Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

Anne Marie Skaar Hasund

A Data-Driven CBR and Clustering Method for Identifying Physical Activity Phenotypes

Master's thesis in Computer Science
Supervisor: Kerstin Bach
Co-supervisor: Aleksej Logacjov
November 2023

NTNU
Norwegian University of
Science and Technology

Anne Marie Skaar Hasund

A Data-Driven CBR and Clustering Method for Identifying Physical Activity Phenotypes

Master's thesis in Computer Science
Supervisor: Kerstin Bach
Co-supervisor: Aleksej Logacjov
November 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

Abstract

This master thesis investigates innovative methods for identifying physical activity phenotypes, primarily utilising Case-Based Reasoning (CBR) and clustering. The study explores the effects of varying the number of clusters, global similarity measures, and data representations in the pursuit of more precise and actionable results. CBR has advantages over other AI approaches in its transparency, which makes it ideal for interdisciplinary work, such as between computer science and public health research.

Physical inactivity remains a pressing global health concern, contributing significantly to healthcare expenditure and straining healthcare systems. While public health recommendations exist, they often follow a one-size-fits-all approach, neglecting the unique needs and activity patterns of individuals. To provide tailored guidance on physical activity, it is essential to identify and explore population clusters characterised by similar activity patterns.

The results indicate that a 4-cluster solution may be optimal for identifying meaningful physical activity phenotypes. Data-driven global similarity measures are found to have little impact on clustering when local similarity measures already account for attribute distribution. In conclusion, this research contributes a generalised method for identifying physical activity phenotypes, offering a template for future work. The combination of CBR and clustering provides a promising avenue for addressing the complexities of physical inactivity and enhancing personalised guidance to promote healthier lifestyles.

Samandrag

Denne masteroppgåva undersøker nyskapande metodar for å identifisere fenotypar for fysisk aktivitet, hovudsakleg ved bruk av Case-Based Reasoning (CBR) og gruppering. Studien utforskar effektane av å variere talet på grupper, globale likskapsmål og datarepresentasjonar for å oppnå meir presise og nyttige resultat. CBR har fordelar framfor andre tilnærmingar innan kunstig intelligens på grunn av si openheit, noko som gjer CBR ideelt for tverrfagleg samarbeid, som til dømes mellom datateknologi og folkehelse.

Fysisk inaktivitet er framleis eit presserande globalt helseproblem som bidreg betydeleg til auka helseutgifter og press på helsetenesta. Sjølv om offentlege helsetilrådingar eksisterer, følgjer dei ofte ei "one-size-fits-all"-tilnærming, som neglisjerer individuelle behov og aktivitetsmønster. For å kunne gi skreddarsydde råd om fysisk aktivitet, er det avgjerande å identifisere og utforske grupper som er kjenneteikna av liknande aktivitetsmønster.

Resultata indikerer at ei løysing med 4 grupper kan være optimal for å identifisere meiningsfulle fysiske aktivitetsfenotyper. Datastyrte globale likskapsmål har avgrensa innverknad på grupperinga når lokale likskapsmål alt tek omsyn til attributtdistribusjonen i datasettet. Avslutningsvis bidreg denne studien med ein generalisert metode for å identifisere fysiske aktivitetsfenotyper og gir ein mal for framtidige studier. Kombinasjonen av CBR og gruppering gir ei spanande tilnærming for å handtere kompleksitetane knytt til fysisk inaktivitet og å forbetre skreddarsydde råd for å fremje ein sunnare livsstil for individet.

Preface

This thesis is submitted as a prerequisite for obtaining the master's degree in Computer Science from the Department of Computer Science at the Norwegian University of Science and Technology (NTNU). The research conducted throughout this work was carried out under the supervision of Professor Kerstin Bach with the additional co-supervision of Aleksej Logacjov.

I would like to extend my deepest gratitude to my supervisor for her constant support and patience during this research journey. Her commitment to my academic development and her generous guidance have meant a lot to me. I also want to thank my partner, family and friends for their unwavering support, encouragement and delightful distractions.

Anne Marie Skaar Hasund
Trondheim, November 2, 2023

Contents

List of Figures

List of Tables

1 Introduction

Case-based reasoning (CBR) is a widely employed AI methodology in domains characterised by abundant historical data and domain-specific expertise. Its versatility extends its utility to various fields and to be used in combination with other techniques, such as clustering. In comparison to other AI methods, CBR is transparent, which makes it ideal for interdisciplinary work where it can be crucial to understand the reasoning behind the model's decisions. Nevertheless, the prerequisite for both domain experts and engineers in its application might render it less appealing for industrial and scientific purposes. To mitigate these challenges, there is a compelling need for systematic data-driven approaches. Verma et al. [1] endorse the adoption of such an approach in the modelling of similarity measures for identifying population clusters with the use of a CBR system, which combines the advantages of CBR and clustering.

With the arrival of objective physical behaviour measurements, new research possibilities have been unlocked in the fields of public health and computer science. The use of body-worn accelerometers in combination with machine learning methodologies, facilitates the recording and prediction of physical activities [2][3]. Human Activity Recognition (HAR) specialises in recognising activities based on sensor data. Leveraging the predictive capacities of HAR, an array of techniques can be implemented for the purpose of identifying physical activity phenotypes.

Physical inactivity stands as a predominant contributor to premature mortality on a global scale and represents a substantial challenge in the domain of public health [4]. These concerns significantly intensify healthcare expenditure and the burden on national healthcare systems. On average in Canada, an inactive individual spends 38% more days in the hospital compared to an active one [5]. The World Health Organisation has recommendations regarding physical activity

but reports that more than 25% of the world's adult population is insufficiently active. There has been no improvement in these levels since 2001 [6].

Presently, activity recommendations follow a largely uniform template for the general population, with variations restricted primarily to different age categories. This one-size-fits-all strategy falls short of accommodating the distinct requirements and activity routines of each individual. To provide more tailored and accurate guidance, while still operating on a population-wide scale, it becomes necessary to identify and explore groups characterised by similar activity patterns, namely physical activity phenotypes.

This thesis endeavours to build on the CBR system and clustering methods proposed by Verma et al. This is done by utilising new and more extensive data from the HUNT4[1] population study, with an overall goal of investigating contemporary and innovative ways of using CBR for the identification of physical activity phenotypes. In contrast to the work of Verma et al. with 9000 participants, this thesis works with the full HUNT4 dataset which includes data from 38 000 participants for up to 6 days.

The objectives of this study are defined by the following research questions:

RQ 1: What is the state-of-the-art in data representation of objectively measured health data, similarity-based clustering, and identifying phenotypes in the field of Artificial Intelligence?

RQ 2: How, and employing which data representation, can CBR and clustering be used to find phenotypes in the HUNT4 dataset?

RQ 3: What are the optimal cluster sizes for forming phenotypes with this data and method?

RQ 4: How does adjusting the global similarity measure influence the clustering?

The thesis will be structured as follows: Chapter 2 will provide the foundational background necessary for a nuanced understanding of the research done. The exploration of the first research question within Chapter 3 will encompass related works. In Chapter 4, a detailed exposition of the methodologies employed for data processing, CBR and clustering within this thesis will be presented, which covers the second research question. Chapter 5 presents the experiments and subsequent results, and thus provides answers to the last two research questions. Chapter 6 will engage in discussion and analysis of the work. The culmination of the thesis and prospects for future research will be presented in Chapter 7.

[1] `https://www.ntnu.no/hunt/hunt4`

2 Background

This chapter presents the background of methods, data, and tools used for the experiments in this thesis, and is heavily based on the contents of the preceding project thesis [7].

2.1 HUNT4 data set

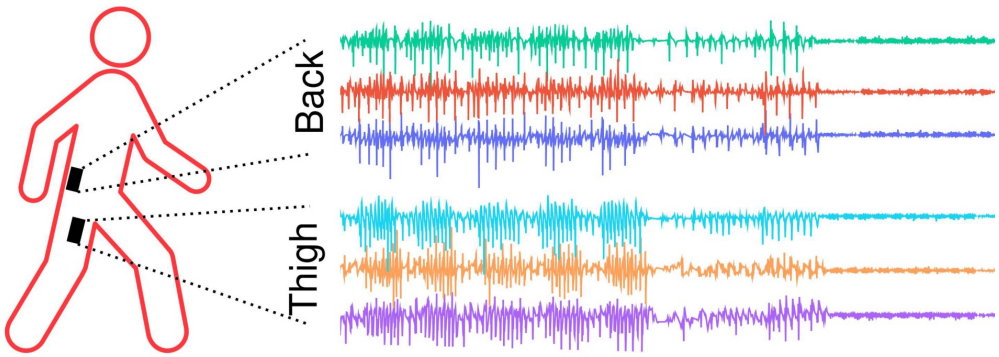
Between 2017 and 2019, an extensive population health study known as HUNT4 was conducted in Nord-Trøndelag, Norway, involving 56000 of its inhabitants. The part of it used for this master thesis consists of the objective monitoring of physical activities for 38000 participants.

To capture this data, two 3-axial accelerometers were affixed to each participant's lower back and right thigh respectively, continuously recording their activities over a seven-day period. Subsequently, machine learning models were employed for Human Activity Recognition (HAR) to classify the accelerometer data stream into six activity categories: lying, sitting, standing, walking, running and cycling [2][3], described in Table 2.1.

Table 2.1: Description of activity categories.

Activity	Description
Lying	Person is lying down horizontally
Sitting	Person is in a seated position
Standing	Person is upright on their feet
Walking	Person is moving with strides
Running	Person is moving with both feet off the ground during a stride
Cycling	Person is riding a bicycle

Figure 2.1 shows an example of the processing of the data, from raw accelerometer data via a machine learning model to 5-second windows.



MACHINE LEARNING MODEL

id	timestamp	weekday	activity
4430292	2018-04-13 13:11:15	4	sitting
4430292	2018-04-13 13:11:20	4	sitting
4430292	2018-04-13 13:11:25	4	walking
4430292	2018-04-13 13:11:30	4	standing

Figure 2.1: Process from raw accelerometer data via a machine learning model to 5-second windows of physical activities.

The process begins with the capture of raw data from the accelerometers, which are in three dimensions, in response to the orientation and movement of the accelerometer. The HAR machine learning model operates on this raw data and segments it into discrete windows, each spanning 5 seconds. It then performs the task of classifying these 5-second windows into the aforementioned activity categories, based on labeled patterns. In the final processing step, the system labels the weekday for all of the windows. Any incomplete days, not adding up to 1440 minutes, are removed from the data set.

2.2 Bouts from Physical Activity Data

Using the 5-second windows to compare patterns is highly granular and will lead to high computation time and cost as well as difficulties finding complex patterns. It is therefore need to explore ways to represent physical activity data. Bouts are a way to perceive consecutive activity types as well as summarise days, as presented by Diaz and Yacef [8].

In physical activity data, a *bout* is a continuous period of a specific length for a specific activity. By organising a day of physical activity into bouts of different lengths it is possible to not only look at each participant's amount of activity but also how this activity is accumulated throughout the day. As an illustrative example, it can differentiate between extended walking activities and more sedentary indoor movements.

2.3 Case-Based Reasoning

Case-Based Reasoning (CBR) has its roots in cognitive science, describing how humans reason when solving problems. The field is founded on the premise that similar problems also possess similar solutions. CBR sets itself apart from other artificial intelligence approaches by employing a unique data representation, as presented by Aamodt in 1994 [9].

Figure 2.2 illustrates the core structure of CBR, where each experience is represented as a *case* that encompasses both problem and solution descriptions. The problem description is comprised of a collection of attributes. Cases are systematically organized within a *case base*, and although they all share the same

set of attributes, they differ in the values assigned to those attributes and their respective solutions.

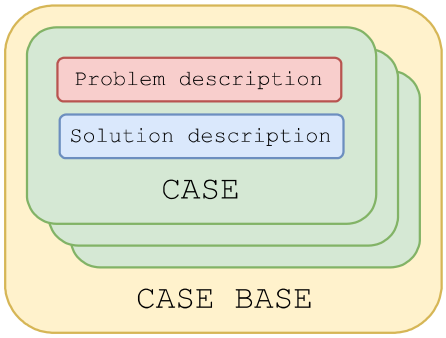


Figure 2.2: Structure of a case base containing cases with problem and solution descriptions.

CBR finds a familiar application in human reasoning, such as when a doctor relies on their memory of a past patient's medical history while diagnosing and treating a new patient who presents similar symptoms. In such a scenario, relevant attributes for problem comparison might encompass age, gender, and specific test results, while relevant solutions can be a specific medicine or an operation. The CBR process adheres to four distinct steps, often referred to as the CBR cycle or the four R's:

1. RETRIEVE: In this first phase, the system seeks out the most similar case within its case base.
2. REUSE: Knowledge and solutions from that similar case are reutilised to address the new problem at hand.
3. REVISE: The suggested solution is subject to revision and adaptation as needed to fit the nuances of the current problem.
4. RETAIN: The case, along with its updated solution, is retained within the case base. This preserves it for potential use in future situations.

When confronted with a problem, the CBR cycle initiates by creating a new case, primarily derived from the problem description, as illustrated at the top of Figure 2.3. This freshly formed case serves as a means to retrieve one or more pertinent previous cases from the case base. The selection of these prior cases is based on a similarity measure, ensuring their relevance to the current problem. The solution

contained within the retrieved case can then be suggested as a potential solution for the new case. However, this proposed solution is not blindly accepted. If a solution does not fit it can be adapted to function better for the new case. Finally, the case, now enriched with its revised solution, is stored in the case base. This practice ensures its availability for addressing future problems.

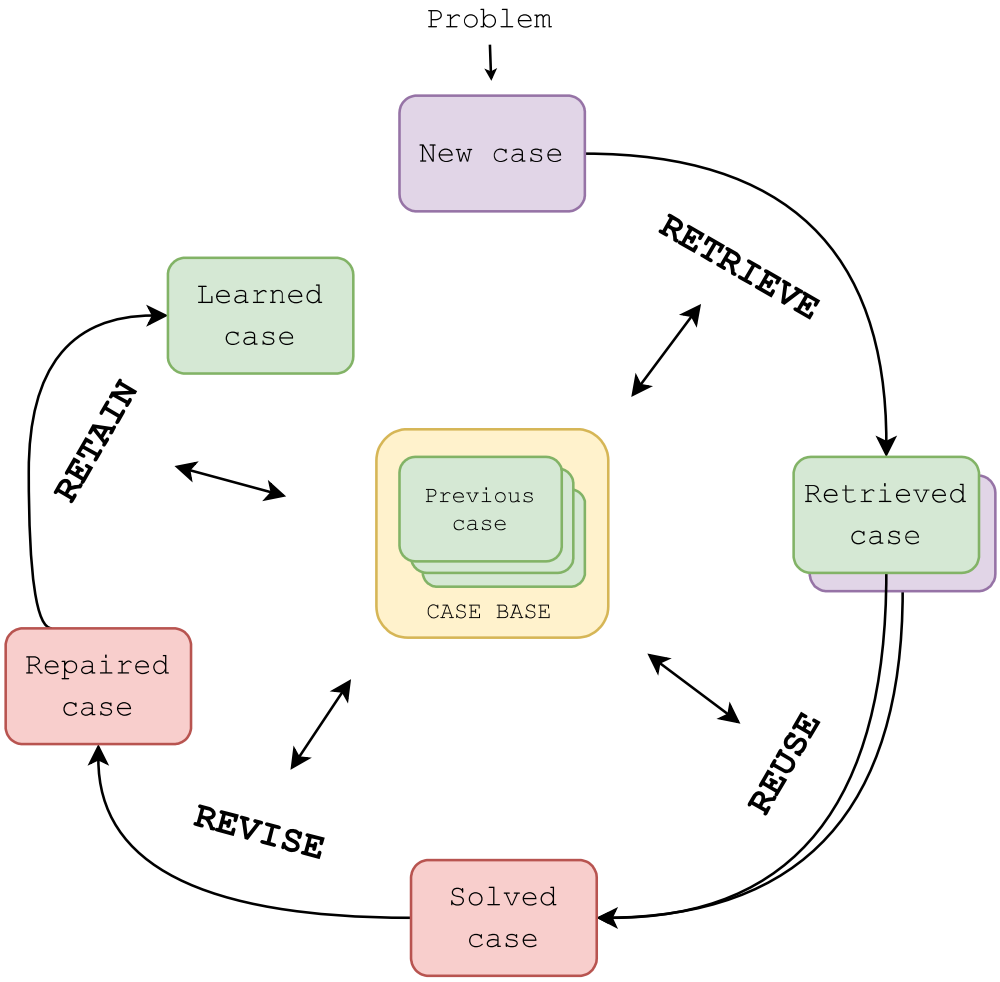


Figure 2.3: CBR cycle displaying the four R's.

2.3.1 Similarity Measures

A case is deemed similar to a new case if its solution can be used to resolve the new problem. This similarity is predicated on the utility of a case for addressing the

new problem, and it is determined by a *similarity measure*, an essential element in the retrieval phase of the CBR cycle.

The similarity measure comprises two parts that operate together, following the local-global principle. Cases are evaluated from both a local, atomic perspective and a global, conceptual standpoint. The local similarity measure scrutinises the resemblance between cases in each of their individual attributes. For instance, in the context of a doctor comparing patients, distinct age ranges might be established as similar, acknowledging that age differences bear greater significance early in life than in later stages.

In contrast, the global similarity measure operates at a conceptual level, taking into account the relative importance of each attribute when comparing cases. For instance, when making a diagnosis for a new patient, a similar test result often carries more weight than factors such as the patient's gender or age. Consequently, attributes related to test results might be assigned a higher significance in the global similarity modelling.

One way of modelling local similarity measures for numerical data in CBR, as presented by Verma et al.[10], is to look at the distribution of values for each attribute in the dataset to be analysed. The local similarity measures are polynomial functions whose degrees need to be determined. Figure 2.4 shows the method of how the range of values for an attribute in the dataset can determine the local similarity measure. It uses the interquartile range (IQR), which is defined as the difference between the 75^{th} and 25^{th} percentiles of the data, and the min-max range to determine the polynomial function.

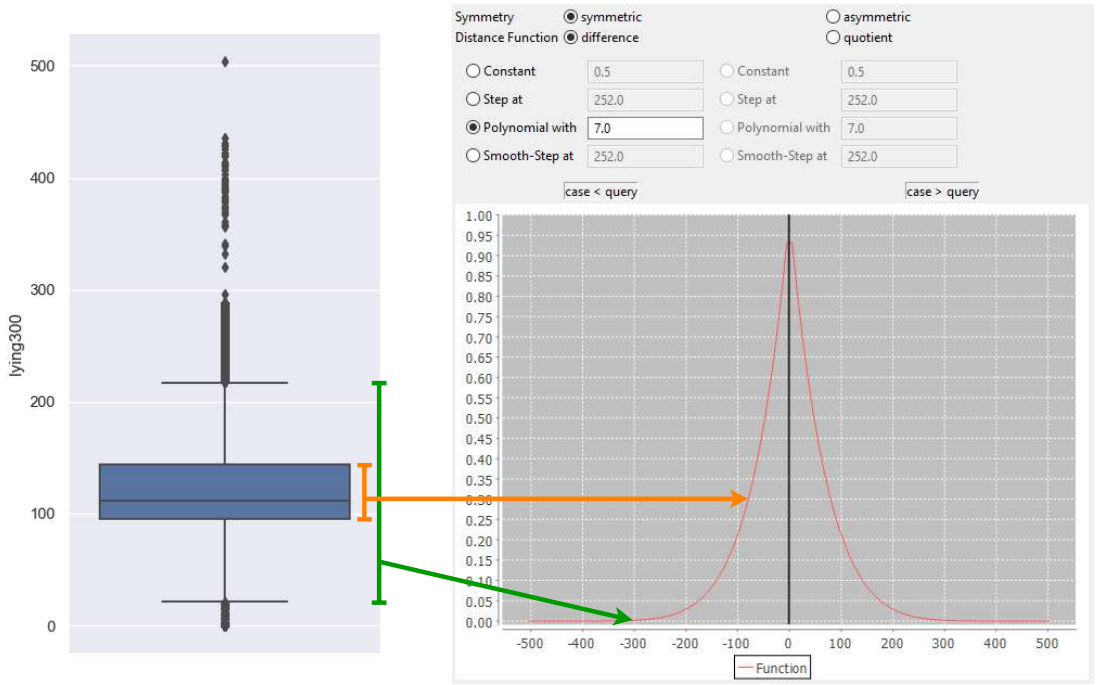


Figure 2.4: Visualisation of how the local similarity measure is based on the attribute's distribution.

The global similarity measure is determined by calculating weighted arithmetic mean, that is the weighted sum (WS) of all attributes' local similarity functions (SMF) divided by the sum of weights, as represented in Equation 2.1, where n corresponds to the number of attributes, while w denotes the weight assigned to each attribute.

$$\overline{WS} = \frac{\sum_{i=1}^{n} SMF_i \cdot w_i}{\sum_{i=1}^{n} w_i}, \tag{2.1}$$

2.3.2 myCBR

myCBR is an open-source Case-Based Reasoning (CBR) tool that's hosted by the Competence Centre for Case-Based Reasoning at the German Research Centre for

Artificial Intelligence (DFKI). It was developed in collaboration with the Centre for Model-Based Software Engineering and Explanation-Aware Computing at the University of West London (UWL).

This tool is designed with two key components. The Graphical User Interface (GUI), called the myCBR workbench, provides users with a visual interface for modelling similarity measures. It allows users to define and customise how cases are compared and evaluated for similarity. Figure 2.5 shows a screenshot of the GUI. Downloads and more information are found on the myCBR website[2].

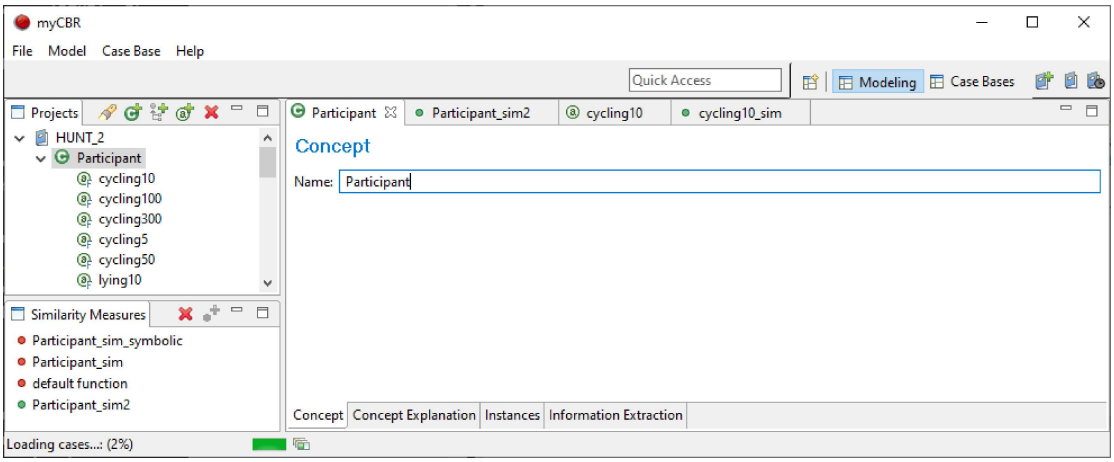


Figure 2.5: Screenshot of the myCBR workbench.

In addition to the GUI, myCBR offers a RESTful Application Programming Interface (REST API) that enables users to interact with the tool through programming. Figure 2.6 shows the Swagger UI for the REST API used via localhost and gives a visualisation of some of the functions provided by the REST API. With this REST API, users can perform tasks like updating the model and retrieving information from it. The REST API is accessed via an open-source GitHub repository[3].

[2] http://mycbr-project.org/
[3] https://github.com/ntnu-ai-lab/mycbr-rest/

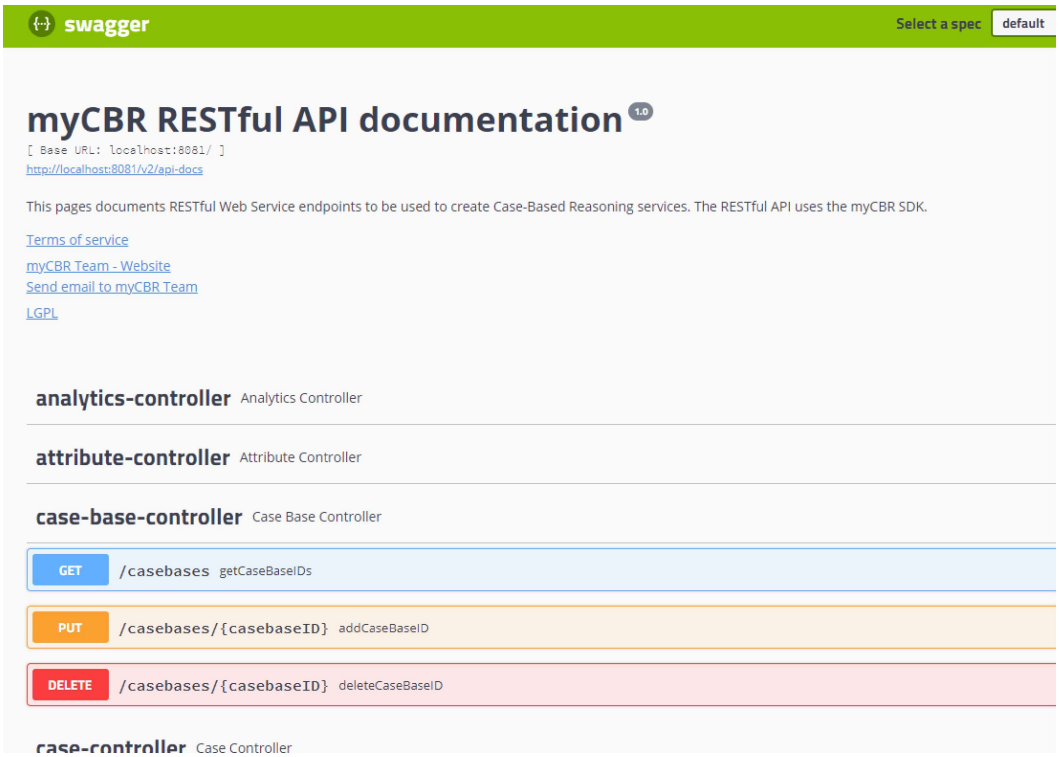


Figure 2.6: Screenshot of the Swagger UI for the REST API.

2.4 Clustering

Clustering refers to the process of organising data points into groups or clusters in a way that maximises the similarity among data points within the same cluster while minimising the similarity with those in other clusters, as discussed by Jain et. al [11]. It falls under the category of unsupervised learning, which means that the data points do not require pre-existing labels for the algorithm to begin learning. This characteristic makes clustering particularly advantageous when dealing with extensive datasets where manual labelling would be impractical or unfeasible.

One commonly employed method for clustering is known as k-means clustering [12], where the parameter k represents the number of clusters and is predetermined. In this approach, k centroids are chosen initially. Then the process consists of two loops. In the first one, new data points are compared to the centroids and assigned to the closest one. In the second loop, the centroids are

updated to become the mean of the data points within their respective clusters. Figure 2.7 illustrates an instance of k-means clustering with three clusters.

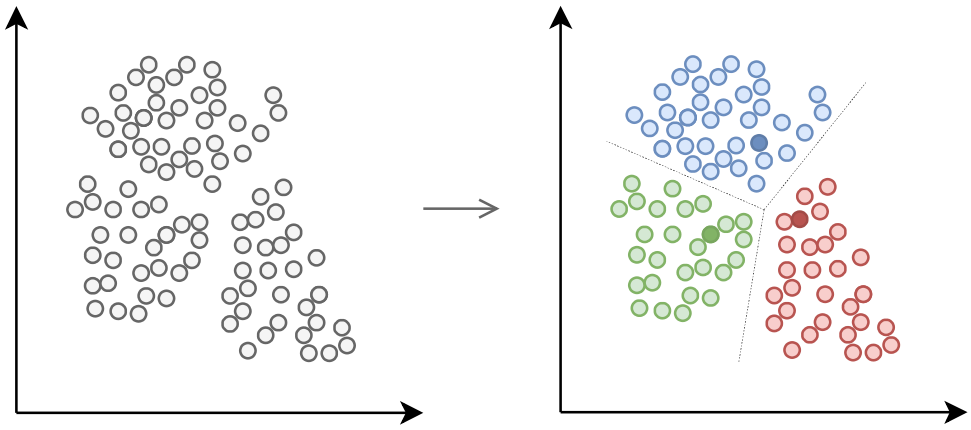


Figure 2.7: K-means clustering. Unclustered data on the left and clustered with k=3 on the left. Centroids are the darker points.

The comparison between data points and centroids typically involves measuring the Euclidean distance, which is the length of the shortest line between the two points. However, other similarity measures can also be used. It's important to note that the final centroids can vary depending on the initial centroids chosen. Consequently, different initialisation techniques are often employed to optimise the clustering process and achieve better results.

2.4.1 Evaluating Clustering Methods

Different methods can be used to evaluate the quality of the clustering results. It is however somewhat more difficult without having a ground truth to compare to. Two intrinsic methods when having no labelled data are testing cohesion and separation. Cohesion refers to the similarity between data points intra-cluster, while separation refers to the inter-cluster similarity. The goal is to have higher similarity within the clusters and lower similarity between the clusters.

The most commonly used intrinsic technique is the Silhouette Coefficient, as presented by Rousseeuw in 1987 [13]. It serves the purpose of detecting densely

populated and distinctly separated clusters, by using the average inter- (meanInterDistance) and intra-cluster distances (meanIntraDistance). Equation 2.2 shows the calculation. When using inter- (meanInterSimilarity) and intra-cluster (meanIntraSimilarity) similarity measures instead of distance, the subtraction is reversed and it changes from dividing on the maximum to dividing on the minimum.

$$\begin{aligned} SC &= \frac{meanInterDistance - meanIntraDistance}{max(meanInterDistance, meanIntraDistance)} \\ &= \frac{meanIntraSimilarity - meanInterSimilarity}{min(meanIntraSimilarity, meanInterSimilarity)} \end{aligned} \tag{2.2}$$

The highest achievable score is 1, while the lowest is -1. Scores approximating 0 suggest clusters that overlap, while negative scores indicate errors in cluster assignment. Figure 2.8 shows a visual representation of the same calculations, where each line shows the Silhouette coefficient of a data point in a coloured cluster. Shorter or negative lines indicate that the data points overlap with another cluster.

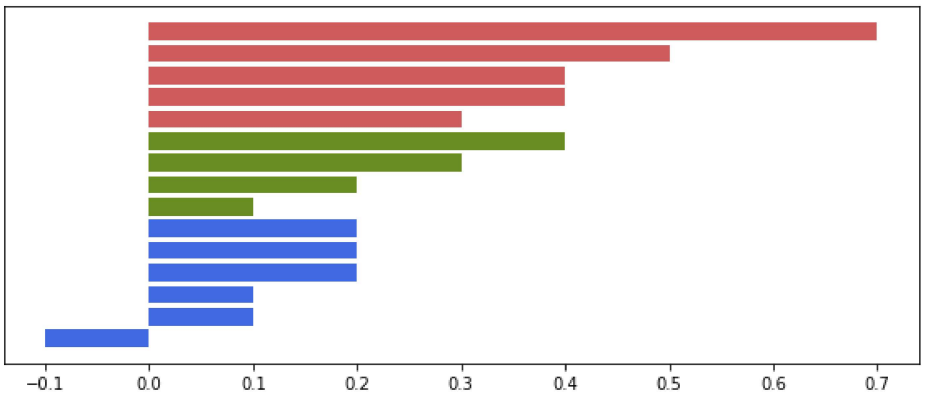


Figure 2.8: Silhouette plot where each line represents the Silhouette coefficient of a data point and each colour a different cluster.

Another method, the Dunn Index, introduced by Dunn in 1974 [14], involves computing the ratio between the smallest inter-cluster distance (minInterDistance) and the largest intra-cluster distance (maxIntraDistance). The goal is to

maximise the Dunn index. However, like the previous method, the prefix gets switched, as shown in Equation 2.3. The ratio is then between the largest inter-cluster similarity (maxInterSimilarity) and the smallest intra-cluster similarity (minIntraSimilarity), and the goal becomes to minimise the Dunn index. While the Silhouette coefficient focuses on the average data points in a cluster, the Dunn index looks at the outliers.

$$\begin{aligned} DI &= \frac{minInterDistance}{maxIntraDistance} \\ &= \frac{maxInterSimilarity}{minIntraSimilarity} \end{aligned} \tag{2.3}$$

2.5 Visualising by Dimensionality Reduction

When clustering using similarity measures on high-dimensional data it is difficult to visualise the clusters. Visualising is helpful to both understand and optimise the cluster compositions. Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbour Embedding (t-SNE) are two techniques employed for visualising high-dimensional datasets by dimensionality reduction in the context of clustering. Figure 2.9 illustrates a simple example of dimensionality reduction, transforming three-dimensional data to two dimensions and further to one dimension.

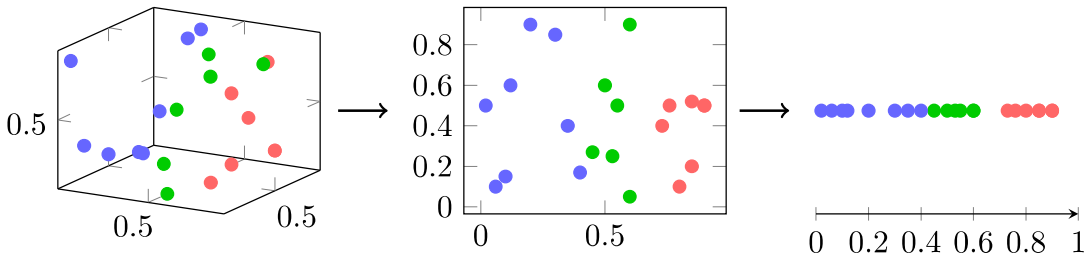


Figure 2.9: Dimensionality reduction from three dimensions on the left, via two dimensions in the middle, to one dimension on the right. The three coloured clusters stay the same.

PCA, a linear dimensionality reduction method, seeks to capture the most important variations within a dataset by identifying its principal components. By projecting the data into a lower-dimensional space while preserving as much variance as possible, PCA simplifies the representation of complex datasets. This reduction in dimensions facilitates visual exploration and interpretation of data, particularly in cases where the original dataset is high-dimensional.

On the other hand, t-SNE is a nonlinear dimensionality reduction technique specialised in visualising high-dimensional data while preserving local similarities. Unlike PCA, t-SNE focuses on maintaining the relative relationships between data points, emphasising the clustering of similar points and the separation of dissimilar ones. In clustering tasks, t-SNE's visualisations provide a means to uncover intricate cluster structures within the data. It is particularly useful when dealing with datasets where clusters exhibit complex and nonlinear patterns.

3 Related Work

This thesis is largely inspired by the work done by Verma et al. [1]. They present an approach that combines Case-Based Reasoning (CBR) and k-means clustering to identify phenotypes within a population. This chapter explores similar research problems in the literature and examines various solutions proposed by other researchers in the field. Hence it provides an overview of the state-of-the-art organised in relevant topics, as requested by RQ1. Some of the related work is reapplied from the preceding project thesis [7].

3.1 Retrieval Strategies

The retrieval is crucial in CBR systems since it finds the most similar, relevant cases. Retrieval strategies focus on correct and fast retrieval, and address three CBR components: case representation, case base size and similarity measures. In Verma's work, case retrieval hinges on a similarity score determined by both local and global similarities, mirroring the principles employed by the k-means algorithm in finding the nearest neighbouring centroid for each data point.

In contrast, Wess et al. [15] and Bergmann et al. [16] advocate for the use of k-d trees for retrieval, a strategy aimed at enhancing efficiency. This retrieval technique can be seen as a form of clustering, as it works like a binary search tree and groups similar cases together for faster retrieval. Another paper by Bergmann and Stromer [17] presents the use of MAC/FAC as a retrieval strategy for semantic workflow cases, and reports significantly reduced retrieval time without notable impact on the quality. The acronym MAC/FAC stands for "many are called, few are chosen", representing a two-stage process where a cheap initial

filter is employed to generate a subset, which is then subjected to more precise selection.

3.2 Case Representation and Similarities

Establishing an effective CBR system entails selecting appropriate attributes for case representation, which can be derived from machine learning or domain expertise. Additionally, defining the similarities within and between these attributes is crucial. Another paper by Verma et al. [10] exemplifies this by transforming accelerometer data into six distinct activity classes and using the average daily minutes spent in each activity as attributes, with domain experts contributing to the definition of these classes. They propose a data-driven approach for developing similarity measures based on target attribute ranges.

Verma et al. [1] demonstrate an intriguing approach by utilising feature importance derived from machine learning models used on patient-recorded outcomes to select features for predicting outcome measures in the healthcare domain. This not only reduces case base complexity but also maintains overall performance. Veites and Bach [18] support these findings by showcasing the effectiveness of expert-selected features and data-driven local similarity measures in their SupportPrim application. Their work involves attributes comprising both symbolic and numerical values, offering a hybrid approach to attribute selection and similarity definition.

Mathisen et al. [19] introduce a framework for learning similarity measures from data, subsequently creating novel similarity measure designs. Instead of applying the local-global principle, they employ a Siamese neural network to compute similarity between cases. Their results underscore the efficacy of using machine learning classifiers as a foundation for similarity measures, with data-driven approaches yielding superior performance. Stahl [20] investigates diverse methods for adjusting similarity measures, utilising user feedback to refine these measures as the case base expands. Additionally, Abdel-Aziz et al. [21] introduce preference-based learning, a novel approach involving pair-wise comparisons to learn similarities based on feedback from users on which of the two in a pair they prefer. This more closely mirrors the origin of CBR, where the learning strategy is inspired by human experience.

3.3 Clustering with CBR

CBR and clustering are used together for various purposes, with the choice of clustering algorithm significantly influencing outcomes and performance. The selection process is closely linked to feature selection within the case representation. Zhu et al. [22] propose a hybrid approach that combines feature selection and clustering to optimise feature utilisation using a clustering method to divide the case base into subsets with a hierarchical structure, showcasing enhanced performance.

In their paper, Ahmed et al. [23] look into using the combination of CBR and cluster analysis for health monitoring of elderly participants with a 90% match between expert and CBR clustering. They propose a time-series clustering strategy to pre-group cases, which accelerates the retrieval process by narrowing the search to specific clusters. Müller and Bergmann [24] also suggest using a cluster-based approach to improve the retrieval phase in Process-Oriented CBR with extensive similarity measures. These approaches serve as alternatives to the MAC/FAC retrieval technique.

Clustering is also employed to improve data presentation and user understanding. Yang and Wu [25] adopt a density-based clustering algorithm, DBSCAN, to merge similar cases, aiding users in interactive CBR tasks. This allows for the utilisation of a large case base while also making the output to the user comprehensible. Fullen et al. [26] use DBSCAN to detect and prevent alarm floods, a situation where an overwhelming number of alarms can exceed a user's cognitive capacity, potentially causing them to overlook critical alarms.

3.4 Use of Visualisations

Generally, AI researchers have a challenge in providing explanations for their solutions. Visualisation techniques can be helpful for enhancing the clarity and comprehensibility of data. It can also be a tool for feature extraction in CBR, either to use as attributes in the system or to improve similarity measures.

Massie et al. [27] propose the use of a visual output tool FormuCaseViz for their CBR problem solution to explain the solution to its user. The tool was by domain experts deemed more helpful than the previous textual output. Three novel visu-

alisation approaches are presented by Schultheis et al. [28] for helping knowledge engineers better understand and model similarities in Process-Oriented CBR. In their research, Martin et al. [29] present an explanation framework for supporting network engineering experts in explaining solutions to non-expert staff in a telecommunication organisation. Cantu et al. [30] argue the necessity of a visualisation tool for correlation detection that can work on both categorical and numerical data and came up with the solution of a Parallel Assemblies Plot (PAP).

Leal et al. [31] use PCA to extract features for use as attributes in a CBR system, with the purpose of detecting correct and incorrect measurements in glucose monitoring systems in an intensive care unit. In their paper, Ruiz et al. [32] suggest the use of Multiway PCA to reduce the high-dimensionality of process monitoring data from a wastewater treatment plant, by summarising the information into fewer variables to be used in a CBR system.

3.5 Identifying Phenotypes

Researchers utilise various techniques to identify phenotypes from accelerometer data in population studies. Marschoellek [33] suggests using the ATLAS index and x-means clustering to identify groups based on activity attributes such as regularity, duration, and intensity. In their research, Diaz and Yacef [8] use the extraction of bouts of physical activity as features for clustering and thus detecting behavioural changes in school children's activity levels. The use of latent class analysis (LCA) is presented by Gupta et al. [34] to identify four activity profiles. Howie et al. [35] propose sex-specific LCA to cluster individuals into five activity phenotypes based on activity intensity. Meisingset et al. [36] employ LCA to categorise musculoskeletal patients into five distinct phenotype groups. In another approach, Willetts et al. [37] advocate for statistical machine learning techniques to identify physical activity phenotypes and sleep behaviour.

CBR's utility extends to optimising activity plans for enhanced physical performance based on time series data. Bergman [38] explores the prediction of optimal finish times for speed skaters based on their past races and external conditions. Smyth and Cunningham [39] aim to recommend race plans to marathon runners, helping them achieve predicted personal best race times by drawing on the race history of similar runners. These applications illustrate the versatility of CBR across diverse problem domains, from phenotype identification to performance optimisation.

4 Method

This chapter provides a thorough account of the procedures, tools, and techniques employed throughout the research process. It addresses the second research question regarding how CBR and clustering can be used to find phenotypes in the HUNT4 dataset. The method is outlined in Figure 4.1, showing the different steps of the process and how they are connected.

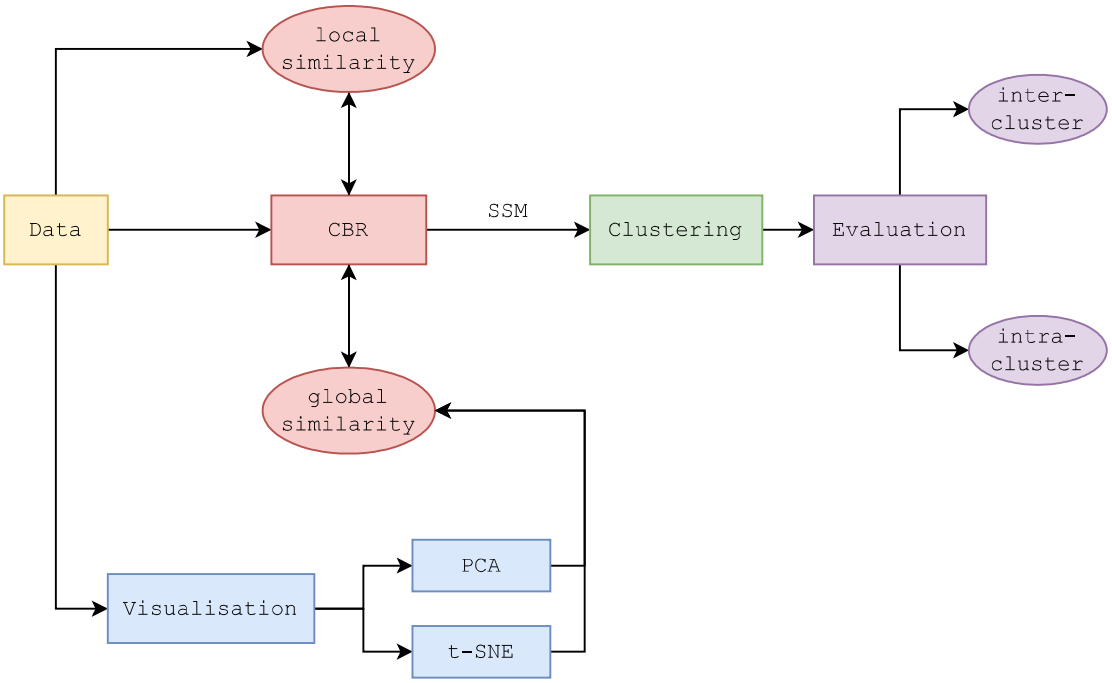


Figure 4.1: Process architecture showing the different steps of the methodology.

Initially, the data is used for two things. Firstly it is processed and used for the case representation and population of the CBR system. Simultaneously, the data is used to model the local similarity measures by utilising the attribute distribution. Visualisation of the data set is used to change the global similarity measure, which starts a series of iterative experiments within the CBR system. From the CBR system, a self-similarity matrix (SSM) is used as a means for clustering based on the similarity between the cases given by the similarity measures. The clusters are later evaluated based on their inter- and intra-cluster similarities. This iterative process starts over by selecting an alternative global similarity measure.

Throughout this methodology chapter, the real-world HUNT4 dataset will be used to explain the different steps in the method. Most of the methodology is however applicable to different domains and datasets.

4.1 Data Pre-processing

The process of making the data ready for use starts with the raw 5-second windows of physical activity data derived from the HUNT4 accelerometer data. Every participant has at least 6 full days of activity and each window is labelled with one of the six activities: lying, sitting, standing, walking, running and cycling. The weekdays are then labelled for every window, before structuring the bouts.

Each participant's daily data is used to compute activity bouts with durations of 300, 100, 50, 10, and 5 seconds, for each day and for every activity. This process initiates by filtering out the longest bouts initially, specifically, all uninterrupted sequences of the same activity lasting 300 seconds or more. Subsequently, the procedure continues to the next most prolonged bouts and goes on in a similar fashion. In the end, all the 5-second windows should be accounted for. Figure 4.2 gives a brief example of how bouts are structured from the 5-second windows.

id	timestamp	weekday	activity
4661641	2018-05-02 18:12:29	2	sitting
4661641	2018-05-02 18:12:34	2	standing
4661641	2018-05-02 18:12:39	2	standing
4661641	2018-05-02 18:12:44	2	walking
4661641	2018-05-02 18:12:49	2	walking
4661641	2018-05-02 18:12:54	2	walking
...	...	...	...
4661641	2018-05-03 06:42:24	3	standing
4661641	2018-05-03 06:42:29	3	walking
4661641	2018-05-03 06:42:34	3	walking
4661641	2018-05-03 06:42:39	3	standing

id	weekday	sitting5	standing10	standing5	walking10	walking5
4661641	2	5	10	0	10	5
4661641	3	0	0	10	10	0

Figure 4.2: Example of how bouts are calculated from the 5-second windows.

As a response to challenges encountered when processing the extensive dataset with diverse weekdays, a simplification of the weekday labelling occurred, reducing it to a binary categorisation of *weekend* (true/false). Subsequently, the activity bouts within these two categories were aggregated from the corresponding weekdays. This results in a data set with 65 340 entries. Table 4.1 gives an overview of the structure of the data to go into the case base.

Table 4.1: Example of data to go into the case base. For visual purposes, the dots (...) represent the rest of the attributes, like *lying10*, *lying5*, *sitting300*, etc.

id	weekend	lying300	lying100	lying50	...	cycling5
5143580_1	1	98	4	2	...	134
4802235_1	1	204	8	3	...	4
4636595_0	0	118	12	9	...	32
4445692_1	1	152	20	14	...	177

4.2 CBR System

The use of the CBR system can be divided into four phases: the creation of a myCBR project with a populated case base, the modelling of the local and global similarity measures, and the pre-processing of the linear retrieval among each case in the form of a self-similarity matrix. To facilitate this work the myCBR REST API is used to ease the programming and retrieval from the myCBR project.

4.2.1 Case Representation and Case Base Population

The project is initialised by declaring all the attributes of the chosen concept. For each activity category, five bouts are declared as *float* attributes: 5, 10, 50, 100 and 300. In addition, the weekend attribute is set at an *integer* due to issues with *boolean* attributes. Then the case base is populated via the REST API from the data, with the case ID and the attributes. As shown in Figure 4.3, the attribute ranges must be entered manually in myCBR for the set-up to be complete.

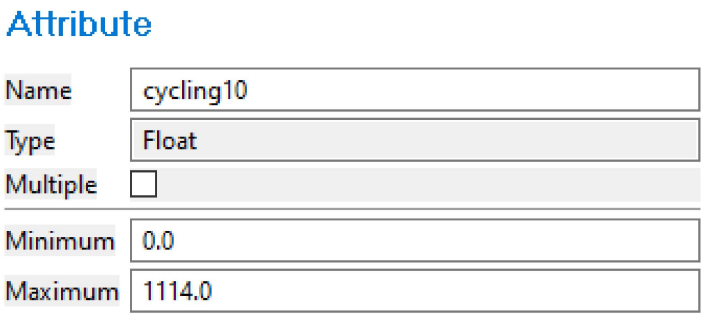


Figure 4.3: Attribute declaration for *cycling10* in myCBR.

Figure 4.4 shows the case representation in myCBR for one case in the case base, with the case ID as its name.

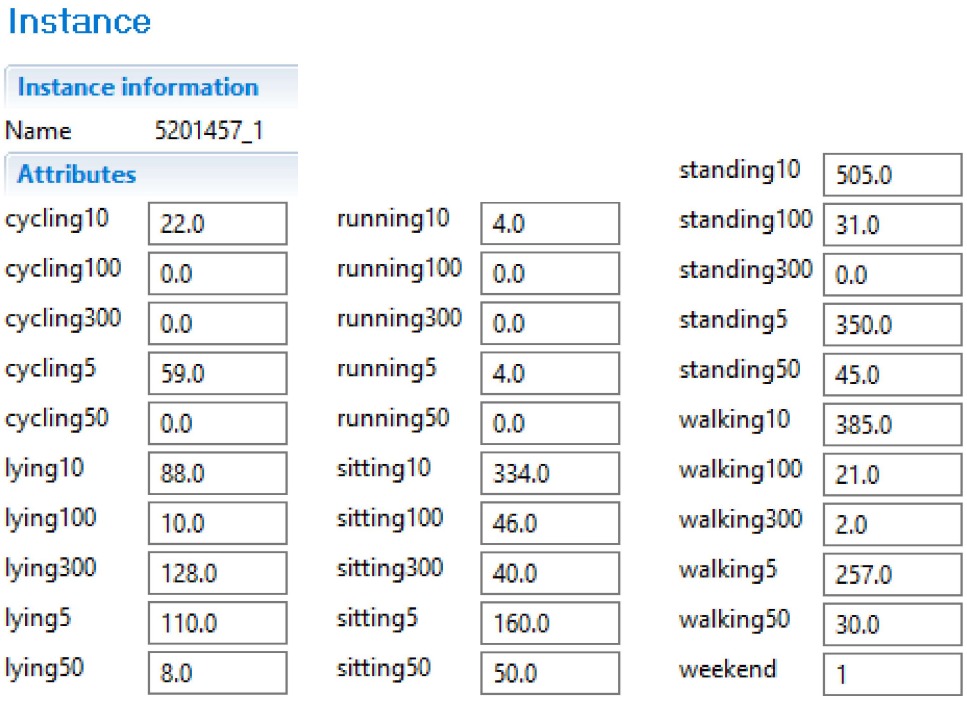


Figure 4.4: Representation of a *case* in myCBR. The name is the ID of the case.

4.2.2 Local Similarity Measures

The second phase is the calculation and set-up of the local similarity measures for all of the attributes. First, the ranges of values for all the numerical attributes are established, including the maximum and minimum as well as the interquartile range (IQR). Verma et al. [10] suggest $y = 0.3$ as the targeted similarity for the IQR on the polynomial function and $y = 0$ for the max-min range. Hence, the polynomial function is obtained when the approximations below are true.

$$y(max - min) \approx 0$$
$$y(IQR) \approx 0.3$$

Figure 4.5 shows multiple possible polynomial functions with polynomial values from 0 to 10. For each of the functions, the red dot shows the target of $y(IQR) = 0.3$. The polynomial function for each numerical attribute will be the one with the red dot closest to the attribute's IQR.

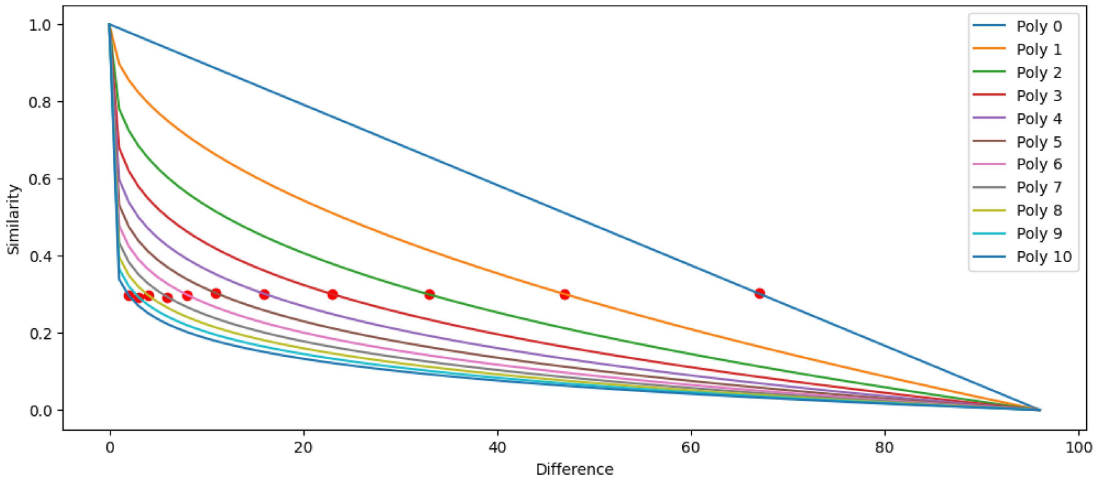


Figure 4.5: Multiple polynomial similarity functions. The red dots mark the target value where $y(IQR) = 0.3$.

A Python function, as shown in Listing 4.1, is made to facilitate the adjusting of the polynomial to fit the targeted values for each of the attributes' distributions. The inputs are the attribute's IQR, a target similarity and the difference between the maximum and minimum value for the attribute.

Listing 4.1: Python function to get polynomial values for local similarity measures

```
def find_polynomial(iqr, target_sim, minmax_diff):
# Iterate through polynomial values until target sim is reached
for poly in np.arange(0.01, 100, 0.01):
    degree = 1/poly
    sim = 1 - (iqr / minmax_diff) ** degree
    if sim <= target_sim:
        return poly
```

The function is run for each of the numerical attributes in the case base. Table 4.2 shows the polynomial values for all the numerical attributes.

Table 4.2: Polynomial value for local similarity measures for all attributes, listed with minimum and maximum values and interquartile range (IQR).

Attribute	Polynomial	Min	Max	IQR
lying300	7.0	0	504	49.0
lying100	10.0	0	299	9.0
lying50	11.0	0	324	7.0
lying10	13.0	0	3948	52.0
lying5	12.0	0	2825	44.0
sitting300	6.0	0	365	47.0
sitting100	7.0	0	305	33.0
sitting50	7.0	0	360	32.0
sitting10	8.0	0	3934	246.0
sitting5	9.0	0	2726	118.0
standing300	11.0	0	286	6.0
standing100	7.0	0	266	27.0
standing50	7.0	0	404	39.0
standing10	7.0	0	3863	356.0
standing5	8.0	0	2788	211.0
walking300	12.0	0	66	1.0
walking100	8.0	0	103	6.0
walking50	8.0	0	149	11.0
walking10	6.0	0	1445	220.0
walking5	5.0	0	984	178.0
running300	NaN	0	28	0.0
running100	NaN	0	32	0.0
running50	NaN	0	33	0.0
running10	NaN	0	325	0.0
running5	15.0	0	302	2.0
cycling300	NaN	0	41	0.0
cycling100	NaN	0	71	0.0
cycling50	13.0	0	96	1.0
cycling10	10.0	0	1114	33.0
cycling5	8.0	0	1004	59.0

Local similarity measures are then set up in myCBR as polynomial functions. The polynomial values from the aforementioned code and table are entered into the polynomial functions as shown in Figure 4.6. For the *NaN* values, 1 is used as a polynomial value in myCBR. A constant 1 is given as the local similarity measure for the binary *weekend* attribute.

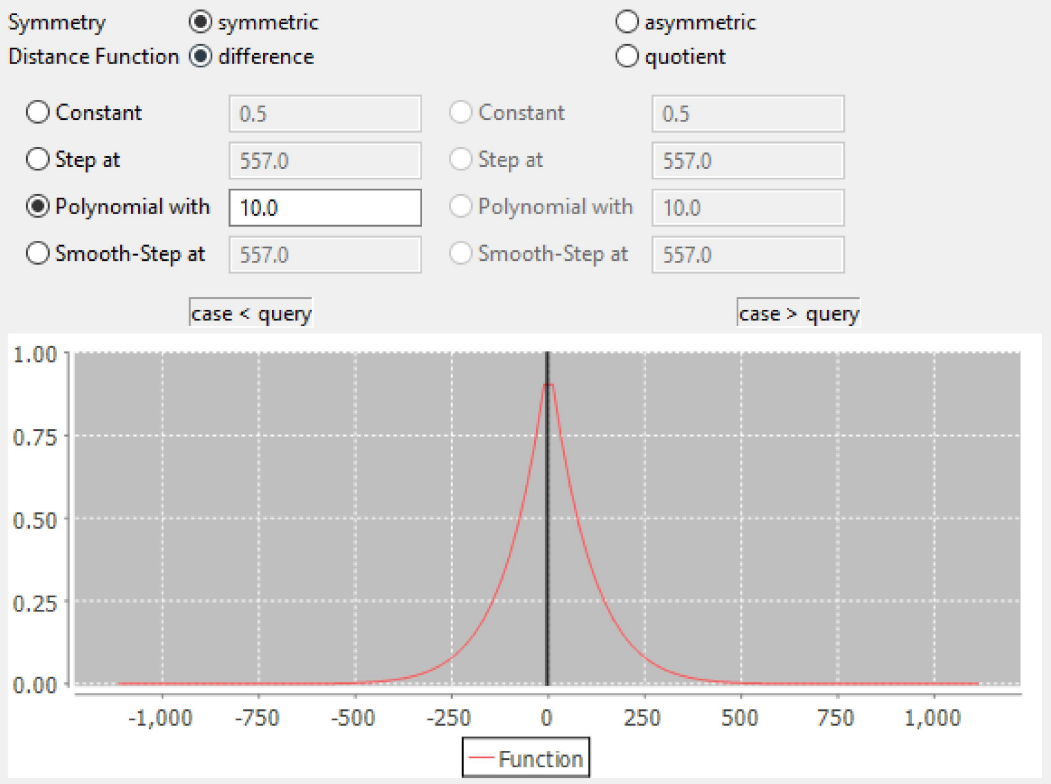
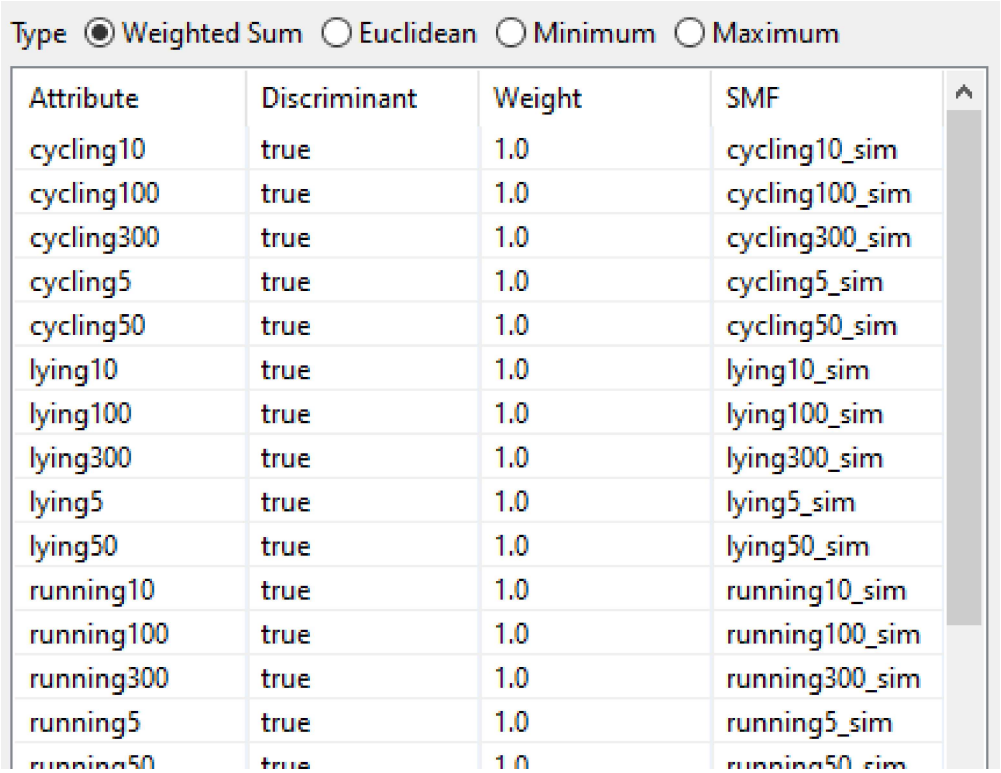


Figure 4.6: Setup of a polynomial function as the local similarity measure for attribute *cycling10* in myCBR with 10 as the polynomial value.

4.2.3 Global Similarity Measures

As shown in Figure 4.7, the global similarity measure is configured by choosing weights for the different local similarity measures before adding them together, referred to as the *weighted sum*. Changes in the weights are made to experiment with cluster populations and are motivated by patterns shown in the visualisation with PCA and t-SNE from the high-dimensional data set as well as by domain expertise.



Attribute	Discriminant	Weight	SMF
cycling10	true	1.0	cycling10_sim
cycling100	true	1.0	cycling100_sim
cycling300	true	1.0	cycling300_sim
cycling5	true	1.0	cycling5_sim
cycling50	true	1.0	cycling50_sim
lying10	true	1.0	lying10_sim
lying100	true	1.0	lying100_sim
lying300	true	1.0	lying300_sim
lying5	true	1.0	lying5_sim
lying50	true	1.0	lying50_sim
running10	true	1.0	running10_sim
running100	true	1.0	running100_sim
running300	true	1.0	running300_sim
running5	true	1.0	running5_sim
running50	true	1.0	running50_sim

Figure 4.7: Configuration of weights for the global similarity measure in myCBR.

4.2.4 Self-Similarity Matrix

To facilitate the comparison of cases during the clustering process, it is essential to establish similarities among them. This is achieved by acquiring a self-similarity matrix (SSM) through the REST API. Due to memory constraints in the REST API, given the substantial size of the SSM (approximately 80 GB), the retrieval process was divided into multiple batches. A shell script was developed to address this, as shown in the pseudocode in Listing 4.2.

Listing 4.2: Pseudocode of shell script

```
for i in seq (first, last, step):
    start REST API
    PID = get last process ID
    sleep 30
    ssm = retrieve ssm(i) via REST API
    save ssm to file
    kill PID
    sleep 5
```

The retrieval of *ssm(i)* gives a matrix consisting of *step* number of columns from case i to case $i + step$. The rows are all the cases in the case base. An example of an SSM with four cases can be found in Table 4.3.

Table 4.3: Example of a self-similarity matrix with 4 cases

	4809449_0	5086788_1	4467236_1	4534570_1
4809449_0	1.000	0.735	0.652	0.798
5086788_1	0.735	1.000	0.710	0.798
4467236_1	0.652	0.710	1.000	0.620
4534570_1	0.798	0.798	0.620	1.000

The matrix values are the corresponding similarities between the cases as given by the similarity measures. When merging all these matrices together, the final SSM can be used to look up the similarity between two arbitrary cases. A visualisation of this through an example with fifteen cases is shown as a heat map in Figure 4.8. The SSM is generated to eliminate the necessity for later similarity retrieval from the CBR system during the clustering process, to enhance computational efficiency and expedite run time.

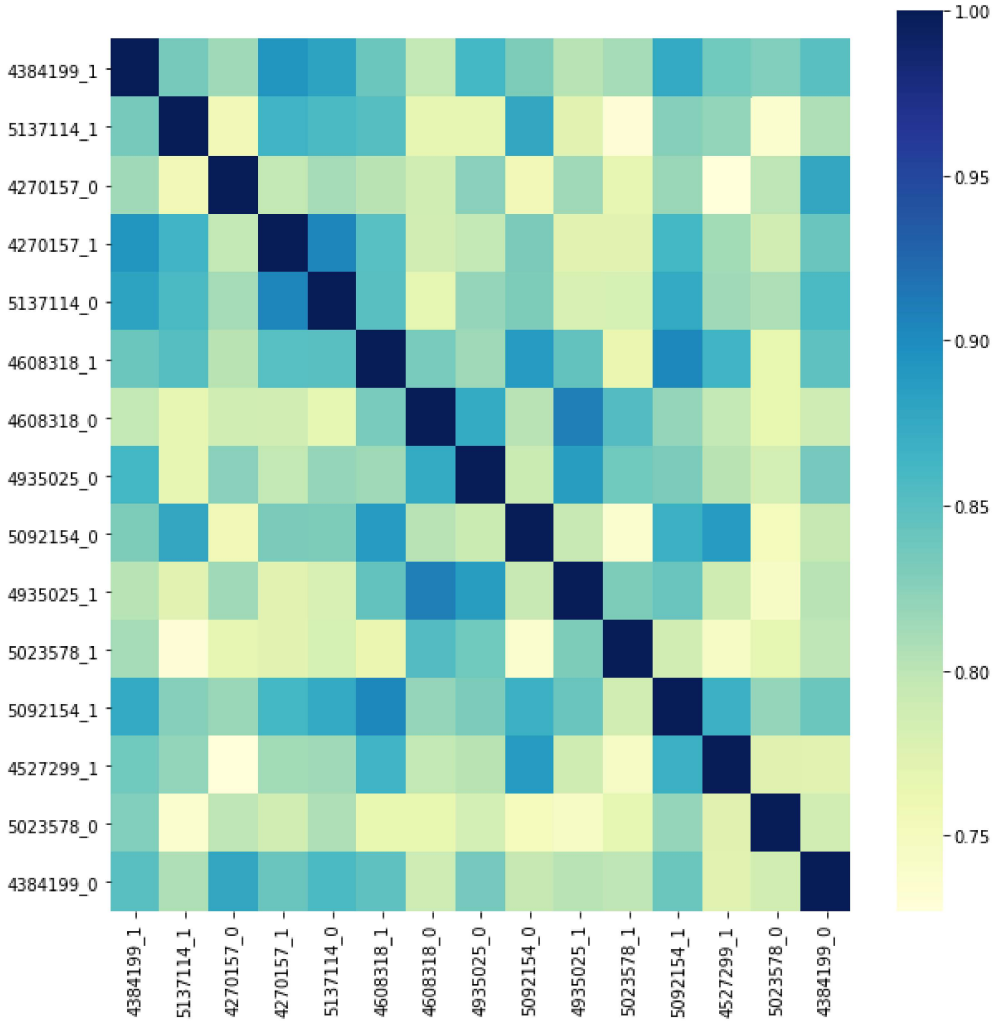


Figure 4.8: Heatmap of an example self-similarity matrix with 15 cases. Darker colours indicate a higher similarity between cases.

4.3 Visualisation

Creating diverse visualisations for various attributes can provide valuable insights into which attribute weights should be adjusted when refining the global similarity measure. This endeavour aims to optimise the global similarity measure with the ultimate objective of improving the quality of the resulting clusters.

PCA and t-SNE plots are used for visual clustering, with the goal of revealing attributes that contribute to cluster distinction, also referred to as attribute importance. Before the plotting, the bout attributes in the data set are normalised and the weekdays are removed from the set. In order to distinguish which attributes contribute to the clusters formed in the plots, the data points are coloured as above or below average for the 300-second bouts for each activity.

Examples of the PCA and the t-SNE plots are shown in Figures 4.9 and 4.10 respectively, coloured for above and below the average of the attribute *walking300*. The Python package *scikit-learn*[4] is used for both plots.

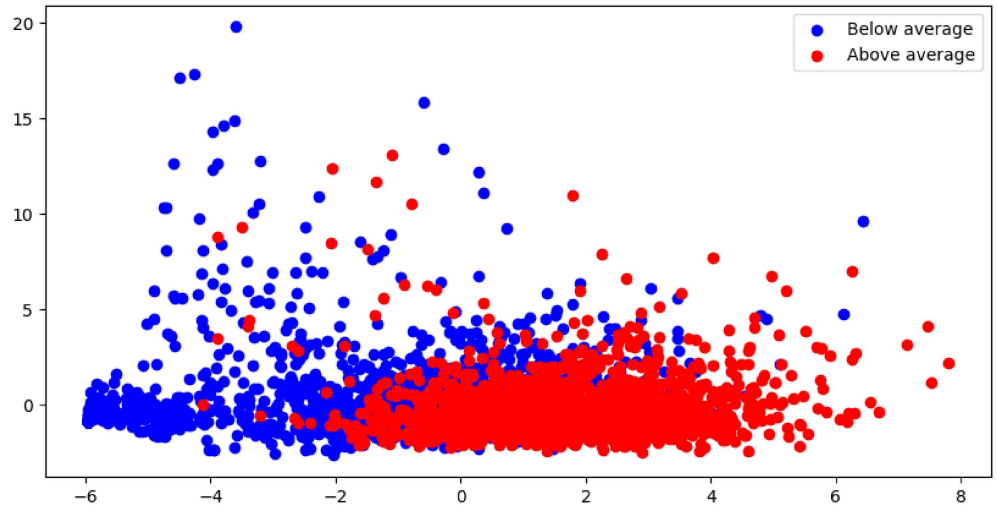


Figure 4.9: PCA plot, coloured for above/below average for an attribute.

[4]https://scikit-learn.org/

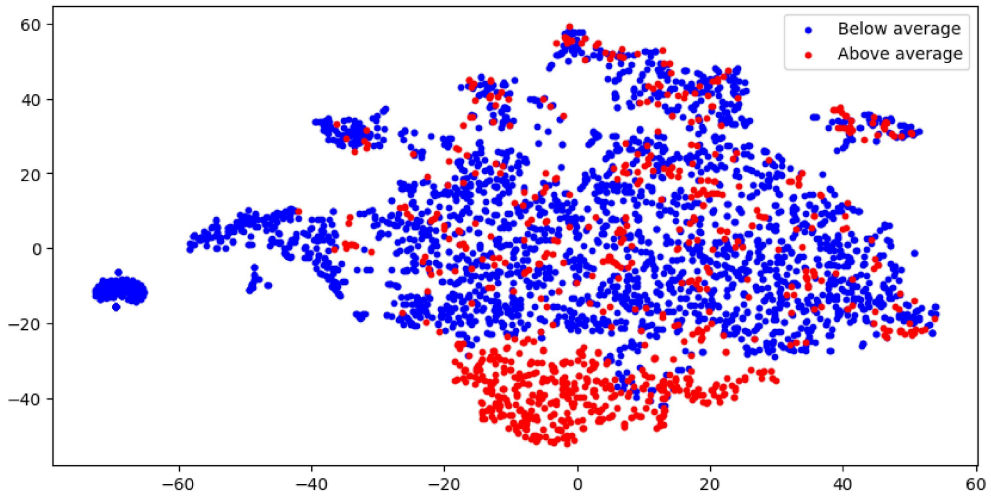


Figure 4.10: t-SNE plot, coloured for above/below average for an attribute.

4.4 Clustering Algorithm

The clustering algorithm used for this thesis is based on a pseudocode found in Verma's paper [1], which is mainly based on k-means clustering but differs in the way the data points are compared. While the Euclidean distance is used for k-means, this algorithm uses the similarity between cases as presented in the SSM. Consequently, the nearest centroid is the one with the highest similarity rather than the shortest distance. The algorithm written in Python can be found in Listing 4.3.

The procedure begins by selecting k random cases as centroids. Subsequently, each case is allocated to its most similar centroid, as determined by the self-similarity matrix. A cluster evaluation is then undertaken to identify the case possessing the highest average similarity within each cluster, which is designated as the new centroid. This iterative process persists until a predefined stopping criterion is met: when the same centroids have been selected for the preceding three iterations or after the tenth iteration. To find the optimal number of clusters, the algorithm is run for $k = 3, 4, ..., 10$. Recent research on activity clustering [34; 35; 36] typically employs 4 or 5 clusters, suggesting that testing up to 10 clusters should provide an adequate range to identify the optimal clustering solution.

Listing 4.3: Clustering Algorithm in Python

```
def clustering_algorithm(ssm, caseIDs, k):
    # random initialisation from list of caseIDs
    centroids = random.sample(caseIDs, k)
    cases = ssm.index.to_numpy()
    hist_mean_sim = [[] for x in range(k)]

    for i in range(0, 10):
        clusters = [[] for x in range(k)]

        for case in cases:
            # get similarity between case and each of the centroids
            similarities = ssm[centroids].loc[case]
            # find most similar centroid to case
            most_similar_centroid = similarities.idxmax()
            # assign case to most similar centroid
            clusters[centroids.index(most_similar_centroid)].append(case)

        for cluster in clusters:
            # find mean case in each cluster
            similarities = ssm[cluster].loc[cluster]
            # find highest average similarity within cluster
            mean_sim = similarities.mean(axis=1)
            mean_case = mean_sim.idxmax()
            # set mean case as new centroid for cluster
            centroids[clusters.index(cluster)] = mean_case
            # save mean case for comparison later
            hist_mean_sim[clusters.index(cluster)].append(mean_sim.max())

        # stop if the last three clusters are the same
        if i > 3 and [np.round(x[-1],5) for x in hist_mean_sim]
          == [np.round(x[-3],5) for x in hist_mean_sim]:
            break

    return centroids, clusters, [np.round(x[-1],5) for x in hist_mean_sim]
```

4.5 Cluster Evaluation Methodology

To assess the quality of the performed clustering, both metric and visual evaluation methods are conducted. The list `hist_mean_sim` in the clustering algorithm saves the average intra-cluster similarity for all the clusters, while the inter-cluster similarity is retrieved from filtering the SSM on the centroids afterwards. The average inter- and intra-cluster similarities are used to calculate the Silhouette Coefficients. In addition, the highest inter-cluster similarity and the lowest intra-cluster similarity are used to calculate the Dunn Index.

Boxplots are used to visualise the distribution of values in the clusters for particular attributes. The plots give a better perspective on the phenotypes the clusters can represent. Figure 4.11 shows an example of boxplots of an attribute for four clusters, where the red line represents the median for the cluster cases and the blue line the population median.

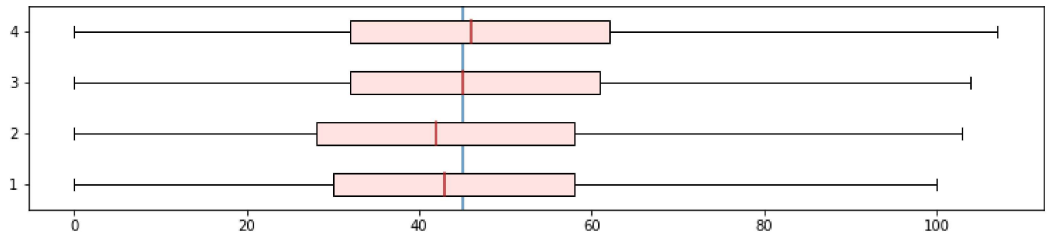


Figure 4.11: Boxplot of an attribute for four clusters. The red line marks the median for the clusters and the blue line the population median.

The use of bar charts is another way of visualising the clusters. This method shows the composition of bouts for an average day for the cases comprising each cluster and for the cluster centroids, which is helpful for distinguishing the clusters. Each day consists of only 24 hours which means that if one activity increases another must decrease.

Figure 4.12 gives an example of such a bar chart, with each of the colour nuances representing the different bout lengths.

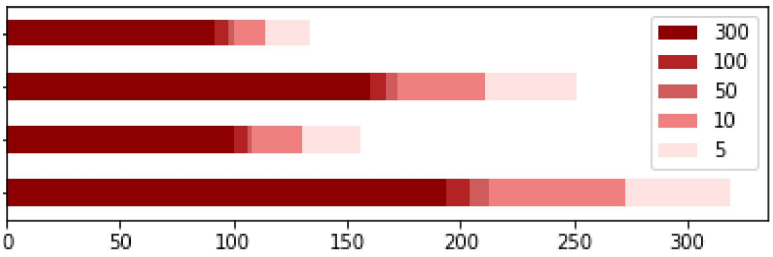


Figure 4.12: Bar chart showing the composition of bouts for 4 clusters

5 Experiments

In this chapter, the experiments and their corresponding results, conducted as part of this thesis work, are presented. The goal of the experiments is to address the third and fourth research questions by exploring different cluster sizes and adjusting the global similarity measure. Both of these changes contribute to optimising the results.

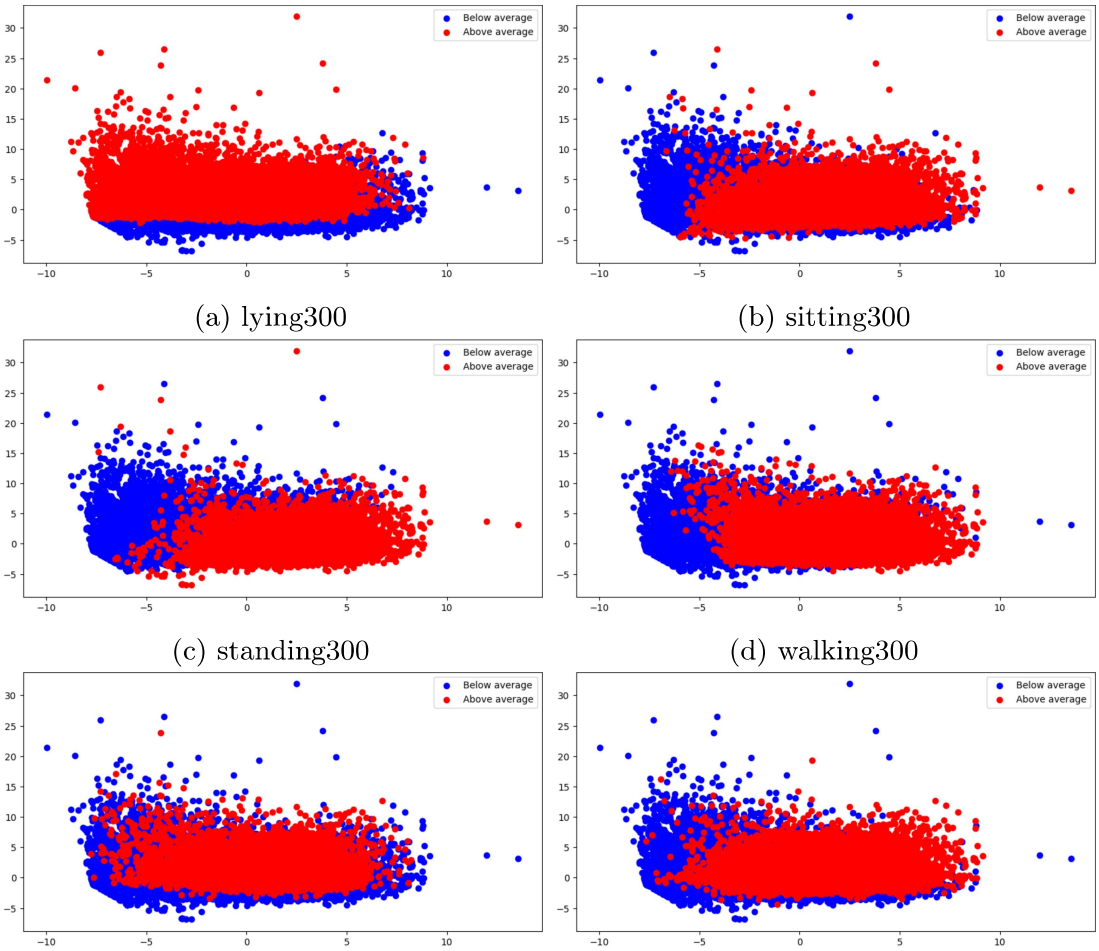
5.1 Iterations

The experiments are carried out in three iterations, each involving a modification of the global similarity measure by changing the weights. Initially, a baseline experiment is performed, wherein all the weights in the global similarity measure (GSM1) are equally set to 1, as shown in Table 5.1. Subsequently, in the second global similarity measure, these weights are adjusted through visualisation techniques. Finally, in the third iteration, the weights are determined by domain expertise.

5.1.1 Visualisation

For the second iteration, the weights in the global similarity measure are changed based on attributes found through visualising the dataset. Table 5.2 displays the similarity weights for the second global similarity measure (GSM2). These alterations are made by insights derived from PCA and t-SNE plots.

The PCA plots for the attributes *lying300, sitting300, standing300, walking300, running300* and *cycling300* are shown in Figure 5.1, marked for above and below average in the data set for each of the attributes.



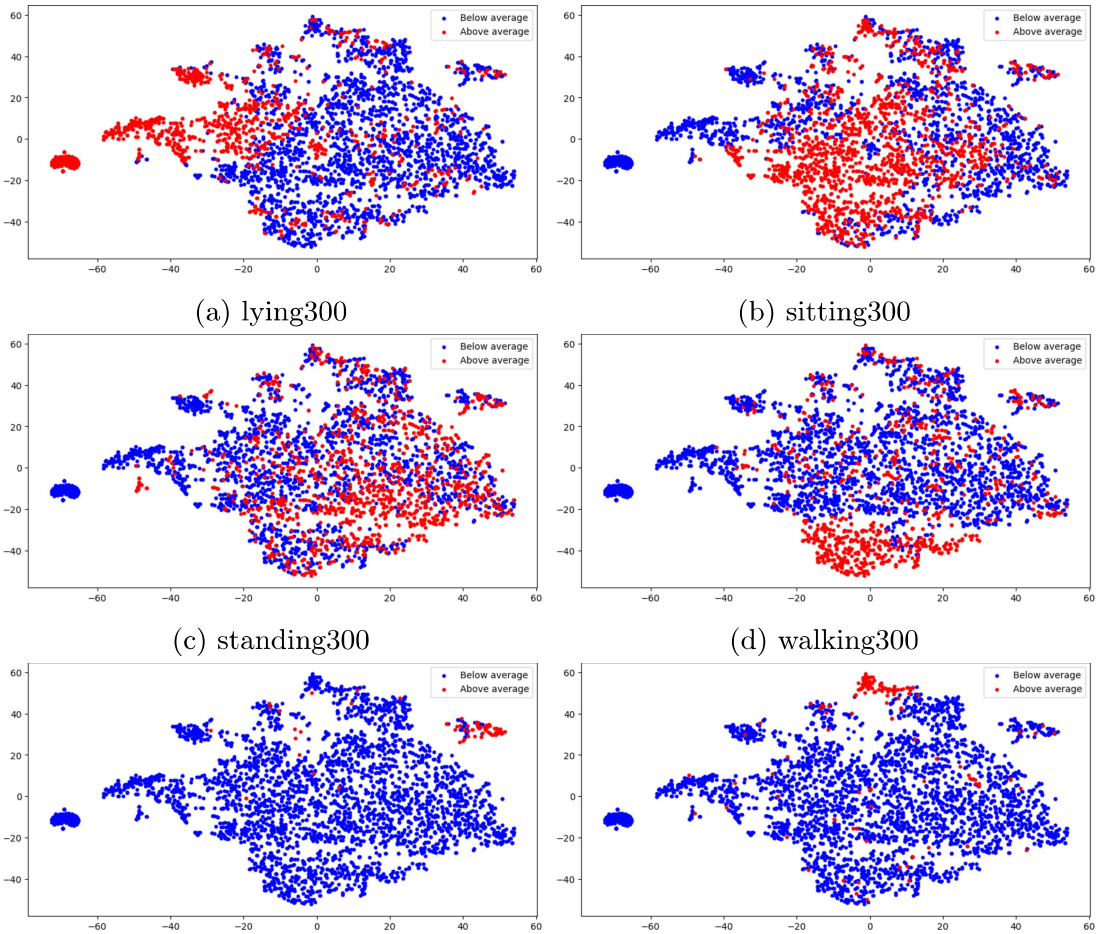
(a) lying300 (b) sitting300

(c) standing300 (d) walking300

(e) running300 (f) cycling300

Figure 5.1: PCA plots of above and below average for the 300-second bouts for the six activity categories.

The corresponding t-SNE plots are found in Figure 5.2, also here marked for above and below average for all the 300-second attributes.



(a) lying300 (b) sitting300

(c) standing300 (d) walking300

(e) running300 (f) cycling300

Figure 5.2: t-SNE plots of above and below average for the 300-second bouts for the six activity categories.

From the PCA plot, it looks like the attributes *lying300*, *running300* and *cycling300* are overlapping, while the others give some separation. From the t-SNE plots, however, the three aforementioned attributes show more clear clustering. This underscores the possible significance of the 300-second bouts in accentuating the clustering of phenotypes in this dataset. Consequently, these attributes make the adjustment of weights in this iteration, where the weights for all the 300-second bouts are set to 2 while the rest remain at 1.

Table 5.1: Global similarity measure 1

Attribute	Weight
cycling10	1.0
cycling100	1.0
cycling300	1.0
cycling5	1.0
cycling50	1.0
lying10	1.0
lying100	1.0
lying300	1.0
lying5	1.0
lying50	1.0
running10	1.0
running100	1.0
running300	1.0
running5	1.0
running50	1.0
sitting10	1.0
sitting100	1.0
sitting300	1.0
sitting5	1.0
sitting50	1.0
standing10	1.0
standing100	1.0
standing300	1.0
standing5	1.0
standing50	1.0
walking10	1.0
walking100	1.0
walking300	1.0
walking5	1.0
walking50	1.0
weekend	1.0

Table 5.2: Global similarity measure 2

Attribute	Weight
cycling10	1.0
cycling100	1.0
cycling300	2.0
cycling5	1.0
cycling50	1.0
lying10	1.0
lying100	1.0
lying300	2.0
lying5	1.0
lying50	1.0
running10	1.0
running100	1.0
running300	2.0
running5	1.0
running50	1.0
sitting10	1.0
sitting100	1.0
sitting300	2.0
sitting5	1.0
sitting50	1.0
standing10	1.0
standing100	1.0
standing300	2.0
standing5	1.0
standing50	1.0
walking10	1.0
walking100	1.0
walking300	2.0
walking5	1.0
walking50	1.0
weekend	1.0

Table 5.3: Global similarity measure 3

Attribute	Weight
cycling10	1.0
cycling100	1.0
cycling300	1.0
cycling5	1.0
cycling50	1.0
lying10	1.0
lying100	1.0
lying300	1.0
lying5	1.0
lying50	1.0
running10	1.0
running100	1.0
running300	1.0
running5	1.0
running50	1.0
sitting10	1.0
sitting100	1.0
sitting300	1.0
sitting5	1.0
sitting50	1.0
standing10	1.0
standing100	1.0
standing300	1.0
standing5	1.0
standing50	1.0
walking10	1.0
walking100	1.0
walking300	1.0
walking5	2.0
walking50	1.0
weekend	1.0

5.1.2 Domain Expert Influence

The third iteration (GSM3) incorporates domain expert input. The specific weightings for this global similarity measure are detailed in Table 5.3. The rationale behind giving prominence to the *walking5* bout lies in the desire to investigate whether individuals with a higher frequency of short walking bouts can constitute a distinct phenotype cluster. An example of such a group could comprise individuals who engage in frequent short walks throughout their workdays, often transitioning between tasks.

5.2 Results

For the three global similarity measure iterations, the average intra-cluster and inter-cluster similarities for each of the cluster sizes are listed in Table 5.4 and presented visually in Figure 5.3. This shows an increase in intra-cluster similarity with the increase in cluster size for all three global similarity measures. The inter-cluster similarity does not show a consistent trend with the increase in cluster size, but the fluctuation follows the same pattern for all three global similarity measures. The second global similarity measure shows the highest average similarity among the three, while the first has the lowest.

Table 5.4: Average intra- and inter-cluster similarities for the three iterations of global similarity measures (GSM). Low inter-cluster values are marked yellow.

	GSM 1		GSM 2		GSM 3	
cluster size	intra	inter	intra	inter	intra	inter
3	0.837	0.839	0.873	0.864	0.864	0.840
4	0.846	0.797	0.876	0.842	0.871	0.822
5	0.848	0.819	0.881	0.866	0.874	0.846
6	0.844	0.806	0.882	0.846	0.874	0.839
7	0.848	0.794	0.885	0.831	0.877	0.817
8	0.853	0.817	0.886	0.843	0.881	0.837
9	0.853	0.815	0.888	0.850	0.881	0.843
10	0.852	0.799	0.889	0.847	0.879	0.831

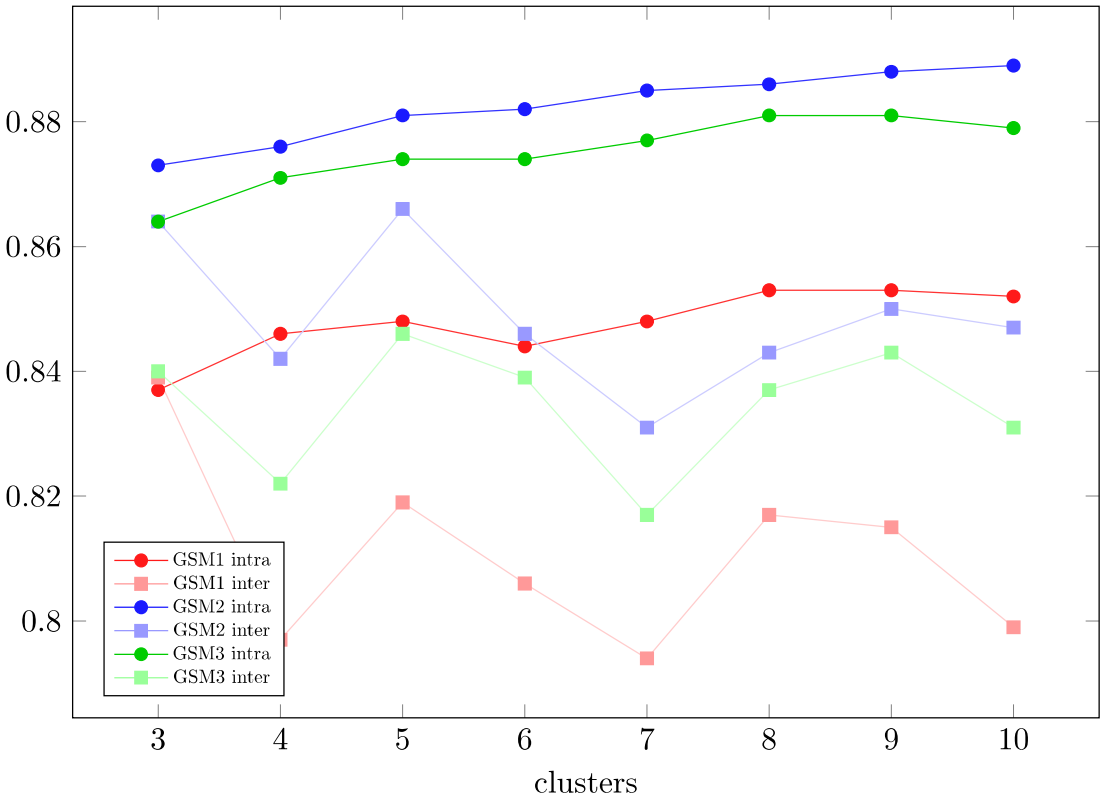


Figure 5.3: Average intra- (circle) and inter-cluster (square) similarity for the three different global similarity measures (GSM), shown in red, blue and green, respectively.

5.3 Evaluation of the Results

Both metric and visual evaluation methods are conducted to evaluate the results.

5.3.1 Silhouette Coefficient and Dunn Index

The evaluation first incorporates the application of two metric methodologies: the Silhouette coefficient and the Dunn index. Detailed numerical data for all cluster sizes and global similarity measures can be found in Table 5.5 and Table 5.6.

Table 5.5: Silhouette coefficients for the three GSMs for all cluster sizes. The highest values are marked yellow.

cluster size	3	4	5	6	7	8	9	10
global sim 01	-0.003	0.062	0.035	0.047	0.068	0.045	0.046	0.066
global sim 02	0.010	0.040	0.018	0.042	0.065	0.052	0.046	0.050
global sim 03	0.029	0.060	0.033	0.042	0.073	0.052	0.045	0.058

Table 5.6: Dunn indexes for the three GSMs for all cluster sizes. The relative lowest values are marked yellow.

cluster size	3	4	5	6	7	8	9	10
global sim 01	1.068	1.049	1.065	1.108	1.065	1.084	1.084	1.055
global sim 02	1.011	1.012	1.050	1.052	1.035	1.045	1.044	1.044
global sim 03	1.014	1.016	1.059	1.055	1.050	1.057	1.071	1.047

Additionally, the outcomes of these two evaluation techniques are graphically represented in Figure 5.4 and Figure 5.5, respectively.

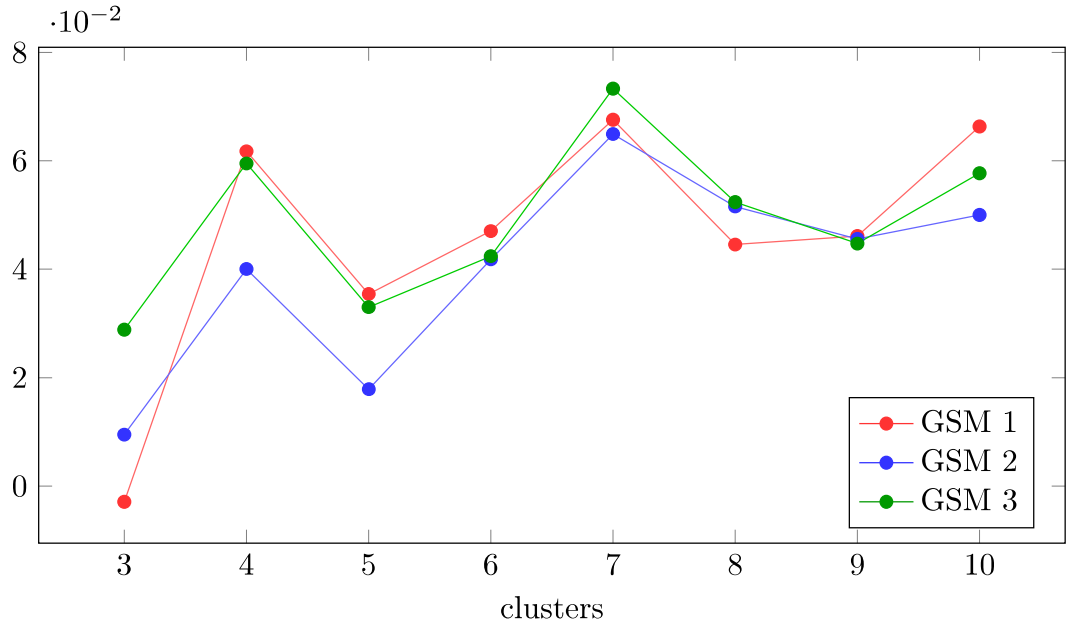


Figure 5.4: Silhouette coefficients for the three GSMs, shown in red, blue and green, respectively, for all cluster sizes.

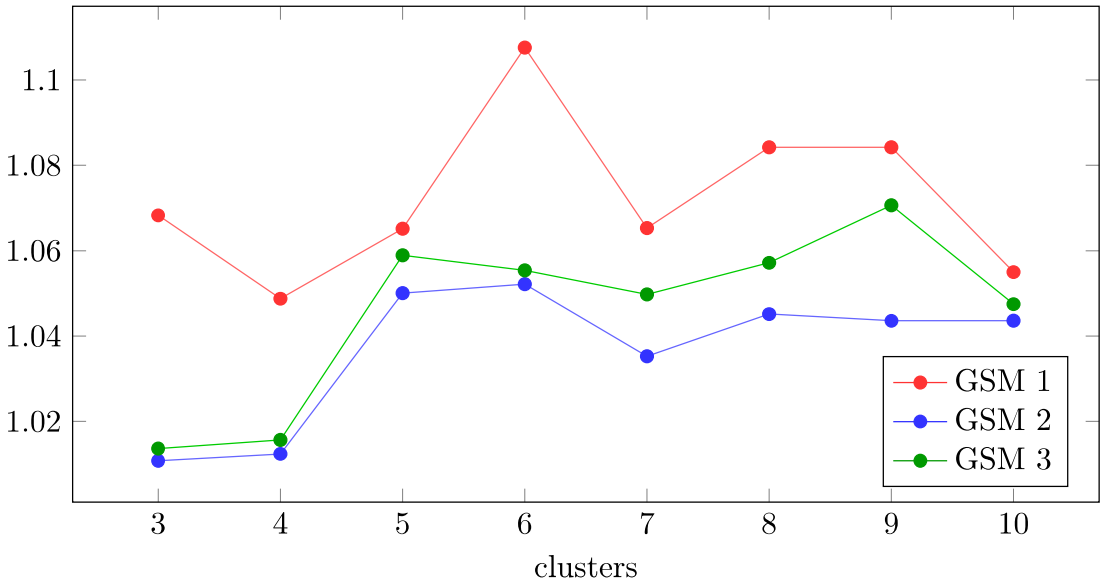


Figure 5.5: Dunn indexes for the three GSMs, shown in red, blue and green, respectively, for all cluster sizes.

Given the Silhouette coefficients, the setup with 7 clusters of GSM3 gives the highest and best score of 0.073. From the Dunn index, the setup with GSM2 for 3 clusters achieves the lowest and best score of 1.011. However, 3 clusters also have the lowest Silhouette coefficients. Therefore, the relative values are more relevant in choosing the best cluster size.

The patterns in the graphs show that cluster sizes 4, 7 and 10 are the most interesting to look at, as these have higher Silhouette coefficients and generally lower Dunn indexes than their neighbours. Generally, the results present a trend of higher scores for increasing cluster sizes for both methods. For the rest of this chapter, 4 clusters will be used to show the results, which is supported by both the evaluation methods and related work [34].

GSM3 has the overall highest average Silhouette coefficient, even though GSM1 for some cluster sizes has higher values. GSM2 has the lowest Silhouette coefficient values. For the Dunn index, GSM1 has the highest values for all cluster sizes and GSM2 has the lowest for all cluster sizes.

5.3.2 Distribution and Composition

Figure 5.6 and Figure 5.7 show the distribution boxplots of the *lying300* and *walking5* attributes respectively, for 4 clusters of the three global similarity measures. The remaining boxplots for 4 clusters, showing all the different activities and bouts, can be found in Appendix A.

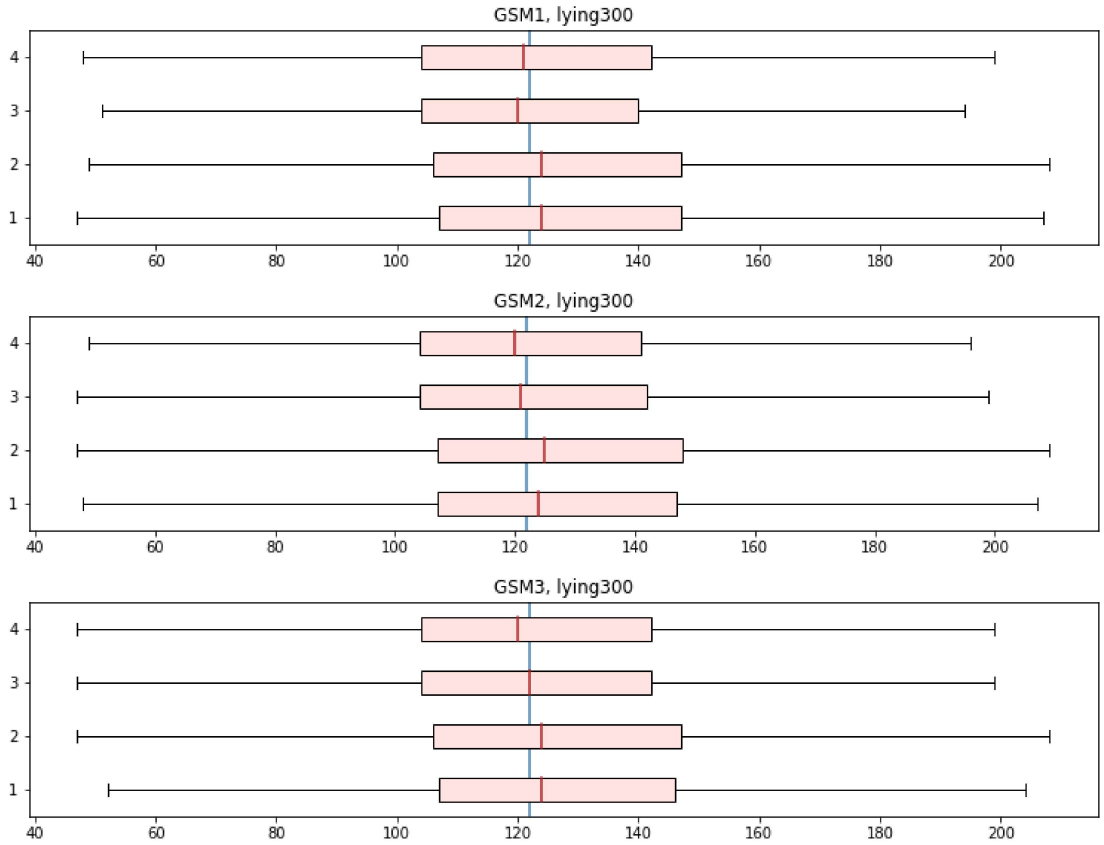


Figure 5.6: Distribution boxplot of 4 clusters for attribute *lying300* for the three global similarity measures. The red line marks the median for the clusters and the blue line the population median.

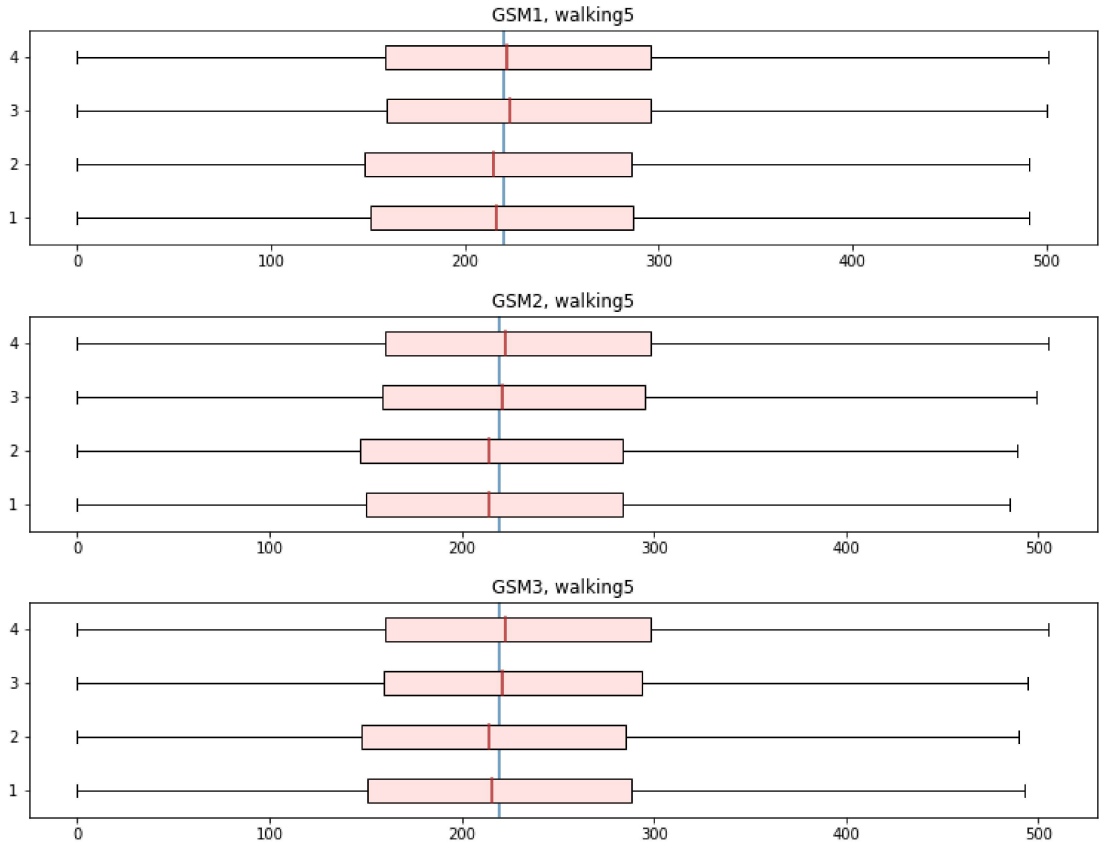


Figure 5.7: Distribution boxplot of 4 clusters for attribute *walking5* for the three global similarity measures. The red line marks the median for the clusters and the blue line the population median.

Both the two figures show close to similar cluster distributions for all three global similarity measures, but the clusters differ in whether or not the cluster medians, depicted by the red lines, are above or below the population median, depicted by the blue line. The same is found for all cluster sizes from 3 to 10.

A day consists of only 24 hours, which means that an increase in one activity entails a reduction in another. Thus, it is advantageous to examine the cluster composition for a comparative analysis of the clusters.

Bar charts showing the bout compositions of the centroids in 4 clusters, can be found in Figure 5.8, which displays the case in the centre of a cluster. Figure 5.9 presents the average bout composition of all the cases in each of the 4 clusters. The cluster compositions for 7 and 10 clusters can be found in Appendix B.

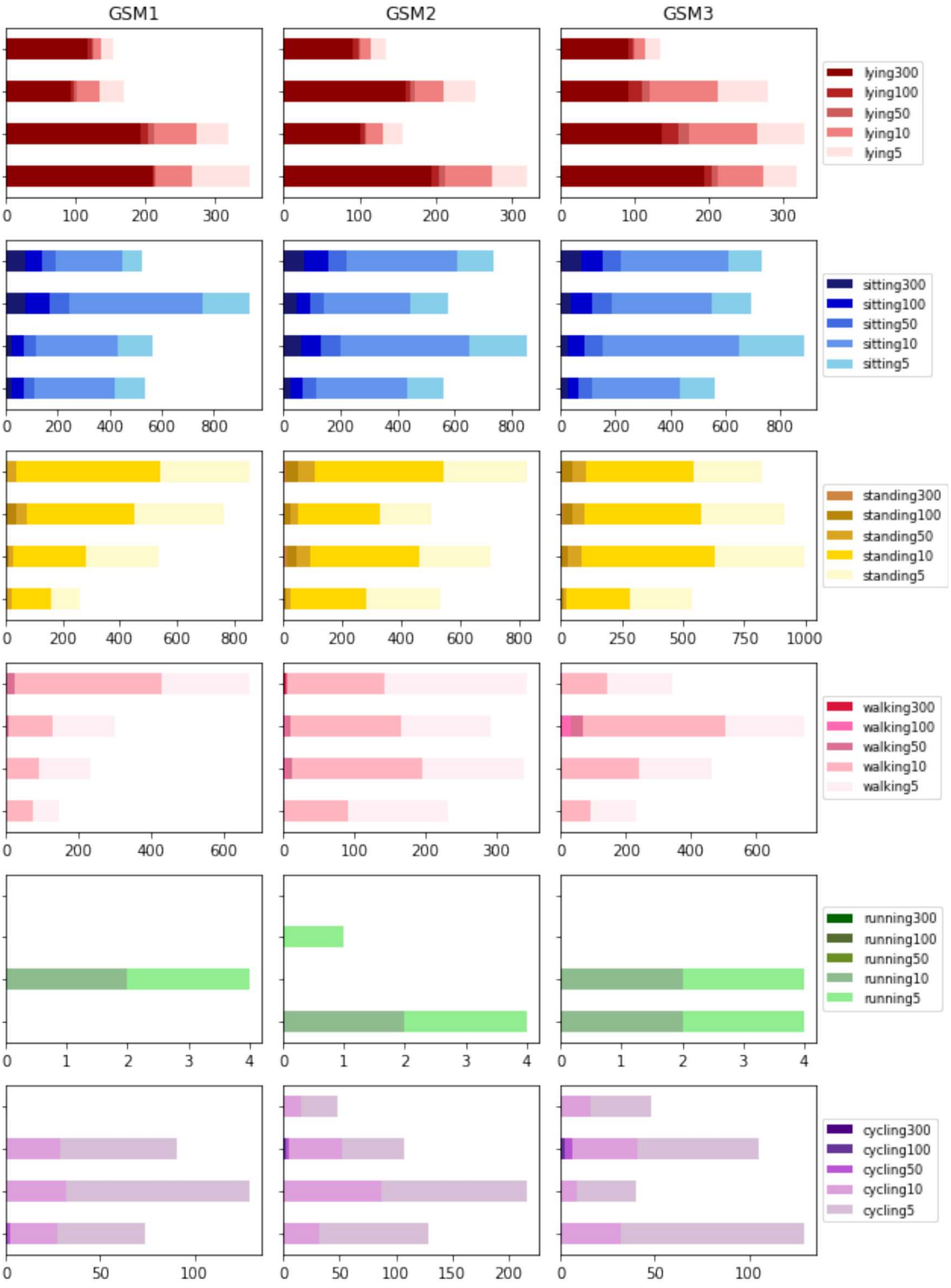


Figure 5.8: Bar charts showing the attribute composition of the 4 cluster centroid cases for the three different global similarity measures (GSM).

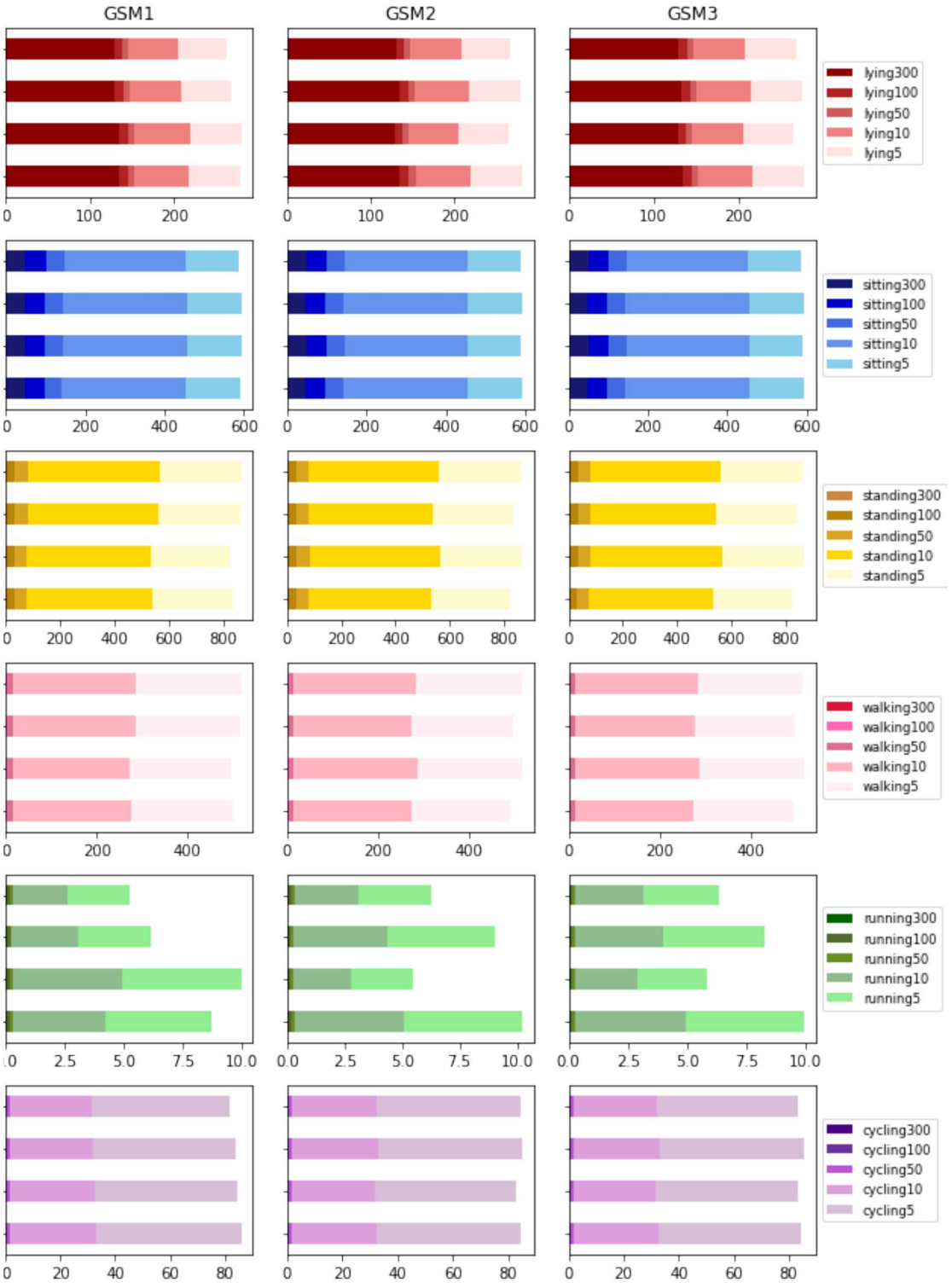


Figure 5.9: Bar charts showing the average attribute composition of the cases in 4 clusters for the three different global similarity measures (GSM).

The centroid bar charts show distinct differences between clusters and GSMs. This applies to all attributes. For the averaged compositions, more similar compositions are seen across all clusters and GSMs, they are however distinguishable. While most of the attributes are similar, larger differences are found in the attributes *cycling5* and *cycling10*.

6 Discussion

This research has been motivated by the objective of exploring existing and innovative approaches to identify physical activity phenotypes. This chapter aims to deliver an assessment of the strategies employed, achieved results and encountered challenges, in accordance with the research questions listed in the first chapter.

6.1 Clustering

The results indicate a correlation between the number of clusters and the average intra-cluster similarities, which aligns with the intuition that more clusters lead to a closer proximity of centroids to their cluster objects which increases cohesion. However, it is imperative to contextualise the number of clusters within the domain of study. While an increase in clusters may yield higher average similarity, it is not inherently beneficial for the specific task of identifying physical activity phenotypes, where the emphasis is on promoting favourable physical activity patterns. The usefulness of the phenotypes for healthcare providers should also be taken into account when discussing cluster sizes.

With the Silhouette coefficient, as calculated in the results, approaching values near zero, it suggests a considerable degree of overlapping within the clusters. It is however important to note that this overlap does not inherently render the clusters invalid or uninformative. Given the extensive dataset encompassing 65 340 data points, such a degree of overlap is within expectations. Study participants display a wide array of activity patterns, which presents a challenge in achieving distinct and well-separated clusters.

The results do however indicate a higher separation for cluster sizes 4, 7 and 10 for all three iterations, which is shown as peaks in the Silhouette Coefficients for these three cluster sizes. In the Dunn index, this is shown as lower values for the three sizes for some of the global similarity measures, but not clearly. Given that four or five clusters are the predominant choices in related research, these findings suggest that a 4-cluster solution might be optimal for identifying meaningful physical activity phenotypes, as this aligns with the conventions in the field. This answers the third research question on what cluster sizes are optimal for forming phenotypes with this data and method.

6.2 Global Similarity Measures

The interpretation of the results is intricate. The Dunn index indicates a preference for GSM2 as the optimal choice, whereas the Silhouette coefficient highlights the merits of GSM3. Interestingly, GSM1 surpasses GSM2 in the Silhouette coefficient but significantly lags behind in the Dunn index. It is noteworthy that, on the whole, the values exhibit close proximity to each other. The visual analysis through distribution boxplots suggests that no considerable changes have been introduced to the cluster compositions due to the different global similarity measures. The bar charts show differences in the centroid compositions, but the average cluster compositions remain quite consistent. This prompts the question of whether the variations in global similarity measures bear any significant impact on the identification of activity phenotypes.

In GSM2, the weighting of the 300-second bouts is doubled. However, for attributes such as *cycling300* and *running300*, the distribution boxplot reveals a median value of zero. This observation raises questions regarding the extent to which this modification in the global similarity measure influences cluster composition. Similarly, in the case of GSM3, the attribute *walking5* exhibits minimal variations in cluster composition. Furthermore, the marginal variations in the evaluation metrics suggest that the adjustments made to the global similarity measure may not have had a substantial impact on the recognition of activity phenotypes. This provides the response to the fourth research question on how adjusting the global similarity measure influences the clustering results.

Maintaining equal weight for the global similarity measure may be the appropriate decision. Given that local similarity measures are determined through a data-driven approach, the attribute distribution is already taken into account. Utilising the distribution to inform weight selection may therefore become redun-

dant. In addition, if the dataset exhibits any bias, that bias will be transferred to the model using a data-driven approach. There could be merit in fine-tuning the weights based on in-depth domain expertise. With a deeper understanding of the characteristics of a phenotype, the weight selection process can be more meticulous and devoid of any bias carried over from the dataset.

6.3 Structuring in Bouts

The chosen data structure for this study is bouts, which was intended to capture the temporal aspect of the data while reducing granularity and serving as an intermediary between the 5-second windows and average minutes per day. Nevertheless, the findings suggest that the utilisation of bouts may not effectively achieve the objective of identifying activity phenotypes within the dataset, as originally intended.

Moreover, the utilisation of bouts in encoding information into the case base results in the loss of important details present in the raw data. Most notably, a lot of the temporal aspects of the data representation are neglected in the current configuration. In its current state, the cases do not adequately encompass the wide spectrum of physical activity patterns existing within the dataset, highlighting a limitation in the chosen approach. Although the chosen approach effectively addresses the diverse durations of activities, it does not consider the timing of activities during the day.

The visualisations of attribute distribution and composition reveal a notable lack of diversity in attribute usage throughout the day within the dataset. For a significant portion of the day, individuals within the population tend to spend their time lying down in bed. Furthermore, for many individuals in the working population, their typical day involves a substantial amount of sitting, interspersed with standing and short walks. This indicates that it is essential to recognise that even small variations in daily activity patterns can be of significant importance when it comes to identifying distinct phenotypes.

6.4 Large Dataset

As previously mentioned, the extensive size of the dataset can influence the degree of overlap observed in the clustering results. While a substantial dataset is valuable for capturing a wide spectrum of diverse activity patterns, it may not necessarily facilitate the clustering process when looking for a smaller amount of phenotypes. In addition, the large number of data points provides a challenge in the technical execution of the research, due to limitations in memory capacity and efficiency.

The literature review reveals that much of the research on activity phenotype clustering has been based on smaller datasets, from small experimental groups of thirty-five children [8] to surveys of around four thousand adults [33]. This prevalence might be attributed to the relatively uncommon availability of extensive datasets like HUNT4 with around thirty-eight thousand participants, which can account for the widespread use of smaller datasets among researchers.

7 Conclusion

In conclusion, this thesis presents the development of a versatile methodology for the identification of physical activity phenotypes, through the integration of case-based reasoning (CBR) and clustering techniques. The methodology is characterised by its data-driven approach to model local similarity measures within the CBR system, complemented by a clustering algorithm inspired by k-means. This clustering process utilises both local and global similarity measures derived from the CBR system to form the clusters. The global similarity measures are modelled through the visualisation of the dataset as well as domain expertise. Notably, the adaptability of the ready-to-use method allows it to be applied across various data representations and domains, making it a robust template for future explorations in the field.

In experimental findings, this study demonstrates a preference for a configuration with four distinct physical activity phenotypes. Furthermore, it elucidates that employing data-driven methods to determine weights for global similarity measures does not significantly impact clustering, particularly when local similarity measures already account for attribute distribution. The visualisations reveal a notable lack of diversity in attribute usage throughout the day with the data representation chosen.

7.1 Future Work

Future work encompasses potential improvements in data representation. The utilisation of bouts, which currently discards a significant portion of the time series data, necessitates the exploration of alternative data representation ap-

proaches. Collaboration with public health experts facilitates the refinement of the global similarity measure.

The clustering algorithm holds the potential to help establish well-functioning CBR systems from unlabelled data. Effectiveness in clustering is indicative of the quality of similarity measures within the CBR system, as the clusters are made directly from the similarities from the system.

One of the primary objectives of identifying physical activity phenotypes is to enable the development of a decision support system for the population. Such a system aims to assist individuals in identifying which changes they must make to their physical activity to end up in a healthier cluster or to avoid being a part of an unhealthy one. For the HUNT4 dataset, the potential for advancement goes even further, with the possibility of incorporating additional health information and health outcomes into the clusters at a later stage.

Bibliography

[1] D. Verma, K. Bach, and P. Mork, "Clustering of physical behaviour profiles using knowledge-intensive similarity measures," pp. 660–667, 01 2020.

[2] K. Bach, A. Kongsvold, H. Bårdstu, E. M. Bardal, H. S. Kjærnli, S. Herland, A. Logacjov, and P. J. Mork, "A machine learning classifier for detection of physical activity types and postures during free-living," *Journal for the Measurement of Physical Behaviour*, vol. 5, no. 1, pp. 24 – 31, 2022.

[3] A. Logacjov, K. Bach, A. Kongsvold, H. B. Bårdstu, and P. J. Mork, "Harth: A human activity recognition dataset for machine learning," *Sensors*, vol. 21, no. 23, 2021.

[4] H. Kohl, C. Craig, E. Lambert, S. Inoue, J. Alkandari, G. Leetongin, and S. Kahlmeier, "The pandemic of physical inactivity: Global action for public health," *Lancet*, vol. 380, pp. 294–305, 07 2012.

[5] N. Sari, "Physical inactivity and its impact on healthcare utilization," *Health economics*, vol. 18, pp. 885–901, 08 2009.

[6] World Health Organization, "Global recommendations on physical activity for health." `https://apps.who.int/iris/handle/10665/44399`, 2010.

[7] A. M. S. Hasund, "Using Case-Based Reasoning to Identify Physical Activity Phenotypes," 2023. Project Thesis NTNU.

[8] C. Díaz and K. Yacef, "Detecting behaviour changes in accelerometer data," in *KHD@IJCAI*, 2018.

[9] A. Aamodt and E. Plaza, "Case-based reasoning: Foundational issues, methodological variations, and system approaches," *AI Commun.*, vol. 7, p. 39–59, mar 1994.

[10] D. Verma, K. Bach, and P. J. Mork, "Modelling similarity for comparing physical activity profiles - a data-driven approach," in *Case-Based Reasoning Research and Development* (M. T. Cox, P. Funk, and S. Begum, eds.), (Cham), pp. 415–430, Springer International Publishing, 2018.

[11] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surv.*, vol. 31, p. 264–323, sep 1999.

[12] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297, Oakland, CA, USA, 1967.

[13] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[14] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *Journal of Cybernetics*, vol. 4, no. 1, pp. 95–104, 1974.

[15] S. Wess, K.-D. Althoff, and G. Derwand, "Using k-d trees to improve the retrieval step in case-based reasoning," in *Topics in Case-Based Reasoning* (S. Wess, K.-D. Althoff, and M. M. Richter, eds.), (Berlin, Heidelberg), pp. 167–181, Springer Berlin Heidelberg, 1994.

[16] R. Bergmann and A. Tartakovski, "Improving kd-tree based retrieval for attribute dependent generalized cases.," 01 2009.

[17] R. Bergmann and A. Stromer, "Mac/fac retrieval of semantic workflows," 02 2013.

[18] P. Marín-Veites and K. Bach, "Explaining cbr systems through retrieval and similarity measure visualizations: A case study," in *Case-Based Reasoning Research and Development* (M. T. Keane and N. Wiratunga, eds.), (Cham), pp. 111–124, Springer International Publishing, 2022.

[19] B. M. Mathisen, A. Aamodt, K. Bach, and H. Langseth, "Learning similarity measures from data," *Progress in Artificial Intelligence*, vol. 9, pp. 129–143, oct 2019.

[20] A. Stahl, "Learning similarity measures: A formal view based on a generalized cbr model," in *Case-Based Reasoning Research and Development* (H. Muñoz-Ávila and F. Ricci, eds.), (Berlin, Heidelberg), pp. 507–521, Springer Berlin Heidelberg, 2005.

[21] A. Abdel-Aziz, M. Strickert, and E. Hüllermeier, "Learning solution similarity in preference-based cbr," in *Case-Based Reasoning Research and Development* (L. Lamontagne and E. Plaza, eds.), (Cham), pp. 17–31, Springer International Publishing, 2014.

[22] G.-N. Zhu, J. Hu, J. Qi, J. Ma, and Y.-H. Peng, "An integrated feature selection and cluster analysis techniques for case-based reasoning," *Engineering Applications of Artificial Intelligence*, vol. 39, pp. 14–22, 2015.

[23] M. U. Ahmed, H. Banaee, and A. Loutfi, "Health monitoring for elderly: An application using case-based reasoning and cluster analysis," *International Scholarly Research Notices*, vol. 2013, 2013.

[24] G. Müller and R. Bergmann, "A cluster-based approach to improve similarity-based retrieval for process-oriented case-based reasoning," in *ECAI 2014*, pp. 639–644, IOS Press, 2014.

[25] Q. Yang and J. Wu, "Enhancing the effectiveness of interactive case-based reasoning with clustering and decision forests," *Applied Intelligence*, vol. 14, pp. 49–64, 01 2001. Copyright - Kluwer Academic Publishers 2001; Last updated - 2011-07-28.

[26] M. Fullen, P. Schüller, and O. Niggemann, "Semi-supervised case-based reasoning approach to alarm flood analysis," 10 2017.

[27] S. Massie, S. Craw, and N. Wiratunga, "Visualisation of case-base reasoning for explanation," 01 2004.

[28] A. Schultheis, M. Hoffmann, L. Malburg, and R. Bergmann, "Explanation of similarities in process-oriented case-based reasoning by visualization," in *Case-Based Reasoning Research and Development* (S. Massie and S. Chakraborti, eds.), (Cham), pp. 53–68, Springer Nature Switzerland, 2023.

[29] K. Martin, A. Liret, N. Wiratunga, G. Owusu, and M. Kern, "Evaluating explainability methods intended for multiple stakeholders," *KI - Künstliche Intelligenz*, vol. 35, 02 2021.

[30] A. Cantu, M. E. Micó-Amigo, S. Del Din, and S. J. Fernstad, "Parallel assemblies plot, a visualization tool to explore categorical and quantitative data: application to digital mobility outcomes," in *2023 IEEE 16th Pacific Visualization Symposium (PacificVis)*, pp. 21–30, 2023.

[31] Y. Leal, M. Ruiz, C. Lorencio, J. Bondia, L. Mujica, and J. Vehi, "Principal component analysis in combination with case-based reasoning for detecting

therapeutically correct and incorrect measurements in continuous glucose monitoring systems," *Biomedical Signal Processing and Control*, vol. 8, no. 6, pp. 603–614, 2013.

[32] M. Ruiz, G. Sin, X. Berjaga, J. Colprim, S. Puig, and J. Colomer, "Multivariate principal component analysis and case-based reasoning for monitoring, fault detection and diagnosis in a wwtp," *Water science and technology : a journal of the International Association on Water Pollution Research*, vol. 64, pp. 1661–7, 10 2011.

[33] M. Marschollek, "A semi-quantitative method to denote generic physical activity phenotypes from long-term accelerometer data – the atlas index," *PloS one*, vol. 8, p. e63522, 05 2013.

[34] N. Gupta, D. M. Hallman, D. Dumuid, A. Vij, C. L. Rasmussen, M. B. Jørgensen, and A. Holtermann, "Movement behavior profiles and obesity: a latent profile analysis of 24-h time-use composition among danish workers," *International Journal of Obesity*, vol. 44, no. 2, pp. 409–417, 2020.

[35] E. Howie, A. Smith, J. McVeigh, and L. Straker, "Accelerometer-derived activity phenotypes in young adults: a latent class analysis," *International Journal of Behavioral Medicine*, vol. 25, 04 2018.

[36] I. Meisingset, O. Vasseljen, N. Vøllestad, H. Robinson, A. Woodhouse, K. Engebretsen, M. Glette, C. Øverås, A. Nordstoga, K. A. Evensen, and E. Schjelderup Skarpsno, "Novel approach towards musculoskeletal phenotypes," *European Journal of Pain*, vol. 24, 02 2020.

[37] M. Willetts, S. Hollowell, L. Aslett, C. Holmes, and A. Doherty, "Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 uk biobank participants," *Scientific Reports*, vol. 8, 05 2018.

[38] O. A. Bergman, ""prediction of personalized speed skating results using case-based reasoning"," Master's thesis, NTNU, 2019.

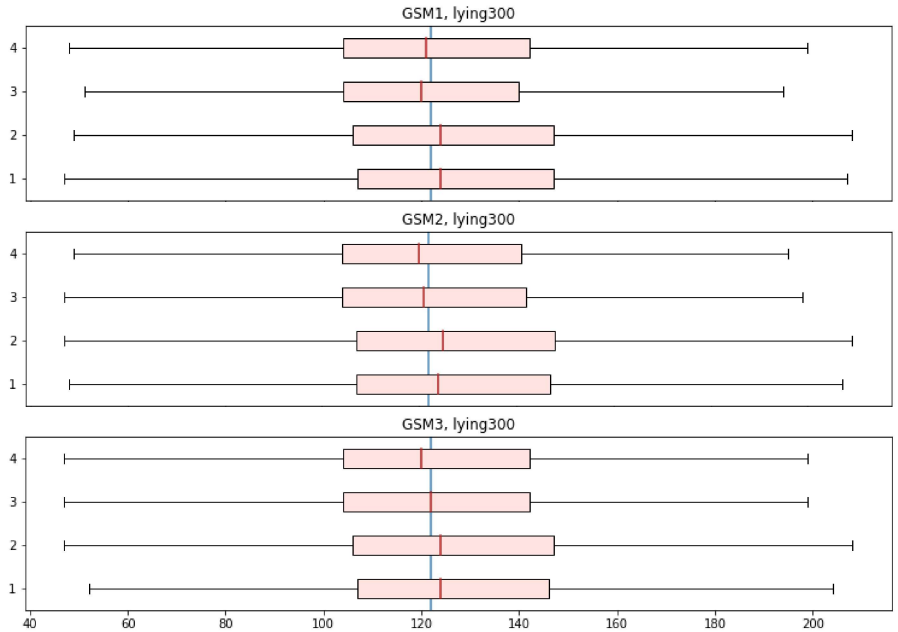
[39] B. Smyth and P. Cunningham, "Running with cases: A cbr approach to running your best marathon," in *Case-Based Reasoning Research and Development* (D. W. Aha and J. Lieber, eds.), (Cham), pp. 360–374, Springer International Publishing, 2017.

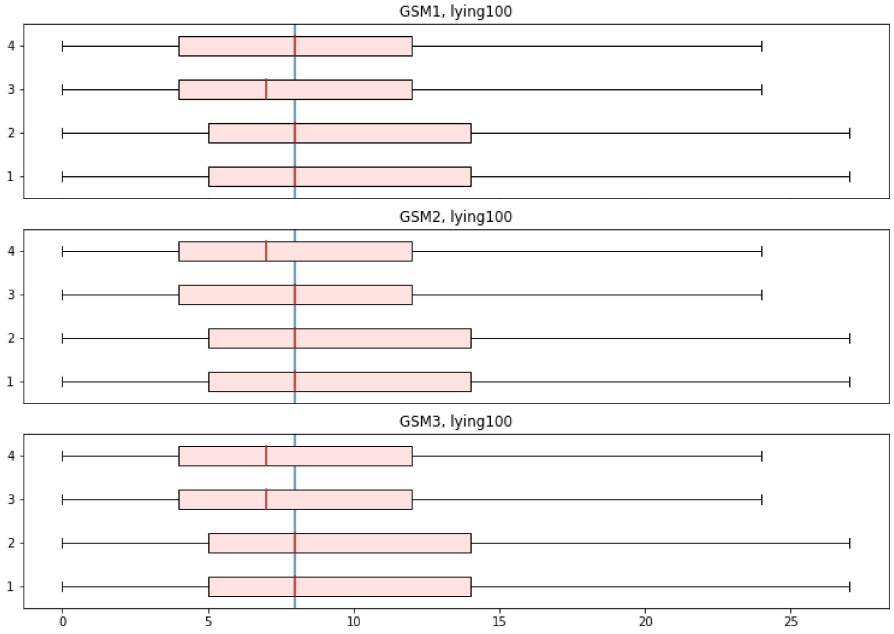
Appendices

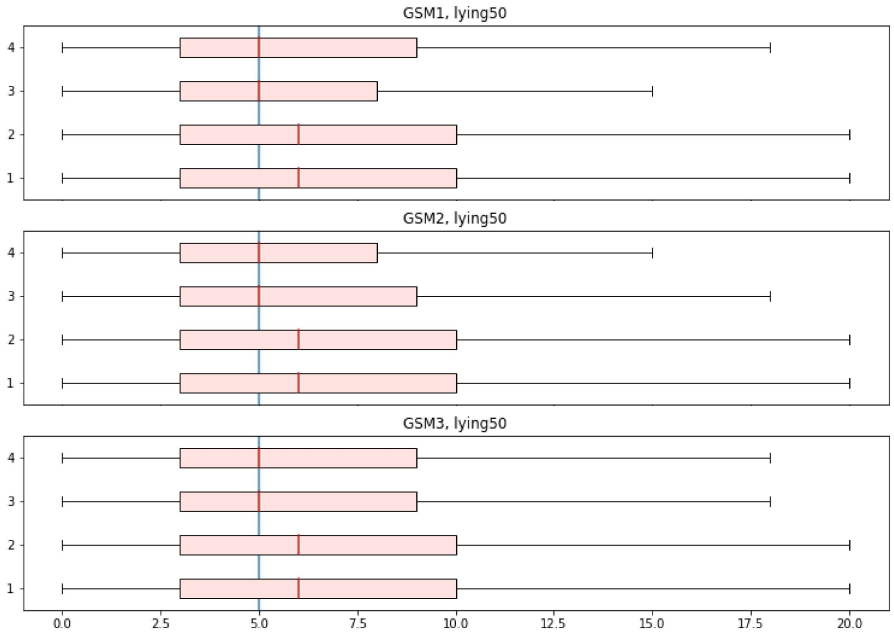
A Boxplots

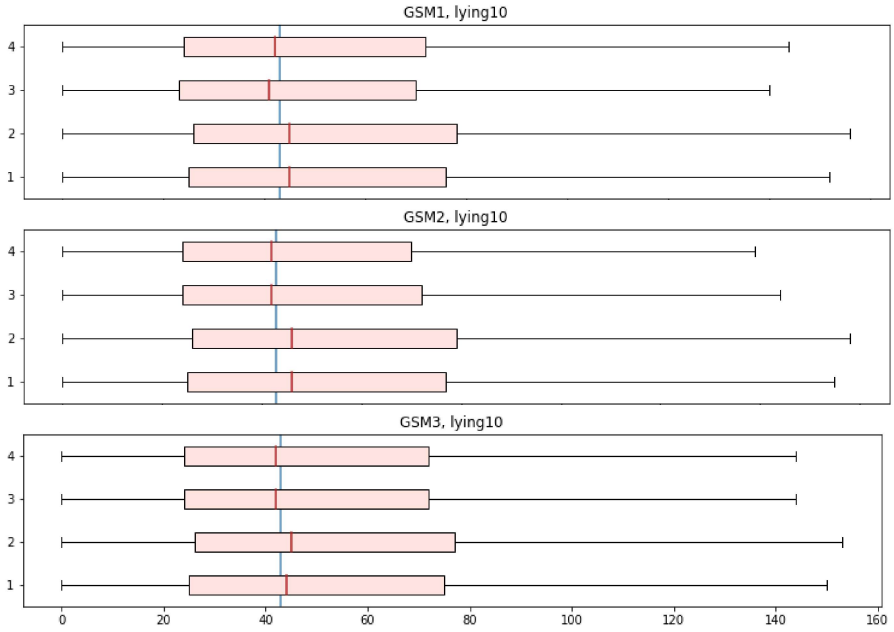
Distribution boxplots for all activity bouts, for 4 clusters of the three global similarity measures.

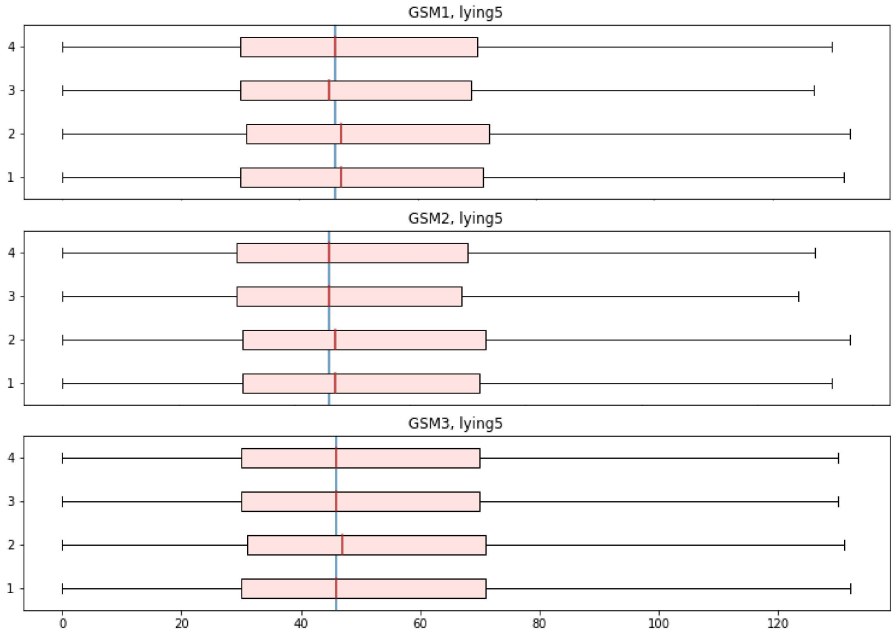
A.1 Lying



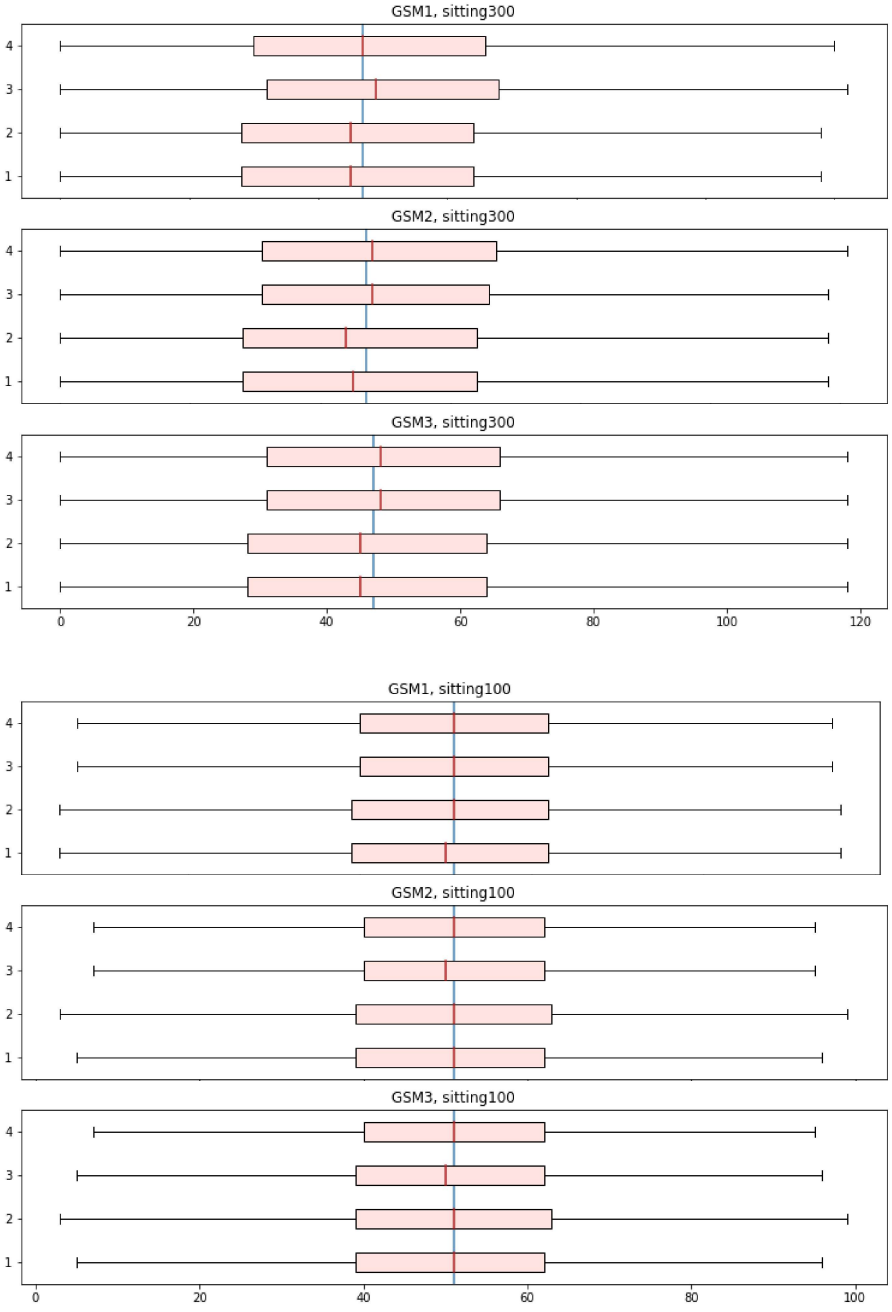

GSM1, lying100
GSM2, lying100
GSM3, lying100
4
3
2
1
0
5
10
15
20
25

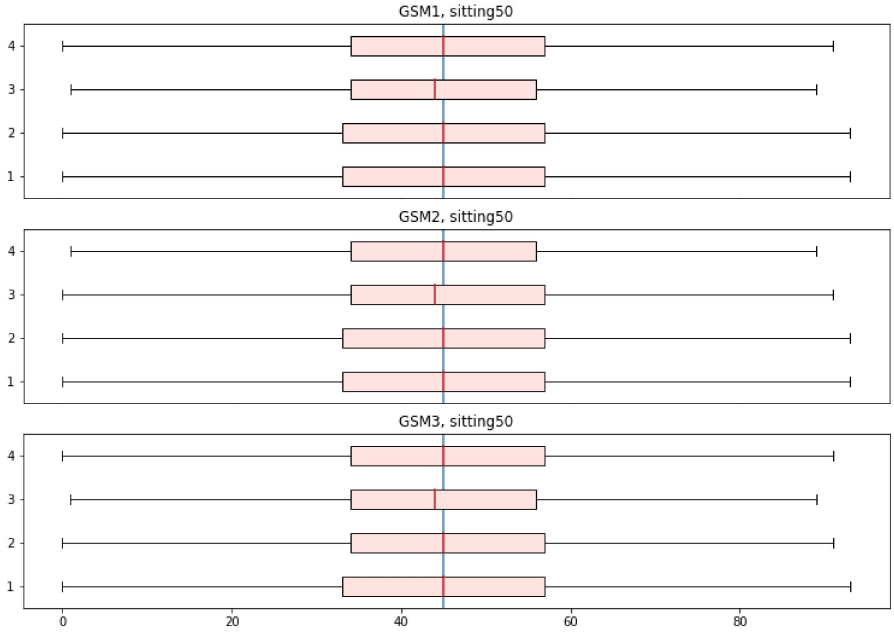

GSM1, lying50
GSM2, lying50
GSM3, lying50
4
3
2
1
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

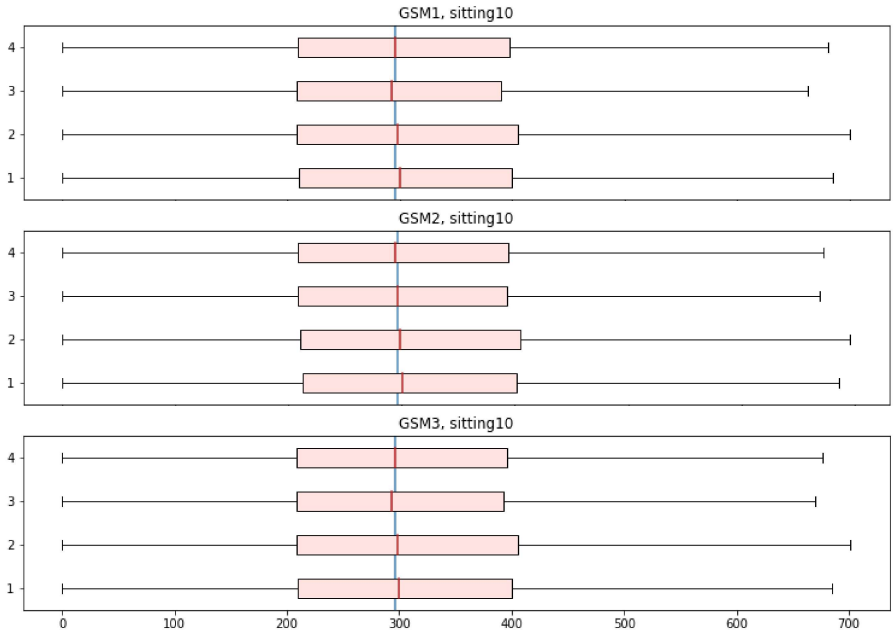

GSM1, lying10
GSM2, lying10
GSM3, lying10
4
3
2
1
0
20
40
60
80
100
120
140
160

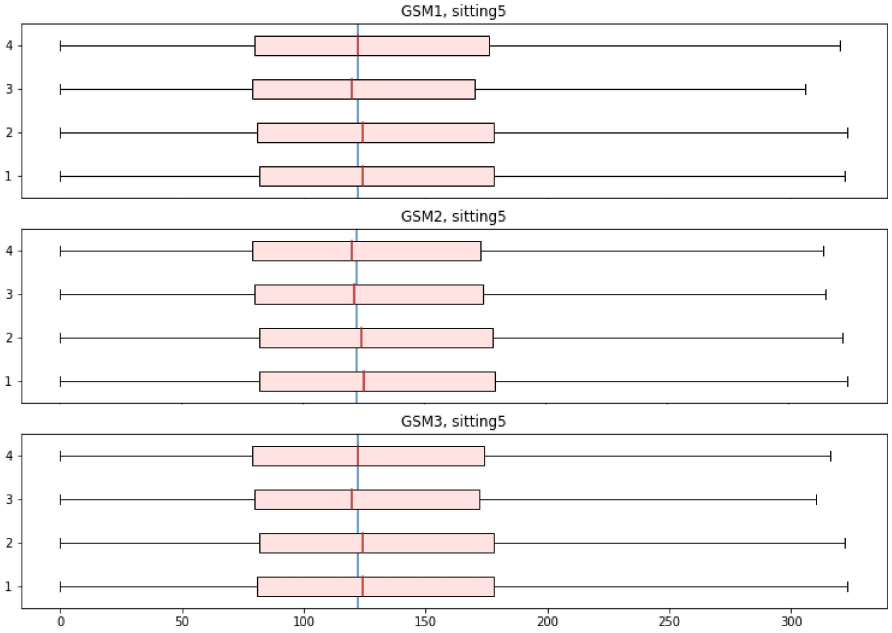

GSM1, lying5
GSM2, lying5
GSM3, lying5
4
3
2
1
0
20
40
60
80
100
120

A.2 Sitting

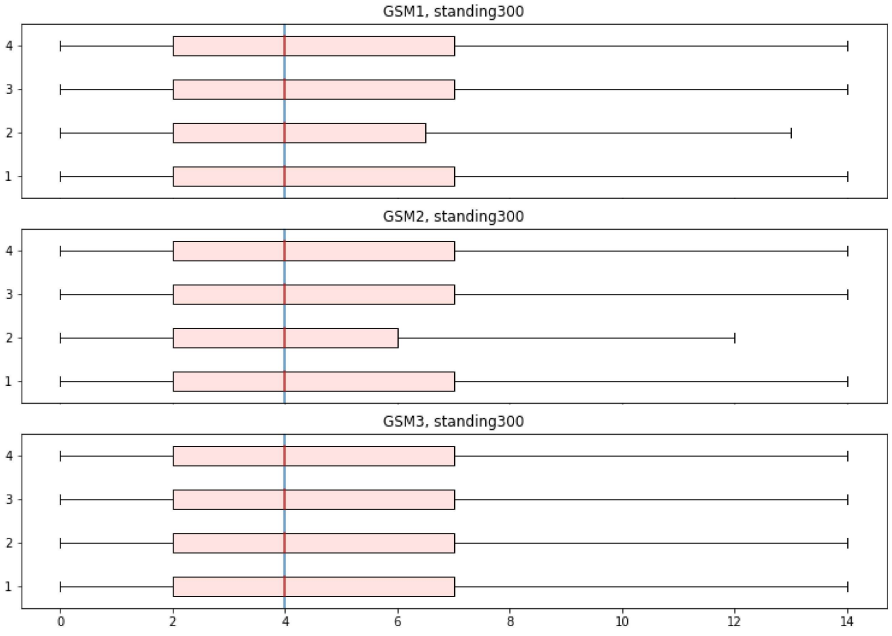

GSM1, sitting300
GSM2, sitting300
GSM3, sitting300
GSM1, sitting100
GSM2, sitting100
GSM3, sitting100

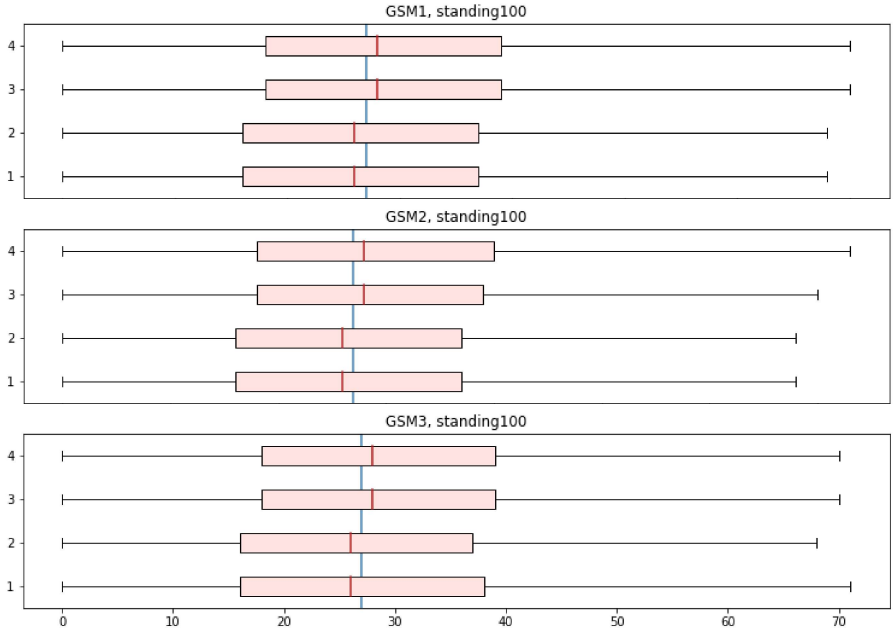

GSM1, sitting50
GSM2, sitting50
GSM3, sitting50
4
3
2
1
0
20
40
60
80

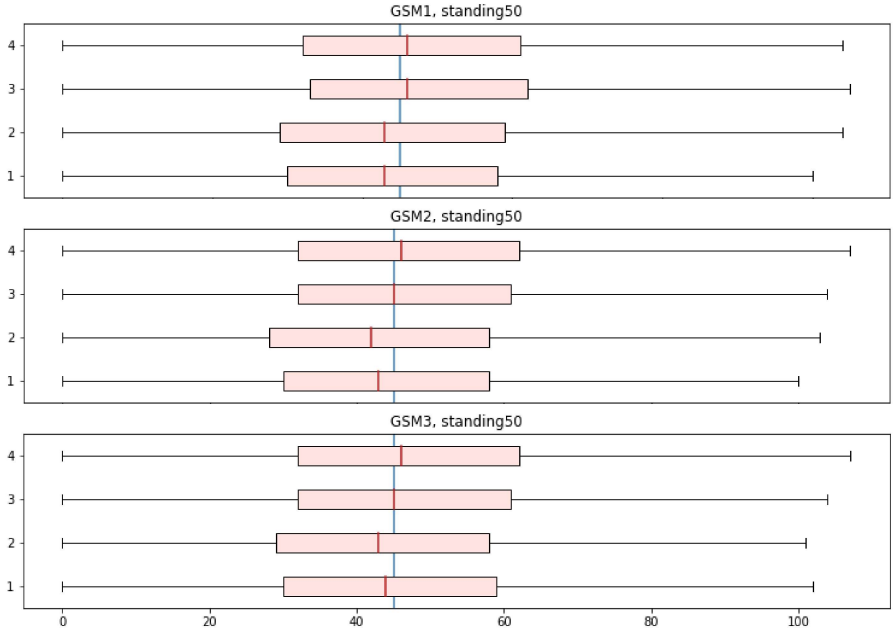

GSM1, sitting10
GSM2, sitting10
GSM3, sitting10
4
3
2
1
0
100
200
300
400
500
600
700

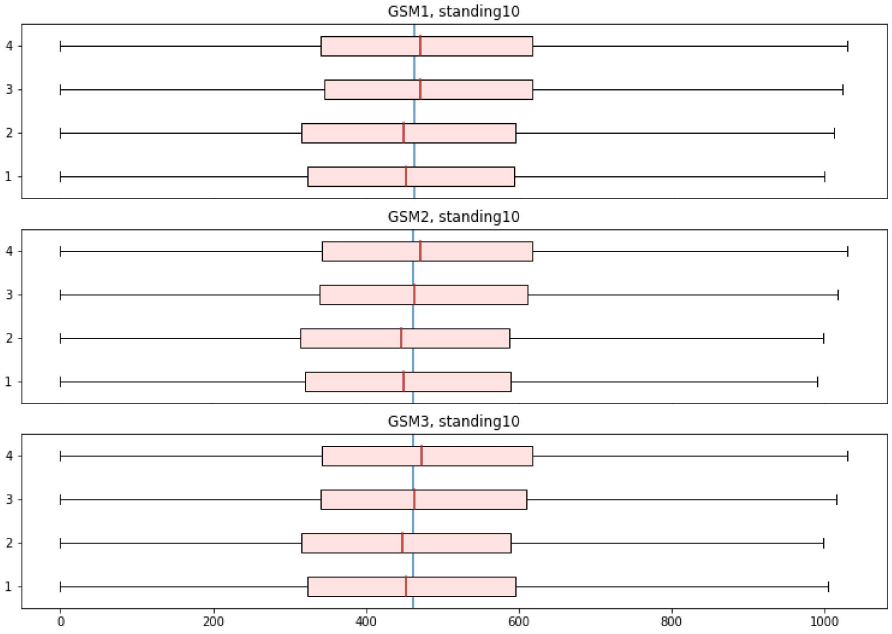

GSM1, sitting5
GSM2, sitting5
GSM3, sitting5
4
3
2
1
0
50
100
150
200
250
300

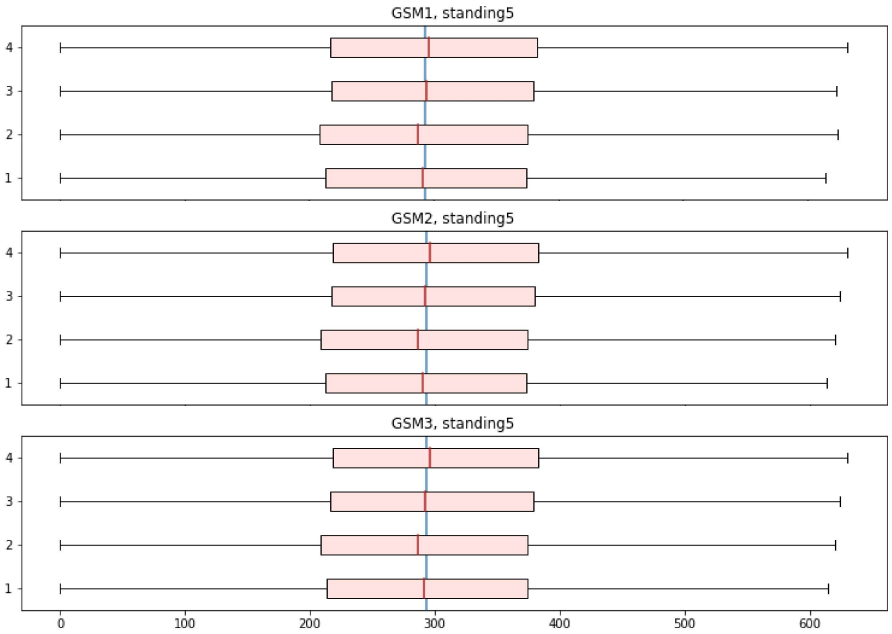
A.3 Standing


GSM1, standing300
GSM2, standing300
GSM3, standing300
4
3
2
1
0
2
4
6
8
10
12
14

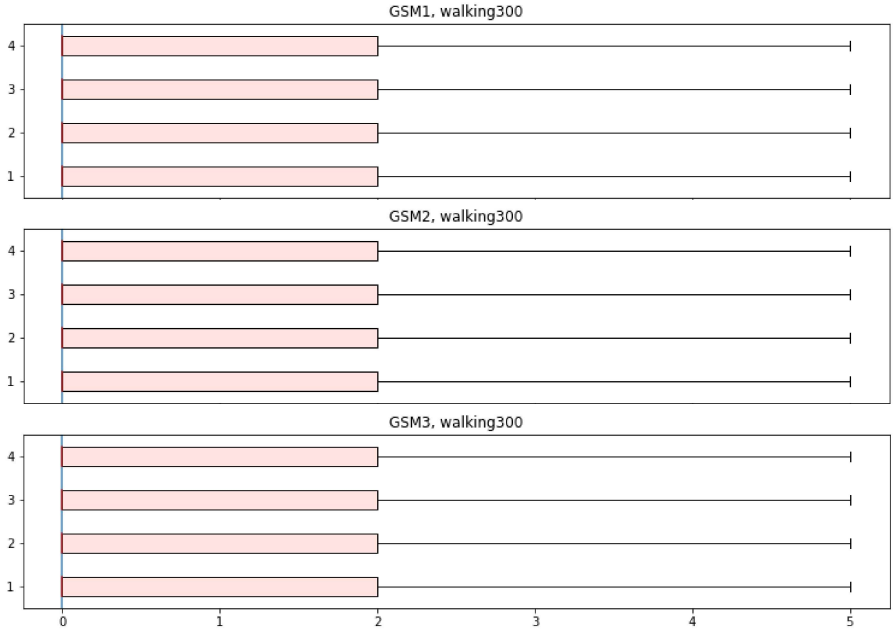

GSM1, standing100
GSM2, standing100
GSM3, standing100
4
3
2
1
0
10
20
30
40
50
60
70

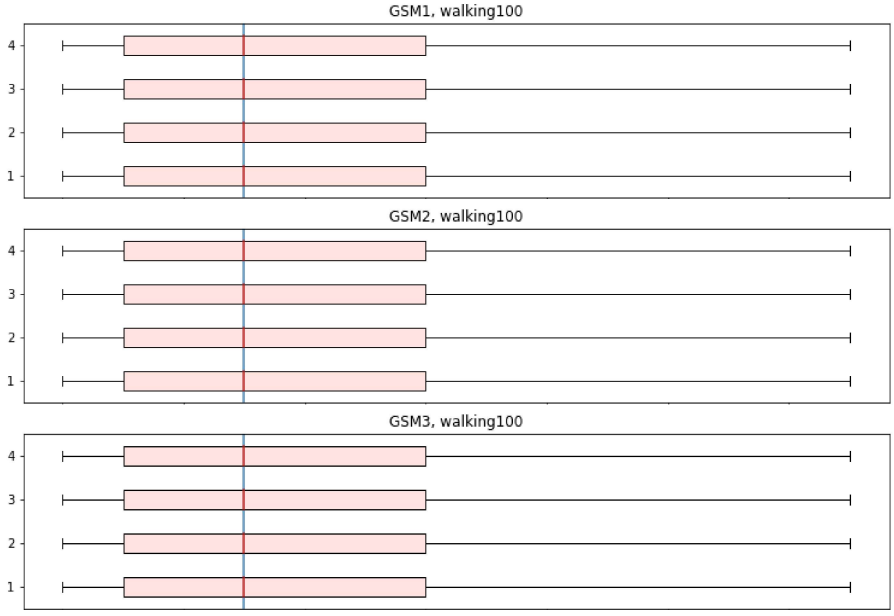

GSM1, standing50
GSM2, standing50
GSM3, standing50
4
3
2
1
0
20
40
60
80
100

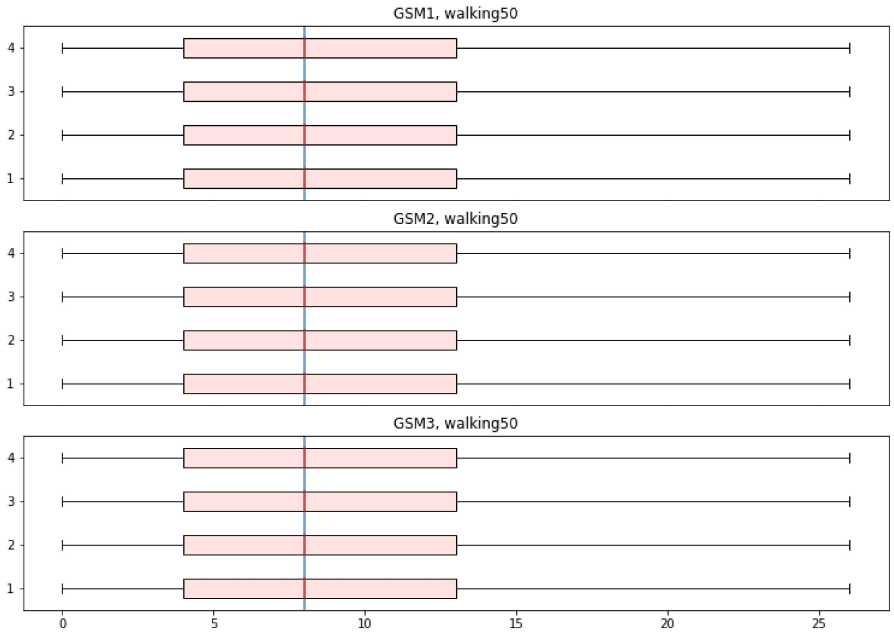

GSM1, standing10
GSM2, standing10
GSM3, standing10
4
3
2
1
0
200
400
600
800
1000

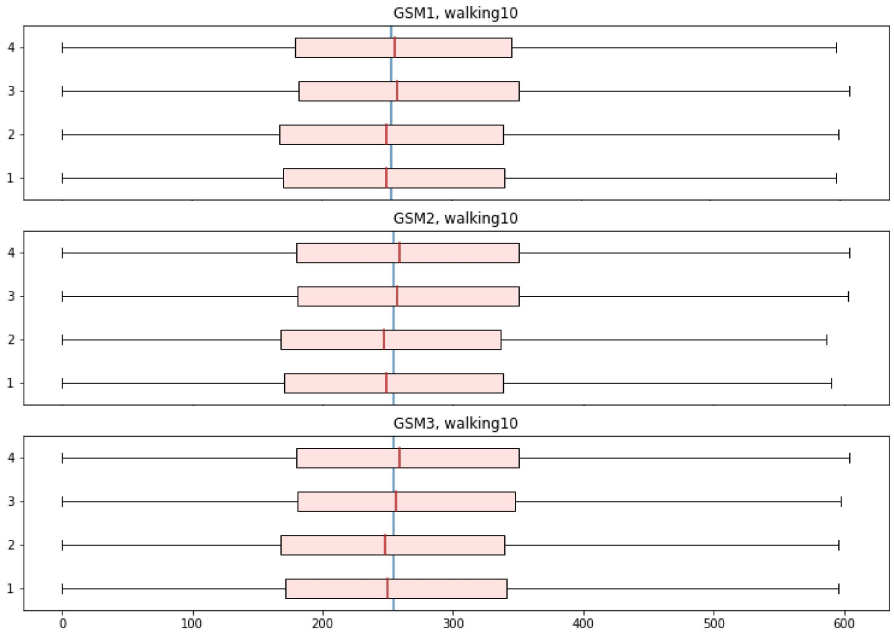

GSM1, standing5
GSM2, standing5
GSM3, standing5
4
3
2
1
0
100
200
300
400
500
600

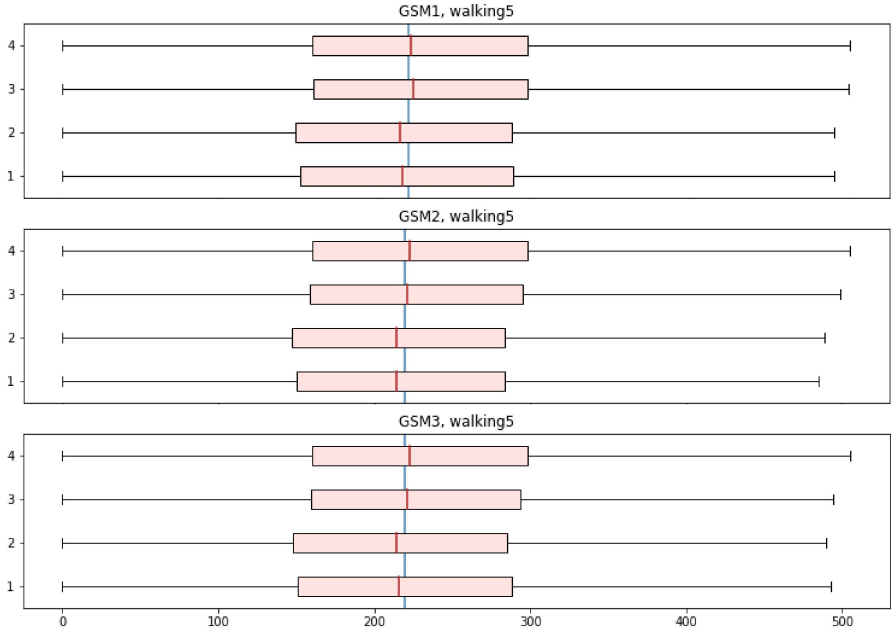
A.4 Walking


GSM1, walking300
GSM2, walking300
GSM3, walking300
4
3
2
1
0
1
2
3
4
5

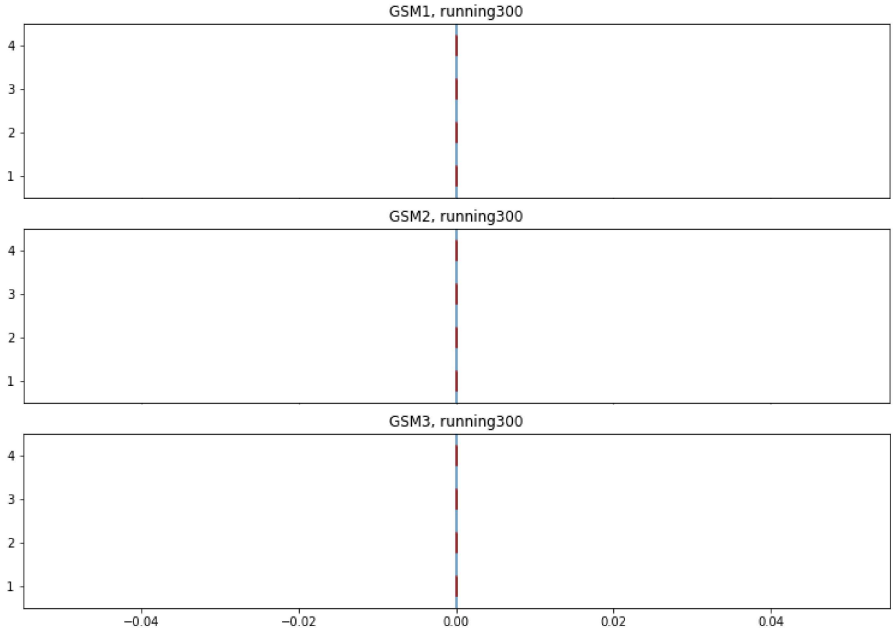

GSM1, walking100
GSM2, walking100
GSM3, walking100
4
3
2
1

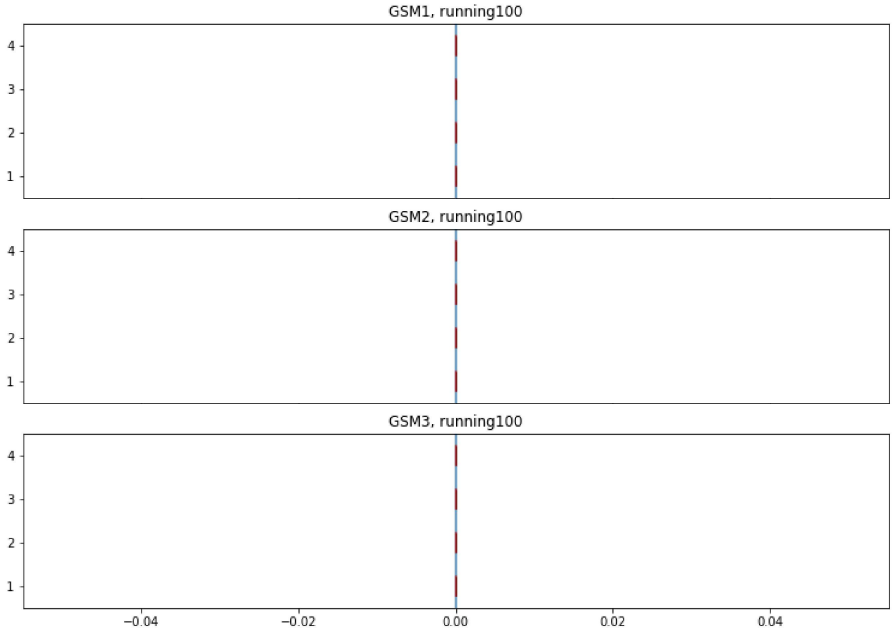

GSM1, walking50
GSM2, walking50
GSM3, walking50
4
3
2
1
0
5
10
15
20
25

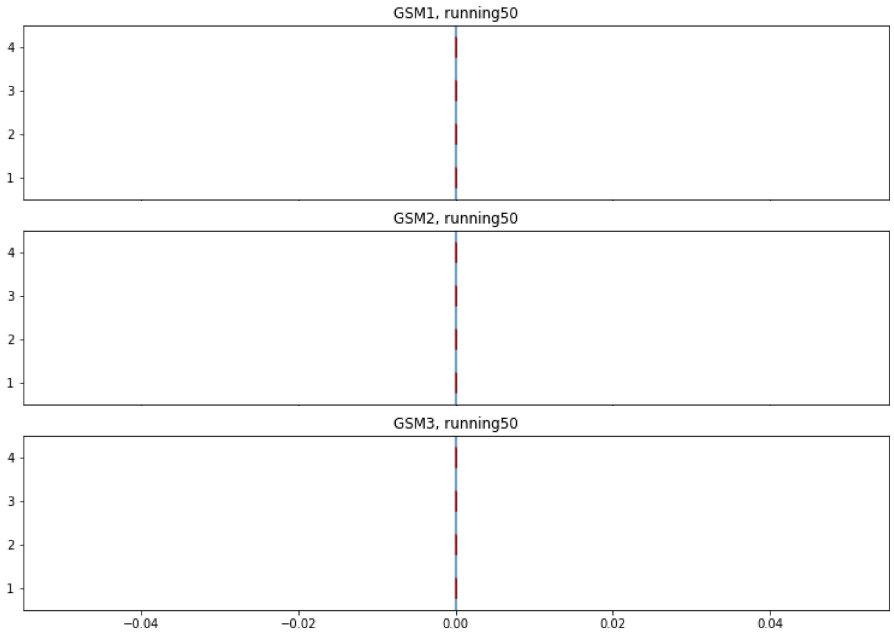

GSM1, walking10
GSM2, walking10
GSM3, walking10
4
3
2
1
0
100
200
300
400
500
600

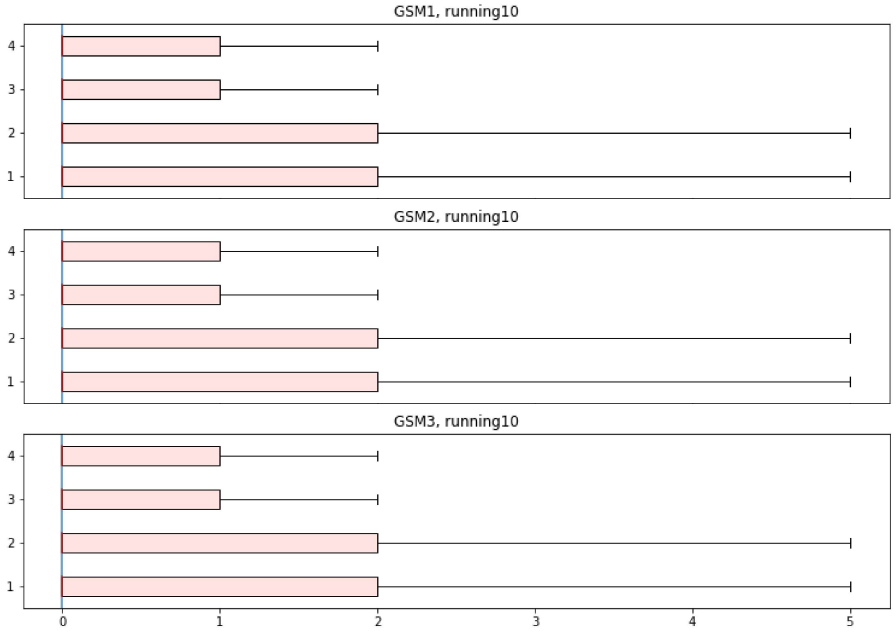

GSM1, walking5
GSM2, walking5
GSM3, walking5
4
3
2
1
0
100
200
300
400
500

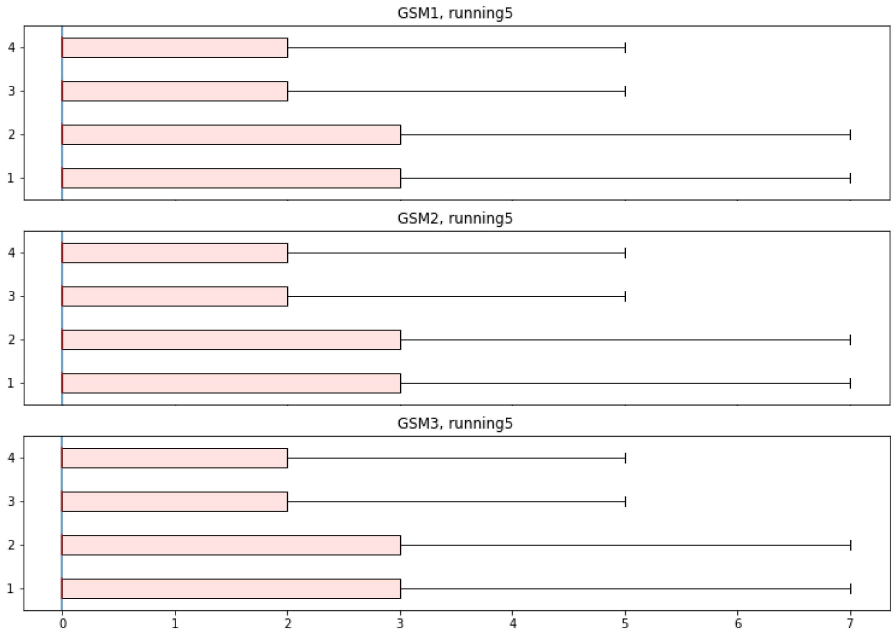
A.5 Running


GSM1, running300
GSM2, running300
GSM3, running300
4
3
2
1
−0.04
−0.02
0.00
0.02
0.04

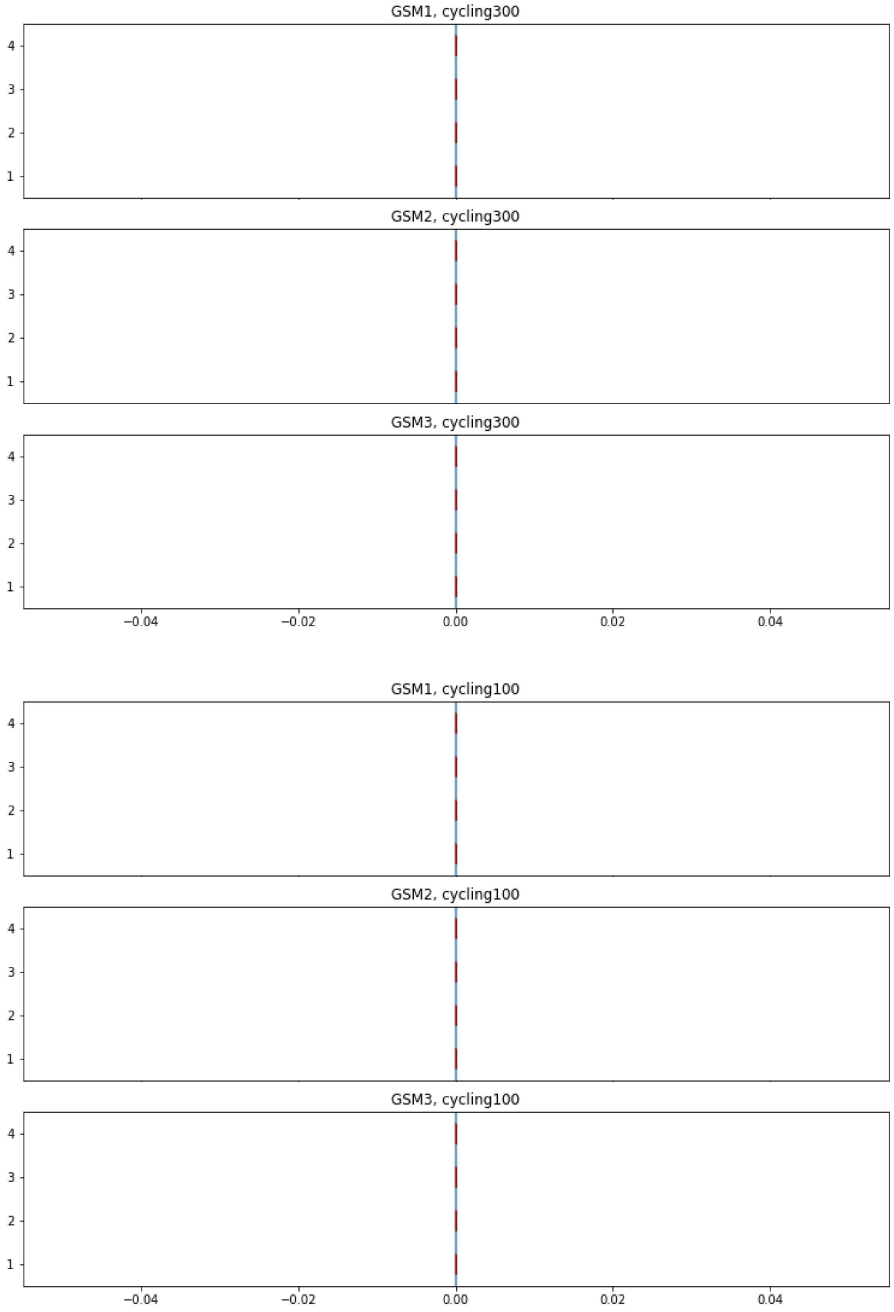

GSM1, running100
GSM2, running100
GSM3, running100
4
3
2
1
−0.04
−0.02
0.00
0.02
0.04

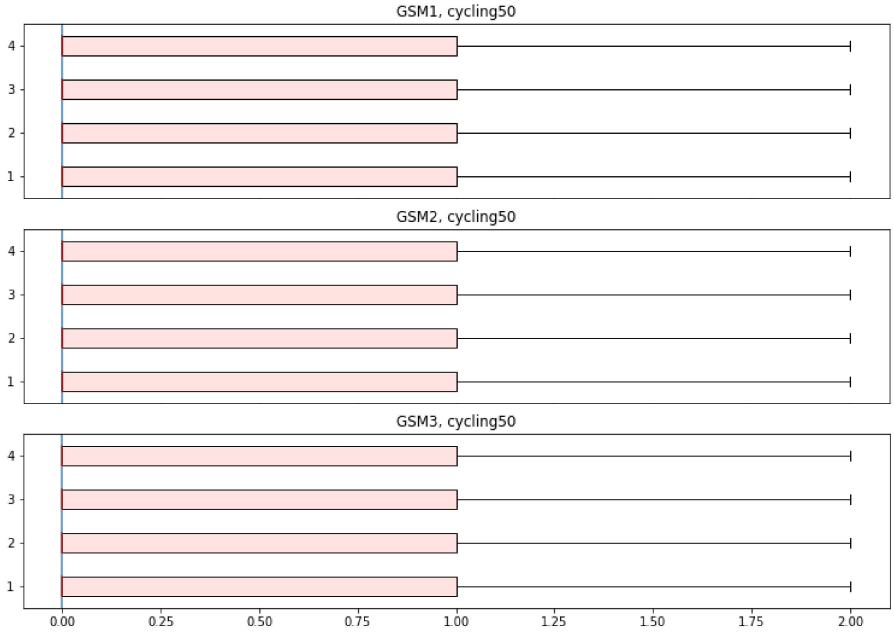

GSM1, running50
GSM2, running50
GSM3, running50
4
3
2
1
−0.04
−0.02
0.00
0.02
0.04

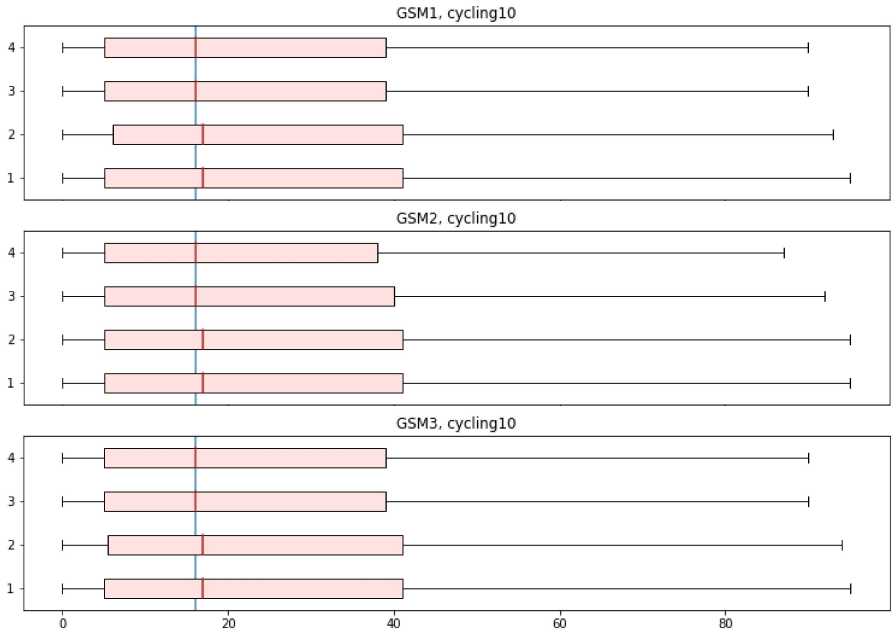

GSM1, running10
GSM2, running10
GSM3, running10
4
3
2
1
0
1
2
3
4
5

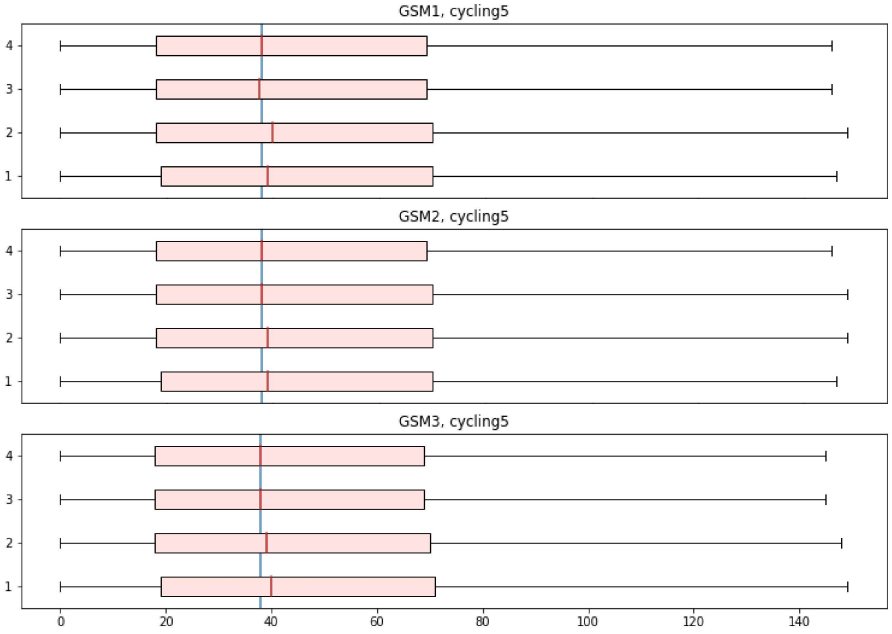

GSM1, running5
GSM2, running5
GSM3, running5
4
3
2
1
0
1
2
3
4
5
6
7

A.6 Cycling

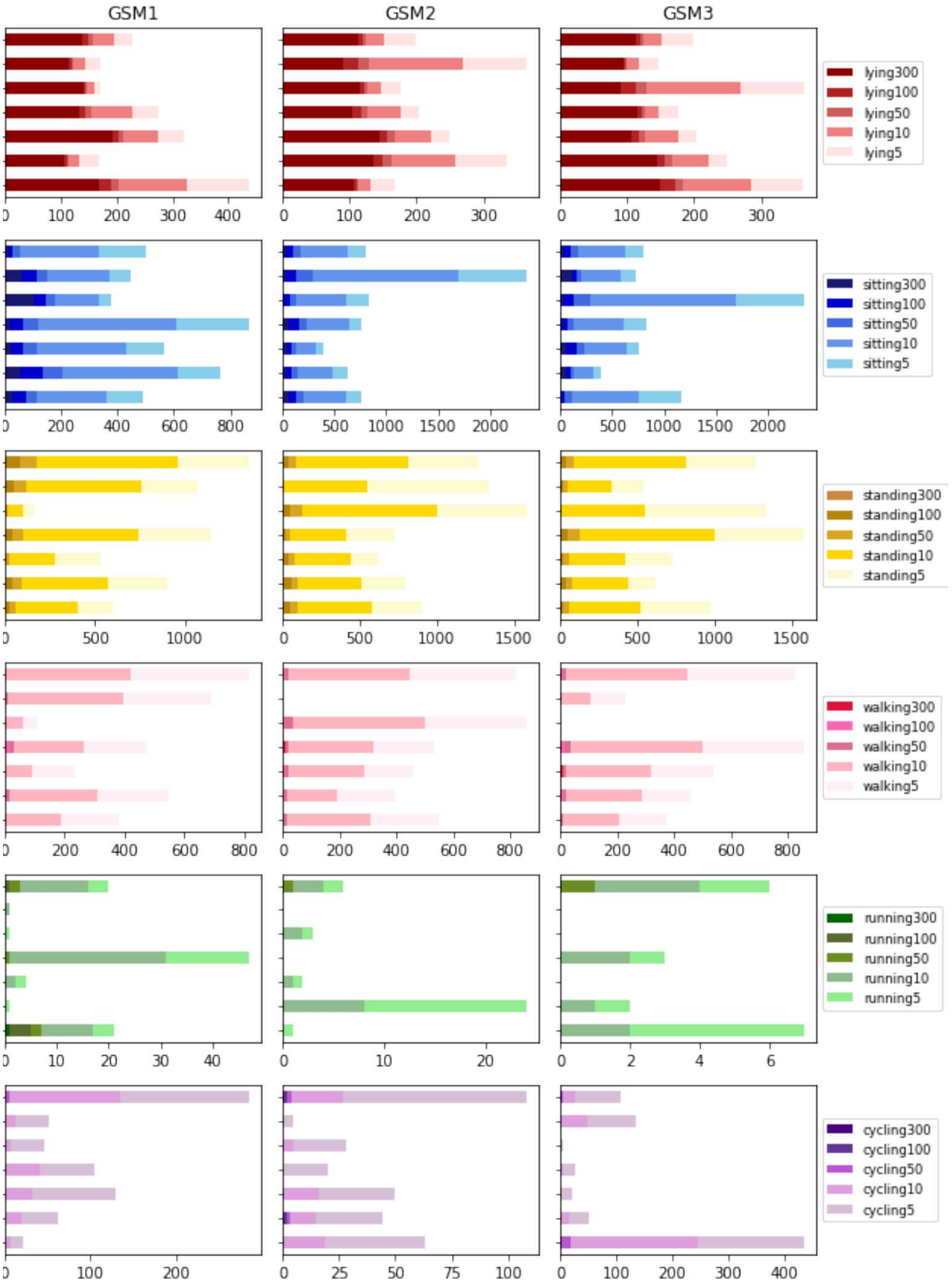

GSM1, cycling300
GSM2, cycling300
GSM3, cycling300
GSM1, cycling100
GSM2, cycling100
GSM3, cycling100
4
3
2
1
−0.04
−0.02
0.00
0.02
0.04


GSM1, cycling50
GSM2, cycling50
GSM3, cycling50
4
3
2
1
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00

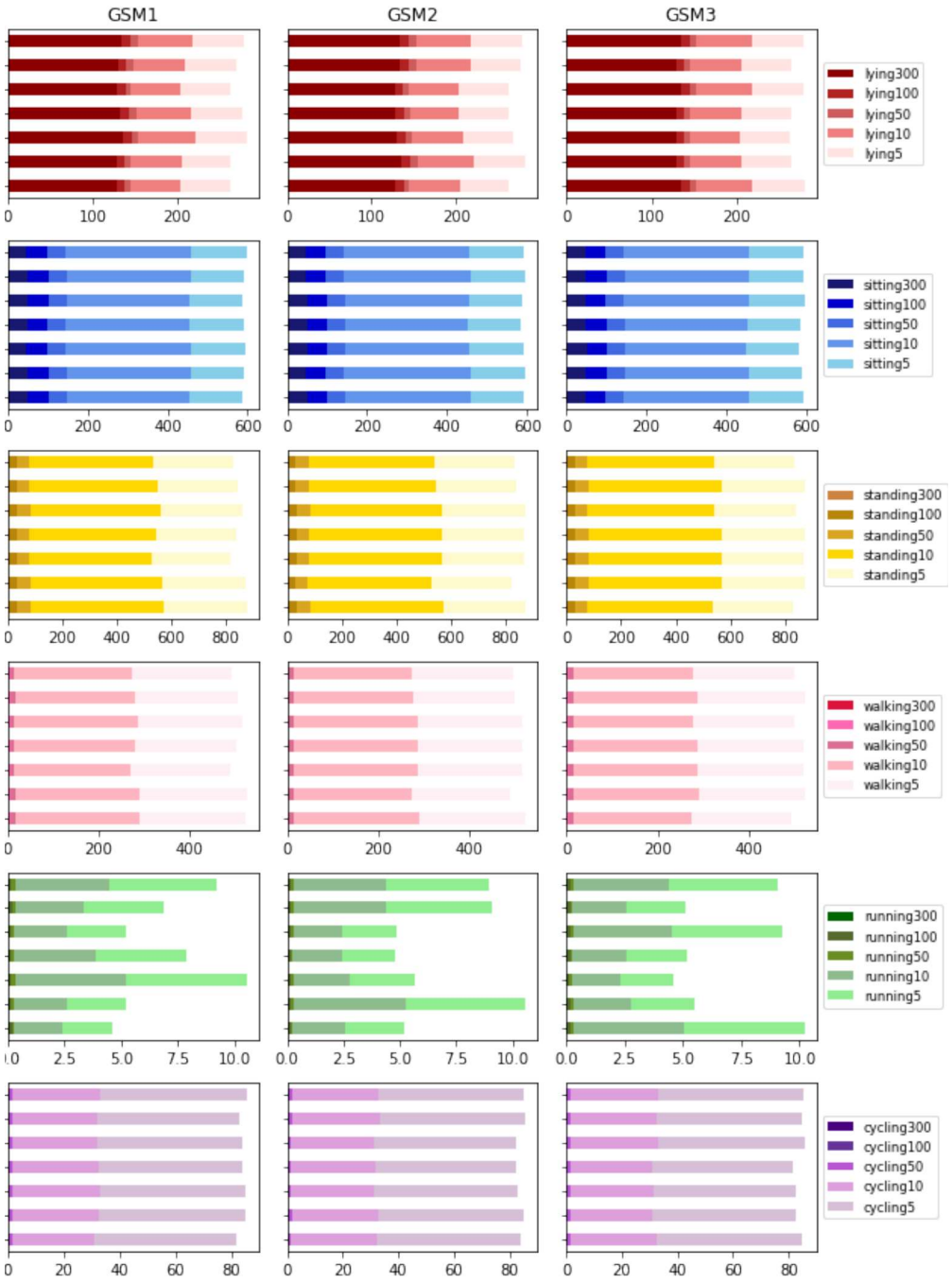

GSM1, cycling10
GSM2, cycling10
GSM3, cycling10
4
3
2
1
0
20
40
60
80


GSM1, cycling5
GSM2, cycling5
GSM3, cycling5
4
3
2
1
0
20
40
60
80
100
120
140

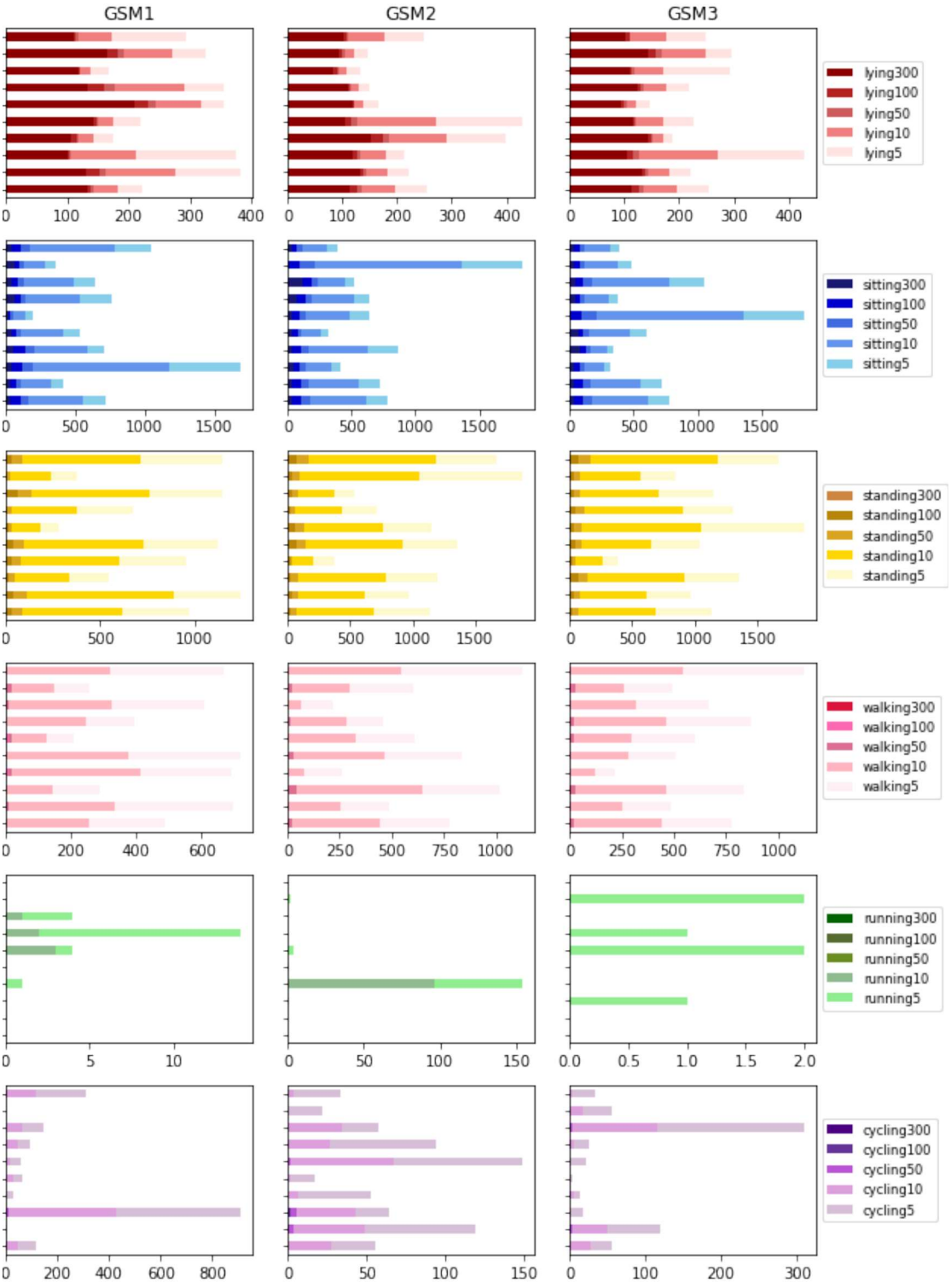
B Bar charts



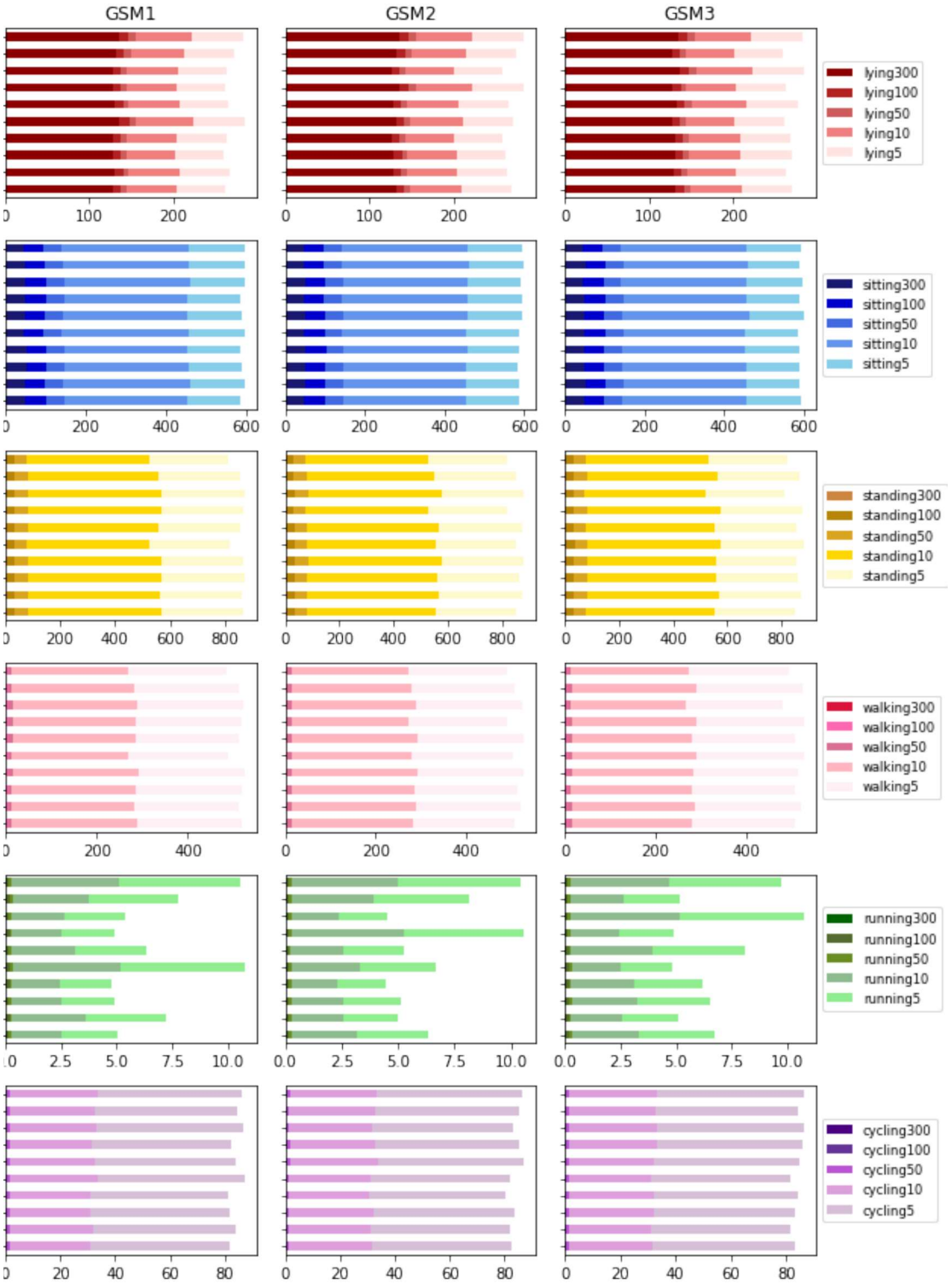
Centroid distribution for 7 clusters



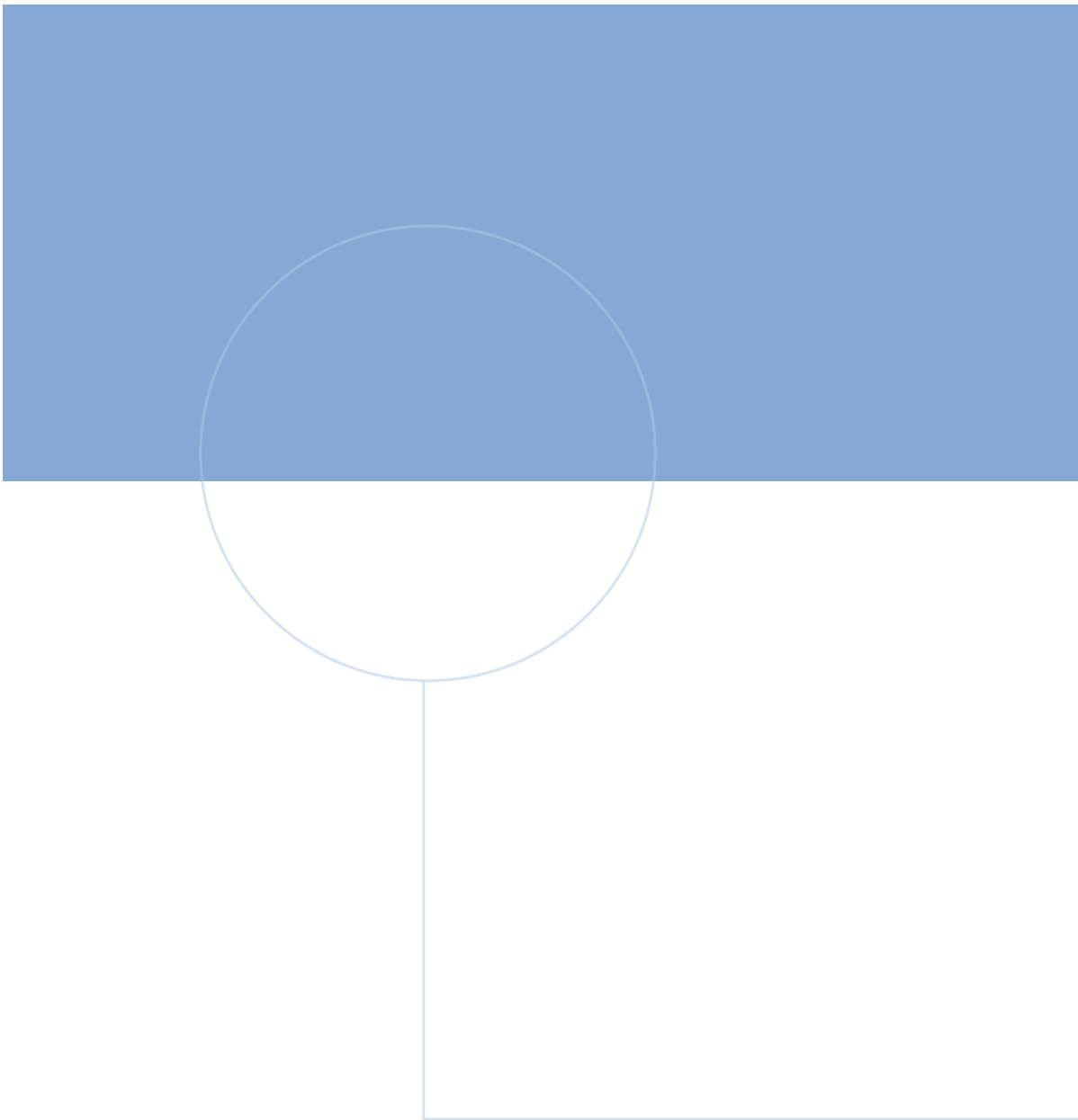
Average distribution for 7 clusters



Centroid distribution for 10 clusters



Average distribution for 10 clusters



NTNU

Norwegian University of
Science and Technology