



GRAMMAR OF TIME TIME SERIES DATA MINING IN STEROIDS

HOW TO TRANSLATE TIME?

JOHN VERY LONGNAME DOE
Master/BSc in Name of Previous Degree

DOCTORATE IN STUDY PROGRAM NAME
NOVA University Lisbon
month, year



GRAMMAR OF TIME TIME SERIES DATA MINING IN STEROIDS

HOW TO TRANSLATE TIME?

JOHN VERY LONGNAME DOE
Master/BSc in Name of Previous Degree

Adviser: Mary Doe Adviser Name
Full Professor, NOVA University Lisbon

Co-advisers: John Doe Co-Adviser Name
Associate Professor, NOVA University Lisbon
John Doe other Co-Adviser Name
Full Professor, NOVA University Lisbon

Examination Committee

Chair: Name of the committee chairperson
Full Professor, FCT-NOVA

Rapporteur: Name of a rapporteur
Associate Professor, Another University

Members: Another member of the committee
Full Professor, Another University
Yet another member of the committee
Assistant Professor, Another University

DOCTORATE IN STUDY PROGRAM NAME
SPECIALIZATION IN SPECIALITY NAME
NOVA University Lisbon
month, year

Grammar of Time Time Series Data Mining in Steroids

Copyright © John Very Longname Doe, NOVA School of Science and Technology, NOVA University Lisbon.

The NOVA School of Science and Technology and the NOVA University Lisbon have the right, perpetual and without geographical boundaries, to file and publish this dissertation through printed copies reproduced on paper or on digital form, or by any other means known or that may be invented, and to disseminate through scientific repositories and admit its copying and distribution for non-commercial, educational or research purposes, as long as credit is given to the author and editor.

Dedicatory lorem ipsum.

ACKNOWLEDGEMENTS

Acknowledgments are personal text and should be a free expression of the author.

However, without any intention of conditioning the form or content of this text, I would like to add that it usually starts with academic thanks (instructors, etc.); then institutional thanks (Research Center, Department, Faculty, University, FCT / MEC scholarships, etc.) and, finally, the personal ones (friends, family, etc.).

But I insist that there are no fixed rules for this text, and it must, above all, express what the author feels.

"You cannot teach a man anything; you can only help him discover it in himself." (Galileo)

ABSTRACT

Regardless of the language in which the dissertation is written, a summary is required in the same language as the main text and another summary in another language. It is assumed that the two languages in question are Portuguese and English.

The abstracts should appear first in the language of the main text and then in the other language. For example, if the dissertation is written in Portuguese the abstract in Portuguese will appear first, then the abstract in English, followed by the main text in Portuguese. If the dissertation is written in English, the abstract in English will appear first, then the abstract in Portuguese, followed by the main text in English.

In the L^AT_EX version, the NOVAtesis template will automatically order the two abstracts taking into account the language of the main text. You may change this behaviour by adding

```
\abstractorder(<MAIN_LANG>) := {<LANG_1>, ..., <LANG_N>}
```

to the customization area in the document preamble, e.g.,

```
\abstractorder(de) := {de, en, it}
```

The abstracts should not exceed one page and, in a generic way, should answer the following questions (it is essential to adapt to the usual practices of your scientific area):

1. What is the problem?
2. Why is this problem interesting/challenging?
3. What is the proposed approach/solution?
4. What results (implications/consequences) from the solution?

Keywords: Keyword 1, Keyword 2, Keyword 3, Keyword 4, Keyword 5, Keyword 6, Keyword 7, Keyword 8, Keyword 9

RESUMO

Independentemente da língua em que a dissertação esteja redigida, é necessário um resumo na mesma língua do texto principal e outro resumo noutra língua. Pressupõe-se que as duas línguas em questão sejam o português e o inglês.

Os resumos devem aparecer primeiro na língua do texto principal e depois na outra língua. Por exemplo, se a dissertação for redigida em português, o resumo em português aparecerá primeiro, seguido do resumo em inglês (*abstract*), seguido do texto principal em português. Se a dissertação for redigida em inglês, o resumo em inglês (*abstract* aparecerá primeiro, seguido do resumo em português, seguido do texto principal em inglês.

Na versão L^AT_EX o template NOVAthesis irá ordenar automaticamente os dois resumos tendo em consideração a língua do texto principal. É possível alterar este comportamento adicionando

```
\abstractorder(<MAIN_LANG>) := {<LANG_1>, ..., <LANG_N>}
```

à zona de customização no preâmbulo do documento, e.g.,

```
\abstractorder(de) := {de, en, it}
```

Os resumos não devem ultrapassar uma página e, de forma genérica, devem responder às seguintes questões (é essencial adaptá-los às práticas habituais da sua área científica):

1. Qual é o problema?
2. Porque é que é um problema interessante/desafiante?
3. Qual é a proposta de abordagem/solução?
4. Quais são as consequências/resultados da solução proposta?

Palavras-chave: Palavra-chave 1, Palavra-chave 2, Palavra-chave 3, Palavra-chave 4

CONTENTS

List of Figures	ix
List of Tables	x
Glossary	xi
Acronyms	xii
Symbols	xiii
Chemical Symbols	xiv
1 Introduction	1
1.1 Motivation	1
1.2 Context and Relevance	3
1.3 Research Questions	3
1.4 Thesis Structure	3
2 Theoretical Concepts	4
2.1 Time Series Fundamentals	4
2.2 Sensing Human Posture, Motion and Physiology	4
2.3 Linguistic Nature of Time Series	4
3 State of the Art	5
3.1 Information Retrieval from Time Series	5
3.2 Occupational Health Sensing and Problems	5
4 Data Description and Management	6
4.1 Public Datasets	6
4.1.1 Classification Benchmark - UCR	6
4.1.2 UCI Machine Learning Repository	6
4.1.3 Physionet	6

4.1.4	CPD Benchmark	6
4.2	Acquired Datasets	6
4.2.1	Office Job Dataset	6
4.2.2	Industrial Job Dataset	6
5	Detection of Events and Summarization of Time Series	7
5.1	7
5.1.1	Feature Representation	7
5.1.2	Self-Similarity	7
5.1.3	Novelty Search	7
5.1.4	Periodic Search	7
5.2	Time Series Profiling	7
5.2.1	Elements with Relevance	7
5.2.2	Minimalist Design	7
5.2.3	Summarize Time Series	7
5.3	Further Developments	7
6	Text Mining Time Series	8
6.1	Synthetic Search on Time Series	8
6.1.1	Time Series Representation	8
6.2	Towards Natural Language for Pattern Search	8
6.3	Classification of Time Series Documents	8
7	Data Description and Management	9
7.1	Public Datasets	9
7.1.1	Classification Benchmark - UCR	9
7.1.2	UCI Machine Learning Repository	9
7.1.3	Physionet	9
7.1.4	CPD Benchmark	9
7.2	Acquired Datasets	9
7.2.1	Office Job Dataset	9
7.2.2	Industrial Job Dataset	9

Appendices

A	NOVAthesis covers showcase	11
B	Appendix 2 Lorem Ipsum	12

Annexes

I	Annex 1 Lorem Ipsum	14
----------	----------------------------	-----------

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

This document is incomplete. The external file associated with the glossary ‘main’ (which should be called `template.gls`) hasn’t been created.

Check the contents of the file `template.glo`. If it’s empty, that means you haven’t indexed any of your entries in this glossary (using commands like `\gls` or `\glsadd`) so this list can’t be generated. If the file isn’t empty, the document build process hasn’t been completed.

If you don’t want this glossary, add `nomain` to your package option list when you load `glossaries-extra.sty`. For example:

```
\usepackage[nomain]{glossaries-extra}
```

Try one of the following:

- Add `automake` to your package option list when you load `glossaries-extra.sty`. For example:

```
\usepackage[automake]{glossaries-extra}
```

- Run the external (Lua) application:

```
makeglossaries-lite.lua "template"
```

- Run the external (Perl) application:

```
makeglossaries "template"
```

Then rerun L^AT_EX on this document.

This message will be removed once the problem has been fixed.

ACRONYMS

This document is incomplete. The external file associated with the glossary ‘acronym’ (which should be called `template.acr`) hasn’t been created.

Check the contents of the file `template.acn`. If it’s empty, that means you haven’t indexed any of your entries in this glossary (using commands like `\gls` or `\glsadd`) so this list can’t be generated. If the file isn’t empty, the document build process hasn’t been completed.

Try one of the following:

- Add `automake` to your package option list when you load `glossaries-extra.sty`.
For example:

```
\usepackage [automake] {glossaries-extra}
```

- Run the external (Lua) application:

```
makeglossaries-lite.lua "template"
```

- Run the external (Perl) application:

```
makeglossaries "template"
```

Then rerun L^AT_EX on this document.

This message will be removed once the problem has been fixed.

S Y M B O L S

This document is incomplete. The external file associated with the glossary ‘symbols’ (which should be called `template.slo`) hasn’t been created.

Check the contents of the file `template.slo`. If it’s empty, that means you haven’t indexed any of your entries in this glossary (using commands like `\gls` or `\glsadd`) so this list can’t be generated. If the file isn’t empty, the document build process hasn’t been completed.

Try one of the following:

- Add `automake` to your package option list when you load `glossaries-extra.sty`.
For example:

```
\usepackage[automake]{glossaries-extra}
```

- Run the external (Lua) application:

```
makeglossaries-lite.lua "template"
```

- Run the external (Perl) application:

```
makeglossaries "template"
```

Then rerun `LATEX` on this document.

This message will be removed once the problem has been fixed.

CHEMICAL SYMBOLS

This document is incomplete. The external file associated with the glossary ‘chemical’ (which should be called `template.chs`) hasn’t been created.

Check the contents of the file `template.cho`. If it’s empty, that means you haven’t indexed any of your entries in this glossary (using commands like `\gls` or `\glsadd`) so this list can’t be generated. If the file isn’t empty, the document build process hasn’t been completed.

Try one of the following:

- Add `automake` to your package option list when you load `glossaries-extra.sty`.
For example:

```
\usepackage[automake]{glossaries-extra}
```

- Run the external (Lua) application:

```
makeglossaries-lite.lua "template"
```

- Run the external (Perl) application:

```
makeglossaries "template"
```

Then rerun L^AT_EX on this document.

This message will be removed once the problem has been fixed.

INTRODUCTION

1.1 Time Series and Challenges of a Data-Driven Society

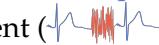
(THIS SHOULD HAVE SOMETHING LIKE: WE BELIEVE THERE IS A LACK OF TOOLS THAT CAN PUT THE INTUITION OF THE USER INTO THE ANALYSIS PROCESS OF TIME SERIES. SEVERAL METHODS ARE ALREADY AVAILABLE BUT WE SHOULD MAKE THEM AVAILABLE AS FUNCTIONAL TOOLS. I HAVE NOT DONE THAT, BUT THIS WOULD BE MY ULTIMATE GOAL AND MOVE TOWARDS THAT.

In recent years, the continuous increase in accessible wearable technology has contributed to a significant amount of data available. The continuous production of data from wearable devices through the usage of mobile phones, smartwatches, hearables, wristbands and other non-invasive wearable sensors has provided a valuable quantity of information. This data often comes as time series, being one of the most common data type in nature [puttinghuman]. As reported in *Tankovska et al.*, the wearable devices usage has more than doubled in the interval between 2016 and 2019, reaching 722 million [tankovska_23_2020] [novathesis-manual], leading to a large volume of time series data being gathered in all possible scenarios, by monitoring patients in healthcare institutions [cpd_medical_1, cpd_medical_2, cpd_medical_3, cpd_medical_4, dataset6, dataset7], tracking everyday activities of humans [cpd_har_1, cpd_har_2, review_1], recording machines in industrial processes or workers motion while performing their tasks [antonio, sara]. It has never been so easy to gather data about any aspect of our life, work, education, society or industry. Of course, having relevant information about a subject is beneficial, but the overwhelming amount of data brings tremendous challenges in the ability to save, process, analyze and retrieve interpretable and meaningful information from which we can act upon[bigdata]. Ultimately, it becomes even harder to have data well structured and labeled, considering that it is a sensitive and time consuming process, which complexity increases with data quantity. This is particularly problematic when developing machine learning applications (remember that Garbage-in Garbage-out - GIGO) [roh2019survey]. In the work of *Roh et al.* is mentioned that data scientists only

rely on a small portion of the available datasets because it is too expensive to label all the data available [roh2019survey], and this is just an example of how much data can be unused.

We believe that we should do more with the data we have and for that, tools should be available to support and help analysts to accelerate the process of information retrieval from time series by making it more expressive and intuitive. In this thesis, we propose several novel methods that contribute to the information retrieval problematic. These methods are designed to help in (1) tell the story behind the time series by means of a visual representation that highlights its structure and organization, (2) make the search of patterns and events with more expressive queries and (3) move towards distance measures that are more human readable. These contributions are part of a general work born with this thesis, called *grammar of time*.

1.2 Linguistic Nature of Time Series

Time Series are a visual domain, from which humans can create a good intuition. It is inherent to our ability to see relevant structures and patterns. The reader can imagine a recurrent shape, such as the **QRS complex** of an electrocardiogram (ECG) signal that is interrupted by a **noisy** segment (). When interpreting this signal, we see that it has 3 representative segments and that the first is very similar to the third one. We could then represent the signal by **A B A**. Some shapes may be harder to distinguish, for instance, consider an accelerometer signal of a subject while **walking** and shifting to **jogging** regime (). Or a change in the shape of the arterial blood pressure (ABP) signal when there is a change in the subject's posture (). In both cases, the signal has 2 structures of a similar representative periodic pattern (**A B**).

This visual intuition is also very clear when a (non-)experience analyst is searching for specific shapes or patterns in time series. The reader may agree that scientists or other professionals often resort to describe the shape they are looking for. For instance, a physician may say "*I am searching for the T-wave, that represents the large peak*" () or "*I am searching for the QRS complex, that looks like a sharp peak followed by a sharp valley*" ().

This visual intuition also happens when analysts are trying to find differences between classes of signals. For instance, the following shapes (1)  and (2)  are different because "*shape 1 has a peak where shape 2 doesn't*".

Time series are carriers of information and the presence of a change in the regimes of a time series or the presence of a specific shape in a segment of a time series may be associated with a specific occurrence in the physical world and be attributed a meaning. This notion of structure and meaning is a good approximation of what represents the foundation of a language: grammar and meaning [grammar].

Grammar is generally defined as the book of rules that constitutes the structure of a

language, and is modeled by the morphology and syntax [grammar]. The first is the structure of words, how these are built or morphed based on context, while the latter consists in organizing words in sequences to form larger linguistic units, such as sentences. Such as a language has morphological and syntax rules that represent its structural information, time series are also organized by a formal structure of ordered subsegments with specific morphological characteristics, organized to build larger segments. This introduces why this thesis is designated *grammar of time* and also introduces the reader further to the problematic that will be explored in this work. We will demonstrate how the developed solutions are helpful to several domains, with a special interest in showing how these can be helpful and meaningful in the context of occupational health.

1.3 Context and Relevance in Occupational Health

Work-related musculoskeletal disorders (WMSDs) prevail as the most common occupational disease in the European Union. These have a global impact on the well being of individuals and their quality of life in a range of working sectors [Irastorza2010], accounting for the second largest responsibility to disability worldwide [Luttmann2003]. These are specially prevalent as upper limb or neck disorder (with 42% of all WMSDs cases reported) [Seidel2019] in several industry sectors, such as textile and automotive, where production processes with pre-defined motions and actions have a repetitive/cyclic nature. This has a negative impact on the risk to develop musculoskeletal disorders, with tremendous consequences to both workers and companies, leading to absenteeism, early retirement and loss of productivity [Trabalhadores, Varandas19].

Several strategies have been implemented to identify, regulate and prevent occupational risk in manufacturing industries, such as (1) the inclusion of job rotation schedules, which promote a variation of the exposure throughout the working day [**jobrotation**] and (2) screening tools, for the assessment of occupational risk exposure, e.g. OCcupational Repetitive Action (OCRA), Rapid Upper Limb Assessment (RULA) or the Ergonomic Assessment WorkSheet (EAWS) [**ocra**, **rula**, **eaws**]. Nevertheless, these strategies are not optimal because they (1) are not automated, relying in observational methods and dedicated personal to inspect video records; (2) are not objective measures; (3) do not take into account differences among the worker's population, as anthropometric, age and experience variability; and (4) present single scores, being insufficient to explain the factors that contributed to this risk. With the advent of Industry 4.0, more companies are using modern strategies that follow digital solutions to provide direct and objective quantitative measures [**romero**]. An example of these incentives is the usage of wearable inertial devices for motion and posture tracking of workers.

Using inertial motion units (IMUs), time series can be collected, and relevant information can be directly measured, e.g. position and velocity of each body segment, postural angles between joints and gait parameters, making these important for ergonomics studies [Caputo2019, Hang19]. There are some limitations of using IMUs, mostly related

with the long term bias (sensor drifting) arising from long acquisitions and the empirical process to fine tune sensor fusion techniques. Other systems can be used for motion capture, such as camera-based methods, but these rely in fixed setup of cameras, which is unmanageable in real industrial scenarios [sara].

SE CALHAR, INCLUIR NA MOTIVACAO QUE EM ALGUMAS EXPERIENCIAS VERIFICAMOS DIFERENCAS ENTRE GRUPO ANTROPOMETRICOS

The usage of time series in this context can play an important role in supporting the decision of ergonomists and other professionals of the industry. In order to develop systems that can use motion and postural data for direct risk assessment and reporting, several challenges arise in the time series data mining domain. For instance, considering the periodic nature of most manufacturing tasks, risk factors are calculated by working cycle. Therefore, methods should be developed to identify working cycles with some variability in their periodicity. In addition, real occupational scenario might have interruptions or changes in the working behavior, due to abrupt production stoppage, shift breaks or even changing to another workspace that has a different motion pattern.

Other questions also arise by ergonomists, such as "*can we find a pattern that has a sharp rise in the IMU from the arm?*" or "*when the worker is using a hand tool to make screwing, can we see a periodic pattern on the IMU from the hand?*", which represent specific patterns with a descriptive shape that can be seen on the signals and are specific of a task. These events can be relevant to study their precise impact on the worker's occupational exposure. Having ways to detect these patterns is of great relevance as well. In this study, we will show how the proposed solutions can have an impact in these problems, and how they contribute to provide relevant visual feedback for information retrieval from the occupational data and make the search of specific patterns more intuitive and expressive, even for non-experienced data analysts, such as ergonomists.

1.4 Research Questions

The previous sections introduced our main motivations related with the development of methods for information retrieval, provided context regarding the grammar of time framework and how the proposed solutions can have a significant contributions in occupational health assesment.

This project addresses all the range of topics of time series, from the moment data is acquired (*sensing*), processed for information retrieval (*analysis*) and how it is used to act upon (*decision making*). The main objectives are related with the development of methods for information retrieval (*analysis*) from time series for better decision making.

1. **Sensing** - Explore in depth the available technology to measure motion and postural variables in occupational scenarios for risk assessment. This will take into account which variables are associated with a risk, based on ergonomic standards. These measures are reaturned as time series, which are processed in the topic *analyzis*;

2. **Analyzis** - In this topic, several research paths are explored. **A** - study (1) how to perform structural information retrieval in time series for segmentation based on change points and periodic points and (2) how are the segments related. For this, we applied a feature-based transformation of the time series and similarity based measures to make a meaningful visual representation, from which the segmentation points can be extracted. **B** - explore symbolic representations of time series, studying how these can be used for more expressive and intuitive pattern search with the help of regular expressions and ultimately natural language. **C** - From the textual representation of time series, study if we can make a higher leveled distance measure, following standard text mining methods. The resulting outputs of these methods can be used to get relevant information to take better decisions, namely in the occupational domain;
3. **Decision Making** - Study meaningful summarization techniques and explore several real-life examples in how the developed methods can help analysts be more aware of the data and move towards a more *democratized* usage of data mining tools for information retrieval in time series.

With this work, we intend to contribute to the state of the art in time series data mining with tools that provide more meaningful representations of time series, from which information can be retrieved with more meaning and at a higher level of abstraction, closer to the human intuition and visual interpretative abilities. This contributes towards more expressive methods and a democratization of these tools to accelerate the analysis process by experts in data mining and make non-experts capable of making high-level analysis.

1.5 Thesis Structure

This thesis provides a detailed description and explanation of the research work developed during the PhD program. It is organized in blablabla...

Figure X illustrates a guideline of the structure of this work, with a short description of each Chapter's content and description.

Chapter 1 introduced the main motivations, goals and context for the development of this thesis. Chapter 2 introduces theoretical concepts necessary to have a complete understanding of the work developed. It covers an introduction to motion and postural sensors used in occupational settings, time series, standard methods for its representation and analysis and text mining concepts. Chapter 3 presents the most recent works related with what we developed, namely in the topics of segmentation, summarization, pattern/event search and dictionary based classification. On Chapter 4 we start describing the data we used, explaining its source for both acquired data and publicly available, for what it was used and how. It includes a detailed description of the protocol used to acquire workers' motion data in real industrial scenarios. The algorithm developed for time series

CHAPTER 1. INTRODUCTION

structural information retrieval is explained in Chapter 5, while Chapter 6 covers the symbolic representation of time series. In this chapter is explained the exploratory path to use this novel representation in query search and classification tasks. Chapter 7 shows the application of the previous methods to an exhaustive set of examples, namely from the occupational scenario, and major results are presented. In addition, this chapter also provides a general discussion about the usage of these methods for decision making. Finally, Chapter 8 gives an overall remark over the outcomes of this thesis and a reflection over the contributions that the developed methods have in making time series preparation and data mining more expressive, quicker and more practicable for an ever increasing number of data available. Each chapter will have a short introduction to situate and contextualize the reader.

SENSING THE PHYSICAL WORLD

2.1 Sensing the Physical World

- Bring an overview of the sensors used for a general set of things - motion, physiological, etc...
- These measure the physical world and retrieve physical changes in what they measure
- These changes can be related with something meaningful and relevant that occurred and can be seen on the data
- In this work, we apply the methods in all kinds of scenarios to show their agnosticism to domains, but focus our attention to a specific context where introducing sensing technology can have a strong impact.

2.2 Occupational Variables in the Industry

- Occupational variables that affect worker's health have long been studied and already defined in several screening tools from standard ergonomic guidelines. We can name EAWS, OCRA, RULA, etc...
- These worksheets are a reference for ergonomists in identifying the variables of interest to measure the risk of each activity performed by a worker.
 - The multiple set of actions of a workstation can be analyzed
 - Typically, these variables are related with motion and posture of body segments. Several scenarios are studied and variables extracted are frequency, intensity and duration of activity. Study specific activities, measure the risk based on vibration from machine or tools (<https://www.cdc.gov/niosh/topics/ergonomics/ergoprimer/default.html>)
 - All these variables, being related with motion, should be studied
 - definition 1 - workstation
 - definition 3 - intensity, duration and frequency
 - definition 2 - vibration
 - hand - is a special case. We should measure the vibration on the grip

2.3 Sensing Worker's Health

The type of variables that are required to perform a risk assessment are related with motion, posture and vibration. These are physical variables that can be quantified by means of inertial sensors, such as accelerometer, gyroscope and magnetometer. These three sensors are used together to compensate limitations of each other in error accumulation from sensor drifting.

- With these sensors, we can measure the orientation of body segments in terms of other body segments or standard body planes (sagittal or frontal plane).

TIME SERIES FUNDAMENTALS

The content of this thesis is diverse and covers several different topics. Therefore, the reader will appreciate that we set the foundations that are necessary to fully capture the essence of this work. For this, we provide an introduction to each of the topics addressed, the global definitions and used notation in this work. We start by explaining occupational domain variables and corresponding sensors used to monitor these. The data of interest in this work is *time series* and the global definitions and notations are provided. Standard pre-processing methods, representation forms and distance measures are also explained. In this chapter, only global definitions will be made. Each further chapter will have additional and more contextualized definitions when needed.

3.1 Global Definitions

The information gathered by sensor are physical quantities that vary with time. These are called *time series* and are the main topic of this work.

Definition 1 - Time Series (T) - A time series is a sequence of real values ordered in time with length $n \in \mathbb{N}$: $T = (t_1, t_2, \dots, t_n)$. Several domains of data rely in the acquisition of multiple time series from multiple axis of the same sensor (e.g. the 3-axis accelerometer) or from multiple sources (e.g. IMU as a fusion of three different sensors), creating a *multi-dimensional time series*.

Definition 2 - Multi-Dimensional T (MT) - A *MT* is a set of $k \in \mathbb{N}$ time series belonging to the same acquisition: $\{T_1, T_2, \dots, T_k\}$.

Segments of interest are often searched inside a *time series*. A segment is called a *subsequence*:

Definition 3 - Subsequence - A *subsequence* is a segment of the time series with size $w \in \mathbb{N}$ and starting from a given position i and ending at position $i+w$ from the *T* or *MT*. A *subsequence* is delimited by two instants in time. This sample that segments a *subsequence* can be considered an *event*.

Definition 4 - Event Following the definitions of [event_def1, event_def2], which state that "*an event is a dynamic phenomenon whose behavior changes enough over time to be considered a qualitatively significant change*" and "*characterized by an interval of measurements that differs significantly from some underlying baseline*", we consider that an *event* is an instant in time e that indicates the presence of a relevant occurrence in the time series. Multiple *events* segment the time series into several *subsequences* of different lengths. Therefore, *event* detection is often considered time series segmentation [cpd_alan].

A common strategy used in time series data mining to find relevant *subsequences* or *events* is the moving window.

Definition 5 - Moving Window - A *moving window* is a process of sliding along a time series T to apply a specific method on each *subsequence* it hovers. The window has, such as the *subsequence* a predefined size $w \in \mathbb{N}$, which starts at a given position i and ends at position $i+w$. The process is iterative and can be made overlapping windows or not. The next window will start at $i+o$, being o the overlapping size.

With this process, each *subsequence* can be filtered, features can be extracted or distances can be measured. We will show several utilities of this technique further when introducing methods used to pre-process a raw time series.

Depending on the context and which conditions the data is gathered, the raw information can contain several sources of disturbance or should be transformed into another dimension to extract the information that matters. The set of tasks taken to prepare the *time series* to enhance information retrieval is called *pre-processing*.

The pre-processing steps we will discuss involve filtering, normalization and transformation.

3.2 Filtering

Time series have multiple sources of disturbance. This disturbance is usually called *noise* and is defined as an unwanted form of energy, but it can have multiple interpretations. It can be caused by internal sources inside a device, such as *white noise*, or be due to external sources, such as motion artifacts, wandering baseline, sensor detachment or the magnetic field from surrounding devices []. Any of these disturbances will affect the analysis stage and should be detected or removed.

3.2.1 Spectral Filtering

Several methods can be used to reduce the influence of noise in the analysis. Standard filtering methods, such as low-pass, band-pass and high-pass filters can be used to reduce the presence of specific frequency bandwidths that are not relevant. There are many configurations for these types of filters, being one commonly used the *Butterworth* filter.

3.2.2 Smoothing

Another often used method that has the purpose of reducing the presence of noise and represents a variation of a low-pass filter is the smoothing technique. Several variations of this technique exist, being the simplest one a moving average, which uses a moving window, calculating the mean in each iteration.

3.2.3 Wandering Baseline

Another type of disturbance on the data that is usually removed is a wandering baseline. An example typically occurs in ECG signals, where the respiration creates a wandering baseline on the signal. This type of disturbance has a very low frequency compared to the meaningful information on the data and can be removed by subtracting a *smoothed* version of the original data.

3.3 Normalization

Normalization of data is an important step in any data mining process. It is essential for data uniformization and scaling, while keeping the morphology and shape of the time series. Several methods can be used for this purpose, namely:

$$\bar{T} = \frac{T}{\max(|T|)} \quad (3.1)$$

the normalized signal (\bar{T}) is scaled by the absolute maximum of T . It is the simplest approach to normalization and guarantees that values are scaled linearly and their modulus cannot be higher than 1.

A variation of this process is the normalization by the range of amplitudes, which is as follows:

$$\bar{T} = \frac{T - \min(T)}{\max(T) - \min(T)} \quad (3.2)$$

here the signal T is normalized to range between [0,1]. Another normalization method, called *z-normalization*, is very commonly used and relies on the distribution of its values:

$$\bar{T} = \frac{T - \mu_T}{\sigma_T} \quad (3.3)$$

where the time series T is subtracted by its mean, μ_T and scaled by its standard deviation, σ_T . The resulting values represent how many standard deviations the signal is away from the mean.

3.4 Transformation

In information retrieval, data has often to be re-scaled, simplified, approximate or represented into another data type. Each can contribute in their own way to capture the most

relevant and meaningful information, or discover a new type of information that once was hidden in the original data. Dozens of methods exist for time series representation, such as Singular Value Decomposition (SVD) or wavelet transform, but only the ones relevant for this thesis will be explained.

3.4.1 Spectral Transformation

One of the first and most well known techniques suggested for time series transformation was the Discrete Fourier Transform (DFT) *fourier*. The idea behind this concept is that any signal, of any complexity, is a decomposition of a finite number of sine waves. Each wave is represented by a complex number, known as the Fourier coefficient, transforming the signal from the time domain to the frequency domain [fourier2]. This transformation allows to see the signal in a different manner, highlighting which frequencies concentrate more or less energy. It unveils the presence of specific types of noise or artifacts, or periodic shapes. Figure ?? shows the transformation of a signal into the frequency domain.

3.4.2 Feature-based Representation

Frequency properties are very relevant to characterize a time series, but others can also be used to get a full characterization of the signal. The process of feature extraction is also a transformation method commonly employed. It is performed by a moving window from which features are extracted. For each feature, f , a feature vector is computed.

Definition 5 - Feature Series - FA *feature series*, F , is a feature representation of a time series with size m that depends on the overlap size $o \in \mathbb{N}$ of the sliding process, making the size of the resulting feature series $m = \frac{n}{w-o}$. Considering the existence of a MT, the *feature series* becomes a *multi feature series* of stacked *feature series*, with size $f_{k,m}$.

When extracting more than one feature, these are grouped into a *feature matrix*.

Definition 6 - Feature Matrix - F_M A *feature matrix*, F_M , is the set of r features extracted for k time series, with size $r \times (k \times M)$.

3.4.3 Piecewise Aggregate Approximation

Another common used transformation method to simplify a time series and reduce its dimension is the piecewise aggregate approximation(PAA) [paa]. The new representation space will have size $1 < N \leq n$, in which N is a factor of the original size n . The searches to keep the average of the N equi-sized subsequences in which the original signal with length n is segmented, which results in $\bar{T} = \bar{t}_1, \bar{t}_2, \dots, \bar{t}_N$, such that [paa]

$$\bar{t}_i = \frac{N}{n} \sum_{j=\frac{n}{N}(i-1)+1}^{\frac{n}{N}i} t_j \quad (3.4)$$

An example is showed in Figure ???. The resulting signal has reduced noise and size, while conserving its trend.

3.4.4 Symbolic Aggregate Approximation

From this method, a new representation technique was born, transforming the signal from the numerical to the symbolic domain. It is called Symbolic Aggregate approXimation (SAX). This method applies PAA to a z-normalized time series and indexes a letter to each sample of the simplified signal based on the distribution of its amplitude values. The signal's amplitude values are separated in bins with equal probability. The number of bins is equal to the size of the *alphabet* chosen. Figure ?? shows an example of the signal transformed into a string with 3 letters in its alphabet. Such as the DFT, SAX opens doors to analyze time series in a completely different manner, profiting from the much acquired knowledge in text mining.

DEFINE A SAX OR SYMBOLIC TIME SERIES HERE

In this thesis we will use feature vectors for several purposes. We also propose a novel symbolic representation technique for time series that is used for expressive pattern search and classification. In order to perform search or classification, we have to be able to calculate the difference/similarity between two time series or *subsequences*.

3.5 Distance Measures

There is an exhaustive number of distance measure for time series, but two of the classical standard measures still provide state-of-the art results in most time series data mining tasks, namely the euclidean distance (ED) and the dynamic time warping (DTW).

3.5.1 Euclidean Distance

The ED is the most straightforward distance measure for time series. Let us consider two time series, Q and C , of length n , so that

$$Q = q_1, q_2, \dots, q_i, \dots, q_n$$

$$C = c_1, c_2, \dots, c_i, \dots, c_n$$

The distance between these two time series under the ED is:

$$ED(Q, C) = \sqrt{\sum_{i=1}^n (q_i - c_i)^2} \quad (3.5)$$

which represents the square root of the sum of the squared amplitude differences between the samples of each signal. Although the distance measure is simple to compute, it is highly susceptible to typical distortions on time series. When using ED, these distortions must be removed, otherwise, other methods, invariant to these distortions, should be used. Examples of distortions are the amplitude and offset distortion, phase distortion, and local scaling ("warping") distortion. The first can be compensated by the z-normalized ED:

$$z_ED(Q, C) = \sqrt{2m\left(1 - \frac{\sum_{i=1}^m Q_i C_i - m\mu_Q \mu_C}{m\sigma_Q \sigma_C}\right)} \quad (3.6)$$

where μ_Q and μ_C are the mean of the time series pair and σ_Q and σ_C are the standard deviation.

The *warping* distortion can be solved with an elastic measure. For this purpose, DTW is typically used.

3.5.2 Dynamic Time Warping

The DTW distance measures the alignment between two time series. Let us consider two time series, Q and C , of length n and m , respectively:

$$\begin{aligned} Q &= q_1, q_2, \dots, q_i, \dots, q_n \\ C &= c_1, c_2, \dots, c_j, \dots, c_m \end{aligned}$$

The alignment is measured by means of a distance matrix with size n -by- m , where the (i^{th}, j^{th}) cell of the matrix contains the $d(q_i, c_j)$ between the two points q_i and c_j , being $d = (q_i - c_j)^2$ [dtw]. Figure ?? shows an example of a distance matrix between two time series. The matrix fully describes the difference between the two time series and maps where these align. The mapping is made by a warping path, W , that represent the set of matrix cells that minimize the warping cost, also defined as the cumulative distance of these cells [dtw]

$$W = w_1, w_2, \dots, w_k, \dots, w_K; \quad \max(m, n) \leq K < m + n + 1 \quad (3.7)$$

$$DTW(Q, C) = \min \sqrt{\sum_{k=1}^K w_k} \quad (3.8)$$

The cumulative distance $\gamma(i, j)$ is calculated as $d(q_i, c_j)$ of the current cell added to the minimum distance adjacent to that cell:

$$\gamma(i, j) = d(q_i, c_j) + \min\{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\} \quad (3.9)$$

When two time series with the same length have a linear warping path, such that $w_k = (i, j)_k, i = j = k$, we have a special case of the ED. DTW has a time and space complexity of $O(nm)$ while the ED has linear complexity ($O(n)$).

3.5.3 Complexity Invariant Distance

A different type of distance measure is also used to cope with complexity invariance. This distance uses a complexity correction factor (CF) with an existing distance measure, such as ED [complexity]:

$$CD(Q, C) = ED(Q, C)(Q, C) \quad (3.10)$$

The CF is defined as [complexity]:

$$CF = \frac{\max\{CE(Q), CE(C)\}}{\min\{CE(Q), CE(C)\}} \quad (3.11)$$

where CE represents the complexity estimate of a time series. This estimate is calculated based on the intuition that if we could "stretch" a time series until it becomes a straight line, this line would be as long as the complexity of the signal. It can be computed as the sum of the $n - th$ discrete differences along the time series[complexity]:

$$CE(Q) = \sqrt{\sum_{i=1}^{n-1} (q_i - q_{i+1})^2} \quad (3.12)$$

These distance measures are performed on the original representation domain of time series. As we showed above, other representation techniques can be employed, creating opportunities for other types of approaches. In this work, we explore other representation techniques to create novel ways of exploring time series. Then, we find that the reader will appreciate that we describe other distance measures employed, namely in the feature-based domain.

3.5.4 Feature-based Distance

As mentioned, a feature series F can be computed from the original time series to represent it based on a specific feature. If the size of the *moving window* is equal to the size of the time series, than F is represented by a single value. Otherwise, each *subsequence* highlighted by the *moving window* is characterized by the feature and the F is computed as an array. When multiple features are extracted, each *subsequence* is characterized by a set of features, creating a feature vector \vec{f} with r feature values. Vector based distance measures can be used with these feature vectors to compare different time series or *subsequences*. There are several vector-based distance measures, including the already mentioned euclidean distance or the manhattan distance, but we will only describe the cosine similarity/distance.

The cosine similarity is a measure of the angle between two vectors determining if these are pointing in the same direction. Consider two feature vectors \vec{f}_A and \vec{f}_B . Their cosine similarity is computed as their normalized dot product [cosine]

$$CS = \frac{\vec{f}_A \cdot \vec{f}_B}{\|\vec{f}_A\| \|\vec{f}_B\|} \quad (3.13)$$

being $\|\vec{f}_A\|$ and $\|\vec{f}_B\|$ the euclidean norm of each feature vector, defined as $\sqrt{\sum_{i=1}^r f_{Ai}}$ and $\sqrt{\sum_{i=1}^r f_{Bi}}$, respectively [cosine].

3.6 Applying Distance Measures

Measuring distances between any time series gives the ability to compare them. It is the fundamental instrument for most time series data mining tasks. With a distance measure, we are able to compare groups of time series for classification purposes or compare *subsequences* with a query template and find if it occurs in the time series. Another relevant application of distance measures is its usefulness to retrieve relevant structural information of a time series by comparing each of its *subsequences* to all other *subsequences*. In this subsection, relevant methods applied with the help of the presented distance measures are explained to retrieve information from a time series. We will start with distance/similarity matrices.

3.6.1 Self-Distance Matrices

A time series can reveal relevant information when each *subsequence* is compared to all the other *subsequences* of the same time series. The result is a pairwise distance matrix that unveils *homogeneity*, *repetition* and *novelty* on the time series. Each are relevant assets for segmentation and summarization tasks.

Let X be a sequence with size N that can be a time series or a representation of a time series in the PAA or feature space, such that $X = (x_1, x_2, \dots, x_i, \dots, x_j, \dots, x_N)$. Each element of X can either be a single value or a vector with r features. Independently of that, a matrix S with size $N \times N$ can be computed, such that:

$$S(i, j) = d(x_i, x_j) \quad (3.14)$$

being d a distance measure between elements $x_i, x_j \in X$ for $i, j \in [1 : N]$. $S(i, j)$ represents a cell of S that contains the distance value. When $i = j$ the distance should be zero, therefore the diagonal of S has the lower values. Besides the main diagonal, other relevant structures can be found in S . These include *homogeneous blocks* and *paths* [**fmp**, **muller**].

Areas with lower distance are highlighted as *homogeneous* structures. These give an indication of *homogeneity* and *novelty*. *Homogeneity* because a *block* along the diagonal means that the time series has a constant behavior during the segment delimited by the *block*. *Novelty* because when S has multiple *blocks* along the diagonal, it shows that the time series shifted its behavior/regime. The moment there is a transition between *blocks* is a potential segmentation point.

When the time series has repeating *subsequences*, *paths* show up on S . The reason for it can be illustrated with the mentioned DTW measure. With DTW, the z-normalized euclidean distance matrix between two time series is computed and the optimal path is used as the final cumulative distance. This *path* is a perfect diagonal if the time series are exactly the same, but can be slightly distorted if these are slightly different. The same type

of *paths* appear in S indicating a low distance between two different *subsequences* of the time series.

3.6.2 Matrix Profile

As mentioned from the previous subsection, measuring all the distance pairs of a time series provides the ability to retrieve relevant structural information. Recently, a strategy was proposed to compute a one dimensional distance profile for a time series based on a z-normalized euclidean distance matrix. By keeping the *nearest neighbor* of each *subsequence*, we retrieve the *matrix profile*. The result gives the minimal distance pair of each *subsequence*, meaning that minimum values are *motifs* and maximum values are *discords*.

3.6.3 Template-based Search

The presented distances can also be used to retrieve a distance profile from a *template*. This type of mechanism belongs to the class of query-based search problems. The process to compute this distance profile involves sliding the template along the signal and applying a distance measure to each iteration. The result should indicate which *subsequences* are more (dis)similar to the used template.

In this thesis, we will propose other methods to perform query-based searches on a time series. Instead of using a *subsequence* template, we will be using text patterns.

3.7 Text Mining on Time Series

The representation of a time series into the symbolic domain makes possible the usage of text mining methods to analyze time series. The reader will appreciate that an association regarding time series and text is made, and are also introduced relevant methods from the text mining domain, used in this work.

3.7.1 Time Series Textual Abstraction

In SAX, the signal is transformed into a sequence of symbols. For this, each sample of the PAA representation is converted into a character, which can then form *words* and *sentences*. Other symbolic representation techniques exist on the literature and this work proposes a novel symbolic representation. In that sense, it is relevant to give the general background that makes this association between the original time series, a symbolic time series and text notation.

Definition 8 - Character - c A *character* is an unit symbolic element that represents a sample or *subsequence* of a time series. Sequences of characters form *words*.

Definition 9 - Word - w A *word* is the concatenation of a sequence of *characters*, giving a textual representation of a *subsequence*. Putting *words* together forms a *sentence*.

Definition 10 - Sentence - A *sentence* represents a group of *subsequences*. It is formed by joining sequences of symbolic *words*.

Definition 11 - Document - d The set of *sentences* in a time series are called a *document*. It represents the entire time series.

Definition 12 - Corpus The *corpus* is a collection of text material (group of documents). It represents the higher level of textual information. This collection is typically annotated and used for machine learning tasks. In this case, a corpus will be represented by the set of documents that describe a time series dataset.

Definition 13 - Vocabulary The *vocabulary* comprehends the set of all different words present in all time series.

3.7.2 Text Features

From the textual representation, text mining methods can now be applied. Here are introduced traditional methods applied for feature extraction of text data, namely the Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF).

Definition 14 - A BoW is a feature matrix representation of a corpus, being the feature the number of occurrences of each *term*, called the term-frequency (*tf*):

$$bow(t, d) = tf_{t,d} = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}, \quad (3.15)$$

being *t* the term that exists in a document, *d* the document and *t'* the term that belongs to document *d*. Here *t* can be a single *word* or an *n-grams*.

Definition 15 N-grams are a span of followed *words* that are counted in the BoW/TF-idf. An example of 2-gram from Figure ?? would be... This method resembles a *moving window* on time series, but for text. It makes the BoW/TF-idf model more robust since it makes it rely in more than single *word* statistics. Regarding time series, this method is relevant because it takes into account time dependencies between *words*, which reflects the time dependency seen in time series between *subsequences*.

The BoW is commonly used to vectorize the textual representation of each time series, but there is common knowledge in the text mining community that if a *term* occurs in all *documents* it is less relevant. To counteract this limitation, the TF-idf matrix is used.

Definition 15 - TF-idf The *Tf-idf* matrix increases the relevance of *t* by means of the *tf*, while reducing its importance in proportion to the number of *documents*, *d* that contain the term *t*. The model is defined by being a ratio between the *tf* and the *inverse document frequency* (idf), which is calculated as follows:

$$idf(t, D) = \log \frac{L}{|\{d \in D : t \in d\}|} \quad (3.16)$$

MC	Description	Example of Match
*	The preceding item will be matched zero or more times	$a^* \rightarrow$
+	The preceding item will be matched one or more times	$a^+ \rightarrow$
?	The preceding item is optional and will be matched, at most, once	$?a \rightarrow$
.	Matches any character	$a.b \rightarrow$
, &	Boolean operators - or, and	$a b \rightarrow$
(?=<)	Positive lookbehind - The string matches the item that is preceded by the pattern inside the lookbehind without making it part of the match	$a(?=<) \rightarrow$
(?<!)	Negative lookbehind - The string matches the item that is not preceded by the pattern inside the lookbehind	$?<!a \rightarrow$
(?=)	Positive lookahead - The string matches the preceding item that is followed by the pattern inside the lookahead without making it part of the match	$(?=)a \rightarrow$
(?!)	Negative lookahead - The string matches the preceding item that is not followed by the pattern inside the lookahead	\rightarrow

L is the total number of documents ($L = |D|$). The final equation of the *tfidf* model is the following:

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D) \quad (3.17)$$

Both *bow* and *tfidf* are matrices that have a vector representation of each *document*, where each element of the vector is the relevance of the *term*. That means that the cosine distance can be used to compute the difference between *documents*.

Due to the probabilistic nature of the *bow* and the fact that it contains discrete features, it is suitable to use in naive bayes classifiers. In the other end, the *tfidf* is typically used with linear support vector machine (SVM) classifiers.

3.7.3 Text Pattern Search

When writing or reading a document, we often have to use the *ctrl+F* key to search for specific words or expressions. This option let us search for direct matches, characters that belong to a word or expression, or even, use regular expressions (REGEX).

The last method is a parsing technique that is convenient to write text patterns, being more flexible than direct matches. It is based on regular languages, following a specific set of rules, and contains a set of meta-characters.

In order to understand the way that regular expressions work, some of the most used characters and REGEX primitives are presented as follows regex2:

STATE OF THE ART

State of the art in the topics mentioned previously as well as what has been done considering the advances in time series data mining.

4.1 Information Retrieval from Time Series

4.1.1 Event Detection

Most of the works available in event detection are focused in change point detection or segmentation. The found strategies are categorized based on (1) their ability to be used online or offline, (2) being univariate or multivariate, (3) based on a model or non-parametric and (4) being unsupervised or supervised [cpd_alan, review_1, review_2]. Regarding supervised methods, there are multi-class, binary and virtual classifiers, optimized for the purpose of detecting change points [review_cpd_1]. The advantage of supervised methods is to not only detect the change point, but give the nature of the change as well. Another example uses neural networks with transfer learning for segmentation [pedromatias]. However, supervised methods rely in very brittle training sets and class imbalance, since there are more in-state sequences than change point sequences [review_cpd_1]. Additionally, a problem reported by [cpd_alan] is that most algorithms were validating the performance of their algorithms in synthetic data, which given the nature of the application was not optimal. In that sense, a benchmark is now available for change point detection [cpd_alan], where methods can be compared on real-data. The proposed work uses this benchmark to compare itself with other non-supervised and offline methods.

Existing non-supervised methods include older but with state of the art performance in change point detection, such as the *Baysian Online* method (BOCPD) [bocpd], the *binary segmentation* (BINSEG) method [binseg] and the *segmentation neighborhoods* (SEGNEIGH) method [segneigh]. These methods have been reported successful in several domains [cpd_alan], however, the BOCPD only achieved good results when parameters were hypertuned, and the BINSEG and SEGNEIGH are not used in multidimensional domains. In addition, these methods are not reported to cope with a multi-time scale change

[cpd_alan]. An available repository provides an implementation of some of these offline methods [review_2], but these lack a visual output that might give the user an intuition over where a change point might be.

Another method, called FLOSS [eamonn1], relies in searching change points based on the nearest neighbors of subsequences, being very successful in real data domains. As it searches for nearest neighbors, the similarity between segments might be compared and used for summarization, but nothing is reported regarding multi-dimensional time series.

The ?? has been used for change point detection in the audio domain, based on a feature representation of the audio signal [MuellerZ19_FMP_ISMIR]. The advantage of using the ?? is the amount of information it provides for a specific time scale. In this work, we profit from these ideas applied in the audio domain, but extend its usage to other time series domains. The tool we propose can be used to detect events with context, associating the estimated events with patterns, (dis)similarities, periodicity and novelty. In addition, if being able to extract the information available in the ??, this tool can be extended to summarization tasks. Finally, although the search mechanism is based on a specific time scale, the process can be made recursive to perform multi-time scale searches recursively.

The proposed method highlights itself for being domain agnostic, work with both uni and multidimensional time series, give events with context by means of the visual information available, but also by the similarity measures in the matrix, that help in associating an event as a change or a periodic segment, and how similar are the segmented subsequences. It is unsupervised and works offline. It can be extended to work in multi-time scale problems with a special interest in time series summarization. We will demonstrate in this work how this method can bring novelty to the problematic of event detection, with a direct application to labelling and time series summarization.

The problems regarded in this work involve essentially the identification of cyclic information and anomalies. Typically, algorithms developed for these purposes may resort to (1) supervised machine learning (ML) methods, which require a certain level of annotation beforehand and (2) unsupervised methods, which are based on the similarity analysis of the signals or their features, without any prior information. Several methods found, employed in the analysis of inertial data, are used in the context of human activity recognition (HAR). The list of supervised ML methods is extensive and promising works are found to achieve this purpose. The application of neural networks [Lara2013], hidden Markov models [Zhu2009], decision trees [Jatoba2008], bayesian networks [Jatoba2008], and semi-automatic process [duarte1], among others, are algorithms capable of detecting and classifying various human actions. Nonetheless, most of the work done in this context only looks to identify previously defined actions like lying, standing, sitting down, move upstairs, etc., that might not be cyclic and rely on a significant amount of labelled data.

Several works that use unsupervised methods for the identification of cyclic information and anomalies are also found. The most simple method of cycle detection is the use of point references on the workplace to describe when a cycle starts and ends. Which is usually considered a system subject to flaws with a requirement for further adjustments

steps [Bauters2014, Bauters2018]. Other more reliable alternatives analyze features of the signal and search for periodic motion in those. An automated algorithm of segmentation was able to separate complex and multidimensional data into smaller segments that can be described through harmonic models. This algorithm revealed to be significantly useful to identify cyclic movement without any *a priori* knowledge of the input data, using a combination of a recursive least squares segmentation algorithm, a model fitting of damped harmonics, and in the end, a clustering analysis to classify the events [Lu2004, Lu2003]. The usage of features is of great relevance in unsupervised works, and methods are found to select adequate features for detection and classification tasks, such as in [machado2015]. Another example is the use of four-pass UKF (unscented Kalman filter) to produce an unified model with kinematic parameters. These may then be segmented by analyzing the parameter's zero crossing velocity and in the end uses a clustering algorithm to identify repetitive segments [Wang2015a].

Other methods rely on a self-similarity approach, namely [neuza], where cyclic information is segmented by searching for minimums, in the convolution of a segment of the signal with itself. The *Matrix Profile (MP)*, which is a method that compares all sub-sequences of a given time series with themselves through an euclidean distance, has also revealed promising results. In the end, it returns the minimum value distance for each segment, highlighting the moments of the time series which are similar within themselves [Yeh2018]. Additionally, autocorrelation revealed itself an useful tool, as the search over maximum values can infer the cyclic nature of the data [Bauters2014]. Finally, for anomaly detection in industrial scenarios, an interesting work applies an unsupervised method based on the clustering of time series segments to detect the execution of improper movements [duarte2].

The following work is inspired over an algorithm for the detection of musical structures on audio signals [Foote2000, audiolabs1, audiolabs2] by means of a *Self-Similarity Matrix (SSM)*. This sort of analysis of self-similarity to collect information about the periodicity has also been performed over video datasets. This type of analysis usually consists on a framework where a Fourier analysis is performed on an *SSM* to characterize and highlight the periodicity of the data from the video [Cutler2002, Cutler2000, Cutler1999].

4.1.2 Summarization

Very few strategies are found to make compact and meaningful representation of time series. The works that can be highlighted refer to time *snippets* and time series *bitmaps* [snippets, bitmap]. The first highlights the limitation of current methods in providing a satisfactory solution to time series *summarization*. It proposes a method that is able to segment the k most *representative* sub sequences of a time series, and use these elements as the summary. This strategy answers several of the discussed demands aforementioned in Section ??, namely the segmentation and similarity. Regarding the time series *bitmap* representation, the strategy is able to provide a coded bitmap with information on cluster,



Figures/keogh_examples.pdf

Figure 4.1: Strategies for time series summary found on the literature. These images are taken from the works from [snippets, bitmap]

anomaly and other regularities on data collection. These bitmaps were used as folder icons, and also answer several of the aforementioned characteristics, such as *similarity* and *events*. An example of both strategies can be seen on Figure ??.

Time series *shapelets* are also a method that could provide interesting results. However, the strategy is *supervised*, and the point of the proposed method is to have *no apriori* knowledge about the structure of the data, except the time scale in which the summarization is performed.

Other interesting strategies provide a transformation of time series into text and could be used for time series summarization, but are not able to suitably summarize a time series from the textual representation [sst, sax].

Strategies that are typically used to present information in a compact way are found in several domains. In text analysis, for instance, the relationship between repeating sequences is illustrated with arc diagrams [bitmap, arcplots]. These show where repeating sequences occur in a very concise way. This has a range of applications that include, for example text and DNA sequence analysis.

One domain that has a particular relevance in data visualization is genomics. Graphical genome maps are found to concatenate a significant amount of information in a very

compact way. Genome features and sequence characteristics are assessed with this visual strategy. An example can be found on Figure ???. This visualization strategy can provide increasing circular layers of information. Although we are used to look at time series from left to right, a circular representation can have benefits to concatenate the information we want to include.

In the musical domain, strategies have also been developed that summarize audio time series with segmentation techniques. One of the strategies that is common to be used involves detecting novelty instances on a similarity matrix representation of the audio signal, called *Self-Similarity Matrix* (SSM). This data structure provides a significant range of information that can be used to retrieve structural information, such as block and periodic structures [fmp1, fmp2, audiolabs1, audiolabs2]. This method will inspire our visualization strategy, which will be explained further.

4.1.3 Classification

Current available methods for time series classification are categorised as shape-based and structure-based. Existing approaches until the last decade were focused in shape-based similarity methods, while during the mid 2010's, methods that would seek the analysis of higher-level features started to be developed [Keogh2004].

Shape-based methods focus their attention in performing local comparisons between time series. Examples of well-known methods are the Euclidean distance (ED) or Dynamic Time Warping (DTW) [jlin2013]. Although both work well with short-length time series, the first has the inconvenient of needing time series with the same length, while also being

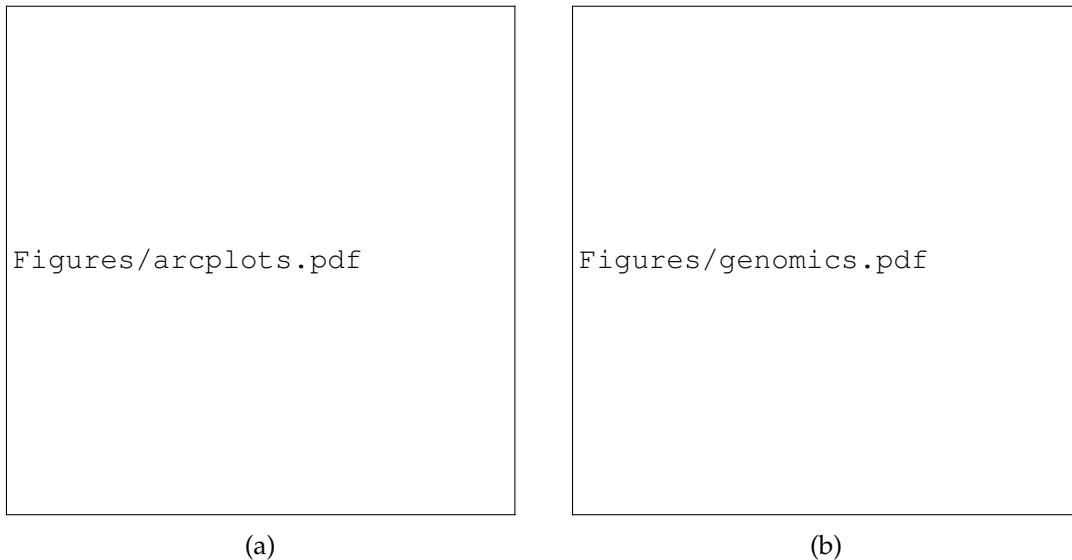


Figure 4.2: A - Diagram for string association. This image is taken from the works from [arcplots]; B - Circular plot by OmicCircos. Several layers (Circular tracks) identify genome position, expression heatmaps, correlation between expression and CNV, among other features. The image is taken from the works from Ying Hu, et al. [genomics].

sensitive to time misalignments. The latter is able to counteract this problem by means of determining the best alignment between two time series [Keogh2004, jlin2013]. These distance measures are usually combined with a k-Nearest Neighbour (k-NN) classifier to solve TSC tasks. The limitations of these techniques come with problems that include the presence of noise or long time series with characteristic sub-structures [BOSS].

In the other end, structure-based methods rely on broader aspects of time series such as the presence of specific morphological structures or patterns, being useful to classify long and noisy time series [BOSS]. Dictionary based methods fit into this category and have recently been used with great success. These techniques rely in a transformation of the time series into a symbolic feature vectors by means of a specific method, such as the *Symbolic Aggregate approXimation* (SAX) [SAX] or the *Symbolic Fourier Approximation* (SFA) [SFA]. The first approach proposed for TSC with symbolic representations was the work of Jessica Lin *et. al* with the *Bag of Patterns* (*BoP*) [jlin2013]. Further proposed methods were conceptually inspired on the *BoP*, using the same reasoning. Techniques such as *Bag of SFA Symbols* (BOSS) and Word ExtrAction for time SEries cLassification (WEASEL), from the same authors, use a similar reasoning but employ the SFA instead [BOSS, weasle].

Using syntactic methods has already been successful for several time series data mining tasks, mostly related with query search and classification. Besides, these methods, being dictionary-based, can be used to show similarity between subsequences by looking into the distribution of word counts. However, current methods rely mostly in incomprehensible sets of characters, such as *aaa*, which are hard to associate with a specific subsequence of the time series, therefore providing limited interpretability. In this work, we propose a method that literally translates the time series into sentences, such as that if a human was to describe a time series with text, it should be possible to separate these time series with the written words. We have seen natural language being used to include the human in the loop for more intuitive and meaningful query searches in time series [hil_naturallanguage]. Such as with SSTs, the purpose is to increase the expressiveness. This kind of descriptive power can be used to provide more intuitive feedback and increase interpretability to understand why a time series is different than others.

There is an existing method that is capable of providing visual interpretability of differences between time series, which is a structure-based method called *shapelets* [shapelets]. Shapelets are representative subsequences of the time series, which characterize a specific class. The advantage of this method is the higher interpretability because relevant shapes from the class can be highlighted [shapelets].

All the mentioned methods are a reference in TSC tasks with innovative concepts that merge ideas from the text-mining domain into TSC domain. One of the advantages of structure-based methods that rely in a dictionary-based concept is to use the words extracted as an interpretable model to differentiate time series. The histogram of words generated gives the user an understanding of which patterns better represent the time series and give an intuition over patterns that differ between classes of time series. This

provides a feedback and explanation over why a class is different than the other. However, dictionaries can be confusing, and the words generated are not intuitively associated with the patterns these represent in the time series. One method that went beyond the previously mentioned methods in that aspect is the SAX-Vector Space Model (SAX-VSM). This method used a weighted word vector representation of the time series and showed which are the relevant words for the classification process and what patterns these represent in the time series, demonstrating that the classification process can be interpretable by measuring the importance of the patterns found for each class of signals [sax_vsm].

The proposed method is built upon the same ideas as the BoP method but uses the SSTS Tool to promote the inclusion of the human reasoning in the classification process and provide more interpretable representations, as inspired by the work of SAX-VSM.

The method has been conceptually designed focusing in providing a solution that copes with (1) enabling the human intuition in the classification process, (2) be invariant to size, (3) have awareness of the order at which structures appear on the time series, (4) be domain agnostic, (5) have a flexible pre-processing to increase the representational power and (6) increase the readability. This method brings novelty by using literal natural language sentences to perform classification of time series, which can be customized by an analyst and moves towards a more readable output on distinguishing time series both visually and with keywords.

4.1.4 Search by Query

There is a large literature on time series similarity search, see [26] and the references therein. However, in most cases it is assumed that the query comes from a downstream algorithm, not a human. As such, there has been relatively little attention paid to the ability of humans to formulate meaningful queries. In principle one could do “query-by-sketching” and invite the user to draw the pattern she is interested in finding [15,16]. The recent “Qetch” system is a prominent example of this approach [15]. However, there are two possible limitations to such an approach: First, it is not clear that most people have the ability to sketch their query. For example, many people cannot even draw an accurate circle [25]. Secondly, as Figure 1 hinted at, classic distance measures may be too literal and limited in expressiveness to retrieve the desired pattern. As a simple example, suppose that a user wishes to retrieve all highly symmetric patterns. There is simply no way to do this with Euclidean distance or similar distance measures. Other researchers have noted these issues and proposed more flexible queries languages for time series. For example, the SDL (shape definition language) of [11]- allows the user to formulate “blurred” queries. However, we believe that most such systems are not accessible for the typical user. For example, in our proposed system, a 3-point-turn can be successfully queried by noting that the surge axis will exhibit three consecutive “bumps” and formulating the query Surge: [peak peak peak]. In contrast, SDL would require: (Shape triplespeak (width

ht1 ht2 ht3) (in width (in order spike (ht1 ht2 ht3) spike (ht1 ht2 ht3))). Several similar query systems based on regular expressions or SQL-like languages have been proposed, but none seem suitable for general use [17,20]. There have also been a handful of other attempts at natural language querying for time series [6,7]. None of these works seem to have been adopted by practitioners. We feel that this is because they probably suffer from too broad an ambition, proposing completely domain independent search. While domain independence is a worthy ambition (and our eventual research goal), it is clearly challenging. Even the word “spike” can have a different meaning for neuroscientists, economists, epidemiologists, and astronomers. In this work we take advantage of the fact that driving is a familiar, even quotidian, activity for most people, and therefore a domain for which most people have strong intuitions for. Moreover, this domain has a near unique property that allows a user to model the behavior they wish to find. We found that, in many cases we could glean intuition as to how a driver’s behavior would reveal itself in telemetry by simply “puppeteering” a smartphone equipped with an app that shows its acceleration and gyroscope readings. For example, by modeling a 3-point turn by sliding the smartphone on a desk, we can see that this behavior best revels itself on the surge axis as three consecutive bumps.

4.2 Occupational Health Sensing and Problems

DATA DESCRIPTION AND MANAGEMENT

Explain our sources of data to explore the methods developed in this work. Show why you chose this type of data and which purpose it has been used.

5.1 Public Datasets

- 5.1.1 Classification Benchmark - UCR
- 5.1.2 UCI Machine Learning Repository
- 5.1.3 Physionet
- 5.1.4 CPD Benchmark

5.2 Acquired Datasets

- 5.2.1 Office Job Dataset
- 5.2.2 Industrial Job Dataset

UNVEILING THE GRAMMAR OF TIME SERIES

In this chapter is described the process to unveil the structure of a time series. The next sections will start by giving an explanation of the relevant information that has to be retrieved and demonstrates how to perform it. The main method is inspired by what is done in *music structure analysis* for segmentation and audio-thumbnailing. The process follows the steps of building a similarity matrix by means of a feature-based representation of a time series. While having already been extensively studied for audio signal structure analysis [**Mueller15_FMP_SPRINGER**, **audiolabs1**, **audiolabs2**, **cpd_audio**], this knowledge has not yet been extended to other types of *time series* domains, which could greatly benefit from it [**muller_music_health**].

6.1 The Problem

Defining what is relevant in a time series highly depends on the context and purpose of the analysis, but globally, for any type of time series, there is a general interest in understanding how the signal is structured, specially for tasks related with data annotation/labeling. The structure of a time series is built of *segments* delimited by *events*. The problem is the search for *events* that are significant.

From definition 4 of Chapter ??, we highlight two primary considerations for the detection of events: (1) an event is a change in the behavior of the time series, and (2) it has to be significant both *statistically* and *qualitatively*. The *qualitative* aspect indicates subjectivity from the analyst because of the domain or context of the problem. Considering this, we will start by explaining the dimensions of the problem: (1) search and (2) type of significance.

6.1.1 Search Dimension

Figures ?? and ?? are an illustrated summary of the two dimensions of the problem. Regarding the *search* dimension, it is formed by three layers: (1) *dimensionality*: the search can be made in one or multiple time series. In the multidimensional space, events can occur simultaneously in several time series, but other events can be specific for each of them

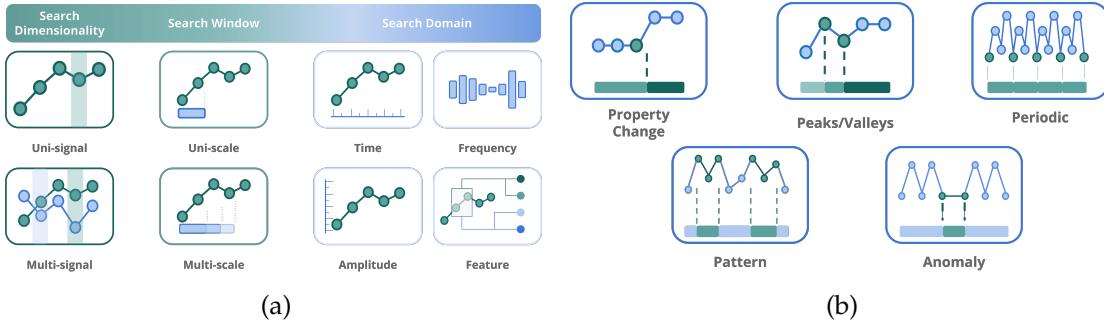


Figure 6.1: (a) Categories of search of events. In this case are shown Dimension, Window and Domain. (b) Examples of different type of events that can be considered significant in a time series.

(e.g. an accelerometer signal has 3 dimensions, but some gestures might be noticeable in only one of them); (2) *time scale*: *events* might occur in different time scales (e.g. when looking in a *TS* of 1 hour long, we might see some relevant events, but when looking for a *subsequence* of 10 minutes (zooming-in), other events are revealed); (3) *domain*: the search procedure might be made directly on the time series by means of time properties, a distance measure (e.g. Euclidean distance), or can be made on the representation level, such as the feature domain.

6.1.2 Type Dimension

In what regards the *event* type, we show in Figure ?? examples of events that are considered significant in a time series: (1) *Property change*: when the change of a property or set of properties is greater than a threshold, such as changes on the mean (FIND THUMBNAIL IMAGE) or frequency (M₁M₂M₃M₄M₅M₆M₇M₈M₉M₁₀M₁₁M₁₂M₁₃M₁₄M₁₅M₁₆M₁₇M₁₈M₁₉M₂₀M₂₁M₂₂M₂₃M₂₄M₂₅M₂₆M₂₇M₂₈M₂₉M₃₀M₃₁M₃₂M₃₃M₃₄M₃₅M₃₆M₃₇M₃₈M₃₉M₄₀M₄₁M₄₂M₄₃M₄₄M₄₅M₄₆M₄₇M₄₈M₄₉M₅₀M₅₁M₅₂M₅₃M₅₄M₅₅M₅₆M₅₇M₅₈M₅₉M₆₀M₆₁M₆₂M₆₃M₆₄M₆₅M₆₆M₆₇M₆₈M₆₉M₇₀M₇₁M₇₂M₇₃M₇₄M₇₅M₇₆M₇₇M₇₈M₇₉M₈₀M₈₁M₈₂M₈₃M₈₄M₈₅M₈₆M₈₇M₈₈M₈₉M₉₀M₉₁M₉₂M₉₃M₉₄M₉₅M₉₆M₉₇M₉₈M₉₉M₁₀₀M₁₀₁M₁₀₂M₁₀₃M₁₀₄M₁₀₅M₁₀₆M₁₀₇M₁₀₈M₁₀₉M₁₁₀M₁₁₁M₁₁₂M₁₁₃M₁₁₄M₁₁₅M₁₁₆M₁₁₇M₁₁₈M₁₁₉M₁₂₀M₁₂₁M₁₂₂M₁₂₃M₁₂₄M₁₂₅M₁₂₆M₁₂₇M₁₂₈M₁₂₉M₁₃₀M₁₃₁M₁₃₂M₁₃₃M₁₃₄M₁₃₅M₁₃₆M₁₃₇M₁₃₈M₁₃₉M₁₄₀M₁₄₁M₁₄₂M₁₄₃M₁₄₄M₁₄₅M₁₄₆M₁₄₇M₁₄₈M₁₄₉M₁₅₀M₁₅₁M₁₅₂M₁₅₃M₁₅₄M₁₅₅M₁₅₆M₁₅₇M₁₅₈M₁₅₉M₁₆₀M₁₆₁M₁₆₂M₁₆₃M₁₆₄M₁₆₅M₁₆₆M₁₆₇M₁₆₈M₁₆₉M₁₇₀M₁₇₁M₁₇₂M₁₇₃M₁₇₄M₁₇₅M₁₇₆M₁₇₇M₁₇₈M₁₇₉M₁₈₀M₁₈₁M₁₈₂M₁₈₃M₁₈₄M₁₈₅M₁₈₆M₁₈₇M₁₈₈M₁₈₉M₁₉₀M₁₉₁M₁₉₂M₁₉₃M₁₉₄M₁₉₅M₁₉₆M₁₉₇M₁₉₈M₁₉₉M₂₀₀M₂₀₁M₂₀₂M₂₀₃M₂₀₄M₂₀₅M₂₀₆M₂₀₇M₂₀₈M₂₀₉M₂₁₀M₂₁₁M₂₁₂M₂₁₃M₂₁₄M₂₁₅M₂₁₆M₂₁₇M₂₁₈M₂₁₉M₂₂₀M₂₂₁M₂₂₂M₂₂₃M₂₂₄M₂₂₅M₂₂₆M₂₂₇M₂₂₈M₂₂₉M₂₃₀M₂₃₁M₂₃₂M₂₃₃M₂₃₄M₂₃₅M₂₃₆M₂₃₇M₂₃₈M₂₃₉M₂₄₀M₂₄₁M₂₄₂M₂₄₃M₂₄₄M₂₄₅M₂₄₆M₂₄₇M₂₄₈M₂₄₉M₂₅₀M₂₅₁M₂₅₂M₂₅₃M₂₅₄M₂₅₅M₂₅₆M₂₅₇M₂₅₈M₂₅₉M₂₆₀M₂₆₁M₂₆₂M₂₆₃M₂₆₄M₂₆₅M₂₆₆M₂₆₇M₂₆₈M₂₆₉M₂₇₀M₂₇₁M₂₇₂M₂₇₃M₂₇₄M₂₇₅M₂₇₆M₂₇₇M₂₇₈M₂₇₉M₂₈₀M₂₈₁M₂₈₂M₂₈₃M₂₈₄M₂₈₅M₂₈₆M₂₈₇M₂₈₈M₂₈₉M₂₉₀M₂₉₁M₂₉₂M₂₉₃M₂₉₄M₂₉₅M₂₉₆M₂₉₇M₂₉₈M₂₉₉M₃₀₀M₃₀₁M₃₀₂M₃₀₃M₃₀₄M₃₀₅M₃₀₆M₃₀₇M₃₀₈M₃₀₉M₃₁₀M₃₁₁M₃₁₂M₃₁₃M₃₁₄M₃₁₅M₃₁₆M₃₁₇M₃₁₈M₃₁₉M₃₂₀M₃₂₁M₃₂₂M₃₂₃M₃₂₄M₃₂₅M₃₂₆M₃₂₇M₃₂₈M₃₂₉M₃₃₀M₃₃₁M₃₃₂M₃₃₃M₃₃₄M₃₃₅M₃₃₆M₃₃₇M₃₃₈M₃₃₉M₃₄₀M₃₄₁M₃₄₂M₃₄₃M₃₄₄M₃₄₅M₃₄₆M₃₄₇M₃₄₈M₃₄₉M₃₅₀M₃₅₁M₃₅₂M₃₅₃M₃₅₄M₃₅₅M₃₅₆M₃₅₇M₃₅₈M₃₅₉M₃₆₀M₃₆₁M₃₆₂M₃₆₃M₃₆₄M₃₆₅M₃₆₆M₃₆₇M₃₆₈M₃₆₉M₃₇₀M₃₇₁M₃₇₂M₃₇₃M₃₇₄M₃₇₅M₃₇₆M₃₇₇M₃₇₈M₃₇₉M₃₈₀M₃₈₁M₃₈₂M₃₈₃M₃₈₄M₃₈₅M₃₈₆M₃₈₇M₃₈₈M₃₈₉M₃₉₀M₃₉₁M₃₉₂M₃₉₃M₃₉₄M₃₉₅M₃₉₆M₃₉₇M₃₉₈M₃₉₉M₄₀₀M₄₀₁M₄₀₂M₄₀₃M₄₀₄M₄₀₅M₄₀₆M₄₀₇M₄₀₈M₄₀₉M₄₁₀M₄₁₁M₄₁₂M₄₁₃M₄₁₄M₄₁₅M₄₁₆M₄₁₇M₄₁₈M₄₁₉M₄₂₀M₄₂₁M₄₂₂M₄₂₃M₄₂₄M₄₂₅M₄₂₆M₄₂₇M₄₂₈M₄₂₉M₄₃₀M₄₃₁M₄₃₂M₄₃₃M₄₃₄M₄₃₅M₄₃₆M₄₃₇M₄₃₈M₄₃₉M₄₄₀M₄₄₁M₄₄₂M₄₄₃M₄₄₄M₄₄₅M₄₄₆M₄₄₇M₄₄₈M₄₄₉M₄₅₀M₄₅₁M₄₅₂M₄₅₃M₄₅₄M₄₅₅M₄₅₆M₄₅₇M₄₅₈M₄₅₉M₄₆₀M₄₆₁M₄₆₂M₄₆₃M₄₆₄M₄₆₅M₄₆₆M₄₆₇M₄₆₈M₄₆₉M₄₇₀M₄₇₁M₄₇₂M₄₇₃M₄₇₄M₄₇₅M₄₇₆M₄₇₇M₄₇₈M₄₇₉M₄₈₀M₄₈₁M₄₈₂M₄₈₃M₄₈₄M₄₈₅M₄₈₆M₄₈₇M₄₈₈M₄₈₉M₄₉₀M₄₉₁M₄₉₂M₄₉₃M₄₉₄M₄₉₅M₄₉₆M₄₉₇M₄₉₈M₄₉₉M₅₀₀M₅₀₁M₅₀₂M₅₀₃M₅₀₄M₅₀₅M₅₀₆M₅₀₇M₅₀₈M₅₀₉M₅₁₀M₅₁₁M₅₁₂M₅₁₃M₅₁₄M₅₁₅M₅₁₆M₅₁₇M₅₁₈M₅₁₉M₅₂₀M₅₂₁M₅₂₂M₅₂₃M₅₂₄M₅₂₅M₅₂₆M₅₂₇M₅₂₈M₅₂₉M₅₃₀M₅₃₁M₅₃₂M₅₃₃M₅₃₄M₅₃₅M₅₃₆M₅₃₇M₅₃₈M₅₃₉M₅₄₀M₅₄₁M₅₄₂M₅₄₃M₅₄₄M₅₄₅M₅₄₆M₅₄₇M₅₄₈M₅₄₉M₅₅₀M₅₅₁M₅₅₂M₅₅₃M₅₅₄M₅₅₅M₅₅₆M₅₅₇M₅₅₈M₅₅₉M₅₆₀M₅₆₁M₅₆₂M₅₆₃M₅₆₄M₅₆₅M₅₆₆M₅₆₇M₅₆₈M₅₆₉M₅₇₀M₅₇₁M₅₇₂M₅₇₃M₅₇₄M₅₇₅M₅₇₆M₅₇₇M₅₇₈M₅₇₉M₅₈₀M₅₈₁M₅₈₂M₅₈₃M₅₈₄M₅₈₅M₅₈₆M₅₈₇M₅₈₈M₅₈₉M₅₉₀M₅₉₁M₅₉₂M₅₉₃M₅₉₄M₅₉₅M₅₉₆M₅₉₇M₅₉₈M₅₉₉M₆₀₀M₆₀₁M₆₀₂M₆₀₃M₆₀₄M₆₀₅M₆₀₆M₆₀₇M₆₀₈M₆₀₉M₆₁₀M₆₁₁M₆₁₂M₆₁₃M₆₁₄M₆₁₅M₆₁₆M₆₁₇M₆₁₈M₆₁₉M₆₂₀M₆₂₁M₆₂₂M₆₂₃M₆₂₄M₆₂₅M₆₂₆M₆₂₇M₆₂₈M₆₂₉M₆₃₀M₆₃₁M₆₃₂M₆₃₃M₆₃₄M₆₃₅M₆₃₆M₆₃₇M₆₃₈M₆₃₉M₆₄₀M₆₄₁M₆₄₂M₆₄₃M₆₄₄M₆₄₅M₆₄₆M₆₄₇M₆₄₈M₆₄₉M₆₅₀M₆₅₁M₆₅₂M₆₅₃M₆₅₄M₆₅₅M₆₅₆M₆₅₇M₆₅₈M₆₅₉M₆₆₀M₆₆₁M₆₆₂M₆₆₃M₆₆₄M₆₆₅M₆₆₆M₆₆₇M₆₆₈M₆₆₉M₆₇₀M₆₇₁M₆₇₂M₆₇₃M₆₇₄M₆₇₅M₆₇₆M₆₇₇M₆₇₈M₆₇₉M₆₈₀M₆₈₁M₆₈₂M₆₈₃M₆₈₄M₆₈₅M₆₈₆M₆₈₇M₆₈₈M₆₈₉M₆₉₀M₆₉₁M₆₉₂M₆₉₃M₆₉₄M₆₉₅M₆₉₆M₆₉₇M₆₉₈M₆₉₉M₇₀₀M₇₀₁M₇₀₂M₇₀₃M₇₀₄M₇₀₅M₇₀₆M₇₀₇M₇₀₈M₇₀₉M₇₁₀M₇₁₁M₇₁₂M₇₁₃M₇₁₄M₇₁₅M₇₁₆M₇₁₇M₇₁₈M₇₁₉M₇₂₀M₇₂₁M₇₂₂M₇₂₃M₇₂₄M₇₂₅M₇₂₆M₇₂₇M₇₂₈M₇₂₉M₇₃₀M₇₃₁M₇₃₂M₇₃₃M₇₃₄M₇₃₅M₇₃₆M₇₃₇M₇₃₈M₇₃₉M₇₄₀M₇₄₁M₇₄₂M₇₄₃M₇₄₄M₇₄₅M₇₄₆M₇₄₇M₇₄₈M₇₄₉M₇₅₀M₇₅₁M₇₅₂M₇₅₃M₇₅₄M₇₅₅M₇₅₆M₇₅₇M₇₅₈M₇₅₉M₇₆₀M₇₆₁M₇₆₂M₇₆₃M₇₆₄M₇₆₅M₇₆₆M₇₆₇M₇₆₈M₇₆₉M₇₇₀M₇₇₁M₇₇₂M₇₇₃M₇₇₄M₇₇₅M₇₇₆M₇₇₇M₇₇₈M₇₇₉M₇₈₀M₇₈₁M₇₈₂M₇₈₃M₇₈₄M₇₈₅M₇₈₆M₇₈₇M₇₈₈M₇₈₉M₇₉₀M₇₉₁M₇₉₂M₇₉₃M₇₉₄M₇₉₅M₇₉₆M₇₉₇M₇₉₈M₇₉₉M₈₀₀M₈₀₁M₈₀₂M₈₀₃M₈₀₄M₈₀₅M₈₀₆M₈₀₇M₈₀₈M₈₀₉M₈₁₀M₈₁₁M₈₁₂M₈₁₃M₈₁₄M₈₁₅M₈₁₆M₈₁₇M₈₁₈M₈₁₉M₈₂₀M₈₂₁M₈₂₂M₈₂₃M₈₂₄M₈₂₅M₈₂₆M₈₂₇M₈₂₈M₈₂₉M₈₃₀M₈₃₁M₈₃₂M₈₃₃M₈₃₄M₈₃₅M₈₃₆M₈₃₇M₈₃₈M₈₃₉M₈₄₀M₈₄₁M₈₄₂M₈₄₃M₈₄₄M₈₄₅M₈₄₆M₈₄₇M₈₄₈M₈₄₉M₈₅₀M₈₅₁M₈₅₂M₈₅₃M₈₅₄M₈₅₅M₈₅₆M₈₅₇M₈₅₈M₈₅₉M₈₆₀M₈₆₁M₈₆₂M₈₆₃M₈₆₄M₈₆₅M₈₆₆M₈₆₇M₈₆₈M₈₆₉M₈₇₀M₈₇₁M₈₇₂M₈₇₃M₈₇₄M₈₇₅M₈₇₆M₈₇₇M₈₇₈M₈₇₉M₈₈₀M₈₈₁M₈₈₂M₈₈₃M₈₈₄M₈₈₅M₈₈₆M₈₈₇M₈₈₈M₈₈₉M₈₉₀M₈₉₁M₈₉₂M₈₉₃M₈₉₄M₈₉₅M₈₉₆M₈₉₇M₈₉₈M₈₉₉M₉₀₀M₉₀₁M₉₀₂M₉₀₃M₉₀₄M₉₀₅M₉₀₆M₉₀₇M₉₀₈M₉₀₉M₉₁₀M₉₁₁M₉₁₂M₉₁₃M₉₁₄M₉₁₅M₉₁₆M₉₁₇M₉₁₈M₉₁₉M₉₂₀M₉₂₁M₉₂₂M₉₂₃M₉₂₄M₉₂₅M₉₂₆M₉₂₇M₉₂₈M₉₂₉M₉₃₀M₉₃₁M₉₃₂M₉₃₃M₉₃₄M₉₃₅M₉₃₆M₉₃₇M₉₃₈M₉₃₉M₉₄₀M₉₄₁M₉₄₂M₉₄₃M₉₄₄M₉₄₅M₉₄₆M₉₄₇M₉₄₈M₉₄₉M₉₅₀M₉₅₁M₉₅₂M₉₅₃M₉₅₄M₉₅₅M₉₅₆M₉₅₇M₉₅₈M₉₅₉M₉₆₀M₉₆₁M₉₆₂M₉₆₃M₉₆₄M₉₆₅M₉₆₆M₉₆₇M₉₆₈M₉₆₉M₉₇₀M₉₇₁M₉₇₂M₉₇₃M₉₇₄M₉₇₅M₉₇₆M₉₇₇M₉₇₈M₉₇₉M₉₈₀M₉₈₁M₉₈₂M₉₈₃M₉₈₄M₉₈₅M₉₈₆M₉₈₇M₉₈₈M₉₈₉M₉₉₀M₉₉₁M₉₉₂M₉₉₃M₉₉₄M₉₉₅M₉₉₆M₉₉₇M₉₉₈M₉₉₉M₁₀₀₀M₁₀₀₁M₁₀₀₂M₁₀₀₃M₁₀₀₄M₁₀₀₅M₁₀₀₆M₁₀₀₇M₁₀₀₈M₁₀₀₉M₁₀₁₀M₁₀₁₁M₁₀₁₂M₁₀₁₃M₁₀₁₄M₁₀₁₅M₁₀₁₆M₁₀₁₇M₁₀₁₈M₁₀₁₉M₁₀₂₀M₁₀₂₁M₁₀₂₂M₁₀₂₃M₁₀₂₄M₁₀₂₅M₁₀₂₆M₁₀₂₇M₁₀₂₈M₁₀₂₉M₁₀₃₀M₁₀₃₁M₁₀₃₂M₁₀₃₃M₁₀₃₄M₁₀₃₅M₁₀₃₆M₁₀₃₇M₁₀₃₈M₁₀₃₉M₁₀₄₀M₁₀₄₁M₁₀₄₂M₁₀₄₃M₁₀₄₄M₁₀₄₅M₁₀₄₆M₁₀₄₇M₁₀₄₈M₁₀₄₉M₁₀₅₀M₁₀₅₁M₁₀₅₂M₁₀₅₃M₁₀₅₄M₁₀₅₅M₁₀₅₆M₁₀₅₇M₁₀₅₈M₁₀₅₉M₁₀₆₀M₁₀₆₁M₁₀₆₂M₁₀₆₃M₁₀₆₄M₁₀₆₅M₁₀₆₆M₁₀₆₇M₁₀₆₈M₁₀₆₉M₁₀₇₀M₁₀₇₁M₁₀₇₂M₁₀₇₃M₁₀₇₄M₁₀₇₅M₁₀₇₆M₁₀₇₇M₁₀₇₈M₁₀₇₉M₁₀₈₀M₁₀₈₁M₁₀₈₂M₁₀₈₃M₁₀₈₄M₁₀₈₅M₁₀₈₆M₁₀₈₇M₁₀₈₈M₁₀₈₉M₁₀₉₀M₁₀₉₁M₁₀₉₂M₁₀₉₃M₁₀₉₄M₁₀₉₅M₁₀₉₆M₁₀₉₇M₁₀₉₈M₁₀₉₉M₁₁₀₀M₁₁₀₁M₁₁₀₂M₁₁₀₃M₁₁₀₄M₁₁₀₅M₁₁₀₆M₁₁₀₇M₁₁₀₈M₁₁₀₉M₁₁₁₀M₁₁₁₁M₁₁₁₂M₁₁₁₃M₁₁₁₄M₁₁₁₅M₁₁₁₆M₁₁₁₇M₁₁₁₈M₁₁₁₉M₁₁₂₀M₁₁₂₁M₁₁₂₂M₁₁₂₃M₁₁₂₄M₁₁₂₅M₁₁₂₆M₁₁₂₇M₁₁₂₈M₁₁₂₉M₁₁₃₀M₁₁₃₁M₁₁₃₂M₁₁₃₃M₁₁₃₄M₁₁₃₅M₁₁₃₆M₁₁₃₇M₁₁₃₈M₁₁₃₉M₁₁₄₀M₁₁₄₁M₁₁₄₂M₁₁₄₃M₁₁₄₄M₁₁₄₅M₁₁₄₆M₁₁₄₇M₁₁₄₈M₁₁₄₉M₁₁₅₀M₁₁₅₁M₁₁₅₂M₁₁₅₃M₁₁₅₄M₁₁₅₅M₁₁₅₆M₁₁₅₇M₁₁₅₈M₁₁₅₉M

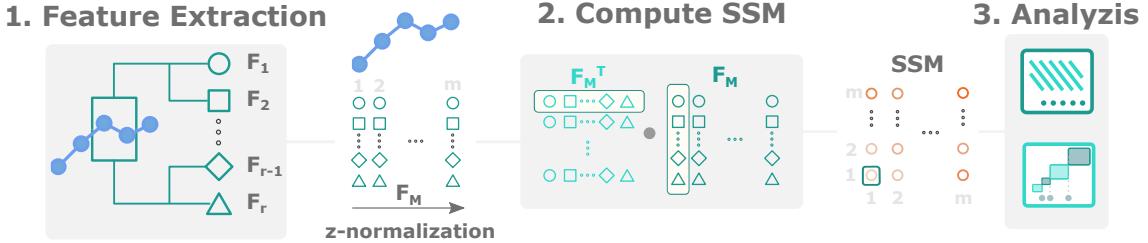


Figure 6.2: Main process to reach the ???. The information needed to calculate the ?? are the record and the input parameters: the window size (w) and the overlapping percentage (o). The first stage involves the feature extraction process, based on w and o values. Features are extracted on each window ($1, 2, \dots, N$), being N the total number of windows. From the first window (w_1), are extracted features (f_1, f_2, \dots, f_K , being K the number of features used). The feature number is also associated with a shape (circle, triangle, etc...). The features can be extracted on multivariate records, being M the number of records used. Each feature is positioned as a row on the F_M . Then follows the ?? computation.

Therefore, we propose an unsupervised methodology that searches for events in (1) uni and multi dimensional space, (2) with a fixed time scale, but with potential to be used in multi-time scale and (3) on a ?? computed by a feature space representation of the time series. The events that will be searched are any relevant change in the matrix related with a change point and/or periodic event.

We provide evidence that the proposed method is reliable for the detection of the mentioned events, supporting our claims with several examples in multiple time series domains (it is type agnostic) and comparing the results with state-of-the-art methods. Besides, we highlight that these events are all extracted from the same source of information (SSM), while also providing some insights in how this could be expanded for multi-time scale search, used for summarization and labelling.

6.2 Building the SSM

In this section, we explain the steps of the proposed method. The extraction of relevant events from time series starts by computing the ???. As explained in Chapter ???, this matrix has relevant structural information to retrieve *events*, namely *blocks* and *paths*. Figure ?? summarizes the steps involved in calculating the ???.

6.2.1 Feature Extraction

The structural information present on the SSM depends on the richness of the set of features into translating the changes and disruptions of the signal. Behavioral changes might be related with a variate set of features. As a feature may be sensitive for a type of change, the type of features should be diverse to identify a multivariate set of events and scan all types of signals. For this purpose, we used the available features from the TSFEL [barandas_tsfel_2020] Python library presented in the Feature Table xxxx.

The features are extracted in a moving window with a size w , specified by the user, with an overlap of size o . These two parameters have a large influence on the results. The first defines the time scale at which features are extracted, therefore the wider the window, the more *zoomed-out* will be the search. The second parameter defines the pixel-resolution of the resulting feature series, decreasing the amount of information (down-sampling) with a smaller overlap.

The extracted features are grouped into a feature matrix (F_M), where the rows represent each feature and the columns the corresponding *subsequence*, described by all features. Features extracted from a multidimensional record are ordered in the F_M as rows as well. The total number of rows can be, at maximum: $r \times k$, being k the number of time series being analyzed and r the number of features extracted, as illustrated in Figure ??.

6.2.2 Feature-based Self Similarity Matrix

After grouping all the features extracted, the next stage applies a distance measure to the feature space and computes the ???. This process consists in comparing each *subsequence* with all the others within the time series record. Since each column of the F_M is the feature characterization of each *subsequence* by the entire set of features, the comparison between segments is achieved by calculating the dot product between the z-normalized transposed F_M and itself:

$$SSM = F_M^T F_M \quad (6.1)$$

The dot product gives a similarity score based on the feature values of each *subsequence*. Cells of the ?? with higher similarity scores indicate that the corresponding *subsequences* have similar feature values [audiolabs1, audiolabs2]. As a result, the ?? provides rich visual information, highlighting structures, such as blocks and paths, that describe the signal's morphological behavior over time and its structure.

In Figure ??, the main structures are illustrated and highlighted in an example of an ?? [audiolabs1]. These structures are illustrated on the time series of example 1 (Figure ??). As mentioned in Chapter 5, the main structures are *blocks* and *paths*. Our proposed methods for annotation takes advantage of these main structures to extract the desired information.

Paths show recurrence of patterns, which is an indication of matching the morphology between corresponding *subsequences*. The example highlights circles in the *sf* layer, indicating recurrence. In *block "C"* are also visible *inverse-paths*, which indicate symmetry, which means that the corresponding *subsequences* are similar in reverse. The cross-path in *block "C"* means that the *subsequences* are periodic and symmetric.

Differently, *blocks* are square shaped structure that indicate homogeneous areas of the ??, which translate as constant behavior in the time series. The change between block structures along the main diagonal indicates a relevant change of morphology and behavior on the time series. In Figure ??, the ?? is segmented into several blocks on layer

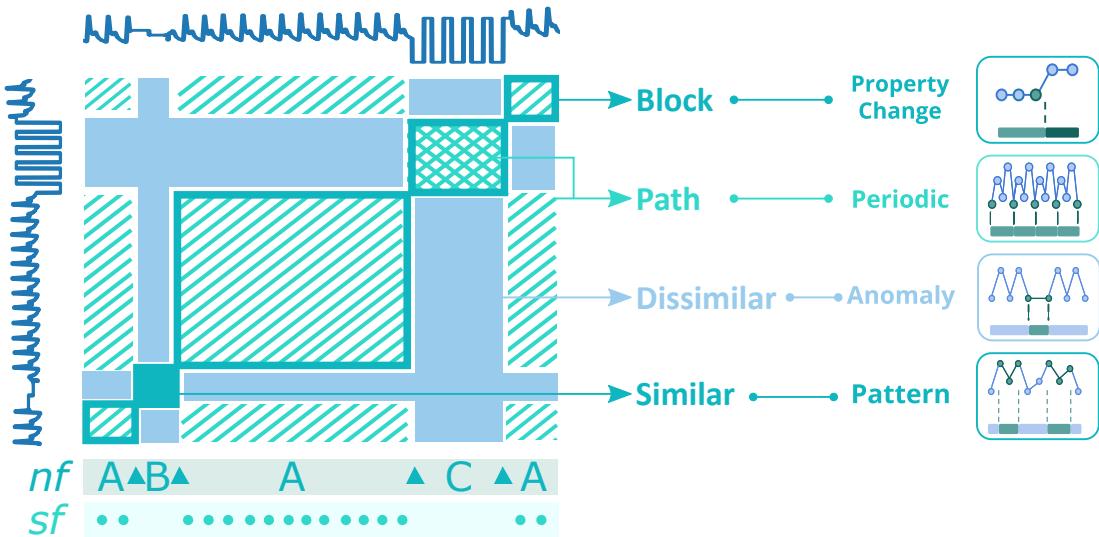


Figure 6.3: Description of informative structures of the SSM. Based on the ?? presented in Figure ??, we show a simplified view with highlights on the relevant structures. The record has 4 main structures: A - homogeneous segment, which corresponds to the BVP periodic signal; B - homogeneous segment of missed data; C - homogeneous segment with detachment of the sensor. The boxes (blue) highlight homogeneous blocks while the sub-diagonals (orange) highlight periodicity in the segment. *nf* and *sf* indicate that the novelty and similarity functions are computed based on this information. Segment C has a cross-pattern, which indicates periodicity and symmetry

nf. The triangular shapes indicate the change point that separate blocks "A", "B" and "C". Besides *paths* and *blocks*, the ?? provides distance measures between *subsequences*, which can be used to highlight dissimilar segments, such as anomalies or highlight very similar subsequences, such as motifs (pattern).

Several strategies were applied on the ?? to extract the mentioned information. Further are explained the approaches used.

6.3 Detection of Events

The ?? is a powerful visual tool *per se*, highlighting relevant information that could be missed if looking at the raw time series. However, being the information on the ??, it should be possible to retrieve it automatically. Here are described the methods used to retrieve *block* transition information and *paths* information.

6.3.1 Block Change Detection

The search for *novelty* is inspired by a method used in musical structure analysis and presented by Foote *et al.* [foote2000]. The process involves searching for transitions between *blocks* using a moving checkerboard square matrix. The result is a 1 dimensional function designated *novelty function* - *nova*.

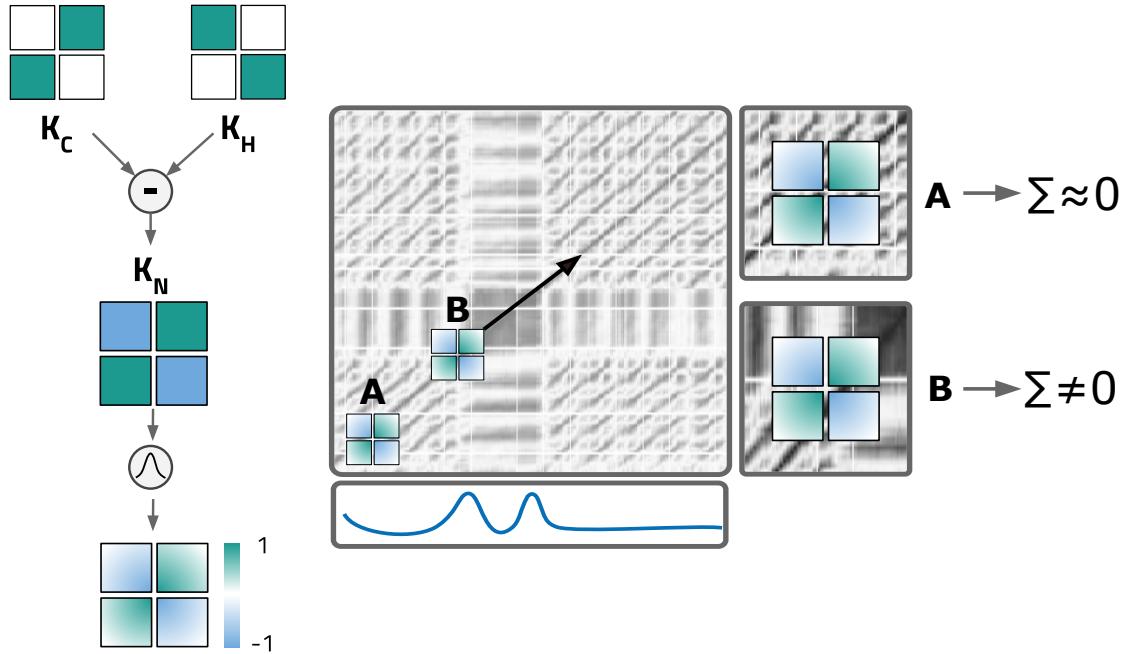


Figure 6.4: (left) Description of the matrix (kernel) used to compute the *novelty function*. The checkerboard pattern is achieved by combining kernel K_H - measure of homogeneity; and K_C - measure of cross-similarity. The resulting kernel (K_N) is combined with a Gaussian function to generate K_G . The Figure is based on the works of Mueller *et al.* [Mueller15_FMP_SPRINGER, MuellerZ19_FMP_ISMIR]; (right) The process to compute the novelty function is described. Kernel K_G is滑动 along the diagonal of the ?? to compute the *novelty function* presented as the bottom sub-plot. Positions A and B show the effect of block transitions on the *novelty function*. Figure based on the works of [Dannenberg2008, Mueller15_FMP_SPRINGER, MuellerZ19_FMP_ISMIR].

As showed on Figure ??, block transitions along the diagonal are represented by a checkerboard pattern. Detecting such patterns can be made by correlating a standard checkerboard matrix with the diagonal of the ???. For this, a sliding squared matrix, designated *kernel*, is used. As illustrated in Figure ??, the kernel has a checkerboard pattern and is combined with a Gaussian function to add a smoothing factor. The kernel, K_N , is a combination of two different kernels: K_H and K_C . The first is responsible for identifying the homogeneity of the ?? in each side of the center point along the diagonal. The higher the homogeneity, the higher will be the values in these sections. The latter measures the level of cross-similarity, returning higher values in cases of high cross-similarity. Therefore, when sliding the kernel K_N along the diagonal, a higher correlation value will be returned when it reaches a segment of the ?? with a similar checkerboard pattern. The result is the mentioned *nova* [Dannenberg2008, Mueller15_FMP_SPRINGER, MuellerZ19_FMP_ISMIR].

As showed in Figure ?? (left), the kernel in position A, which is placed on an area of high homogeneity, returns a value close to 0 when summing the product between it and the section of the ?? it overlaps. In the other end, in position B, the kernel reaches a segment with low cross-similarity and high diagonal similarity, which results in high correlation values with a checkerboard pattern. The *nova* is high in these transition

segments [**Dannenberg2008**, **Mueller15_FMP_SPRINGER**, **MuellerZ19_FMP_ISMIR**].

Each section of the kernel has the same size, L , being the total kernel size configured by $D = 2 \times L + 1$, with $L \in \mathbb{N}$. The kernel has an odd size to adapt zero values in centered points. It also has total size $D \times D$, being K_N defined by the following function [**Mueller15_FMP_SPRINGER**, **MuellerZ19_FMP_ISMIR**]:

$$K_N(i, j) = (a_i) \cdot (b_j) \quad (6.2)$$

being $a, b \in [-L : L]$ and "" representing the sign function, which indicates the sign of the value (1, 0 or -1).

A radially symmetric Gaussian function is used to smooth the Kernel, with the following equation [**Mueller15_FMP_SPRINGER**, **MuellerZ19_FMP_ISMIR**]:

$$\phi(p, u) = \exp\left(-\frac{1}{2L\sigma^2}(p^2 + u^2)\right) \quad (6.3)$$

being σ the standard deviation, equal for both x and y dimensions of the matrix, L the size of each kernel's section, and p and u the position in the x and y dimensions, respectively. The final kernel is computed by point-wise multiplication with the Gaussian function:

$$K_G = \phi \cdot K_N \quad (6.4)$$

The *nova* is calculated by correlating the kernel with the diagonal of the matrix:

$$nova(m) = \sum_{i,j=0}^{2L+1} K_G(a_i, b_j) SSM(m + a_i, m + b_j) \quad (6.5)$$

, being the sample of the novelty function $m \in [0 : N]$ and $a, b \in [-L : L]$.

The change point events are represented by local maxima (peaks) in *nova*, which can be detected by standard peak finding strategies.

6.3.2 Periodic Search

As aforementioned, *paths* indicate the presence of similarity and reoccurring patterns can be visualized on the ???. The moment in time the *paths* start indicates the position at which the period of the pattern begins. In order to find the periodicity, we compute the similarity function, s_f , which is calculated by summing the values of the ?? column-wise (either column-wise or row-wise would work, since the matrix is symmetric), being each element of the s_f calculated by:

$$s_f(x) = \sum_{i=0}^m SSM_{ix} \quad (6.6)$$

where i is the column position for the sum, s_f_j the sample of the function at position j and m the size of one of the dimensions of the ???, which is equal to the feature-series size. As

segments with similar morphology will be similarly described by the extracted features, the columns will have a similar representation, hence a similar value on the *sf*. In cases where the time series is periodic, the similarity function will enhance this behavior by having valleys at the moment the *path* starts. The identification of events related with the periodicity of a time series is then possible by searching for local minima (valleys) on the similarity function.

6.4 Experimental Evaluation in Selected Use-cases

After explaining the process to represent the time series into a feature-based similarity matrix and the methods used to retrieve information from it, we present selected use-cases from multiple domains to exemplify its universal usage.

6.4.1 Use-Case 1 - Human Activity

The example presented in Figure ?? shows the usage of the ?? on a record of Dataset 2.

In this example, the method was applied to all the 3-axis of the accelerometer data. All are showed and described with the sequence of activities as captioned in Figure ??.

The ?? was computed using a window size of 250 samples, and an overlap of 95 %. Blue indicate segments with higher similarity. Along the diagonal, these blocks are visible and change point events are estimated as the transition between these, highlighted by the *nova* function. The kernel used for this detection had a size of 45 samples.

In this example, we can identify that the detected change point events match with the activity transitions. Although all transitions are visible on the novelty function, the ones that correspond to transitions between similar segments of activities are harder to find, namely the transitions between walking activities. This is plausible since the properties of these segments are similar and the morphological difference is not as significant as when shifting between dissimilar activities (e.g. between *Laying* and *Walking*).

Any significant change in properties will be detected by the proposed method. As presented in Figure ??, at the end of the time series, the period in which the subject was performing the *Walking upstairs* activity is affected by other changes in the time series. These are significant and also correspond to *block* transitions, which are also evident in the novelty function. The proposed strategy, being unsupervised, is sensitive to any change, as long as it is observed as a significant change in the signal's properties

When *zooming-in* the *SSM* into segment A, which shows transitions between walking behaviors, the checkerboard pattern that highlights change points is clearer and the three different walking patterns are easily segmented. The two major peaks in the corresponding *nova* function are from these transitions, as presented in Figure ?? (left).

Considering that the signal is a walking behavior, the reader might question the fact that the periodicity of the walking pattern is not exhibited on the matrix. The reason is

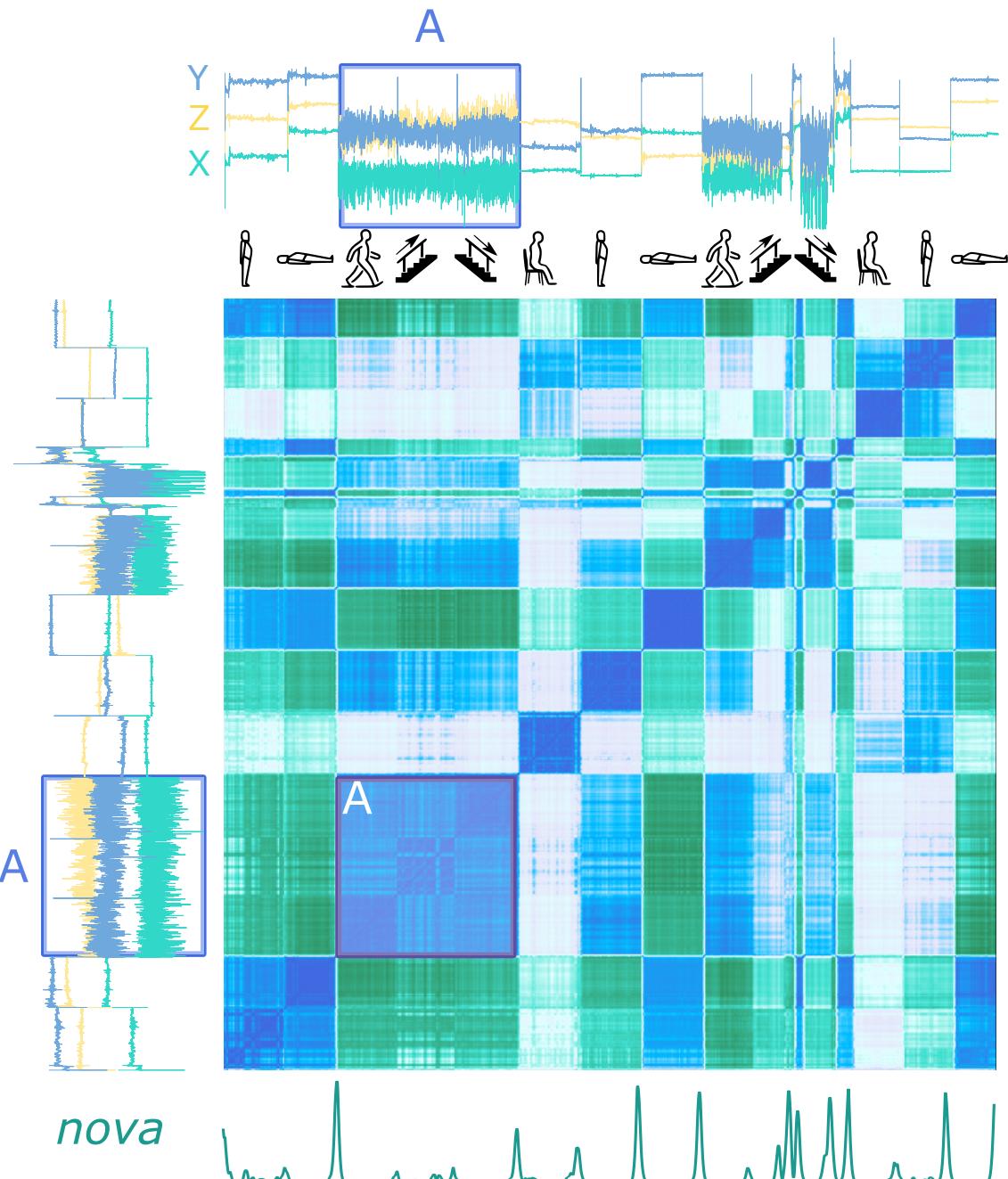


Figure 6.5: Change point event detection strategy applied on the SSM to search for change point events. The sequence of activities is presented as follows: *Sitting* → *Laying* → *Walking* → *Upstairs* → *Downstairs* → *Sitting* → *Standing* → *Laying* → *Walking* → *Upstairs*. The input variables used are $time_{scale}=250$ samples, $kernel_{size}=45$ samples, overlap=95%

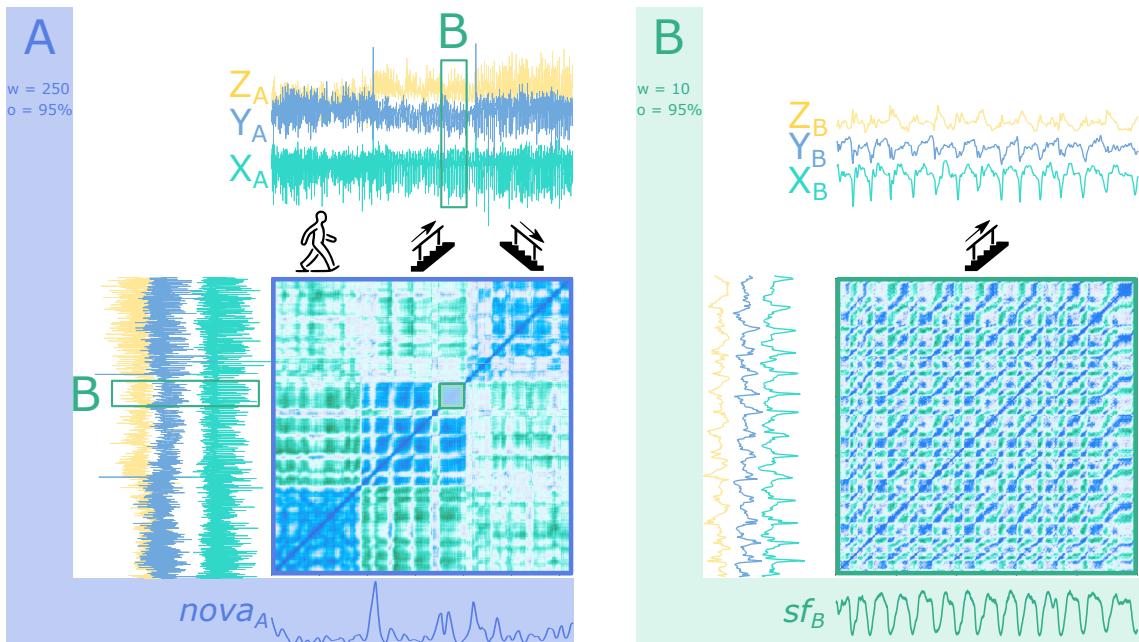


Figure 6.6: ABP signal change point detection. The parameters used were a size of 5000 samples, with an overlap of 75% and a kernel size of 25 samples.

that the window size used to compute the ?? of Figure ?? is too large. If features are extracted with a smaller window size, closer to the walking period, the *paths* that indicate the recurrence of shapes are visible. Figure ?? (right) shows the ?? built from the segment *B* of the original time series, with a window size of 10 samples and an overlap of 95 %. The matrix shows the *paths*, from which it is possible to extract the periods with the similarity function (sf_B).

6.4.2 Use-Case 2 - Medical domain

This is an experiment on arterial blood pressure data from a physionet dataset [tilt, PhysioNet]. There are change points indicated as a modification in the physiological signal due to a change in the posture of the subject. The change point is very well discovered with our proposed method, as signaled by the ground truth (*Posture Change* in Figure ?? left).

The shape of the ?? signal in each regime is very similar, being hard to notice by human eye where the change happened.

The same happens on the ?? signal from Figure ?? right. It displays a condition called pulsus paradoxus...the change point is hardly perceivable by human eye, but the proposed strategy is able to clearly show the difference in both regimes.

In ECG signals it is common to find artifacts [dataset6]. In this case we segmented noise from the ECG signal due to a jump performed by the subject with the novelty function. The segment is showed as a large area of red in the matrix, which indicates a dissimilarity with the rest of the signal, as well other structures from the ECG, which

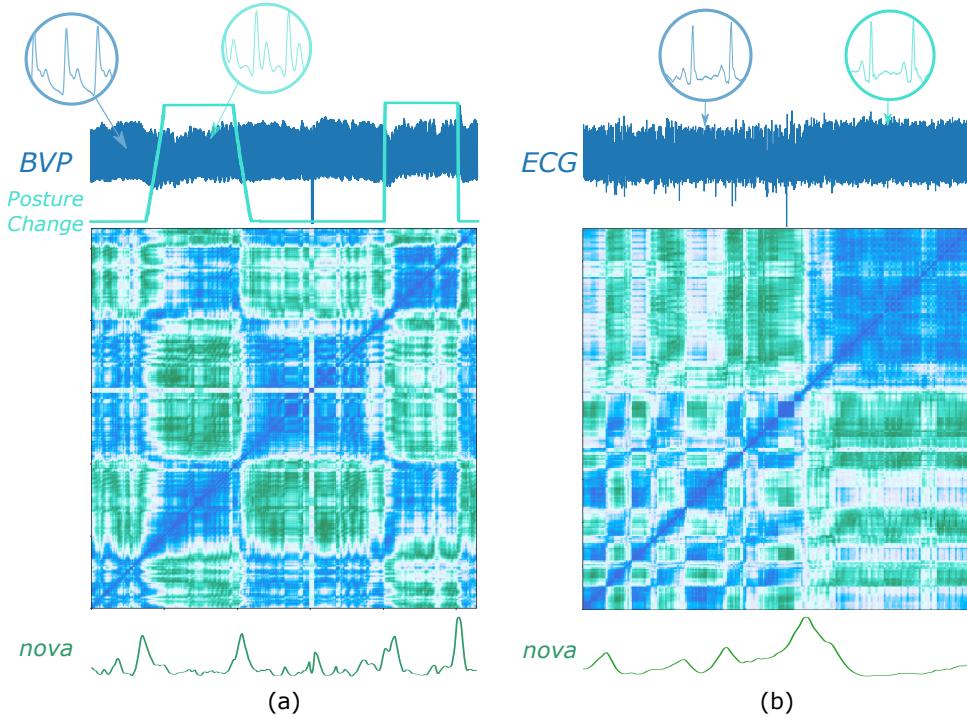


Figure 6.7: (a) ABP signal change point detection. The parameters used were a size of 5000 samples, with an overlap of 95% and a kernel size of 200 samples.

can be segmented with a smaller kernel size. In addition, we can compute the similarity function to segment the periods of the ECG.

6.4.3 Use-case 3 - Multidimensional

The proposed method accepts both single and multidimensional records. The difference regards the number of features extracted. As presented on Figure ??, the same set of features are extracted for each time series of the record and combined in the F_M .

Using a single time series of a multivariate record is optional and depends on the detection's purpose. In some cases, using a single time series from a multidimensional record can lead to missing relevant events undetected. An example of this can be seen on Figure ?? with record "Occupancy" from Dataset ??.

The record is a multi-dimensional time series that measures room occupancy based on temperature, humidity, light and CO₂. All events can only be detected if using several time series of the record [cpd_alan]. On Figure ??, left, a single time series was analyzed by the proposed method to detect relevant events, while Figure ??, right is the result of using all the time series of the record.

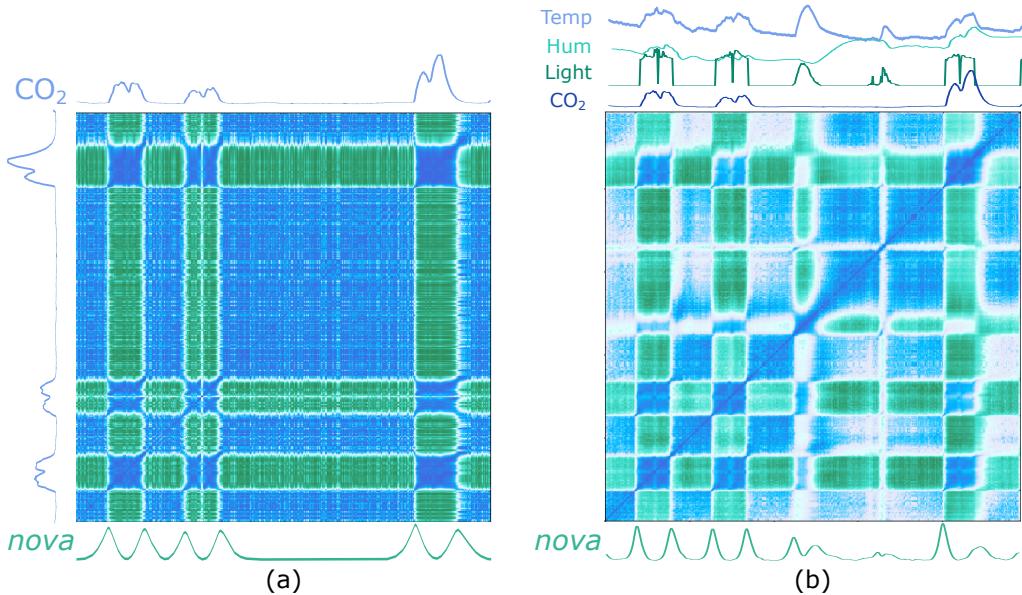


Figure 6.8: ECG signal from Dataset 7. The parameters used were a size of 500 samples, with an overlap of 95% and a kernel size of 50 samples and 10 samples, respectively.

6.5 Time Series Profiling

6.5.1 Elements with Relevance

6.5.2 Minimalist Design

(Find better words to describe this)

6.5.3 Summarize Time Series

LANGUAGE FOR TIME SERIES DATA MINING

Human language is foremost a means of communication, in which the information is represented by sentences, composed by words that can be broken into sequences of symbols. The diversity of possible arrangements of symbols and words gives the versatility in the process of transmitting information. The analysis of how symbols and words can be arranged in order to have a valid structure and comprise meaningful information involves the study of grammar and meaning davidCrystal.

Time series are, in their turns, carriers of information about a certain measure. These comprise sequences of ordered real domain numerical data observed during a given temporal interval, which are typically plotted as variations in amplitude. As aforementioned, the visual perception of the morphological behaviour of these series is, in many cases, enough to solve the problem and find the pattern that is being searched.

In terms of morphology, many attributes can be extracted from the visual perception of the signal, such as rising and falling slopes, concavity, direction, amplitude thresholding, frequency, time and amplitude range of a slope, among others. For instance, we can identify positive peaks by finding a rising slope followed by a falling slope, which is precisely the mechanism developed in Horowitz *et. al.* for peak detection in electrocardiography Horowitz.

We can think of this morphological description in terms of a characterization by means of features. These features can either be used to transform the signal into a symbolic representation, in which each sample is converted to a symbol, or a feature-based representation, in which a *word-feature series* characterizes the time series for a specific property.

In this chapter we present two proposed strategies to use text for pattern search on time series. The first profits of a symbolic characterization of the signal, introducing a novel representation, while the second, uses a feature based representation of the signal, being each feature attributed to one word.

7.1 Syntactic Search on Time Series

7.1.1 Preparing the Data

7.1.2 Connotation - The Symbolic Time Series

- connotation -

7.1.3 Expressive Syntactic Search

- search -

In this study, we propose a tool that focuses in ease simple query search tasks in time series, which we refer as **??**. This is achieved by an innovative methodology, where the user gives a syntactic nature to time series, which turns the search procedure less verbose and more related with the reasoning of the user in recognizing the desired pattern.

In 1943, McCulloch and Walter Pitts made the first theoretical description in a logic interpretation of the physiological events of neuron networks that served as inspiration for Kleene (1956) to create a set of rules that represent a finite state machine. Kleene described the nerve net as an arrangement of a finite number of neurons, where each has a sequence of states/events represented by integers. The state's values are influenced by the sensory response to the environment Kleene, and are said to be equally spaced in time. This algebraic description of neural nets can be extrapolated for time-series, in which the sequence of numbers is abstracted as a sequence of states to which values correspond to the sensory response of the environment.

The set of regular rules is a way of describing a specific sequence of states in the neural net - *a pattern*. This functionality has been extended into the field of text processing, in which this set of rules is able to describe a pattern as a sequence of characters, designated as a regular expression. Using a symbolic representation to characterize the sequence of states of time series in multiple attributes, regular expressions can be extended as a time series's parser to search patterns on it Thompson.

In 1980, Kenneth E. Iverson has discussed the importance of notation, nomenclature and language as tools of thought APL1. A regular expression is a good example of a tool of thought, by expressing the recognition of a pattern into a sequence of characters, but other examples can be given, such as in chemistry, botany and especially in mathematics.

E. Iverson believed that, although mathematical notation is not universal and unambiguous, it provides one of the best-known and best-developed examples of language as a tool of thought. With this in mind, he developed a programming language called APL (A Programming Language), which has the advantages of being universal and unambiguous, and incorporated the principles of mathematical notation.[\[referencia\]](#)

One of the fundamental characteristics of this tool is the provision of graphic symbols for the execution of functions and operations, which are meant to express the thought of the user in solving a problem. The tool presented in this work is inspired by this reasoning

and uses graphical symbols in the pre-processing and symbolic connotation steps. With this, the proposed tool profits of E. Iverson ideas to be intuitive, simple to use yet complex enough to reach the desired end, being a powerful tool of thought for query search in time series.

7.2 Towards Interpretable Time Series Classification with SSTS

Extend the usage of the symbolic mechanism. Having text

7.2.1 Using SSTS to translate Time Series

7.2.2 Vectorization of Time Series Documents

7.2.3 Towards Interpretable Results

7.3 Towards Natural Language for Pattern Search

7.3.1 Mapping Features to Words

7.3.2 Linguistic Operators

7.3.3 Natural Language Query for Time Series

The proposed method, conceptually developed based on text mining techniques, abstracts how a time series can be structured in a linguistic representation, similar to how the human would describe a time series with words. In order to introduce the reader with this abstraction and representation, we explain how we use *SSTS* to make this abstraction.

The transformation from the numerical domain to the textual domain is made using *SSTS* [**ssts**]. This method uses three steps to perform a query search on the time series and finding the corresponding pattern. The steps include (1) the pre-processing; (2) the symbolic connotation and (3) the search:

- Pre-processing: prepare the signal for the translation into the textual domain, removing noise or any disturbance in the signal that affects the pattern search;
- Connotation: transforms each sample of the time series into a character by extracting properties of the signal that are based on a conversion rule either defined by the user, or pre-defined in our vocabulary;
- Search: regular expression query that is matched on the textual pattern and corresponds to a *pattern* on the *time series*.

An example of the detection of shapes with the help of *SSTS* in a set of time series is made in Figure ???. The example shows the potential of this mechanism to create the description made in Figure ??.



Figures/SSTS_example.png

Figure 7.1: (Top) Using SSTS to detect the rising stage of a time series. Each step of the process is written described as follows: (1) pre-processing: Sm is the function *Smooth* with a window size of 25 samples; (2) connotation: $D1$, indicates the first derivate, from which each sample is converted to z - Flat, p - rising and n falling; (3) search - regular expression $p+$ searches for all sequences with 1 or more p characters. (Bottom) Example of sentence generation. Using the other search queries ($p+$, $n+$, $z+$), we can find the derivative patterns and convert it into ordered words.

APPLICATIONS, RESULTS AND VALIDATION

8.1 Validation Metrics

The validation metrics differ depending on the type of events we are searching. The ones used for evaluating the performance when using the *novelty function* are calculating the recall, precision and F1-measure in detecting the ground truth events, as well as calculating the distance at which the detected events are from the ground-truth events. In the other hand, the detection of periodic events by means of the *similarity function* is validated by calculating the number of correctly segmented periods. Each of this validation strategies are explained further.

8.1.1 Validate Events Detection

Each record is compared sample by sample and the performance is evaluated by considering the number of true positives (TP), false positives and negatives (FP, FN). As the ground-truth event is one sample, we added an error margin to have a proper validation. This margin dictates if an event can be TP, FP or FN. In this work, we used the window size as the chosen margin. The estimated events are considered one of the following categories:

- TP - is counted when the estimated event is in the margin around the ground-truth event;
- FP - is counted whenever it is out of a margin around the ground-truth event, or when there is more than one estimated event inside the margin;
- FN - is counted when there is no estimated event inside the margin of the ground-truth events.

From the count of TP, FP and FN, we are able to calculate performance metrics, such as Precision, Recall and F1-measure, which are calculated as follows:

$$Pre = \frac{TP}{TP + FP} \quad (8.1)$$

$$Rec = \frac{TP}{TP + FN} \quad (8.2)$$

$$F1 = 2 \cdot \frac{Pre \cdot Rec}{Pre + Rec} \quad (8.3)$$

Additionally, the distance of the TP events from the ground-truth events is calculated with several distance-based metrics, namely the mean-absolute-error (MAE), the mean-squared-error (MSE) and the mean-signed-error (MsE):

$$MAE = \sum_{i=1}^k \frac{|g_i - e_i|}{k} \quad (8.4)$$

$$MSE = \frac{1}{k} \sum_{i=1}^k (g_i - e_i)^2 \quad (8.5)$$

$$ME = \frac{1}{k} \sum_{i=1}^k (g_i - e_i) \quad (8.6)$$

The precision measure is relevant to indicate if the method is able to only estimate events that belong to the ground-truth category, while the recall measure is an important indication of how many ground-truth events are missed in the estimation of the method. Both measures are combined in the F1-measure. The true negative samples are not considered since most of them would be correctly estimated, which would wrongly improve the accuracy.

The distance-based metrics evaluate how far are the TP from the corresponding ground-truth events (MAE and MSE) and which is the direction of estimation of events (if before or after the ground-truth events - ME).

These are the metrics employed for all datasets except for *Dataset 8*, for which the evaluation was made considering their internal measures, as explained in [cpd_alan].

8.2 Segmentation Performance

In this section, we present several examples in how this method is useful for the segmentation of time series. The reader will appreciate that we also provide a measure of the algorithm's performance considering ground truth events (as presented in Section ?? and ??), while also testing our proposed solution by comparing it with several methods for change point detection from the *Turing Change Point Detection Benchmark* [cpd_alan].

In addition, we give insights about how this method could be used to summarize a time series and assist the labelling process of time series.

We make available all the code and results on the online repository.

This section is divided into three main categories: (1) Validate the usage of the ?? on segmentation with several use-cases; (2) Provide intuition over the parameters used,

Dataset	Signals	# Ch	Task	TP	FP	FN	Prec (%)	Rec (%)	F1 (%)
Dataset 1	ACC	3	HACP	98	16	16	0.860	0.860	0.860
Dataset 2	ACC-GYR	6	HACP	157	18	22	0.897	0.877	0.887
Dataset 3	ACC-GYR	6	HACP	1378	313	263	0.815	0.840	0.827
Dataset 4	ACC-GYR	12	HACP	499	71	38	0.875	0.929	0.902
Dataset 5	EMG	8	Act/Rel	309	0	72	1.000	0.811	0.811
Dataset 6	ECG	1	Noise	132	25	10	0.841	0.930	0.883
Dataset 7	ECG	4	Noise	21	2	3	0.913	0.875	0.894
Total	N.A.	N.A.	N.A.	2629	465	386	0.850	0.872	0.861

Table 8.1: Overall results for the performance of the method on change point detection. The dimension of the records is presented on the column *# Ch*, as well as the types of signals used and the task in which applied (HACP - Human Activity Change Point detection; Act/Rel - Activation/Relaxation of the EMG detection and Noise detection).

explain specific use-cases results, show the difference when using multi-dimensional time series and comments on the scalability and speed (3) comments on how to explore the usage of the ?? for summarization and labelling.

The proposed method has been tested on publicly available datasets from different domains to infer its performance on detecting change point events. These datasets have categorized labels that were used to generate ground-truth events. These include different contexts (HAR, Hand Posture, Noise Detection, etc...) and different types of data (Inertial data, EMG and ECG). More details are given on the problem associated with each dataset on Section ??.

The method has been computed in the same conditions and by following the same procedure for all records of all datasets. The features used have been the same for each record, varying the time scale parameter, the overlap size of the sliding window and the kernel size parameter. The peak detection strategy was the same for all records, which is based on a threshold value. The threshold value varied for each record.

Results for publicly available datasets are presented in Tables ?? and ???. Table ?? indicates the performance in detecting the change point events.

8.2.1 Further Application 1: Search by Example

8.2.2 Further Application 2: Multidimensional Segmentation

The proposed method accepts both single and multidimensional records. The difference regards the number of features extracted. As presented on Figure ??, the same set of features are extracted for each time series of the record and combined in the F_M .

Using a single time series of a multivariate record is optional and depends on the detection's purpose. In some cases, using a single time series from a multidimensional record can lead to missing relevant events undetected. An example of this can be seen on Figure ?? with record "Occupancy" from Dataset 8.

Dataset	T_s (s)	MAE/T_s	MsE/T_s
Dataset 1	5	0.53	-0.12
Dataset 2	10	0.29	-0.07
Dataset 3	1	0.34	-0.04
Dataset 4	25	0.23	-0.00
Dataset 5	1	1	-0.13
Dataset 6	10	0.12	-0.09
Dataset 7	1	0.17	-0.06
Average	N.A.	0.32	-0.07

Table 8.2: Distance error as a ratio of the time scale (T_s) for the detected TP.

The record is a multi-dimensional time series that measures room occupancy based on temperature, humidity, light and CO₂. All events can only be detected if using several time series of the record [cpd_alan]. On Figure ?? .left, a single time series was analyzed by the proposed method to detect relevant events, while Figure ?? .right presents the application of the method to all the time series of the record. The results are different because the information from all the time series is combined, while with the single dimensional record, the detected events resulted from the information available on the single time series.

8.3 SSTS Performance

8.3.1 Expressiveness Measure

8.3.2 Classification Results

8.3.3 Interpretability of Data Representation

8.4 Natural Language Search Performance

8.5 Application to Occupational Scenario

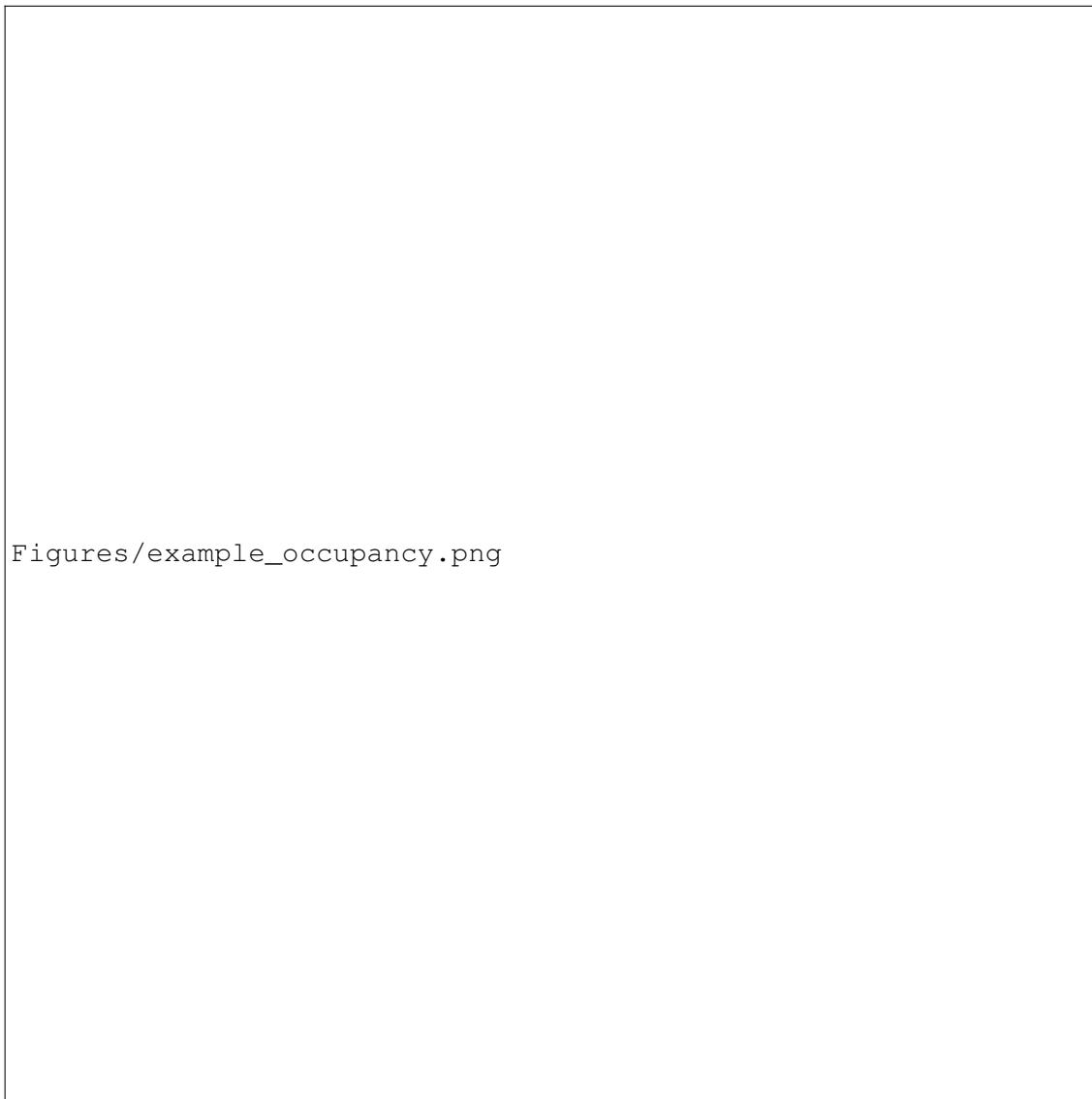
8.5.1 Storytelling Occupational Data

1 - Example of the data we acquired in Volkswagen when we arrived and tried to compare two workstations. We were trying to find specific cyclic moments and labelled it by hand. We can use this tool to make this identification (SSM):

2 - Describe the pattern by means of words or a regular expression

8.5.2 Pattern Search in Occupational Data

8.5. APPLICATION TO OCCUPATIONAL SCENARIO



Figures/example_occupancy.png

Figure 8.1: Proposed method applied on "Occupancy" record of Dataset 7. A single time series of the record is used to extract events.

CONCLUSION

10

FUTURE WORK

A

NOVATHESIS COVERS SHOWCASE

This Appendix shows examples of covers for some of the supported Schools. When the Schools have very similar covers (e.g., all the schools from Universidade do Minho), just one cover is shown. If the covers for MSc dissertations and PhD thesis are considerable different (e.g., for FCT-NOVA and UMinho), then both are shown.

B

APPENDIX 2 LOREM IPSUM

This is a test with citing something [ecoop12-dias] in the appendix.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetur adipiscing elit. In hac habitasse platea

dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetur.

ANNEX 1 LOREM IPSUM

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum

wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetur.



