorithms.

Paper outline: Section 2 describes some application scenarios. Section 3 explains our design decisions. Section 4 summarizes the algorithms evaluated. Section 5 describes the experimental setup and Section 6 presents the results. Section 7 concludes.

2. APPLICATION SCENARIOS

In this section we describe two scenarios with slightly different perspectives on compression and change detection.

2.1 Smart Grid

The Smart Grid is an intelligent communication network which monitors and controls a power network. The integration into such networks of renewable energy producers alters the conventional power flow [1]. These producers are inconsistent and have performance peaks, which in turn demand intelligent power distribution systems.

Consider a company which has to manage a power-distribution network. The company collects, stores and analyzes the data delivered by the many devices (e.g., smart meters, power plants) in its network. The data needs to be analyzed in real-time, thus online change detection is indispensable. To significantly reduce communication and storage costs, the data must be lossily compressed. Now think of a sudden increase in power consumption. The company must react as soon as possible for example by powering up additional power plants. To this end, it must detect the change in the first place, which is not only the consumption measured by one single device, but an aggregate of the entire grid. As a takeaway, we observe that good compression and high-quality change detection are both very important in this scenario.

2.2 Internet of Things

Internet of Things (IoT) refers to large networks of small or embedded devices, which communicate wirelessly. For many IoT entities, energy optimization is a primary constraint, as they are powered by batteries or use energy harvesting methods like micro solar panels. Thus, wireless data transmission often is the biggest factor regarding energy consumption, as the power required to transmit data increases quadratically or even with the power of 4 with the distance between sender and receiver [3]. The power consumption of data compression in turn increases only linearly with the size of the data. Thus, it is reasonable to send data that is lossily compressed over a distance. Detecting changes is often computationally heavy (e.g., overall computational complexity Bayesian Online Change Point Detection is $\mathcal{O}(n^5)$, where n is the length of the sequence under consideration [2]) and should be performed on the central unit; it therefore has to take place on compressed and later decompressed data [17]. Now consider a home automation system, where a central control unit can adapt the heating when several temperature or humidity sensors detect a change in the weather. Online change detection is needed to react in short time. This specific scenario benefits more from a high compression ratio than from better change detection, in contrast to the previous scenario.

3. DESIGN DECISIONS

Designing the comparison study envisioned is challenging; in particular, there are various design decisions that one must address. In the following, we describe the respective alternatives and the rationale behind our choices.

Benchmark Change Points To assess the quality of change-detection methods, it is very common to compare the change points detected to a ground truth. This however has at least two issues. First, ground-truth metadata can diverge from the detectable changes. To illustrate, the heart-rate dataset PAMAP[2] comes together with the information when exactly a test person has changed his activity. The heart naturally takes some time to adapt to new activities. Second, most change-detection algorithms can only detect specific kinds of changes. ADWIN for instance is specialized on changes of mean values. In other words, comparing to a ground truth evaluates the suitability of the change-detection algorithm for the dataset. Thus, rather than comparing to a ground truth, we let the change-detection method identify change points on the specific dataset without any compression, and we use these change points as our benchmark, dubbed *benchmark change points*. See Section 5.4 for details. Compression is only used in the actual comparison study, i.e., when looking for change points on the compressed data. We call the change points identified on the compressed data *comparison change points*.

Parametrization The result quality and performance of change-detection and compression algorithms depend on their input parameters. In particular, setting the parameters of the change-detection algorithm when comparing alternatives is intricate. One option is to use the parameters

[2]http://www.pamap.org/demo.html, May 18, 2015

recommended in the underlying publications. But this ignores characteristics of the data the algorithms run on. An alternative is to use the parameter values that give way to good results on the data currently examined. If so, these values obviously need to be found, and this is not trivial. We for our part pursue this option nevertheless, as follows. We use an optimization technique to find those parameter values. This requires a reference point. Despite the limits mentioned in the previous paragraph this reference point is the ground truth. I.e., that optimization minimizes the distance between it and the result of the change-detection algorithm without compression.

Multi-objective Optimization With a focus on result quality, optimizing change-detection and compression algorithms in combination has two objectives: low error rate of change detection and good compression ratio. In general, there are several kinds of methods to perform optimization with multiple objectives. One approach is to derive a single value, using for example a weighted sum. This is easy to implement, but finding appropriate weights highly depends on the specific use case (see Section 2) and is notoriously difficult. A more sophisticated, but at the same time more costly approach is multi-objective optimization, resulting in a Pareto frontier. We have chosen this second option because it is more informative.

Error Measure Finding good parameter values requires a measure for the change-detection error. One can use a relatively simple measure, such as the number of correctly detected change points. An alternative is to calculate individual errors for paired changes, misses and false positives, and one can further refine this using application-specific weights. While this is markedly more complex, it also provides more insight. Because we aim to compare change points in detail, we choose the latter option. We use a framework providing that functionality, the MILTON distance measure [10].

Training Data The evaluation envisioned can take place on the complete dataset or on a subsequence. We see two advantages in using a subsequence. The first one is that the parameter optimization is quicker. Second, we can do so to validate the hypothesis that it is sufficient to run this optimization on a data subsequence, and the result also performs well on the complete dataset or on any other data of the same kind. This is important, because we focus on online algorithms that do not operate on complete datasets, but on streams of data.

4. FUNDAMENTALS

In this section we review the compression and change-detection algorithms covered in our study (Table 1). We also review two evolutionary algorithms. We then say how we quantify the deviation of change points. Here we can only provide informal summary descriptions of those algorithms, but the publications describing them contain detailed explanations. They also specify the respective parameters. Further information as well as code is available on the project website[3].

4.1 Compression Methods

A broad review of the literature has resulted in the following categories of model-based compression algorithms: constant, straight-line or polynomial model compression. For each category we have chosen at least one representative, typically one which has received a lot of coverage in the scientific literature.

Table 1: Overview of algorithms and their abbreviations

Compression Algorithms	
APCA	Adaptive Piecewise Constant Approx. [16]
SF	Slide Filter [12]
CHEB	Chebyshev Approximation [6, 7, 4]
WAVE	Wavelet Approximation [19]
PPA	Piecewise Compression Algorithm [11]
Change Detection Algorithms	
ADWIN	Adaptive Windowing [5]
ED	Event Detection from time series data [13]
CF	ChangeFinder [18]
OKCD	Online Kernel Change Detection Algorithm [9]
BOCD	Bayesian Online Changepoint Detection [2]
Optimization Algorithms	
SOEA	Single Objective Evolutionary Algorithm
NSGA-II	Non-Dominated Sorting Genetic Algorithm [8]

Adaptive Piecewise Constant Approximation (APCA) [16] adds data points to an adaptive window until the difference between the maximal and minimal value within this window exceeds a given threshold. Each window then is compressed by summarizing the data points as the arithmetic mean of its maximal and minimal value.

Slide Filter (SF) [12] makes use of several functions which approximate a set of data points. It starts by computing the values of these functions for a window consisting of two data points. Then more points are added to the window, while the functions which do not fulfill the error threshold anymore are left aside. This is continued until only one function remains. This remaining function then is the approximation of the window.

Chebyshev Approximation (CHEB) [6, 7, 4] tries to represent fixed size windows by a linear combination of Chebyshev polynomials up to a given dimension. If the approximation deviates more than the given error threshold, it stores the original data instead.

Wavelet (WAVE) [19] uses a discrete wavelet transform (DWT) to compress time series. The data goes through a low-pass filter and a bandpass filter to construct the corresponding continuous wavelet function.

Piecewise Polynomial Algorithm (PPA) [11] proposes a method which combines several compression methods. The algorithm keeps adding data points to the current window until the error threshold does not hold anymore. It then compresses this window using the best compression algorithm out of several ones.

4.2 Change Detection Techniques

Our study covers change-detection techniques of the following important categories: sequential analysis, maximum-likelihood estimation, kernel based techniques and Bayesian analysis techniques. Again, we have chosen one representative for each category.

Adaptive Windowing (ADWIN)[4] [5] uses a sliding window which is partitioned into buckets. Each bucket can contain

[3]http://www.ipd.kit.edu/~efros/EvalCD/

[4]More specifically, we use ADWIN2, which is often referred to as ADWIN.

several data points; it does so by storing their number and an aggregate of their values. Each time a data point is added to the window, it is put into a new bucket. When a certain number of buckets is reached, the two oldest buckets are merged. If the difference of the average values of two neighboring buckets exceeds a dynamic threshold, a change is reported and the last bucket is dropped. This dynamic threshold is computed for each comparison of two buckets. It depends on the difference of the numbers of data points of the two buckets.

Event Detection (ED) [13] is based on maximum likelihood estimation. It examines a data window to which data points are added step by step. In each step, it determines if the window can be split into two significantly different segments. Each segment then is approximated by fitting a model to it, and the error between the model and the data is determined. The point which minimizes this error for both segments is reported as change point. The models used are derived from base classes such as algebraic polynomials, radial, wavelet or Fourier. We for our part have chosen algebraic polynomials, just as in [13].

ChangeFinder (CF) [18] describes a two-stage algorithm which combines outlier detection and change detection. In a first stage, the algorithm learns an auto regressive (AR) model from a given time series. For each data point of the time series, a score is obtained by calculating the loss, be it the logarithmic one or the quadratic one. An outlier results in an isolated high score, while changes manifest themselves as series of high scores. Smoothing the scores removes the outliers. The smoothed values from the first AR model are then used to learn another AR model in the second stage of the algorithm. The scores of the second model describe the probability for data points being change points.

Online Kernel Change Detection (OKCD) [9] uses one-class support vector machines for change detection. For each data point of the time series, the immediate past subset $x_{t,1}$ and the immediate future subset $x_{t,2}$ are mapped into a feature space. A kernel method is used; it ensures that the mapped input space is a subset of a hypersphere with radius one, centered at the origin of the feature space. Support vector classification then finds hyperplanes in the feature space which separate the training vectors $\Phi(x_{t,1})$ and $\Phi(x_{t,2})$ from the center of the hypersphere. To decide whether a change point is present, the authors introduce a dissimilarity measure in feature space:

$$D_H = \frac{\overset{\frown}{c_{t,1}c_{t,2}}}{\overset{\frown}{c_{t,1}p_{t,1}} + \overset{\frown}{c_{t,2}p_{t,2}}}, \tag{1}$$

where $c_{t,1}$ and $c_{t,2}$ are the centers of the hypersphere sections intersected by the hyperplanes, and $p_{t,1}$ and $p_{t,2}$ are two points where the hyperplanes intersect the hypersphere. The arc represents the arc distance between the two points. If the dissimilarity measure exceeds a given threshold, a change point is reported.

Bayesian Online Changepoint Detection (BOCD) [2] uses a Bayesian approach. It divides a time series into partitions and assumes that for each partition there is an i.i.d. probability distribution of the data values. Thus, the change points are the boundaries between the partitions. For each new data point, the algorithm estimates the probability distribution since the last change point and then computes the probability that the new point belongs to this distribution. When this probability drops suddenly, a change is reported.

4.3 Optimization Techniques

On the technical level, some decisions like choosing an optimization algorithm have been necessary as well. We for our part use evolutionary algorithms. NSGA-II [8] is our choice for multi-objective optimization. Calculating benchmark change points needs only single-objective optimization (see *Change Point Baseline* in Section 3), which leads us to the faster SOEA algorithm.

Single Objective Evolutionary Algorithms (SOEA) start with a random set of problem solutions, referred to as initial population. The objective is to identify individuals, i.e., solutions, with low fitness. In each generation step, the individuals are sorted by their fitness, and two parents are randomly selected from among the top τ percent. They create two children by crossing over, and these children are mutated with a certain probability. Additionally a new random individual is introduced in each generation. The three newly created individuals replace the three individuals with the worst fitness. The algorithm terminates after a certain number of generations, or when the fitness falls below a certain threshold.

Non-dominated Sorting Genetic Algorithm [8] is an evolutionary optimization algorithm with multi-objective support (NSGA-II). It approximates a Pareto-optimal frontier over several generations. It starts with an initial, random population, and each generation categorizes the individuals into fronts, sorts the individuals within these fronts and uses the best individuals to create a new population, which then are added to the population of the next generation. More specifically:

1. An individual belongs to a front if there does not exist another individual in this current or in any previous front dominating it. An individual x dominates another individual y if and only if x is never inferior to y in any objective and x is superior to y in at least one objective.

2. For the sorting within the fronts, a so-called density value is assigned to each individual. It quantifies the density of solutions surrounding this individual.

3. The best individuals are chosen based on front and density. They are used to create a new population by means of recombination and mutation.

After several generations, the first front typically is a nearly Pareto-optimal frontier.

4.4 Measure for Quantifying the Impact of Lossy Transformations on Subsequent Change Detection (MILTON)

An important constituent needed for a study such as ours is a measure quantifying the difference of two time series containing change points cp and $\hat{cp}$. d_{MILTON} is such a measure [10]. It categorizes the changes as paired changes (PC), false positives (FP) and misses ($MISS$). Paired changes are changes which occur in both time series and are mapped to each other. False Positives are change points which occur in $\hat{cp}$ but not in cp, while misses are change points occurring in cp but not in $\hat{cp}$. For each of these categories an error is

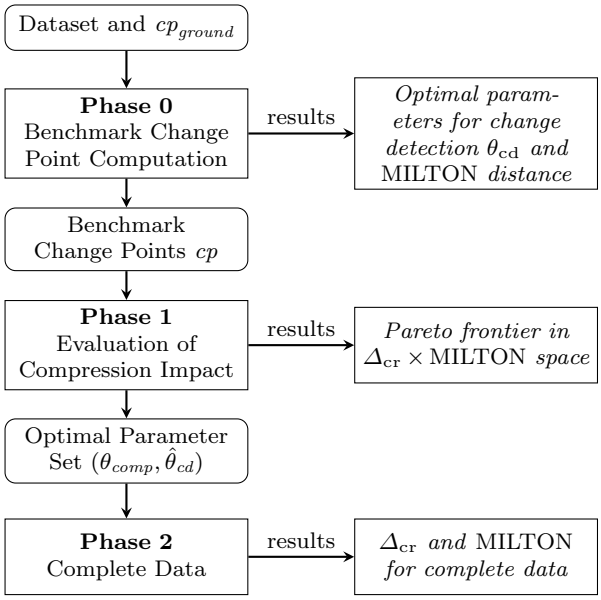

Figure 1: Overview of the experiments

calculated (*errPC*, *errFP*, *errMISS*). These errors then are combined into a total one:

$$d_{MILTON}(cp, \hat{cp}) = \frac{errPC + errMISS + errFP}{|PC| + |MISS| + 1} \quad (2)$$

We explain our parametrization of MILTON in Section 5.

5. EXPERIMENT SETUP AND INITIALIZATION

Our evaluation consists of three phases which build on each other, as shown in Figure 1. Phase 0 finds optimal parameters for a change-detection algorithm on a subsequence of an uncompressed dataset, to provide benchmark-change points (Section 5.4). Phase 1 (Section 5.5) uses the benchmark-change points to find good parameters of the compression and change-detection algorithms for any combination of dataset, compression algorithm and change-detection algorithm. This brings up the question under which circumstances the parameters found on a subsequence are also well suited for the complete dataset. We study this question, i.e., the validity and applicability of good parameters on complete datasets, in Phase 2 (Section 5.6).

5.1 Framework

For the experimental evaluation we have designed and implemented a flexible generic framework which supports the different algorithms and is extensible to test further algorithms. We have integrated existing implementations whenever available. For APCA, SF and CHEB we have used the implementations of [14][5]. The source code for ADWIN[6] and BOCD[7] is publicly available as well. For the wavelet compression we use a method from [19], which is part of the MATLAB libraries. We also reuse existing implementations of PPA and *MILTON*. We have implemented the remaining algorithms (ED, CF and OKCD) in MATLAB following the original publications, and they can be downloaded from our web page. Our framework handles algorithm implementations in C, C++, C#, R, MATLAB or Java.

The framework allows to define jobs for each experiment. A job consists of the algorithms chosen for compression, change detection and optimization, together with their parameters. It also includes the dataset and reference change points. Jobs cover the workflow of the experiments depicted in Figure 1. To distribute the work among several machines, the jobs are stored in a database where any free node can poll an open job. The results are then stored in the database as well.

5.2 Datasets

For the experiments we use artificial datasets as well as real world datasets (see Figure 2). We have generated the artificial datasets so that they contain well-defined changes, in line with earlier work [18, 15]. We use an autoregressive function similar to the one in [15] which generates change points at every 200th point of time. In the *Rising Mean* dataset we increase the mean of normally distributed noise by 1 at every change point. The *Variance Change* dataset alters the variance of the noise between 1 and 3 at a change point. The rationale is to study the behavior of the algorithms under another kind of change. Algorithms 1 and 2 contain the pseudo code generating this data, the companion web page contains them as MATLAB code.

$\mu \leftarrow 0$, $\sigma \leftarrow 1$
$x(0) \leftarrow 0$, $x(1) \leftarrow 0$
for $t \leftarrow 2$ *to* $t = length\_of\_dataset$ **do**
  $x(t) \leftarrow 0.6 \cdot x(t-1) - 0.5 \cdot x(t-2) + N(\mu, \sigma^2)$
  if t *mod* $200 = 0$ **then**
    $\mu \leftarrow \mu + 1$
  end
end

Algorithm 1: *Rising Mean*

$\mu \leftarrow 0$, $\sigma \leftarrow 1$
$x(0) \leftarrow 0$, $x(1) \leftarrow 0$
for $t \leftarrow 2$ *to* $t = length\_of\_dataset$ **do**
  $x(t) \leftarrow 0.6 \cdot x(t-1) - 0.5 \cdot x(t-2) + N(\mu, \sigma^2)$
  if t *mod* $200 = 0 \wedge \sigma = 1$ **then**
    $\sigma \leftarrow 3$
  else if t *mod* $200 = 0 \wedge \sigma = 3$ **then**
    $\sigma \leftarrow 1$
end

Algorithm 2: *Variance Change*

In real-world datasets, defining change points unambiguously is not possible in general. We use datasets from different fields that are annotated with change points as metadata: EEG data, heart rate monitoring and electricity data.

The EEG dataset[8] has been captured while the subject was opening and closing his eyes; this leads to a noticeable

[5]lsirwww.epfl.ch/benchmark/, May 18, 2015
[6]https://github.com/abifet/adwin, May 18, 2015
[7]hips.seas.harvard.edu/content/bayesian-online-changepoint-detection, May 18, 2015
[8]https://archive.ics.uci.edu/ml/datasets/EEG+Eye+State, May 18, 2015

peak. We have removed a one-value outlier at 898 by interpolating the neighboring values to get more stable change-detection results.

The heart-rate dataset comes from the (PAMAP) project. More specifically, we use the outdoor dataset of Subject 2. It contains activities like sitting, walking, running or playing soccer. Since the data has been captured with 100 Hz, which is way above the resolution of the heart-rate monitor, we have reduced the dataset by using every hundredth data point, in order to reduce the data volume. This is because the lighting is relatively independent of other household appliances.

We also use the REDD energy-consumption data[9]. It records the power consumption of a house broken down into different electrical consumers. For our evaluation we have selected Channel 17 of House 1, which is the lighting in one of the rooms.

Table 2 shows the lengths of the subsequences we have selected according to the *Training Data* design decision. For the EEG dataset the segment is slightly larger than for the other data. This is because the ground-truth change points are farther apart.

5.3 Notation

In this paper $x(t) \in \mathbb{R}, t = 1, 2, \ldots, n$ is a real-valued one-dimensional time series. $cd(\cdot|\theta_{cd})$ stands for a change-detection algorithm with parameters θ_{cd}. By applying it to a time series x, we get $cp(t) \in \{0, 1\}, t = 1, 2, \ldots, n$ where

$$cp(t) = \begin{cases} 1 & \text{if } t \text{ is a change point} \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

We further define a transformation $trans(\cdot|\theta_{trans})$ with parameters θ_{trans}. This method takes the input time series x and creates the time series $\hat{x}(t) \in \mathbb{R}, t = 1, 2, \ldots, n$, where each value $\hat{x}(t)$ is the result of a compression and subsequent decompression step:

$$\hat{x} = trans(x|\theta_{trans}) \quad (4)$$

Δ_{cr} is the compression ratio:

$$\Delta_{cr} = \frac{size\_of\_compressed\_data}{size\_of\_original\_data} \quad (5)$$

Note that *size_of_compressed_data* cannot be derived from $\hat{x}$. This is because it does not represent the compressed data.

Table 3: Symbols used and their meaning

Symbol	Meaning
x	Original time series
$\hat{x}$	Compressed time series
cp_{ground}	Gound truth change points
cp	Benchmark change points on x
$\hat{cp}$	Comparison change points on $\hat{x}$
Δ_{cr}	Compression ratio
PC	Number of paired changes
FP	Number of false positives
$MISS$	Number of missed changes
$errPC$	total error of paired changes
$errFP$	total error of false positives
$errMISS$	total error of missed changes

[9]http://redd.csail.mit.edu/, May 18, 2015

Table 4: List of parameters for the optimization algorithms, fitness function and MILTON distance with their corresponding values used for our experiments.

Algorithm	Parameter	Value
SOEA	population size	100
	exit fitness	0.001
	max. generations	500
	mutation rate	0.2
	mutation change	0.4
	selection pressure τ	0.4
NSGA-II	population size	500
	exit fitness	0.001
	max. generations	10
	mutation rate	0.2
	mutation change	0.4
Fitness	weight α	0.5
MILTON	$f_{TIME}(\Delta_t)$	$\lvert\Delta_t\rvert$
	$f_{SCORE}(\Delta_s)$	0
	$f_{MISS}(s)$	$(s+1)^2$
	$f_{FP}(s)$	s

5.4 Phase 0: Benchmark Change Point Computation

Recall that Phase 0 is not an experiment in its own right, but an initialization step. It is described next. As stated under *Benchmark Change Points* in Section 3 on Design Decisions, we do not use the ground-truth change points as reference for our evaluation. Instead we use benchmark change points by minimizing the MILTON distance to the ground truth:

$$\arg\min_{\theta_{cd}} d_{MILTON}(cd(x|\theta_{cd}), cp_{ground}) \quad (6)$$

To this end, we use an SOEA. Table 4 shows its parameterization.

We have executed the further phases only for combinations of change-detection techniques and datasets which lead to acceptable results. We deem a result acceptable if the number of paired changes exceeds the one of false positives and the one of misses, so that most of the original change points are found. To facilitate a comparison, Table 2 lists the MILTON distance against the ground-truth change points. '-' stands for results that have not been accepted. Most algorithms have found a parametrization on Rising Mean and Variance Change. The ADWIN and Variance Change combination does not have a result, because ADWIN only finds changes of mean values [5]. As expected, some change-detection algorithms do not perform well on some real-world datasets. This is because their ground truth is based on secondary observations that do not necessarily cause a change in the data at exactly the same time.

5.5 Initialization of Phase 1: Evaluation of Compression Impact

An important goal of our evaluation is to determine an optimal set of parameters which preserves the change points cp found before compression as much as possible while maximizing the compression at the same time. This is a multi-objective optimization problem with the objectives $d_{MILTON}(cp, \hat{cp})$ and compression ratio Δ_{cr}, where $\hat{cp} =$

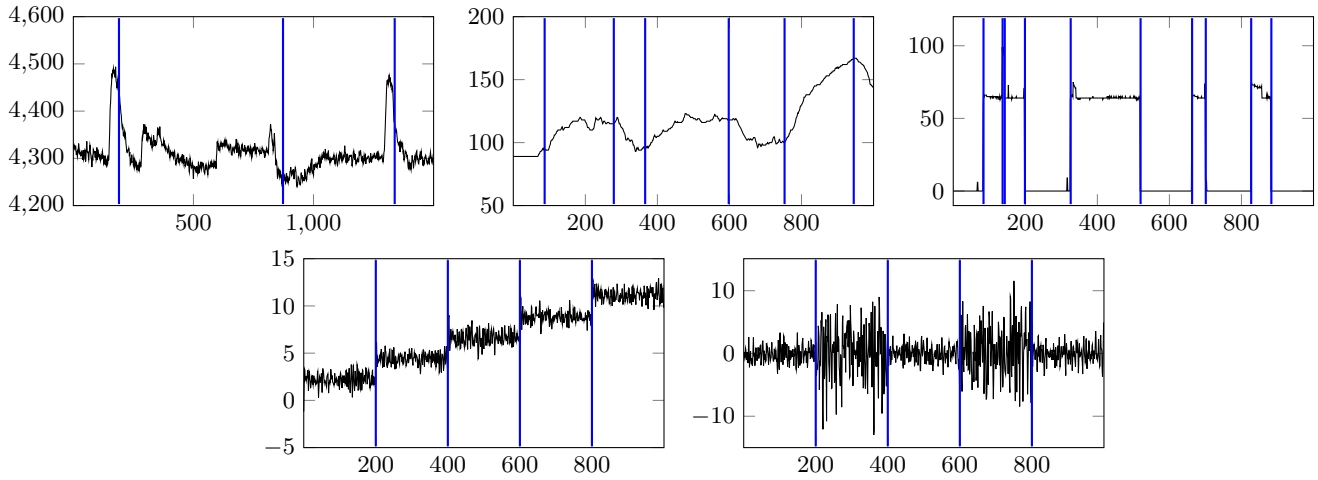

Figure 2: Plots of excerpts of all datasets including ground truth. Top row shows real world datasets: EEG, PAMAP and REDD (from left to right); bottom row shows the artificial datasets: Rising Mean and Variance Change.

Table 2: Overview of datasets with their corresponding MILTON distance of initial parameter calculation, where $|CP|$ is the number of change points in the whole dataset and $|CP_{Seg}|$ in the segment

					d_{MILTON}				
Name	**Length**	$\|CP\|$	**Segment**	$\|CP_{Seg}\|$	**ADWIN**	**ED**	**CF**	**OKCD**	**BOCD**
Rising Mean	10000	49	1000	4	2.006	1.049	1.009	0.202	0.003
Variance Change	10000	49	1000	4	-	2.822	0.833	0.422	0.217
REDD	10000	144	1000	10	2.365	-	0.104	-	0.464
PAMAP	2280	13	1000	6	1.018	-	-	1.041	0.791
EEG	14979	23	1500	3	0.27	0.272	-	1.287	0.017

$cd(\hat{x}|\hat{\theta}_{cd})$. The parameter space consists of the parameters for the compression and the change-detection algorithm: $\theta = (\theta_{trans}, \hat{\theta}_{cd})$. To evaluate the algorithms we study all possible combinations on each dataset.

To find optimal parameter sets we use an adaptation of NSGA-II described in the next paragraph. See Table 4 for the parameterization of NSGA-II. The weighting functions for the errors in the MILTON distance are important as well. See again Table 4, with functions similar to [10].

The parameters of the change-detection and compression algorithms have a range of validity which must be kept during the optimization. Therefore we modify NSGA-II to calculate its random values in the range $[\phi_{min}, \phi_{max})$ during the initialization of the population and for mutations. For some parameters we have reduced this range even further to reduce the search space and to speed up the optimization process. Tables 5 and 6 list the parameters and the ranges we have selected. An additional modification is the distinction between float and integer parameters. Random values for the float parameters are calculated as $\phi' = \phi_{min} + r \cdot (\phi_{max} - \phi_{min})$, where r is an equally distributed random number in $[0; 1)$. For integer parameters this value is then rounded: $\phi' = \lfloor \phi' \rfloor$.

The result of this phase is a Pareto frontier that represents the best possible trade-offs between compression ratio and the preservation of change points. Each individual consists of the MILTON distance, the compression ratio and the root-mean-square error (RMSE) calculated for the corresponding set of compression and change-detection parameters. To select an individual of this frontier suitable for a specific application scenario, a *fitness* is calculated:

$$\textit{fitness} = \alpha \cdot d_{MILTON}(cp, \hat{cp}) + (1 - \alpha) \cdot \Delta_{cr} \qquad (7)$$

where α is a parameter to weigh the addends. Note that α is used only to select an individual in the result set *after* the optimization is finished. This provides a lot of freedom and flexibility during the evaluation.

[10]In our case *MSet* specifies the maximum degree of the polynomials for the approximation (cf. Subsection 4.2).

Table 5: List of parameters for all compression algorithms and the ranges of the optimization

Algorithm	**Parameter**	**Min**	**Max**	**Type**
All methods	threshold ϵ	0	0.3	float
CHEB	segment length	4	$\|x\|$	int
WAVE	max. level	1	10	int
PPA	max degree	2	5	int

5.6 Initialization of Phase 2: Complete Datasets

Section 3 has explained the necessity to evaluate the best performing combination of compression and change detection on a subsequence but also on the complete dataset. The details of the experimental setup when it comes to the

Table 6: List of parameters for all change detection algorithms and the ranges of the optimization

Algorithm	Parameter	Min	Max	Type
ADWIN	M	2	10	int
	δ	0	1	float
ED	δ	0	1	float
	p	1	200	int
	$MSet$ [10]	0	5	int
CF	T	3	10	int
	k	2	10	int
	r	0	0.4	float
OKCD	m_1	2	200	int
	m_2	2	200	int
	ν	0.2	0.8	float
	η	0	1	float
	σ	0	1	float
BOCD	μ_0	0	2	float
	κ_0	0	5	float
	α_0	0	5	float
	β_0	0	5	float
	λ	> 0	500	float

complete dataset are as follows: The parameters for a specific α of the Phase 1 experiment are applied to the complete dataset. For the Rising Mean dataset, we have divided the ϵ threshold by 10, because it depends on the global maximum that is 10 times higher on the complete Rising Mean dataset. As a reference, we use the parameters computed in Phase 0 (Section 5.4) on the complete dataset. Then the MILTON distance d_{MILTON}, compression ratio Δ_{cr} and RMSE are calculated.

6. RESULTS

This section first describes and discusses our results of the Phase 1 experiments and then presents the results of Phase 2 on the complete data. As an initial, exemplary illustration, Figure 3 visualizes the data transformation in the different phases for the combination REDD, BOCD and APCA. The top plot, although not the focus of our study, shows the raw data with ground-truth change points as vertical straight lines. The middle plot shows the benchmark change points calculated in Phase 0. The bottom plot shows the change points on the compressed data with the best result parameters of Phase 1. Comparing the top plot to the middle plot, we can see that BOCD fails to detect two changes and incorrectly identifies two other ones. Taking benchmark change points as a reference when run on the compressed data, i.e., comparing the middle plot to the bottom plot, BOCD fails to detect two changes. Thus, the decrease in the performance of BOCD on compressed data is small in this case, compared to its performance on the original data.

6.1 Phase 1 – Results

The results of the Phase 1 experiments are Pareto frontiers. To illustrate, Figure 4 shows a sample of the Pareto frontiers on the Variance Change dataset. Each plot contains all compression techniques for one change-detection algorithm, except for ADWIN, which is not applicable to this dataset (see Section 5.4). There is not any frontier dominating all other frontiers, therefore no single best solution exists. Dependent on the dataset and parameter α,

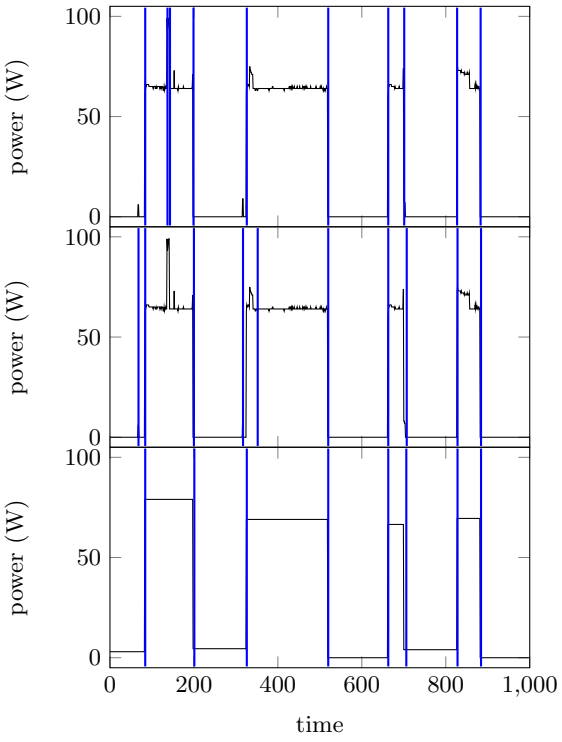

Figure 3: BOCD and APCA on REDD. Top down: raw data with ground-truth change points, result of benchmark-change point computation, best result of change detection with compression ($\alpha = 0.5$).

different combinations of change-detection and compression algorithm yield the best result.

We observe that comparing the large number of Pareto frontiers produced by our experiments is difficult. Thus, we select the individuals with the lowest fitness (see Equation 7) of each Pareto frontier for different values of α and compare their MILTON distance and compression ratio in Figures 5 and 6.

Earlier we have described two scenarios (SmartGrid, IoT) where an approach such as ours is indispensable if one wants to find a good combination of compression and change-detection algorithms. These scenarios have different requirements. We therefore examine the Pareto frontiers for two different α values, $\alpha = 0.5$ for the Smart Grid, and $\alpha = 0.05$ for IoT. For each solution on the Pareto frontier we calculate a fitness value using the respective α value. We have chosen the parametrization with the lowest fitness that indicates the best result for the scenario. We get a triple (*fitness*, MILTON distance, Δ_{cr}) for each experiment. See Figures 5 and 6 for the MILTON distance and Δ_{cr} values, for $\alpha = 0.5$ and $\alpha = 0.05$. The MILTON distance is the value above the horizontal axis, the compression ratio is below. For $\alpha = 0.5$, the best combinations of compression and change-detection algorithm for each dataset are as follows:

- Rising Mean: BOCD with APCA clearly is the best solution, because it achieves a MILTON distance of almost zero and also has the best compression ratio.
- Variance Change: The best fitness is obtained with BOCD and CHEB, closely followed by BOCD and

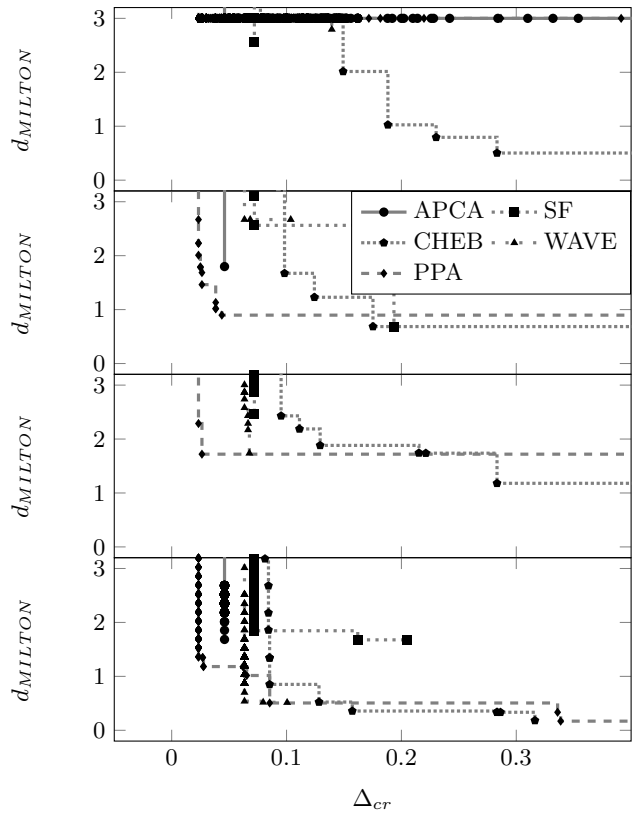

Figure 4: Pareto frontiers for ED,CF,OKCD and BOCD (from top to bottom) on the *Variance Change* dataset.

PPA, although the compression ratio is not best.

- REDD: CF with SF is the best combination, as it has a close to zero MILTON distance and a very good compression ratio.
- PAMAP: The best algorithms are BOCD and CHEB, mainly because of the low MILTON distance. BOCD with WAVE also performs very well.
- EEG: BOCD with SF clearly is optimal.

Overall BOCD performs very well on all datasets and is only beaten once by CF. We have made further noteworthy observations:

- A MILTON distance larger than 3 means that no change points have been found after compression. Thus, CF on Rising Mean and ADWIN on the EEG data do not work at all.
- The Variance Change dataset has relatively high compression ratios. This is because the changes in variance are difficult to compress, especially on combinations OKCD with PPA and ED with CHEB.
- The REDD dataset is easy to compress while keeping the change points, because of its very sharp edges and low signal-to-noise ratio. It therefore has the lowest average fitness of all datasets.

Comparing the results for $\alpha = 0.05$ (Figure 6) to the ones above, the compression ratios are smaller. However, the MILTON distances are higher. Both effects are expected. On the Variance Change dataset, the results with the lowest MILTON distance for $\alpha = 0.5$ has a very high compression ratio. For $\alpha = 0.05$, those combinations yield a much lower compression ratio.

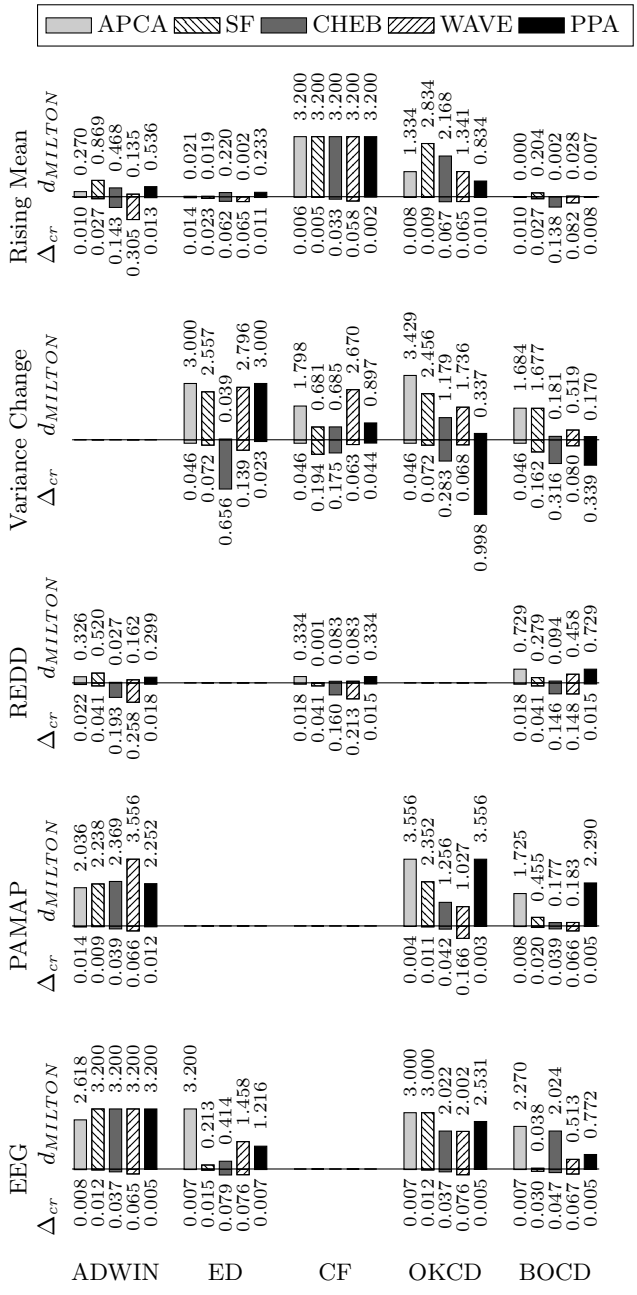

Figure 5: Overview of the best solutions for $\alpha = 0.5$ on each combination of compression algorithm, change detection algorithm and dataset.

On PAMAP and EEG the solutions for $\alpha = 0.5$ and $\alpha = 0.05$ are the same. The best results for $\alpha = 0.05$ from Rising Mean and REDD Data are identical to those

from $\alpha = 0.5$. This is because their low compression ratios are not reduced further. Some of the other results have slightly lower compression ratios. Our takeaway is that our experiments do indeed help to find solutions for specific use cases. The results of this phase also show that the quality differs a lot between different combinations in the same setting. Thus it is very important to be able to study different combinations using a setup as elaborate as ours.

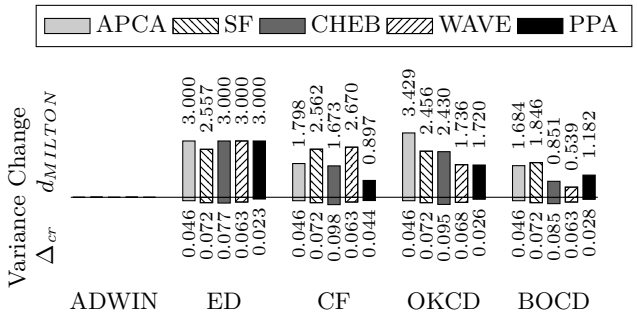

Figure 6: Solutions for $\alpha = 0.05$ on Variance Change.

6.2 Phase 2 – Results

As explained in Section 5.6, we apply the parameters of the results from Phase 1 to the complete data. For every set we choose the combination with the lowest fitness. We expect that those parameter sets will achieve the same quality of results, mainly in terms of stable ratios between PC, FP and MISS, as on the subsequence datasets. Since d_{MILTON} grows quadratically for the number of FP and linearly for the one of MISS (see Table 4), it is hardly comparable on the complete dataset. On the other hand, we expect that the compression ratio stays constant.

Table 7 shows our results. In contrast to our expectation, there are disproportionally more FPs and misses than on small sets. From EEG we conclude that three change points are not sufficient to train the change detection properly. The NSGA-II overfits the parameters on the training data. To avoid this, we have also tested the other combinations which do not show the best but nevertheless good fitness. The lower part of Table 7 shows that these parameters yield results on the complete dataset which are as good as the ones on the excerpts. On REDD for instance, ChangeFinder detects eleven times as many change points but only eight times as many FPs. In contrast to Section 6.1, BOCD performs best in only two out of five cases. In every case the compression ratio differs only slightly. To sum up, we can say that the results from a short subsequence can be used on the complete dataset without losing quality. However, it is necessary to rely on several results from Phase 1 to find the ones adapting best.

7. CONCLUSION

In many situations, compression and change-detection methods must be used in combination. In such a setting however, a number of questions are unclear, e.g.: Which combination is best for a given scenario? How to find a good parameterization of compression and change-detection algorithms when these are used together? How well can we trade compression ratio against change-detection quality? This article has featured a comprehensive experimental evaluation that addresses these questions.

A study such as ours requires a number of non-trivial design decisions. This article has listed the important issues, together with the respective options and our rationale behind the 'winner' alternatives.

An important insight is that the overall picture is very differentiated. Result quality highly depends on the dataset. For instance, the change-detection method ChangeFinder is the best performing algorithm on REDD, but the worst performing one on the Rising Mean dataset. Our platform has turned out to be an appropriate tool to find good parameterizations, at least if the dataset inspected is sufficiently representative and large.

When data is compressed, the intention always is to decompress it later and use it in some way. Change detection is one kind of data usage, but other kinds of usage obviously abound and are important as well. Just think of the plethora of different stream-mining approaches which have been proposed in the recent past. Generalizing the work described here to other kinds of usage is important and is part of our future work.

8. REFERENCES

[1] Communication Networks for Smart Grids. Computer Communications and Networks. Springer London (2014)

[2] Adams, R.P., MacKay, D.J.C.: Bayesian online changepoint detection. arXiv preprint arXiv:0710.3742 (2007)

[3] Akyildiz, I.F., Weilian Su, Sankarasubramaniam, Y., Cayirci, E.: A survey on sensor networks. IEEE Communications Magazine **40**(8) (2002)

[4] Arion, A., Jeung, H., Aberer, K.: Efficiently maintaining distributed model-based views on real-time data streams. In: Global Telecommunications Conference (GLOBECOM 2011), IEEE

[5] Bifet, A., Gavaldà, R.: Learning from time-changing data with adaptive windowing. In: Proceedings of the 2007 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics

[6] Cai, Y., Ng, R.: Indexing spatio-temporal trajectories with chebyshev polynomials. In: Proceedings of the 2004 ACM SIGMOD

[7] Cheng, A.F., Hawkins III, S Edward, Nguyen, L., Monaco, C.A., Seagrave, G.G.: Data compression using chebyshev transform (2007)

[8] Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation **6**(2) (2002)

[9] Desobry, F., Davy, M., Doncarli, C.: An online kernel change detection algorithm. IEEE Transactions on Signal Processing **53**(8) (2005)

[10] Efros, P., Buchmann, E., Englhardt, A., Böhm, K.: How to quantify the impact of lossy transformations on change detection. In: Proceedings of the 27th International Conference on Scientific and Statistical Database Management (2015)

Table 7: Comparison of dataset excerpts (left) and complete datasets (right) using the same parameters.

Best fitnesses							
Combination	d_{MILTON}	Δ_{cr}	**RMSE**	**Fitness**	**|PC|**	**|FP|**	**|MISS|**
Rising Mean: APCA, BOCD	0.000 | 0.674	0.010 | 0.011	0.827 | 0.690	0.010 | 0.342	4 | 48	0 | 3	0 | 9
Variance Change: CHEB, BOCD	0.181 | 1.716	0.316 | 1.031	1.559 | 0.000	0.249 | 1.374	5 | 75	1 | 39	0 | 40
REDD: SF, CF	0.001 | 0.957	0.041 | 0.048	4.449 | 4.479	0.021 | 0.503	11 | 126	0 | 9	0 | 37
PAMAP: CHEB, BOCD	0.177 | 0.476	0.039 | 0.039	4.585 | 4.873	0.108 | 0.258	6 | 13	1 | 3	0 | 1
EEG: SF, BOCD	0.038 | 3.104	0.030 | 0.017	14.95 | 17.305	0.034 | 1.560	3 | 9	0 | 30	0 | 1

Best generalization							
Combination	d_{MILTON}	Δ_{cr}	**RMSE**	**Fitness**	**|PC|**	**|FP|**	**|MISS|**
Rising Mean: APCA, ED	0.021 | 0.432	0.014 | 0.018	0.698 | 0.669	0.035 | 0.225	4 | 43	0 | 1	0 | 5
Variance Change: SF, CF	0.681 | 0.580	0.194 | 0.182	2.153 | 2.422	0.435 | 0.762	7 | 56	2 | 5	1 | 8
REDD: WAVE, CF	0.083 | 0.244	0.213 | 0.207	0.921 | 1.021	0.255 | 0.225	11 | 155	1 | 8	0 | 8
EEG: PPA, BOCD	0.772 | 1.919	0.005 | 0.002	36.01 | 43.65	0.333 | 0.960	3 | 10	3 | 21	0 | 0

[11] Eichinger, F., Efros, P., Karnouskos, S., Böhm, K.: A time-series compression technique and its application to the smart grid. The VLDB Journal (2014)

[12] Elmeleegy, H., Elmagarmid, A.K., Cecchet, E., Aref, W.G., Zwaenepoel, W.: Online piece-wise linear approximation of numerical streams with precision guarantees. Proceedings of the VLDB Endowment **2**(1) (2009)

[13] Guralnik, V., Srivastava, J.: Event detection from time series data. In: SIGKDD (1999)

[14] Hung, N.Q.V., Jeung, H., Aberer, K.: An evaluation of model-based approaches to sensor data compression. IEEE Transactions on Knowledge and Data Engineering **25**(11) (2013)

[15] Kawahara, Y., Sugiyama, M.: Change-point detection in time-series data by direct density-ratio estimation. In: Proceedings of the 2009 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics

[16] Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S.: Locally adaptive dimensionality reduction for indexing large time series databases. ACM SIGMOD Record **30**(2) (2001)

[17] Miorandi, D., Sicari, S., Pellegrini, F.d., Chlamtac, I.: Internet of things: Vision, applications and research challenges. Ad Hoc Networks **10**(7) (2012)

[18] Takeuchi, J., Yamanishi, K.: A unifying framework for detecting outliers and change points from time series. IEEE Transactions on Knowledge and Data Engineering **18**(4) (2006)

[19] Vishwanath, M.: The recursive pyramid algorithm for the discrete wavelet transform. IEEE Transactions on Signal Processing **42**(3) (1994)