Internal Report 2014-01 September 1, 2014

# Leiden University

# Computer Science

## A Missing Value Ignoring Approach for Whole Time Series Clustering of Longevity Corebody Temperature Data

Michael de Winter

s1322524

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

**Acknowledgements**

I highly thank all my academic supervisors for all their time, feedback, effort and guidance throughout this work.

And of course my parents for there constant support.

# A Missing Value Ignoring Approach for Whole Time Series Clustering of Longevity Corebody Temperature Data

Michael de Winter
s1322524
LIACS , LUMC, Leiden University

September 1, 2014

**Abstract**

The SwitchBox project is a study that has collected various physiological time series about longevity. This data is collected from the offspring of long living persons and the partners of the offspring who serve as a normal control group. This thesis mines Core Body Temperature data from the SwitchBox project, to hopefully extract useful information regarding longevity.
As such this thesis work will study/explore the time series data mining task of whole time series clustering. Which requires a different approach when compared to static data clustering. However the current whole time series clustering literature assumes that missing value's can be handled by data preprocessing techniques, such as averaging or interpolation. It is feared here that in case of the Core Body Temperature data these approaches would only distort/bias the dissimilarity of the clusters, instead a missing value ignoring approach for whole time series clustering might be a better approach.
This is thought because although every participants has 84 hours of Corebody Temperature data, a pattern of dissimilarity between the groups is assumed to happen every 24 hours. As such 24 hours of missing value's might not affect the clustering process while there are still 60 hour left of data to form similarity or dissimilarity with.
This thesis work proposes with a missing value's ignoring approach to whole time series clustering, and compares it to interpolation while experimenting on synthetic data. Then finally applies it to the Corebody Temperature data.
The missing value ignoring clustering algorithm had a better performance compared to interpolation, it had no problem finding groups in the synthetic data. However the clusters found in Core Body Temperature had no relation at all with the offspring group or the normal control group.

# Contents

# 1 Introduction

In recent years the growth of the amount of data being collected has drastically increased, it is said that we are moving toward the information age. This means that there is an explosion of data, which changings the way of analyzing and doing research in various fields namely in the form of data mining: the act of finding useful information in large data.

With this data growth there has also been a growth of time series data. With the development of a large number of sensors, telemetry devices and other on-line data collection tools, the time series data being collected is increasing rapidly.

Within this context is SwitchBox project which is a study that has collected various physiological time series data about longevity. The data is collected from the offspring of long living persons and the partners of the offspring who serve as a normal control group. This thesis will use whole time series clustering on the Core Body Temperature data from this project as a exploratory analyses, for the research purposes and hopefully explanation of longevity.

The data mining of time series though requires a different approach than to static data mining. Mainly in reducing time series data high dimensionality, finding similarities measures that account for the chronological nature of time series and better indexing methods.

While there has been a lot of research interest in data mining for time series, most of them have assumed that these time series have no missing value's or that the missing data can be handled by preprocessing techniques modeling techniques such as interpolation or regression [13][17] [50]. However in the case of the Core Body Temperature data, which has a lot of missing value's, it is thought that these techniques would only bias and increase the similarity of the data too much, while a missing value ignoring approach might work better.

This is thought because it is assumed a pattern of dissimilarity occurs every 24 hours, while every participants has 84 hours of Corebody Temperature data. If for example 24 hours is missing, this might not affect the clustering process while there are still 60 hour left of data to form similarity or dissimilarity with. This means that a missing value ignoring approach to whole time series cluster might be a better approach.

This thesis work proposes and experiments with a missing value ignoring approach to time series clustering and compares it to interpolation. First on synthetically made data that is similar to the Core Body Temperature in terms of length,form and missing value's but with 2 groups with distinctive dissimilar patterns.

And then applying it to the Corebody Temperature, to hopefully find a dissimilar pattern in Corebody Temperature between off spring of longevity and the partners of the offspring serving as a normal control group.

This introduction is further split into the following sections: Time Series Data Mining Introduction which will explain how time series data mining is different compared to static data mining, Switchbox Study introduction which will further describe the Switchbox study, the research object and lastly the outline of the thesis.

## 1.1 Time Series Data Mining Introduction

The scientific extracting of useful information from large databases is called *data mining*, the process extracts previously unknown and potentially useful information with various methods of artificial intelligence, machine learning, statistics and database system for various fields. It can be defined as a process of applying computational techniques that, with acceptable computational time , produce patterns of models over the data [11].

However a large amount of the data that is being collected is in the form of time series series i.e. data that was collected with repeated measurements of time, as is the case with the longevity data from SwitchBox. More over time series data is characterized by its numerical, high dimensional, large data size, continuous nature and is always considered as a whole instead of an individual field. This requires a different approach when mining for useful information than with static data i.e. data points values that do not change or change neglect able with time, because classical static data mining algorithms assume relatively low dimensionality, time series algorithms on the other hand must handle dimensionalities in the hundreds or thousands. With this high dimensionality not only are there computation time considerations but the meaning of similar to and clusters become unclear in high dimensional space, a occurs known as the "curse of dimensionality" found in [2].

The heart of time series data mining lies in reducing high dimensionality in the form of data representations and the difficulty of finding a similarity measurements between time series based on human perception.

Some of the typical time series related data mining tasks are:

- **Anomaly Detection**: given a normal time series $X$ and a new time series $Y$ find parts in $Y$ which contain surprising/interesting/unexpected occurrences. [34]
- **Prediction**: given a certain time series $X$ with $n$ data points try to predict $n+1$ data points. [6]
- **Clustering**: find natural groups in the time series database with some similarity/dissimilarity measure. [1]
- **Classification**: giving a unlabeled time series $X$ assign it to a certain class. [18]
- **Query by Content/Indexing** giving a certain query time series $X$ and a similarity measure $SM$, find the most similar in the database. [10]

It is used in various scientific fields ranging from medical to meteorology with examples such as: economic forecasting, intrusion detection, budgetary analysis, inventory management, medical treatments, gene expression analysis, electrocardiogram (ECG) analysis, power consumption analysis etc.

Nowadays time-series data mining techniques are mature although abundant and complex but with the development of a large number of sensors, telemetry devices and other on-line data collection tools, the amount of time series data is increasing rapidly, thus learning how to mine time series data is a valuable skill to have.

In this work we will be focusing on clustering of the time series data of longevity data collected during the SwitchBox Study.

## 1.2 SwitchBox Study Introduction

The Switchbox study is a study across different universities which has the overall objective to gain a better understanding of the homeostatic mechanisms to facilitate maintenance of health from early life through to again, it is a follow up study from the Leiden Longevity Study.

It is done 10 years after the Leiden Longevity Study (LLS), LLS is a longitudinal cohort study consisting of 421 long-living families. Inclusion criteria for long living participants are as follows: men must be aged 89 or above and women 91 or above, at least one living brother or sister who fulfills the first criterion, sibling pairs have to have an identical mother and father and finally the parents of the sibling pairs have to be Dutch and Caucasian. From the LLS it already has been shown that the children of long living parents better regulate their blood-sugar levels and metabolism [53][55].

The SwitchBox study has been setup to study and recruit the children or offspring of the LLS participants. It is a more intensive study on neuroendorine output, metabolism and brain function. The study spans across 5 European countries, with the combined goal of trying to gain a better understanding of the homeostatic mechanisms regulating aging. To hopefully discover and recommend healthy ways of reaching a older age.

But as with most studies various missing value's and noise entered the data, the missing value's where sometimes so large that not be handled by conventional techniques.

## 1.3 Research Objective

The research objective of this thesis work is to do a exploratory data analysis of the Core Body Temperature Time Series data collected during the SwitchBox longevity study, in form of a clustering analyses.

To hopefully find a distinction between the offspring of longevity group and the normal group in regards to Core Body Temperature, which can hopefully explain longevity and recommend healthy ways of reaching a older age.

As such different clustering methods and ways to adjust them for time series are analysed.

Because of the nature of the noise and missing value's in the data of SwitchBox, experimentation with and implementation of a missing value ignoring approach within the clustering of whole times series is done. This approach is compared to more traditional techniques in time series data mining namely interpolation.

## 1.4 Outline

First related work will be discussed in Chapter 2, then it is explained which cluster algorithm, distance measure is chosen and how the problems of the Corebody Temperature data missing values is overcome in Chapter 3, a more detailed description is given about the data and the problems during data collection in Chapter 4, with our approach fully explained it is first tested on synthetic generated data in Chapter 5, It is then applied to the Corebody Temperature in Chapter 6 and finally a conclusion and further work is giving in Chapter 7 and Chapter 8 respectively.

# 2 Related Work

The related research fields to this thesis work involve studies concerning data mining tasks for time series, while focusing mainly on: whole time series clustering, data representation,similarity measurements and missing value's.

For each of these subjects there exists a plethora of different methods and approaches which are suggested in literature, here mostly a brief overview will be made.

## 2.1 Basics Of Time Series Data Mining

Time series data mining requires a different approach compared to static data mining, unlike static data, times series values changed with time and are large in dimension and size.

Work that gives a comprehensive overview of the current existing time series data mining are: [13],[17] and [50].

As the authors of [13] state there are 3 basic components are always present in every time series data mining task:

- **Data representation**: A representation technique should derive the notion of shape by reducing the dimensionality of data while retaining it's essential chronological features.
- **Similarity measurement**: This measurement should establish a notion of similarity based on perceptual criteria, allowing recognition or human vision even though they may not be mathematical identical.
- **Indexing method** : How should a massive set of time-series be organized to enable fast querying? In other words, what indexing mechanism should be applied? The indexing technique should provide minimal space consumption and computational complexity.

For this thesis work, because of the relative small data set size, indexing method are not really required since the space consumption and the computational complexity are both acceptable, although it could still have been improved by a indexing methods.

### 2.1.1 Data Representations

Data representation has to done because time series are essentially high-dimensional data, using a algorithm to work on this raw time series would be too computationally expensive. Added benefits are efficient storage, speed up of processing and noise removal. The authors of [13] list 5 requirements for any representation:

1. significant reduction of the data dimensionality.
2. emphasis on fundamental shape characteristics on both local and globals cales.
3. low computational cost for computing the representation.
4. good reconstruction quality from the reduced representation.
5. insensitivity to noise or implicit noise handling.

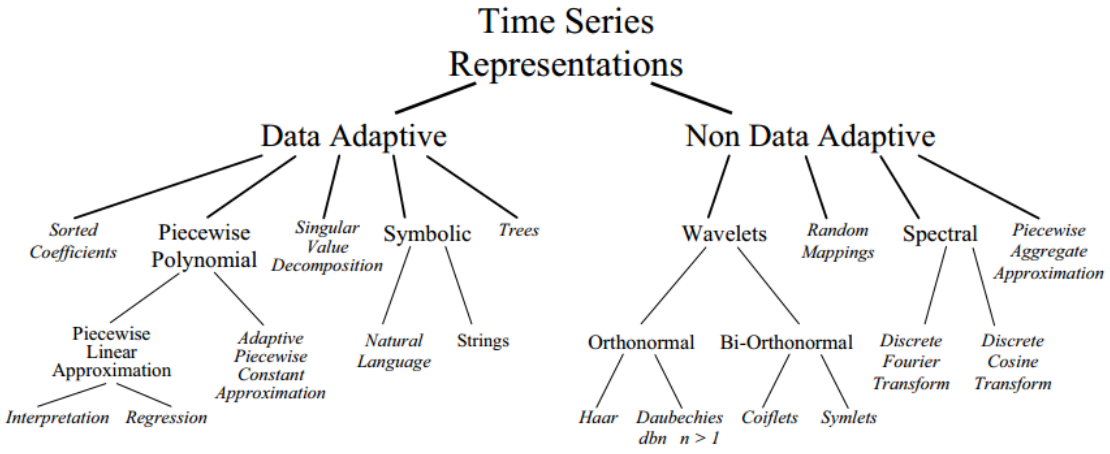
Again a plethora of representation techniques have been researched each of them offering different trade off to the above requirements.

The authors of [36] divide representations into these categories:

**Non Data Adaptive:** which used a fixed number of points to reduce the space in a region. A example would be the Piecewise Aggregate Approximation (PAA), which divides time series into equal-length segments and recording the mean of all the data points that are within that segment [33].

**Data Adaptive:** which change the number of points in a region according to some feature it's exhibits. A example would be the Pieacewise Linear Approximation (PLA), which approximates the time series with straight lines for dimensionality reduction, either by regression [26] or by interpolation [32].

The authors of [50] discuss the various data representations, they display this figure which represent a hierarchy of all the data representations:




### 2.1.2 Similarity Measurements

The next important step is to define a certain notion of similarity between time series when measuring between time series it must account for human abstraction which can see problems as: amplitude, scaling, temporal dependency, noise and outliers.

As the authors of [17] state, there exists difference between whole sequence matching and subsequence matching. In whole sequence matching all of the time series are considered in it's distance function. In subsequence matching different length of time series are matched. The smaller time series has to be placed at every offset within the longer time series. In this thesis work only whole sequence methods are used and discussed.

The authors of [13] propose that a similarity measure provide the following properties:

1. It should provide a recognition of perceptually similar objects, even though they are not mathematically identical

2. It should be consistent with human intuition

3. It should emphasize the most salient features on both local and global scales.

4. A similarity measure should be universal in the sense that it allows to identify or distinguish arbitrary objects, that is, no restrictions on time series are assumed.

5. It should abstract from distortions and be invariant to a set of transformations.

The authors of [13] also classifies similarity measures into four categories:

**Shape based distances:** distances which compare the overall shape of the time series with each other, prominent examples include: Euclidean distance [**?**] and Dynamic Time Warping [5].

**Edit based distances:** distances based on the minimum number of transformations required to transform one series into a other, examples include: Longest Common SubSequence (LCSS) [47].

**Feature-based distances:** approaches which extract features describing aspects of the series that are then compared with any kind of distance function, examples include: Multiresolution Vector Quantized (MVQ)[16].

**Structure based distances:** which finds a higher-level structures in the time series to compare them on a more global scale, examples include: Hidden Markov Models with continuous output values [**?**].

The authors of [13] recommend the choice of a similarity measure depends on the nature of the data as well as application specific properties: if the time series are short and visual perception is meaningful shape-based methods may provide meaningful abstraction, when periodicities are more central and causality is less relevant feature based seems to be more appropriate and finally if time series are long and little knowledge about the structure is available structure based have the advantage of being generic and parameters free. However even with these recommendations it is equally hard to find a appropriate similarity measure as evaluation one, although the tightness of the lower bound appears to be the most appropriate option to evaluate. [35].

## 2.2 Clustering Analysis

Clustering analysis is used to analyse if there is a natural grouping in the data by grouping the data into clusters based on inter cluster similarity and outer cluster dissimilarity.

It is often one of the steps in exploratory data analyses, when there is little prior information available about the data, and the decision makers must make as few assumptions as possible is when clustering is most appropriate.

As the authors of [28] and [14] state, there are a lot of different terminologies , assumptions and data contexts for the clustering analysis in various communities, as such there exits a plethora of different clustering methods. As the authors of [14] state "cluster is in the eye of the beholder, a exact notion of a cluster can not be defined, as such there

is no golden clustering technique". Different clustering algorithms often contain implicit assumptions about the cluster shape, the more information of the data at hand the more likely the true class structure can be assessed [44] [27].

The authors of [7] and [22] classify different clustering algorithm from the data mining perspective into the following methods with general some characteristics and brief description of the general methods below, although it is hard to have crisp categories because in some cases they can overlap:

| **Method** | **General Characteristics** |
|---|---|
| Partitioning methods | Find mutually exclusive clusters of spherical shape.<br>Distance-based.<br>Can use mean or medoid to represent cluster.<br>Used on small to medium size data sets |
| Hierarchical methods | Clustering is a hierarchical decomposition.<br>Cannot correct erroneous merges of splits.<br>May incorporate other techniques like -<br>- microclustering or consider object "linkages". |
| Density-based methods | Can find arbitrarily shaped clusters.<br>Clusters are dense regions in space -<br>-that are separated by low-density regions.<br>Each point must have a minimum number-<br>-of points within it's "Neighbourhood".<br>May filter out outliers. |
| Grid-based Methods | Use a multiresolution grid data structure<br>Fast Processing time |
| Model-based | Tries to best fit distributions or models on the data. |
| **Other (less general) Methods:** | |
| Constraint Based | Incorporate used defined constraints |
| Graph Partitioning | Based on cutting edges from graphs to from clusters |
| Supervised Learning Based | Update clusters based on current assignments<br>serving as the target attribute values supervising the learning. |
| Machine Learning Based | Using gradient descent to optimize objective functions. |
| Evolutionary Based | GA's can be used in the optimisation of the objective function<br>of clusters, although with high computation costs. |

### 2.2.1 Partition methods

Given a set of $n$ unlabeled data, partitioning methods constructs $k$ partitions of the data, where each partition represents a cluster containing at least one object and $k \leq n$. Basic partitioning methods are usually use exclusive cluster separation, in which each object belongs to exactly one cluster but it can be relaxed with fuzzy cluster separation, in which objects can belongs to more cluster with a certain degree.

Two most used exclusive cluster separation partition algorithms are the k-means algorithm [42], where each cluster is defined by a mean, and the k-medoids algorithm [31], where each cluster is defined by the most central object in the cluster.For fuzzy clustering separation there is the fuzzy c-means [8] and fuzzy c-mediods algorithm [38].

These heuristic algorithms use a iterative relocation technique that improve the clus-

tering by moving objects from one group to another in respect to the objects distance to the cluster. The heuristic algorithms are good for find spherical-shaped cluster in small to medium size databases. To find cluster which are non-spherical or other more complex shapes other methods such as density-based methods are needed. Most of the genetic clustering algorithms implement the spirit of partition methods.

### 2.2.2 Hierarchical methods

Hierarchical method algorithms work by grouping data into a tree of clusters, and can be classified as either agglomerative or divisive based on how the hierarchy is formed. The agglomerative approach start by placing objects in it's own cluster and then merging cluster into larger and larger clusters, until a certain termination condition or when there is one large cluster. The divisive approach is exactly the opposite.

Hierarchical method algorithms can be distance,density or continuity based, various extensions also consider clustering in subspace as well.

The downside is to Hierarchical method algorithms is that once a step is done (merge or split), it can never be undone.

There is trend to integrate with other clustering techniques, Chameleon [30] and CURE [21] do a analysis of the object linkages at each hierarchical partition, where as BRICH [59] uses iterative relocation to refine results in hierarchical agglomeration.

### 2.2.3 Density-based methods

The idea of density based methods algorithms is to grow a cluster as long as the number of data points in the number neighbourhood exceeds some threshold, rather then producing a cluster explicitly. As such density methods can find more shapes then only spherical shaped clusters and can filter out outliers, such as in DBSCAN [15].

### 2.2.4 Grid-based methods

Grid based methods algorithms place the object into a finite number of cells that form a grid structure. All the clustering is performed on the grid structure, the quantized space, this has the advantage of having a post processing time, STING [57] is a example.

### 2.2.5 Model-based methods

Model based method algorithm see clusters as different distributions, the task here is calculated the correct variable for each distributions (i.e. mean, standard deviation,..etc), a example is the EM algorithm [45].

## 2.3 Time Series Clustering

The authors of [13] divide time series clustering into 2 sub tasks whole time series clustering and subsequence clustering.

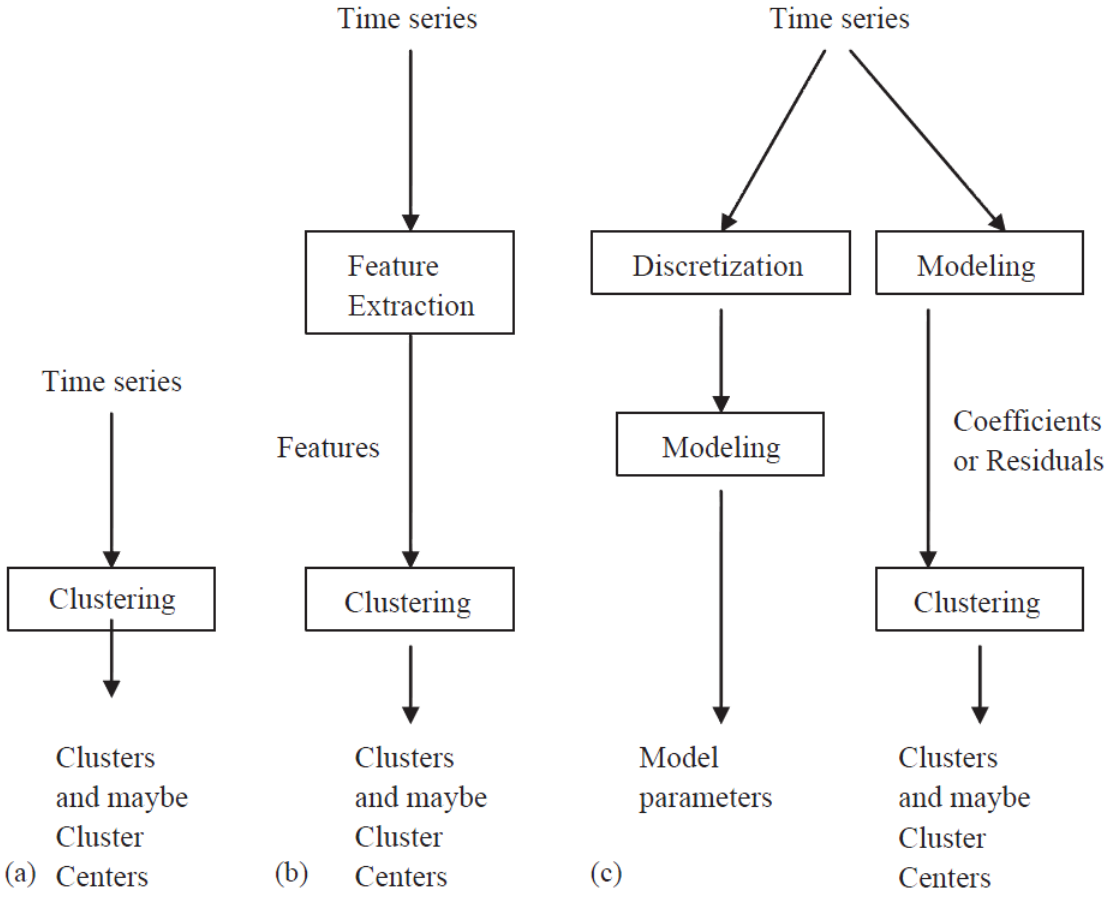
With Whole time series clustering, clustering is applied to each complete time series in the data set. The objective is to regroup the entire time series into clusters so that the time series that are the most similar as possible to each other are placed within the same cluster. With sub sequence clustering, clusters are created by extracting subsequences from a single or multiple longer time series.

For this thesis work will be focusing on whole time series clustering.

The author of [41] give a survey of the current research into time series clustering. Three of the 5 major categories of clustering methods, spoken of in the above section. have been utilized directly with or modified for time series, namely: partitioning methods, hierarchical methods and model-based methods.

They divide 3 approaches for clustering time series:

Figure 1: The authors of [41] divide time series clustering into 3 approaches: (a) Raw Based Approach, (b) Feature Based Approach, (c) Model Based Approach




**Raw Based Approach:** These approaches work directly with the raw time series and the major modification lies in replacing the distance and/or similarity measures for appropriate ones for time series.

The authors of [20] applied the fuzzy c-means algorithm to raw univariate functional MRI time series data, in order to provide functional mappings of human brain activities on the application of a stimulus, with different similarity measures.

**Feature Based Approach:** These approaches try to convert the time series into a form of a lower dimension of static data, so they can be directly used in conventional static data clustering algorithms. However most of the feature based approaches are usually application dependent.

The authors of [52] apply hierarchical clustering on flow velocity wind data, spectra was used as a feature obtained by either the original time series with the mean adjusted to zero or by principal component analysis.

**Model Based Approach:** These approaches consider that the time serie was produced a model or a mixture of different probability distributions. The time series are convert into a form of model parameters,either for direct use as cluster or used in conventional static data clustering algorithms.

A example would be [43], which used an agglomerative hierarchical clustering procedure on the $p$-value 's of a chi square test statistic applied to every pair of time series.

## 2.4 Approaches to Missing Value's

Since the Core Body Temperature data has a lot of missing value's, this thesis work also looks at literature about how to deal with missing value's.

The authors of [23] give various methods to handle missing value's for static data mining:

1. Ignoring the tuple
2. Fill the missing value's manually
3. Use a global constant (e.g.,"Unknown") to fill in the missing value.
4. Use a measure of central tendency for the attribute (e.g.,the mean or the median)
5. The use of the mean or the median for all samples that belong to the same class as the participant
6. The use of the most probable value to fill in the missing value with methods such as regression, inference-based tools using Bayesian formalism or decision based induction. The most popular method.

Although they highlight that methods 3 to 6 bias the data by filling in value's that may not be correct.

In the field of statistics the process of replacing missing data with subsisted value's is called imputation. The authors of [40] highlight different imputation methods as well as other methods for missing value's:

1. **Most Common Attribute Value:** The value of the attribute that occurs most often is selected to be the the value for all the missing value's of the attribute.
2. **Concept Most Common Attribute Value:** A restriction of the Most Common Attribute Value to the concept. Now the missing value's are replaced with the most common value within the class.
3. **Hot deck imputation:** Identify the most similar case to the missing value and substitute it with it.
4. **Cold deck imputation:** Filling in a missing value's with a value's other than the missing value

5. **Regression or classification methods:** They highlight several regression and Machine Learning Methods.

In the field of machine learning the authors of [19] compare nine different approach for missing value's for classification, from these methods the C4.5 algorithm and the missing value ignoring approach performed the best.

In the field of econometrics there exists several prediction models, one of the most prominent being ARIMA [9]. These methods can also be used for predicting missing value's, examples of this method being [25].

The authors of [29] compare different regression methods for time series data mining.

# 3 Clustering of Corebody Temperature Time Series Data

## 3.1 Clustering Analysis Introduction

With clustering analysis or just simply called clustering, the objective is to find a certain natural structure of unlabeled data by putting data into homogeneous subsets based on a within subset similarity minimization and between subset similarity maximization. Each of these subset is a called a cluster, objects in a cluster are similar within to one another and dissimilar to objects in other clusters.

It is a form of data exploration to find groups within data previously unknown.

Clustering is also sometime called automatic classification, to high light the critical difference that clustering automatically finds the groupings, the distinct advantage of clustering. Within machine learning it is called unsupervised learning because information but the classes is not available and it has to learning from observation in contrast to supervised learning where class information is available and has to learn from example.

Although what precisely in a cluster cannot be defined, and as such there are many different clustering produces and algorithms available, there is no "correct" clustering algorithm for every situation [14].

This in order to find groups in the Core Body Temperature which could be explained by from which class the data came from, either offspring from longevity or the control group. For example if a certain found group during clustering only contains data from participants of the off spring

If there groups are somehow related to offspring from longevity or the normal control group, a better understanding of longevity could be obtained.

For this study there we also need to tackle our specific problem of missing data which have occurred during the collection of the data during the switchbox study, more of this will be discussed below.

First the k-medoids algorithm will be explained and why it is chosen to handle missing values, next it is talked about how to adjust Clustering for time series, then in order to make clustering happen for time series we will discuss our picked similarity measure and finally solutions are provided which handle our specific problem which will handle.

## 3.2 K-medoids Clustering Algorithm

The k-medoids clustering algorithm has been chosen for this thesis work because of: the easy adaptability for time series,computational speed,our approach to handle missing values, it's robustness due to it's outlier insensitive characteristic and because our domain experts think that the groups can be found by a median, it was first spoken of by [31].

The objective is to minimize the objective function of the total distance between all patterns from there respective cluster centers, with the number of cluster centers being chosen by the user. The initial cluster centers can be randomly assigned or picked in a manner that is consistent with domain knowledge , then with a iterative process the data points are assigned to the nearest cluster and the cluster are recalculated based on the data points assigned to the clusters. This is repeated until convergence is reached or after a certain number of iterations[24].

It is a partition method based on the k-means algorithm [42], but instead of a mean, a median is taking. The k-mean algorithm is sensitive to outliers which can dramatically distort the mean value of a clusters and thus the assignment of other objects to the cluster.

Instead of using a mean as reference point for a cluster, actual object is chosen to represent a cluster, each object is assigned to the cluster of which the representative object is the most similar.

In this thesis work it is thought that some outliers and noise are still present in the data, because standard data preprocessing techniques do not work on the Corebody Temperature instead filters are used to remove of not physiological possible value's however this does not guaranty noise removal, more on this in section 4, and thus the more robust k-medoids is chosen over the k-means algorithm.

The downside is it's complexity $O(k(n-k^2)$, where $k$ is the number of clusters and $n$ is the number of data points, and the number of cluster centers that needs to chosen by the user compared to other approaches. However calculation for this specific data can be done with acceptable time and since we want to see the difference between 2 groups of offspring of longevity and a control group, 2 cluster centers can be chosen.

Below is the K-medoids algorithm is displayed:

Table 1: **K-mediods algorithm**

| | |
|---|---|
| 1. | Initialize $k$ clusters either by randomization or picked with domain knowledge. |
| 2. | Assign each data point $n$ to the cluster based on calculating the distance between. each cluster and choosing the nearest cluster. |
| 3. | Calculate the new cluster taking a median of all data points assigned to the cluster in the previous step |
| 4. | if there is no changes in data points assigning to new cluster or until a set iterations have been made stop else repeat 2 and 3 |

In this thesis work we do not have the problem of guessing the number of cluster centers, there are 2 groups and thus 2 cluster centers are used.

The 2 cluster center are initialized by respectively taking the highest temperature and lowest temperature for each data points from all the time series. This is because the domain experts think that the distinctive discriminating pattern is based on higher and lower temperature during certain parts of the day.

## 3.3 Adjusting Clustering for Time Series: Raw Based Approach

Most clustering algorithm have been developed for static data, clustering programs developed as an independent or part of a large suite of data mining software only work with static data, but more and more of the current data is in the form of time series. But luckily time series research is becoming more mature.

As the authors of [41] highlight there are 3 different approaches to time series clustering namely: Raw Based Approach, Feature Based Approach and Model Based Approach.

Raw Based Approaches deal directly with the raw time series, major modification lies in replacing the distance/similarity measure from static data with a appropriate one for

time series data, Feature Based Approaches convert raw time series into lower dimensional static data to be used directly for standard static data clustering algorithms and Model Based Approaches assume that time series are generated by some kind of model or distribution.

In this thesis work a raw based approach will be used and thus appropriate similarity measure for time series will have to be chosen. However still a data representation will be used to reduce the dimensionalty from seconds to minutes.

## 3.4 Similarity Measure Used for Time Series Clustering

In order for k-medoids clustering to work with time series, a way is needed to measure the similarity between time series while accounting for there chronological dependency, time series similarity measures solve this, as already discussed above.

The authors of [12] has showed that despite all various/complex proposed similarity measures the 30 years old Dynamic Time Warping (DTW) usually performs better. The authors of [13] also recommends shaped similarity with short time series and visual perception relevance. So in this study is decided shaped based similarity measures will be used, specifically the Euclidean Distance and DTW. In short Euclidean Distance is, within a time series, one datapoint to one datapoint measure while DTW measures by warping to more than one point.

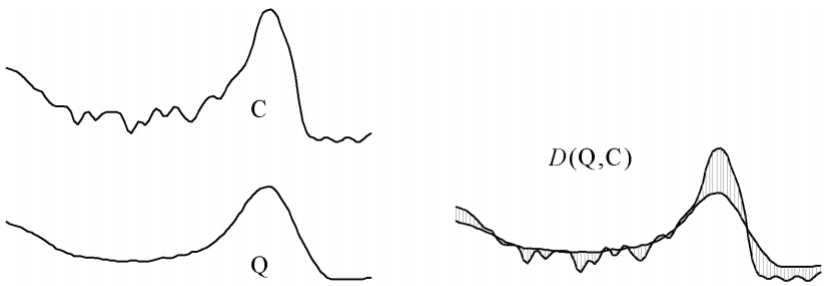
Two similarity measures are used because although the DTW usually performs better because of it's robustness, the fear is that the warping might lead to errors because a distortion in the time could actually be the distinctive factor between the groups and if errors can occur when assigning objects to a cluster because of warping across over a large area of missing value's, so experiments are done with both similarity measures.

### 3.4.1 Euclidean Distance

Euclidean distance is the most simplest and widely used similarity measure as show in [2],[**?**] and [**?**]. Calculated by this formula , where $C$ and $Q$ are two equal length $n$ time series:

$$D(Q,C) = \sqrt{\sum_{i=1}^{n}(Q_i - C_i)^2} \tag{1}$$

Figure 2: The authors of [50] display this figure for the intuition behind Euclidean distance



Despite it's simplistic nature, is still a competitive similarity measure compared to other more complex similarity measures, especially if the size of the database is relatively large [35]. However the downside of the Euclidean distance is that it compares the $i$-th point of one time series to the $i$-th point of another time series i,e, the mapping of points between the two time series is fixed, this means that it is not very robust and can be very sensitive to scaling,outliers, noise and misalignment in time; which could have occurred during data collecting, shown by varies studies [1],[4],[5]and [51].

### 3.4.2 Dynamic Time Warping

Originally used as a speech recognition tool for words which could recognize words spoken despite wide variations in timing and pronunciation, the authors of [5] successfully introduced Dynamic Time Warping (DTW) to the time series data mining community. It overcomes the problems the Euclidean distance has by allowing the comparing of one to many points, with the construction of a warp path which allows for both global and local shifting of the time dimension.

This can decrease the error rate of miss matched similarities, although the authors of [12] have shown that with a very large data set this improvement disappears. . The time warping is stated as follows: Given two time series $X$ and $Y$, of length $m$ and $n$

$$X = x_1, ....., x_m \tag{2}$$

$$Y = y_1, ....., y_n \tag{3}$$

To align them we construct an $m$-by-$n$ matrix where each element in the matrix $(i, j)$ is the Euclidean distance $d(x_i, y_j)$ between the two points $x_i$ and $y_i$. A warping path $W$ is then constructed which is a set of matrix elements with mapping between $X$ and $Y$:

$$W = w_1, ....., w_K \qquad max(m, m) \leq K < m + n - 1 \tag{4}$$

The warping path $W$ is subjected to various constrains: the warping path has to stop in diagonally opposite corner cells of the matrix $w_k = (1, 1)$ and $w_K = (m, n)$, the warping path must only have steps in adjacent and diagonal cells $W_k = (a, b)$ then $W_{k-1} = (a', b')$ where $a - a' \leq 1$ $b - b' \leq 1$ and lastly the points must be monotonically spaced in time $W_k = (a, b)$ then $W_{k-1} = (a', b')$ where $a - a' \geq 0$ $b - b' \geq 0$.

Then with these conditions were are interested in the warp path with the shortest distance:

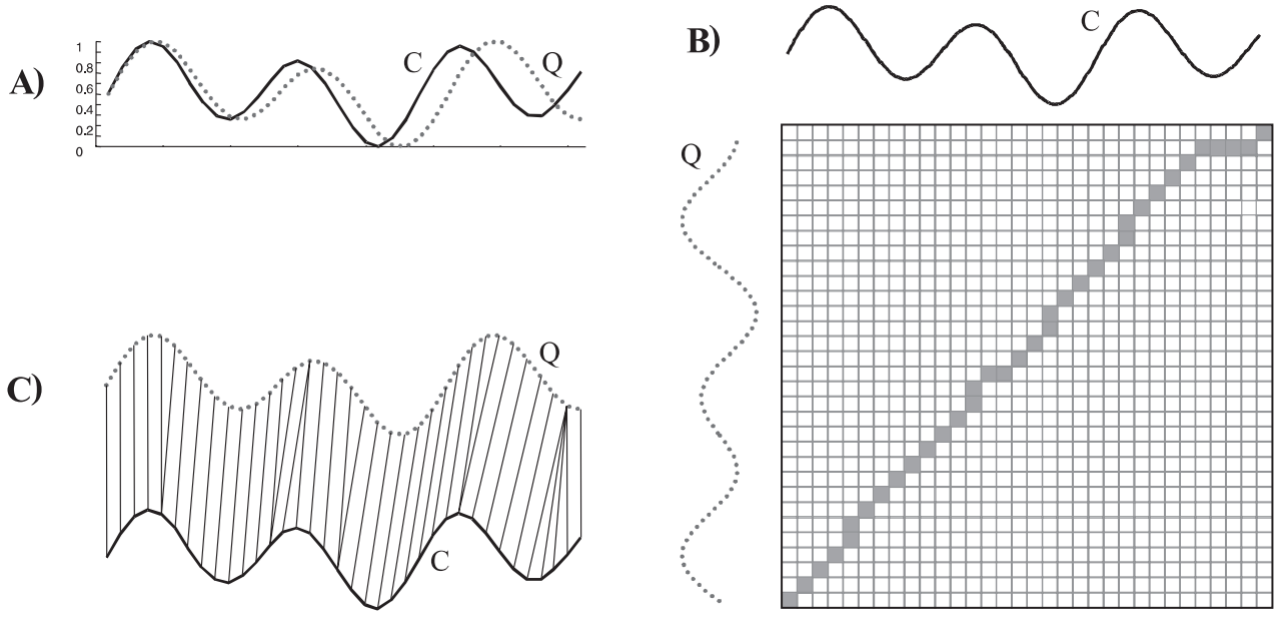
$$DTW(X, Y) = min(\sqrt{\sum_{k=1}^{K} w_k}) \tag{5}$$

With dynamic programming the minimization of the distance is calculated and so the optimal warp path is found:

$$d_{cum}(i, j) = d(y_i, x_j) + min(d_{cum}(i - 1, j - 1), d_{cum}(i - 1, j), d_{cum}(i, j - 1)) \tag{6}$$

The visual representation of the workings of dynamic time warping:

Figure 3: [37] display this figure to visualize Dynamic Time Warping




In Figure 2 **A** represents time series C and Q which are slightly out of sync in the x-axis, **B** is the warping path matrix which finds the optimal path between time series C and Q, shown in grey and **C** is the resulting better alignment.

The downside of DTW is it's computation complexity $O(n^2)$ on the current hardware setup it's takes 6 hours for the k medoids algorithm to convergence with DTW. Although there exists approaches which speed up DTW [37], these are left outside the scope of this work.

This is a rather brief explaining about Dynamic Time warping for a more detailed description is giving in [5],[39] and [48]

## 3.5 Data Representation Used for Time Series Clustering

A data representation should be used on raw data for computational reduction, noise removal and also in our case to obtain a even sampling rate. Here we have chosen for the piece wise aggregate approximation (PAA), because of it's easy of implementation and as the authors of [12] have shown there seems to be no difference in tightness of lower bound between PAA and other more complex data representations

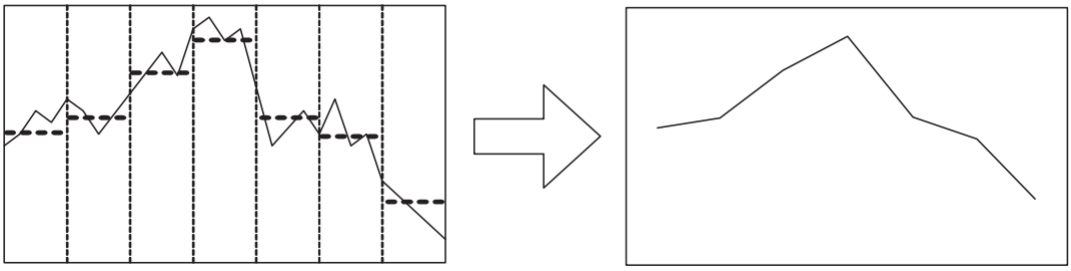
The PAA takes a mean value for a segment to compress the corresponding set of data points i.e. for each interval of the time series data it takes a mean.

If $T = (t1, ...., t_m)$ is a time series then the compressed time series $T_c = (t_1c..., t_nc)$ is made by

$$T_kc = \frac{\sum_{i=y_k}^{x_k} T_i}{x_k - y_k + 1} \tag{7}$$

To visual illustrate how PAA works:

Figure 4: The authors of [17] displays this figure for the visual interpretation of PAA



The interval was set to be 5 minutes and with 3 and a half days i.e. 84 hours of data,this leaves us with 1008 data points in the time series for each participant.

The added problem of the Corebody Temperature is that not all the participants have a the sampling rate, ranging form 15 to 30 seconds. By taking intervals of 5 minutes, this data representation also solves this problem.

## 3.6 Adjusting k-medoids Algorithm for missing values in Time Series

As is the case with most data collection various errors/constrictions caused noise,outliers and missing value's to entered the Core Body Temperature data.

Contributing to the large amount of missing value's is that standard techniques for noise/outlier removal would only bias the data even more, instead filters were created which remove data points which are physiologically not possible or measurements in intervals of which it is certain that it would produce erroneous data, a more detailed explanation in the subsection data pre processing of section 4.

In figure 5 the grey area represents the amount of missing value's from a random participant with the use of filters. As the authors state in [23], realistic options for handling missing value's are using a mean of median to fill the missing value from the same class as the given time series or the use of a model to fill the missing value's such as: regression, inference based tools, decision tree induction, interpolation, ARMA...etc.

However with respect to the mean or median value's filling, it might not be correct, since we are dealing with the class offspring from longevity it is not certain if all of the participants of that class inherited that gene and thus if this approach is used it could generalize and bias the data too much for this specific class.

With respect the use of model to fill the missing value's, it is thought here that the intervals of missing value's are sometimes just too large to model, which would only bias the data. Around 50 % of the participants have missing in there data which are longer than 6 hours, as such large intervals could be filled with generalized and biased data, which could interfere with the convergence of the clustering algorithm i.e. assigning objects to the wrong cluster. Also there exists the problem as with the means and medians, if the inputs of the models are based on classes this could also generalize too much.

A better alternative and easier approach would be the use of a missing value ignoring approach for this specific Core Body Temperature data, which could work because of the following reason:

- A distinctive pattern would realistically occur in a certain time frame equal to or smaller than 24 hour, since we have 84 hours of data for each participant, 6 hours or more missing value's can possible not effect the distinctive discriminating pattern between the groups. I.e. there is large chance that the distinctive discriminating pattern between the 2 groups could not have occurred in the interval that has missing value's or could this have occur in the interval not affected by the missing value's.

This work proposes and experiments with the following fairly straight forward theorem to solve this issue:

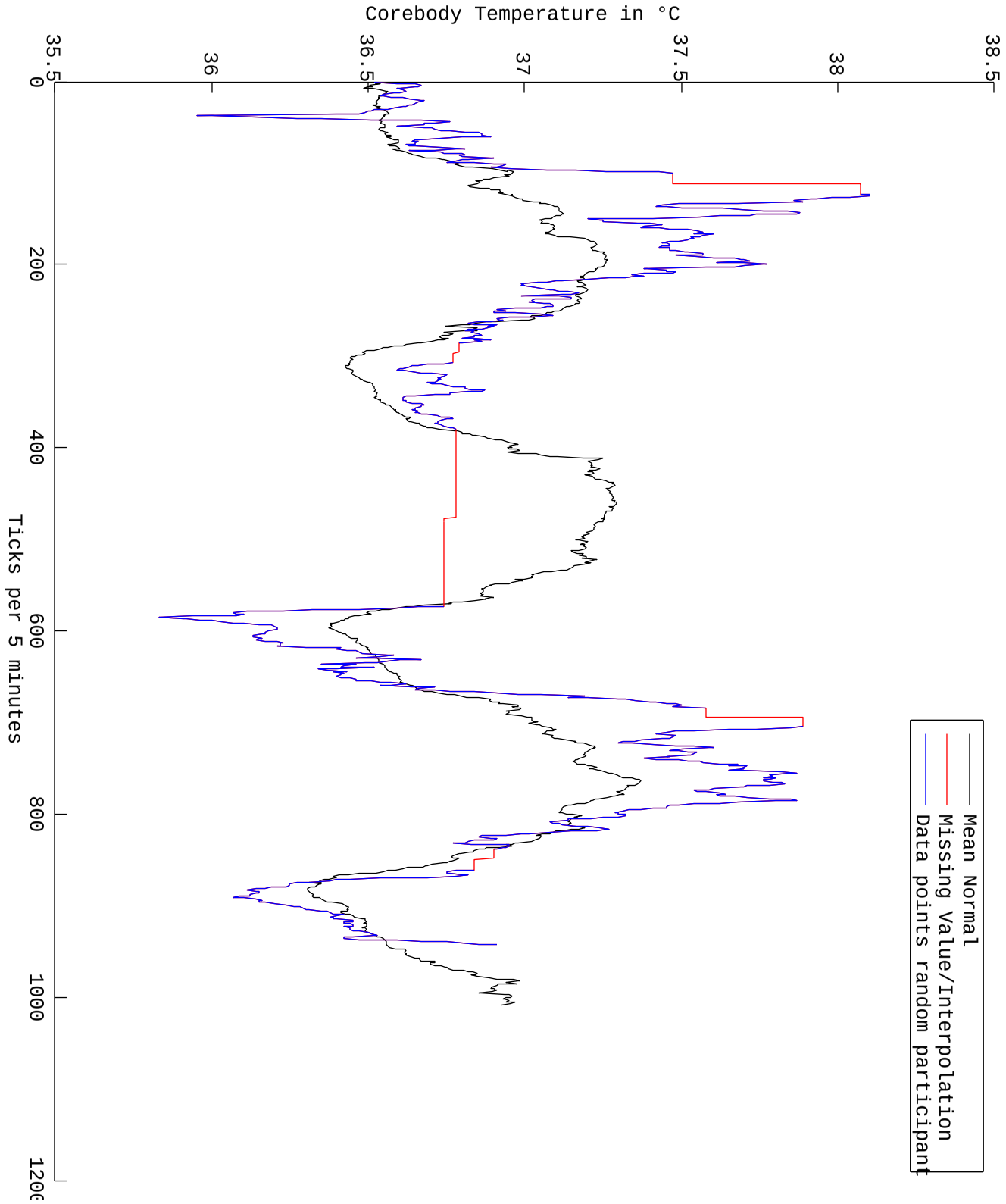
> Let a certain time series cluster $C_1$ with $C_1 = c1_1, ..., c1_{1008}$ in the k-medoids algorithm only be recalculated and similarity measured from the data points a certain participant has. For instance $P_1$ which only has the following points $P_1 = [p1_1, ..., p1_{100}][p1_{200}, ...p1_{350}]$, will only base it's similarity measure and recalculation of clusters on $C_1 = [c1_1, ..., c1_{100}][c1_{200}, ...c1_{350}]$

Figures 5 and 6 display: a random participant of the normal control group, the mean of the normal control group and interpolation of the missing value's. Figure 5 uses the interpolation method of nearest neighbor while Figure 6 uses the piecewise cubic method of interpolation. It can be seen in both these figures that interpolation biases the data.

While the missing value ignoring approach will ignore the red lines and only use the blue lines in it's calculations.

Various experiments will be done in order to see despite the missing value's in the data if this approach can still detect natural synthetic made groups of data with missing value's.

Figure 5: Missing Value's and Interpolation with the nearest neighbor method of a random participant



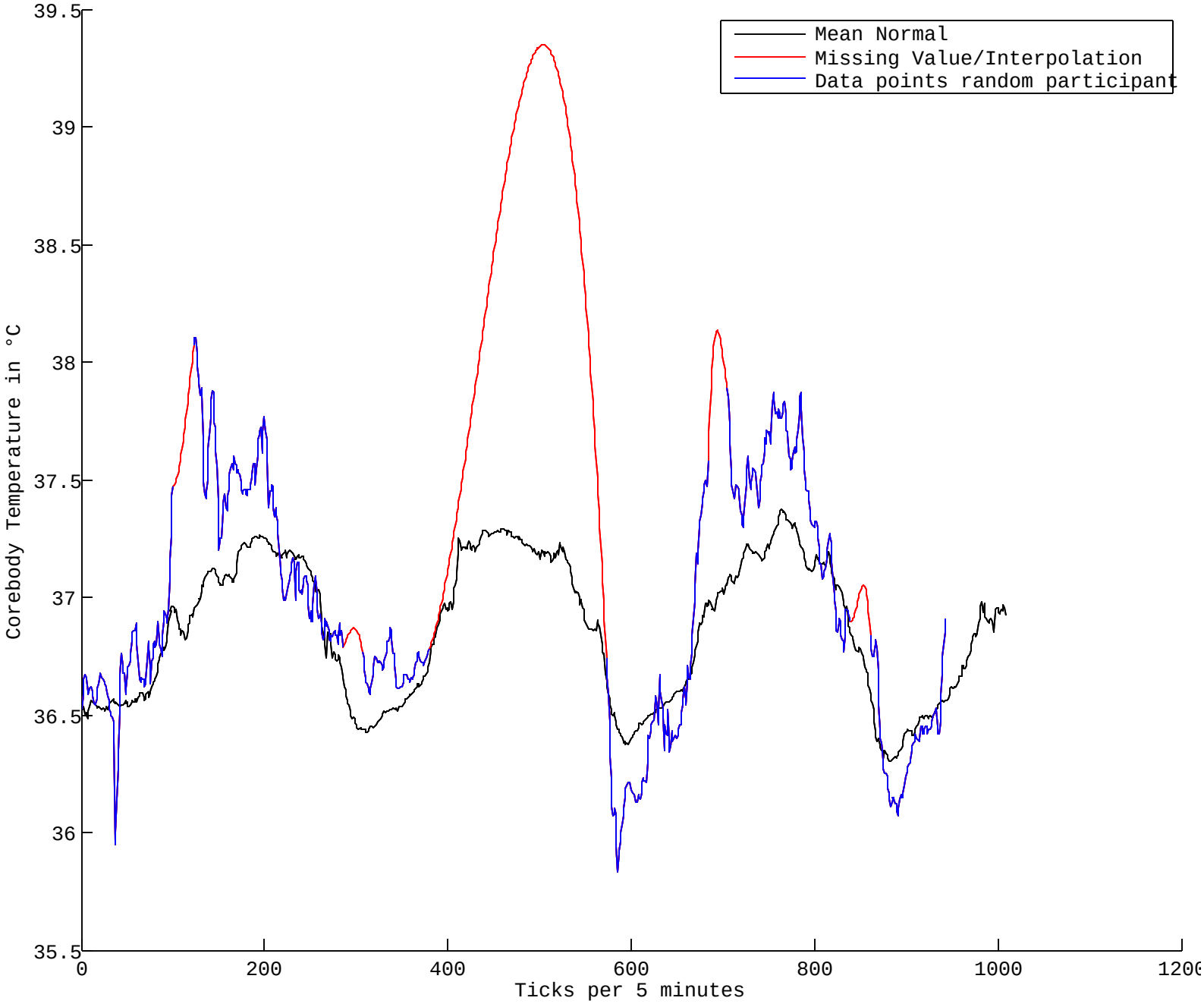



Figure 6: Corebody Temperature for different groups without any data preprocessing.

# 4 SwitchBox Data Description and Preprocessing

## 4.1 Data Description

In this thesis work Core Body Temperature is used which is one of the various longevity data variables collected during the SwitchBox Study, which recruited 70 offspring from the participants from the Leiden Longevity Study and 62 partners of these offspring as a control group, which leaves us with Core Body Temperature from a total of 132 participants.

The domain experts here at the LUMC predicted that the Core Body Temperature would have the most distinction between the groups, this why the Core Body Temperature will be look at first.

Additionally the participants included had to go through a certain criteria:

- aged 55-77 years.
- Have a normal BMI: $19kg/m^2 < BMI < 33kg/m^2$.
- fasting plasma glucose $<= 7mmol/L$.
- no significant chronic disease.
- no hormone therapy.
- smoking or alcohol addiction.
- no 3kg weight gain or lose,with the last 3 months.

Each included participant had to wear an Equivital monitor(Equivital EQ02 SEM,Hidalgo, UK) for 5 constructive days which measured various physiological parameters:heart rate, body position, movement, respiration rate, skin temperature and finally Core Body Temperature.

In order to measure the Core Body Temperature the participant had to shallow three temperature capsules (Capsule REF 500-0100-2, Respironics Inc., Murryville, PA, USA) for three consecutive days, around dinner time. The pills measured temperature data per 250 milliseconds and were connected to a Equivital monitor device by radio emission, with a maximum range of one meter. After a few days the capsules exited the participants body by stool.

Although the other variables were measured for 5 days the capsules could only make reliable Core Body Temperature data for approximately 84 hours or 3 and a half days.

But even with this approach still missing values and noise entered the data during these 3 and half days of data collection for the following reasons:

- It is estimated that it takes 5 hour before each pill is fully incorporated in the participants. For this the reason the first 5 hours are ignored from all the participants and the consecutive pills were instructed to be shallowed 5 hours before estimated time the other pill would exit the system. However these 5 hours are a estimated it could sometimes take longer for these pills to enter the system and lead to noise and missing value's.

- Sometimes the pills could early exit the participant system via stool which limited data collected. It can sometimes happen that the pills exited the system too soon, leading to missing value's.
- The monitors had to recharge for 15 min, which leaves missing value's in the data.
- Although factory setting were limited to 32 °C and 42 °, still measures outside of this range were being measured.

Lucky there were other reasons for inconsistency in the data but these could be fixed:

- The monitor could sometimes incorrectly measure a pill from the partner. However each pill had a serial number associated to a certain participant, so it was a matter of setting each measurement to the correct participant.
- Incorrect date format, months and days could sometimes be reversed.
- The equivital device sometimes had the wrong time zone, the correct time zone had to be adjusted.

The following graphs on the next page displays the data only with the removal of the first 5 hours of each pill, but with the data representation PAA with a interval of 5 minutes:

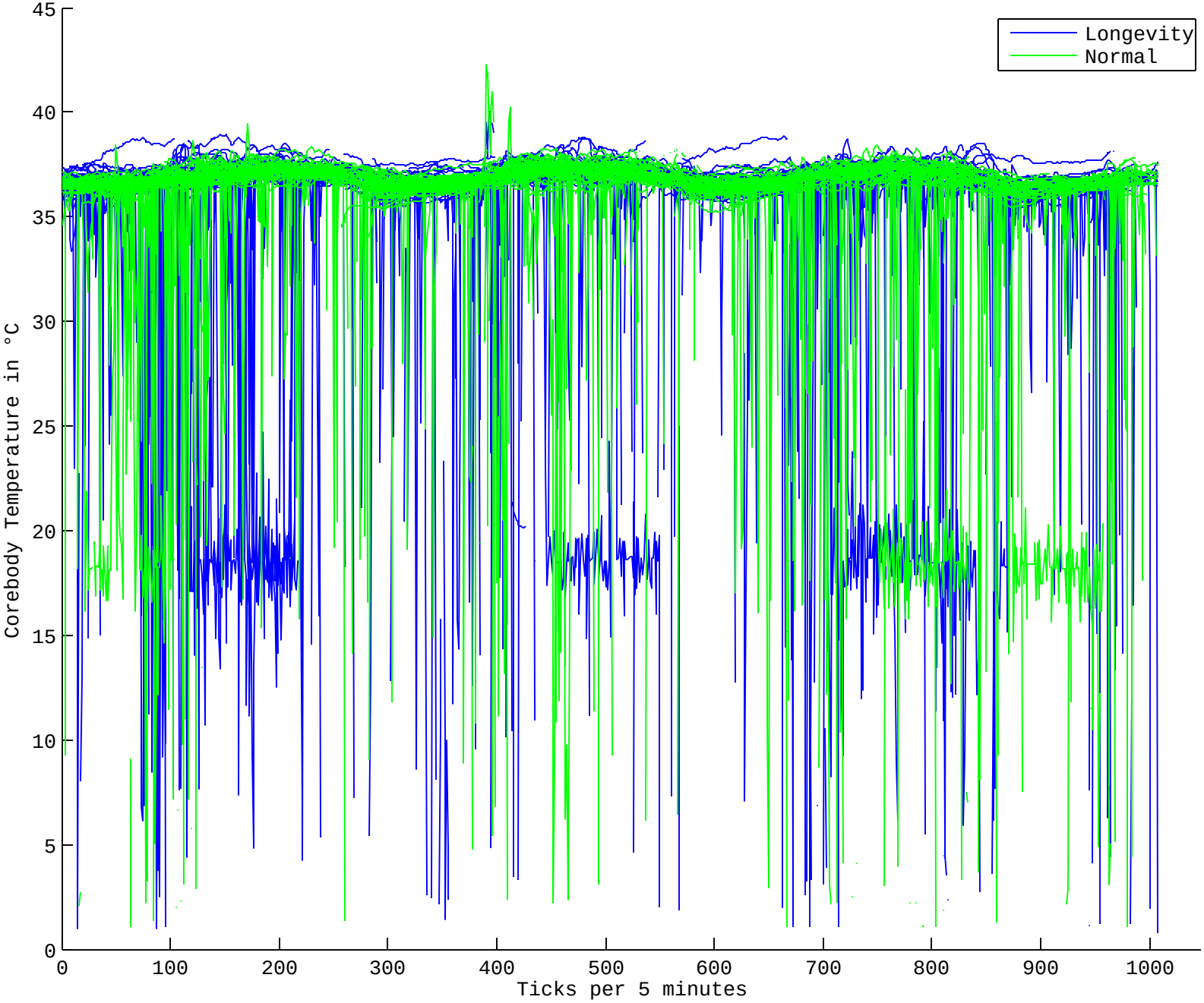



Figure 7: Corebody Temperature for different groups without any data preprocessing.

## 4.2 Data Preprocessing

As with most studies, various noise/erroneous data was picked up by the measurement devices, the same is for the SwitchBox Study.

The human body can physiologically only reach core body temperatures between 32 °C and 42 °, yet as you can see in Figure 6, a lot noise entered the data by having a lot data points under 32 °C, for the reasons discussed in the section above.

Figure 6 also displays that with use of noise/outliers techniques such as a convolution analysis, would require a very strong convolution, which would only help to normalize the data too much to reduce the distinction between the 2 groups of longevity and control.

As such noise outliers removal techniques are strongly believed to be not helpful in the convergence on the clustering algorithm, because of this instead a filter approach is used to filter out physiological not physiologically possible data and data in intervals of which it is certain that it would produce erroneous data.

The filters to remove data are as follows:

1. Temperature data which is below 32 °C and above 42 °, are removed from the data
2. Within 1 minute the temperature can physiological never raise or drop more than 0,3 °C. Although this occurred very infrequent only in 1 % of the data.
3. When the temperature monitors were charging, generated data during this time were removed.
4. The first 6 hours of each pill is removed because this the amount of time it takes until pills are fully incorporated

However with the use of these filters around 50 % of the participants have missing in there data which are longer than 6 hours, but collectively the participants always have data in the 3 and a half days, more specific for each data point in the 3 and a half days of data collection there are always 20 or more from the 132 participants which have data there.

These filters also do not guaranty noise removal since even though the measurements are physiological possible, hence the use of k-medoids.

Figure 8 on the next page displays the means of the 2 group after applying the filters, with the data representation PAA with a interval of 5 minutes.

As you can see in Figure 8 look very good in terms of there physiological nature, however the means of the 2 groups are not very significant apart from each other, a clustering analysis is still preformed because it might not have been so that the off spring of the longevity people inherited the longevity gene, thus there could be a sub group which affects the mean.

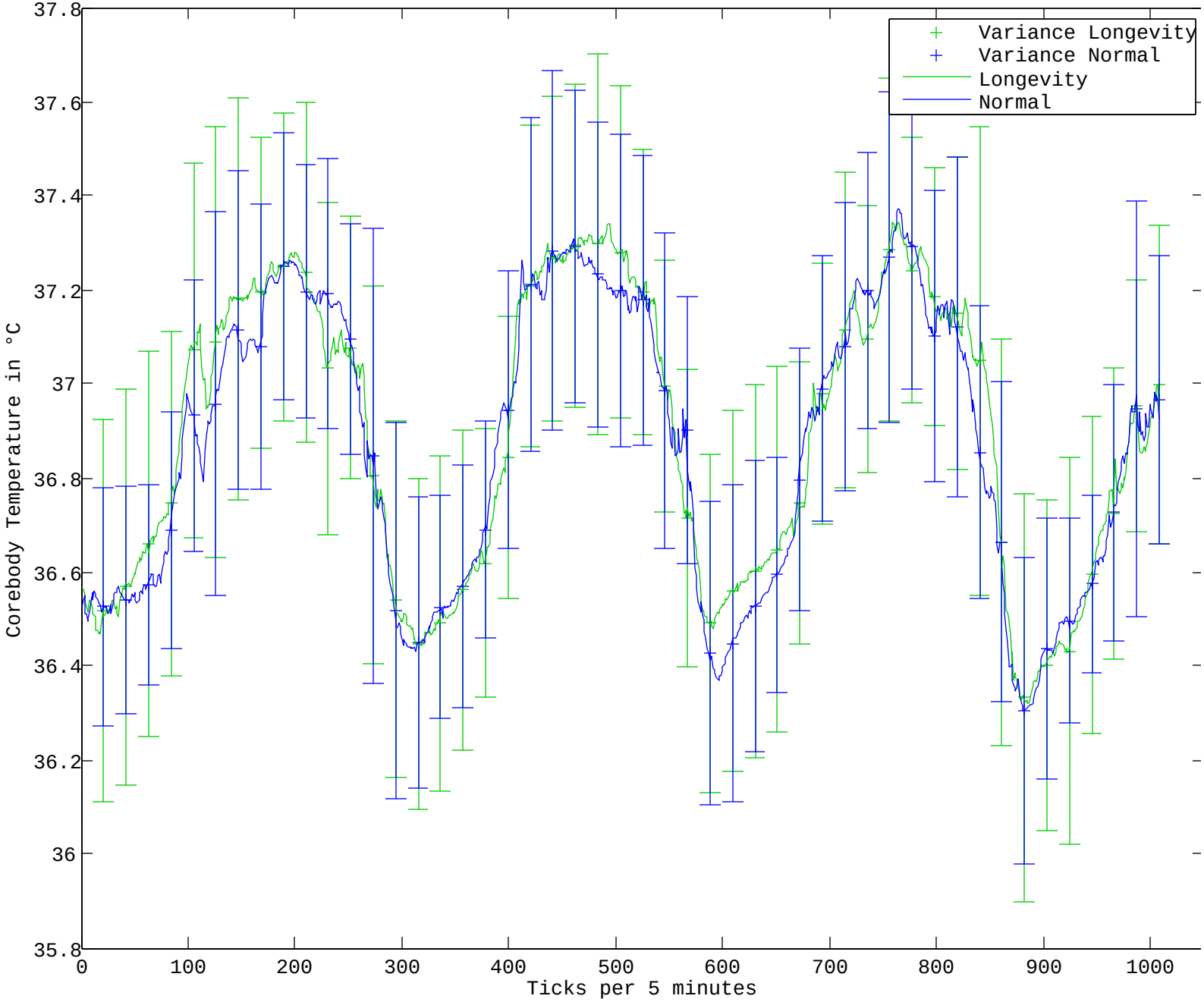



Figure 8: Mean Core Body Temperature for the two classes

## 4.3 Descriptive Statistics about the Corebody Temperature

In this section various descriptive statics will be given about the Core Body temperature.

First to start of a table with the number of participants, amplitude and Mean is given:

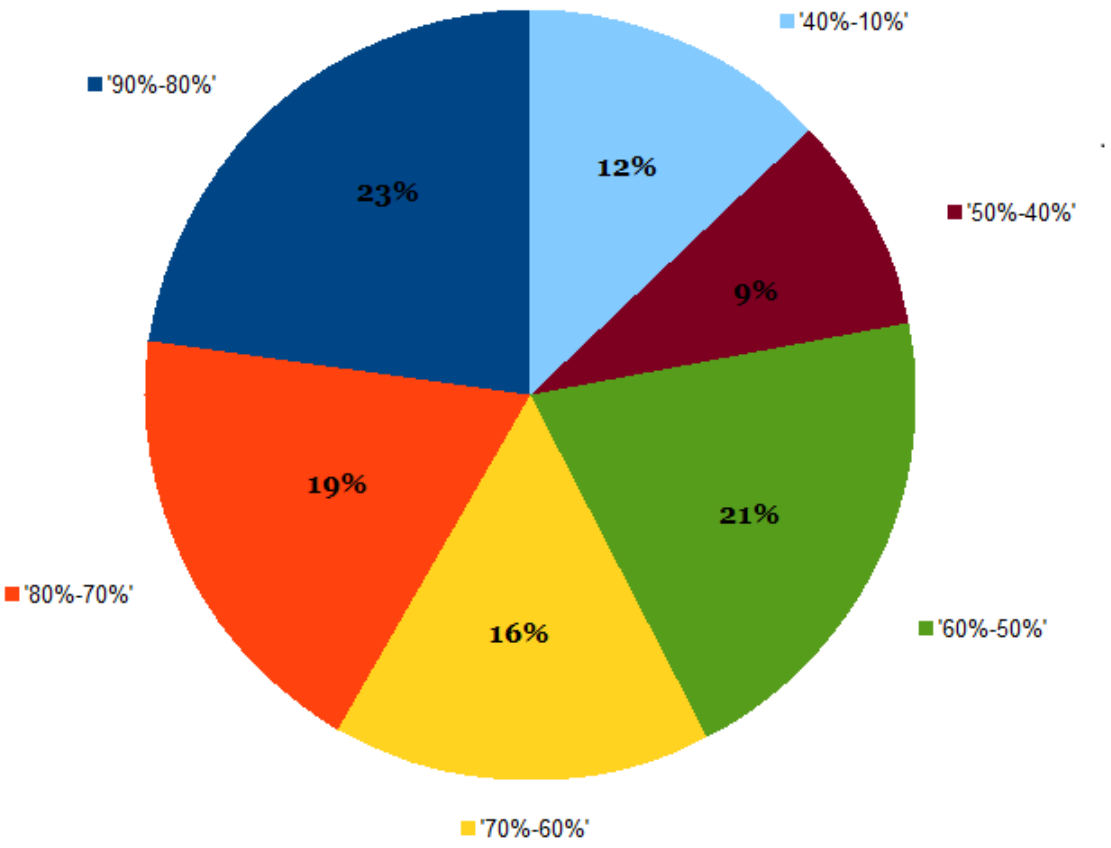
Table 2: The statistics of the CoreBody Temperature for the 2 groups.

| | **Longevity** | **Control** | **Total** |
|---|---|---|---|
| Participants | 70 | 62 | 122 |
| Amplitude | 1.73(0.43) | 1.68(0.44) | 1.71(0.44 ) |
| Mean | 36.89(0.35) | 36.86(0.3) | 36.88(0.33 ) |

As again can be seen from Table 2 there isn't a lot of difference between the groups in terms of Amplitude and Mean, however Amplitude and mean do not give take into the chronological nature of time series and as such do not say everything about the time series.

Next in order to give insight into the amount of missing value's in the Core Body Temperature the following pie chart is shown:

Figure 9: Pie Chart of Missing value's, to get a insight into the amount of missing value's



Here the colors represent the amount of missing value's, while the percentages in the pie charts represent the respective amount of participants which have these amount of missing value's.

# 5 Experiments on Synthetic CBF Data with Missing Value's

First in order to test if our k-medoids cluster algorithm can still convergence on the natural grouping in data with participants with missing value's in there data, it will tested on synthetic made data of which it is certain that they are created from different classes.

The Cylinder Bell Funnel data set has been chosen, it is a known and widely used classification data set [35] [58][54] and [46]. The time series in this data set are of three classes: cylinder,bell and funnel, samples of data are generated as follows:

$$c(t) = (6+n) * X_{[a,b]}(t) + e(t) \tag{8}$$

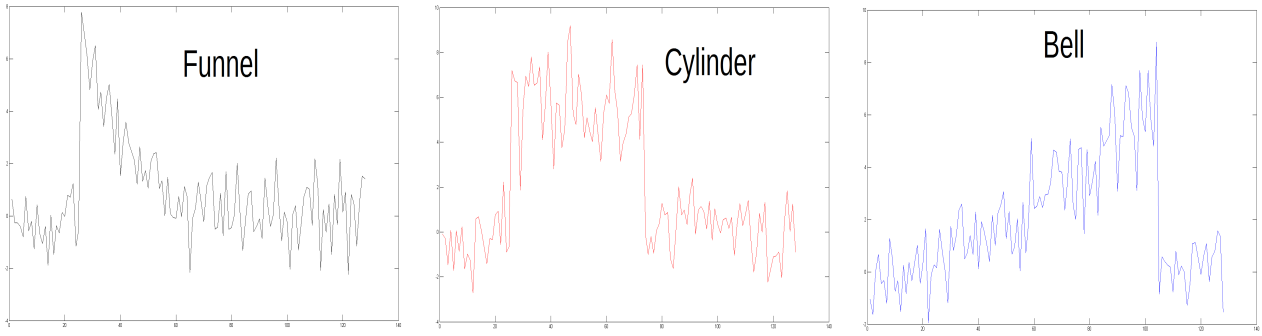
$$b(t) = (6+n) * X_{[a,b]}(t) * (t-a)/(b-a) + e(t) \tag{9}$$

$$f(t) = (6+n) * X_{[a,b]}(t) * (b-t)/(b-a) + e(t) \tag{10}$$

Where $c(t)$ is the cylinder, $b(t)$ is the bell, $f(t)$ is the funnel, $t \in [1, 128]$ , $X_{[a,b]}(t)$ if $a \leq t \leq b$ , 0 other wise, $n$ and $e(t)$ are drawn from a standard normal distribution $N(0, 1)$, $a$ is an integer drawn uniformly from $[16, 32]$ and $b-a$ is an integer drawn uniformly from $[32, 96]$.

This figure shows a example of each class:

Figure 10: The three classes of the CBF data set




The funnel class is characterized by a sudden increase at $a$, and a gradual decrease until $b$.

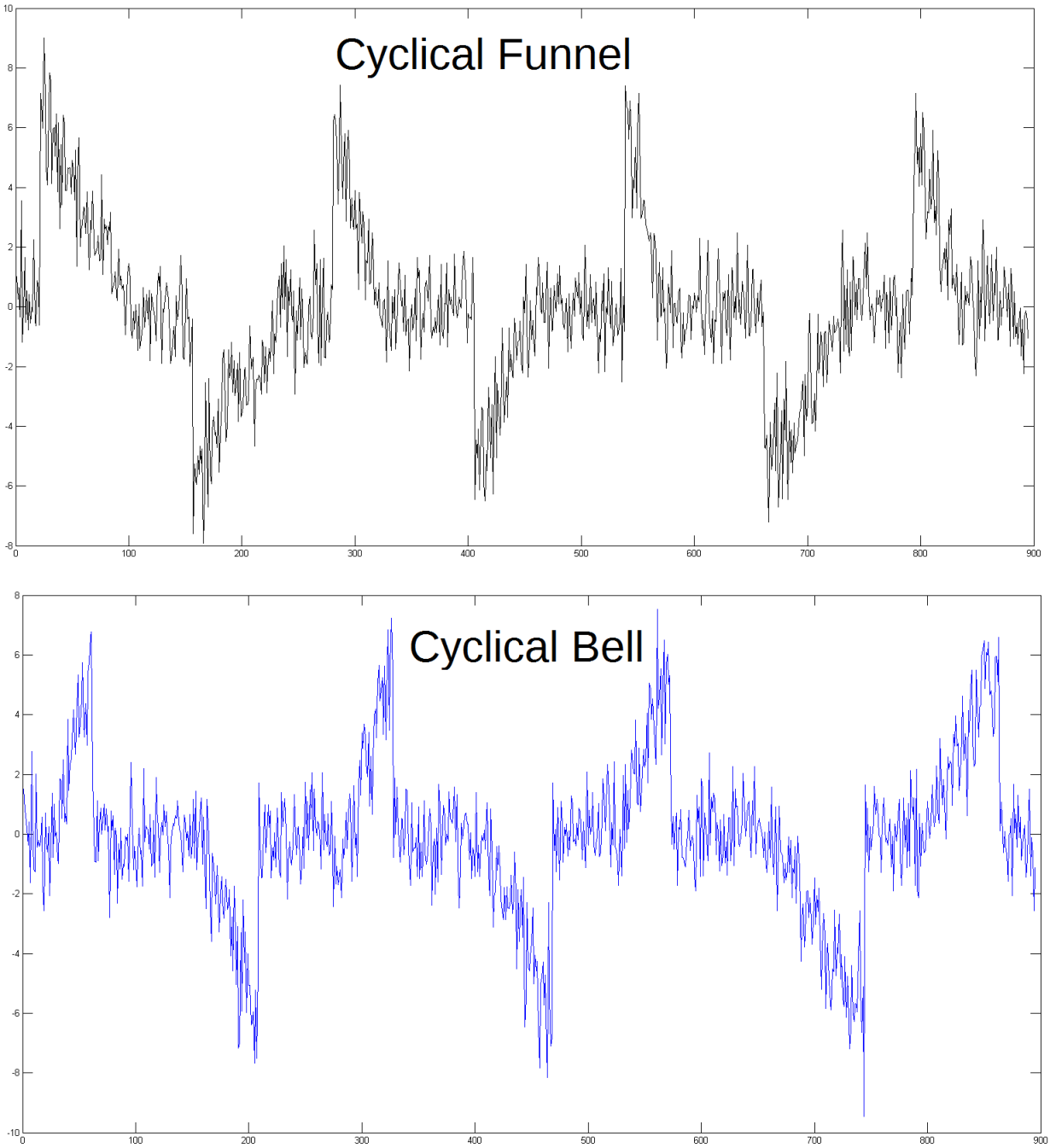
The cylinder class is characterized by a plateau from a time to $b$ time, and by sharp rise and drop - before and the plateau.

The bell class is characterized by a gradual increase from $a$ to $b$, followed by a sharp drop at $b$.

However the data is not the same to our Core Body Temperature in the sense that it is not cyclical and more than 2 classes, in order to adjust to this the formula is repeated 6 times while mirroring in terms of y-axis of the one before them so that they make 3 and a half cycles,only use the bell and the funnel formula's and have the same amount of data points as with the Core Body Temperature.

Which results in the following graphs for the examples of these 2 classes:

Figure 11: Cyclical Funnal and Cyclical Bell examples




.

As you can see in the figure the data now makes a cycles of 3 and a half with in total 1008 data points for each time series, all this in order to make it more similar to our CoreBody Temperature.

70 instances of the bell class were created and 70 instances of the funnel class were created, later they will be adjust to the same ratio as the core body temperature data set, again to match our Core Body Temperature.

With this synthetic data the distinctive discrimination pattern occurs in the peaks every 12 hours, which hopefully the k-medoids algorithm can find.

First to asses the performance of the k-medoids algorithm and the distance measures, it be tested on the Cyclical generated data without missing value's.

## 5.1 Performance without missing value's: Euclidean Distance

The cluster formed in Figure 12 are mostly the same as the figures in Figure 10, thus meaning that the k medoids has converted properly.

Table 3: Confusion Matrix for K-means with Euclidean distance.

| | Bell | Funnel |
|---|---|---|
| $c1$ | 70 | 0 |
| $c2$ | 0 | 70 |

As you can see from the tables with results the k-medoid algorithm with Euclidean Distance had perfectly discriminate between the 2 classes.

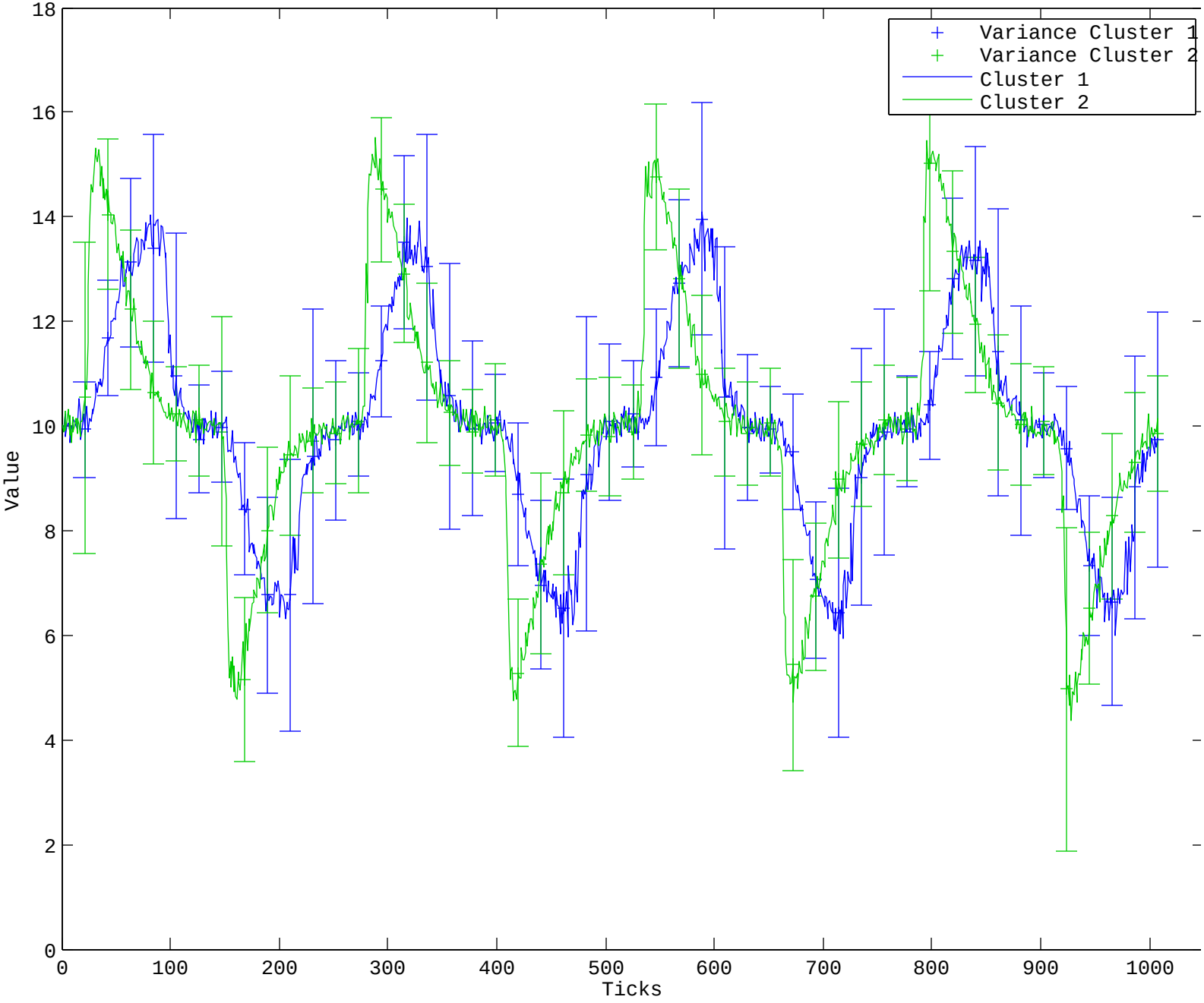



Figure 12: Clusters formed by the k-medoids algorithm with Euclidean Distance

## 5.2 Performance without missing value's: Dynamic Time Waring

The cluster formed in Figure 13 are mostly the same as the figures in Figure 10, thus meaning that the k medoids has converted properly with the Dynamic Time Warping distance measure.

Table 4: Confusion Matrix for K-medoids with Dynamic Time Warping.

| | Bell | Funnel |
|---|---|---|
| $c1$ | 70 | 0 |
| $c2$ | 0 | 70 |

Also as seen in the confusion matrix, the algorithm perfectly converged on the data.

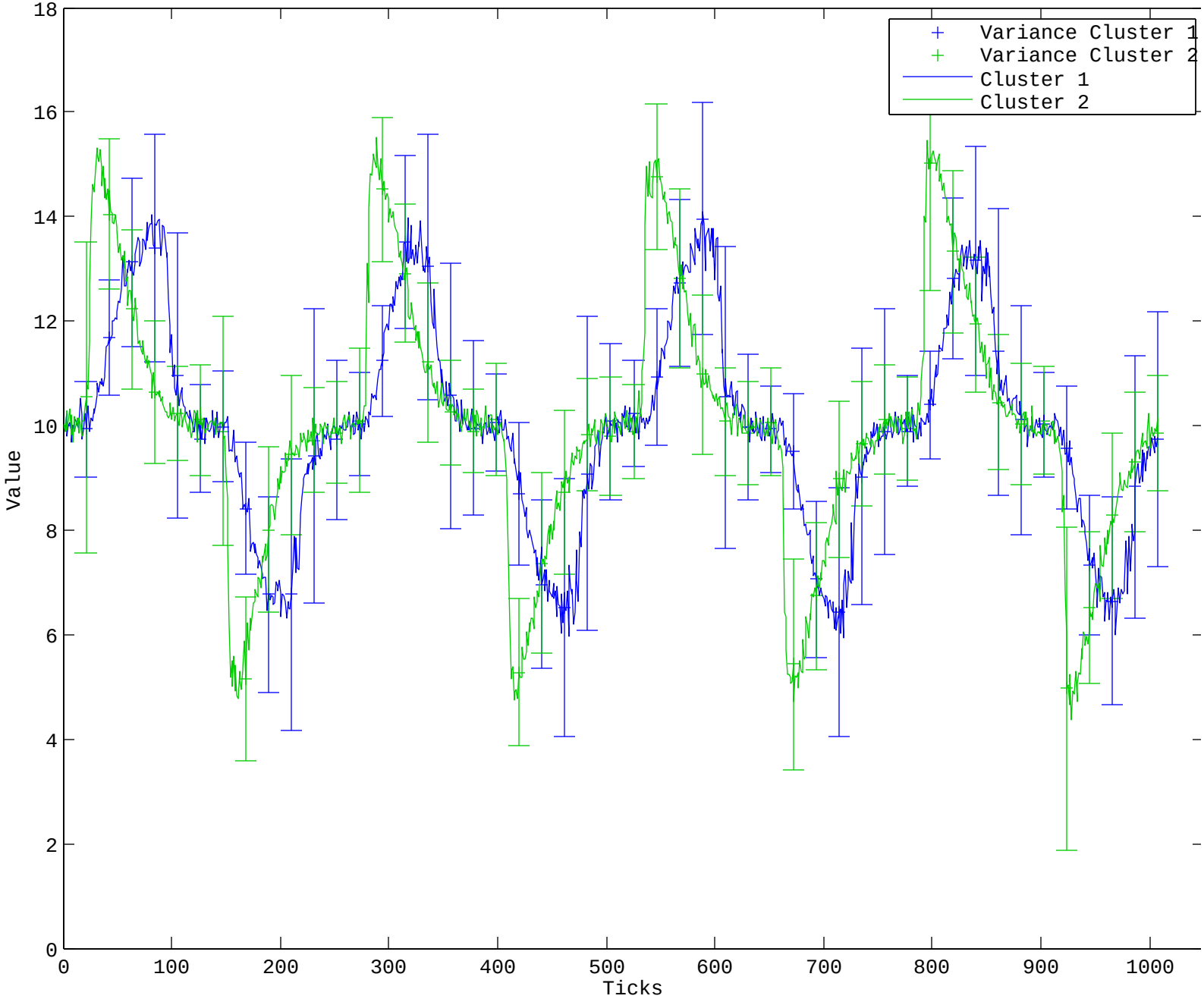



Figure 13: Clusters formed by the k-medoid algorithm with Euclidean Distance

## 5.3 Generating missing value's

Now missing value's are added to the cyclical funnel and bell data set, the missing value's which are added are the same as the missing value's which the core body temperature has. Also the number of instance of each class are now the same as the number of participants of each group.

## 5.4 Performance with missing value's: Euclidean Distance

As you can see the cluster formed here are mostly the same as the figures in Figure 14, thus meaning that the k means has converted properly even with the missing value's.

Table 5: Confusion Matrix for K-medoids with Euclidean distance and with missing value's.

| | Bell | Funnel |
|---|---|---|
| $c1$ | 70 | 0 |
| $c2$ | 0 | 62 |

Again the confusion matrix is perfect, so for this data set the K-medoids algorithm has worked even with missing value's.

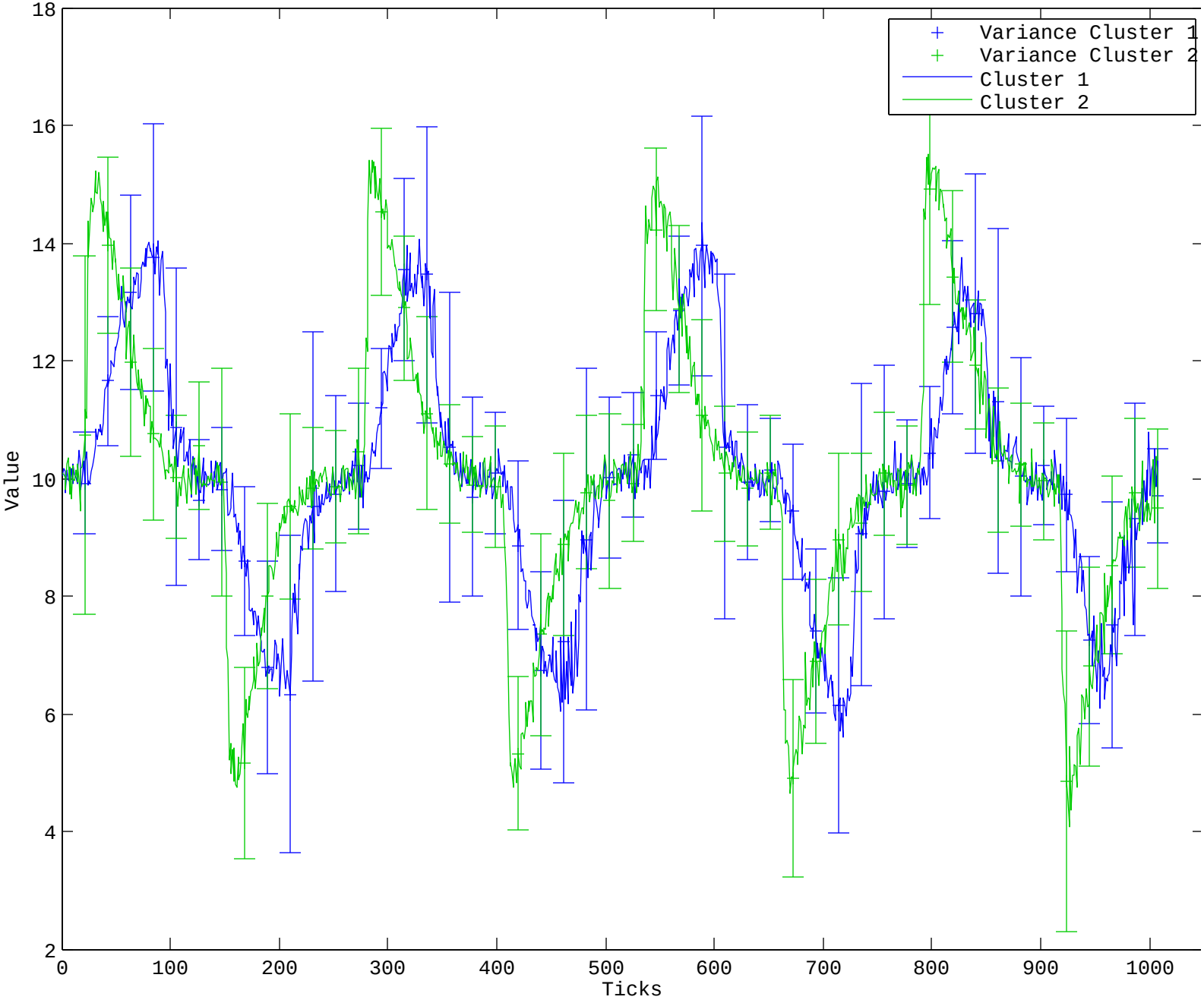



Figure 14: Clusters formed by the k-medoids algorithm with Euclidean Distance

## 5.5 Performance with Missing Value's: Dynamic Time Warping

Also with the use of the similarity measure of Dynamic Time Warping, the cluster algorithm convergenced nicely on finding the 2 classes, even with the amount of missing values.

Table 6: Confusion Matrix for K-medoids with Euclidean distance and with Missing Value's.

| | Bell | Funnel |
|---|---|---|
| $c1$ | 70 | 0 |
| $c2$ | 0 | 62 |

As can be seen in Figure 15 and the table 6 confusion matrix, the algorithm properly converged.

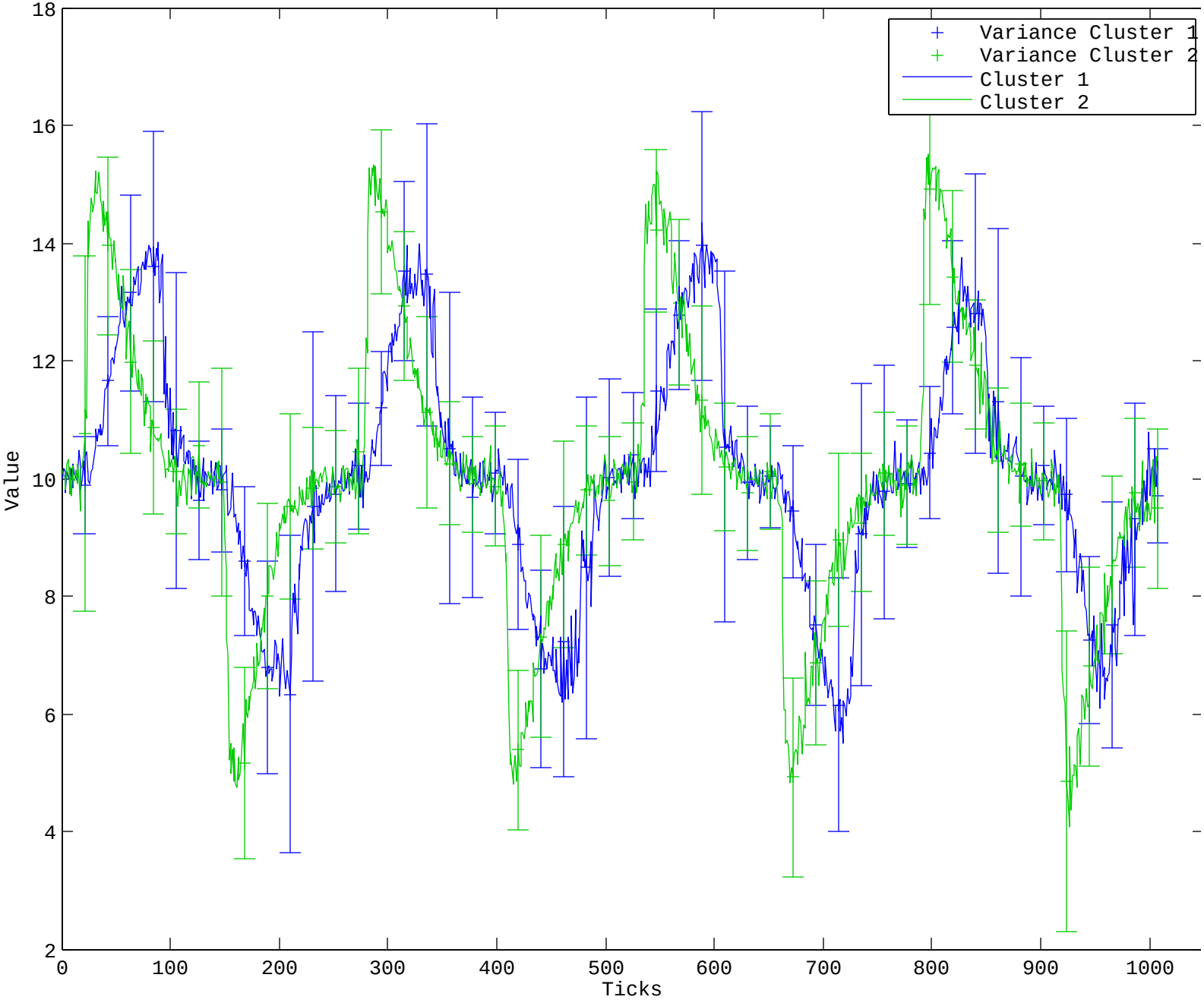



Figure 15: Clusters formed by the k-medoids algorithm with Euclidean Distance

## 5.6 Performance with Piecewise Cubic Interpolation

In Figure 16 and Figure 17 we can see the cluster formed by the clustering algorithm but now with the help of piecewise cubic interpolation done to the missing value's, either with the euclidean distance measure of the dynamic time warping distance measure.

This is the specific confusion matrix for this setup:

Table 7: Confusion matrix for piecewise cubic interpolation with ED:

| | Bell | Funnel |
|---|---|---|
| $c1$ | 10 | 54 |
| $c2$ | 60 | 8 |

Table 8: Confusion matrix for piecewise cubic interpolation with DTW:

| | Bell | Funnel |
|---|---|---|
| $c1$ | 24 | 62 |
| $c2$ | 46 | 0 |

As we can see in Table 7 and 8 piecewise cubic interpolation has a worse performance compared to the missing value ignoring approach with both dynamic time warping and euclidean distance. This interpolation method had generated the best results in this experiment for the sake of clarity the other methods won't be displayed here.

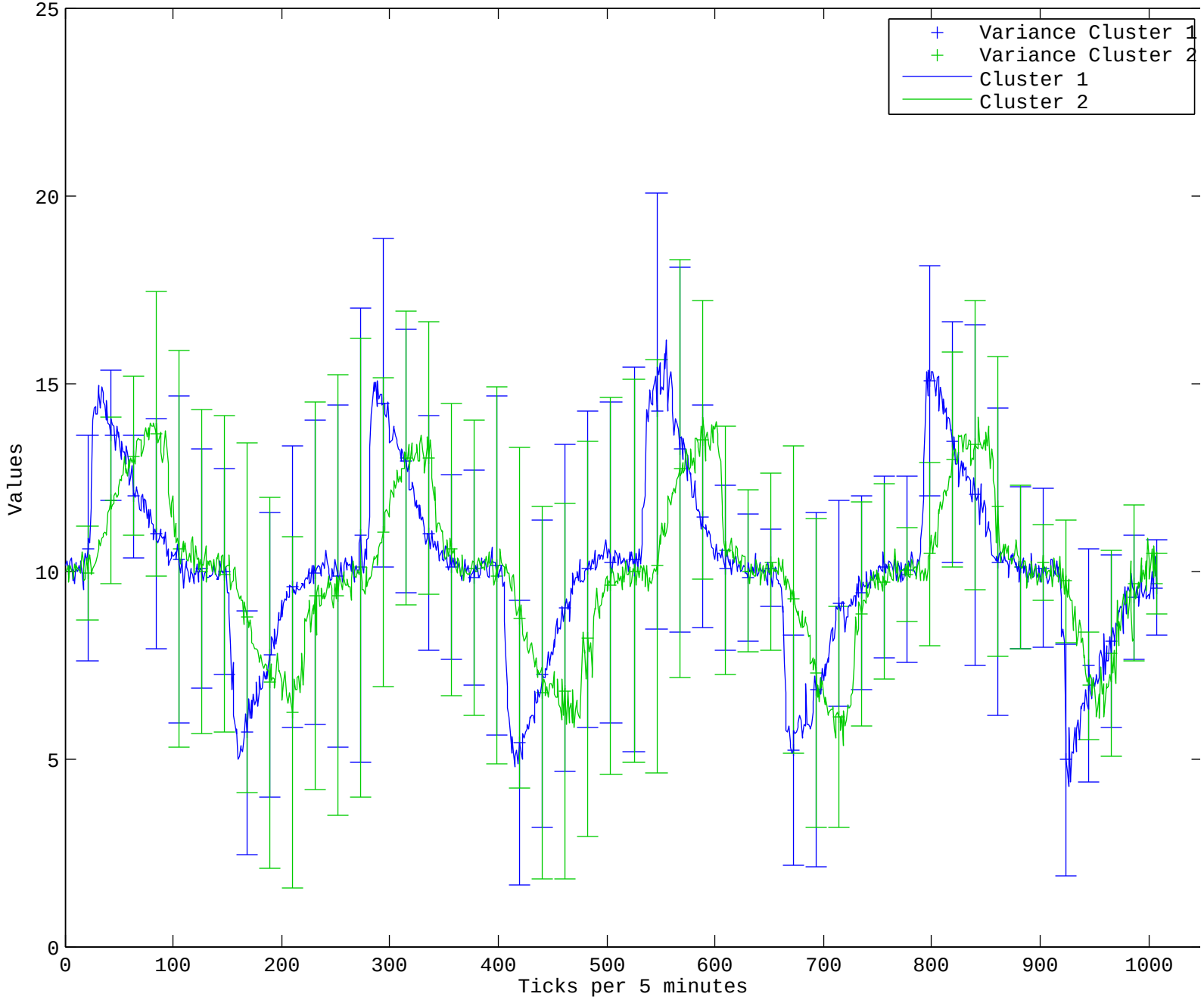



Figure 16: Clusters formed by the k-medoids algorithm with Euclidean Distance with Interpolation.

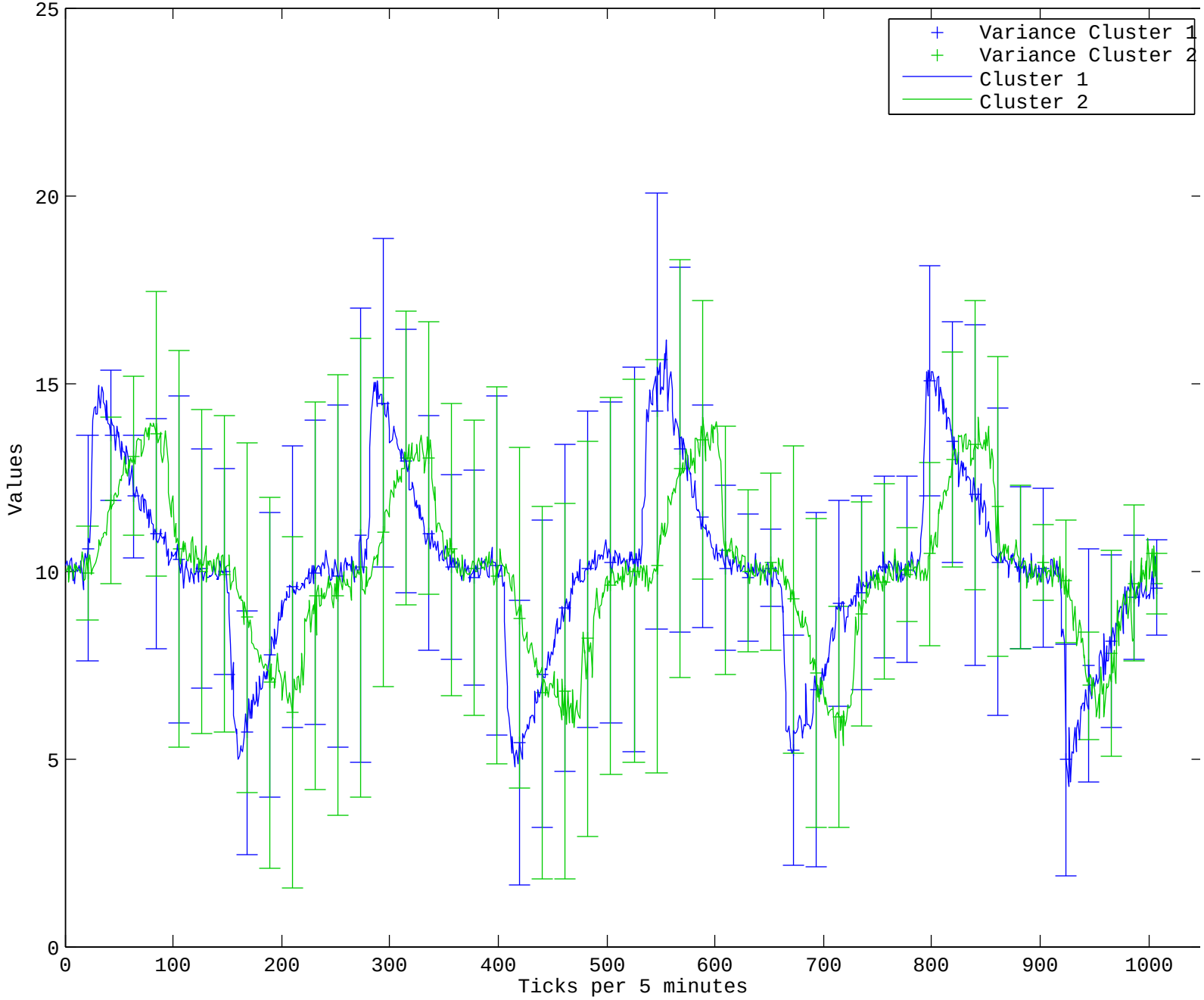



Figure 17: Clusters formed by the k-medoids algorithm with Dynamic Time Warping with Interpolation.

## 5.7 Conclusion on Synthetic Data Clustering with Missing Value's

For this current set up it seemed that the k-medoid algorithm could still convergence on non-missing value's and missing value's data when a certain distinctive discriminating pattern occurs every 12 hours which is assumed to by realistic by the domain experts. This means that it supports the fact that it could also convergence on the Corebody Temperature data.

While Interpolation of the missing value's had a worse performance compared to the missing value's approach, thus it is concluded here that the missing value ignoring approach will be used on the Corebody Temperature.

# 6 Experiment Clustering Results of Corebody Temperature Time Series Data

## 6.1 Clustering Results on Core Body Temperature

### 6.1.1 Euclidean Distance

The are the clusters which were formed with the Euclidean distance and our missing value's adjusted K-medoids algorithm, are displayed in Figure 15 on the next page.

The green and blue lines displayed in the graphs represent the mean line of all the participants which where assigned to that respective cluster. It can be seen that these means differ in the sense that cluster 1 in consistently higher than cluster 2, however it can also been see that the standard deviation spread shows that this difference is not strongly significant.

Table 9: Confusion Matrix for K-medoids with Euclidean distance.

| | Longevity | Normal |
|---|---|---|
| $c1$ | 49 | 40 |
| $c2$ | 21 | 22 |

Seeing the mean graphs and the results of the setup of the experiments, there is a strong indication there is no difference between the two groups

In relation to Longevity Cluster 1 has 89 of the 132 participants in which roughly 50 percent is of longevity, 49 and 40 for respectively Longevity and normal. And Cluster 2 has 43 of the 132 participants in which 21 are longevity and 22 are control. So this means that not a strong relation has been found between these clusters and longevity.

No other explanation for these clusters have been found, i.e. a relation with sex or season could also not explain these clusters.

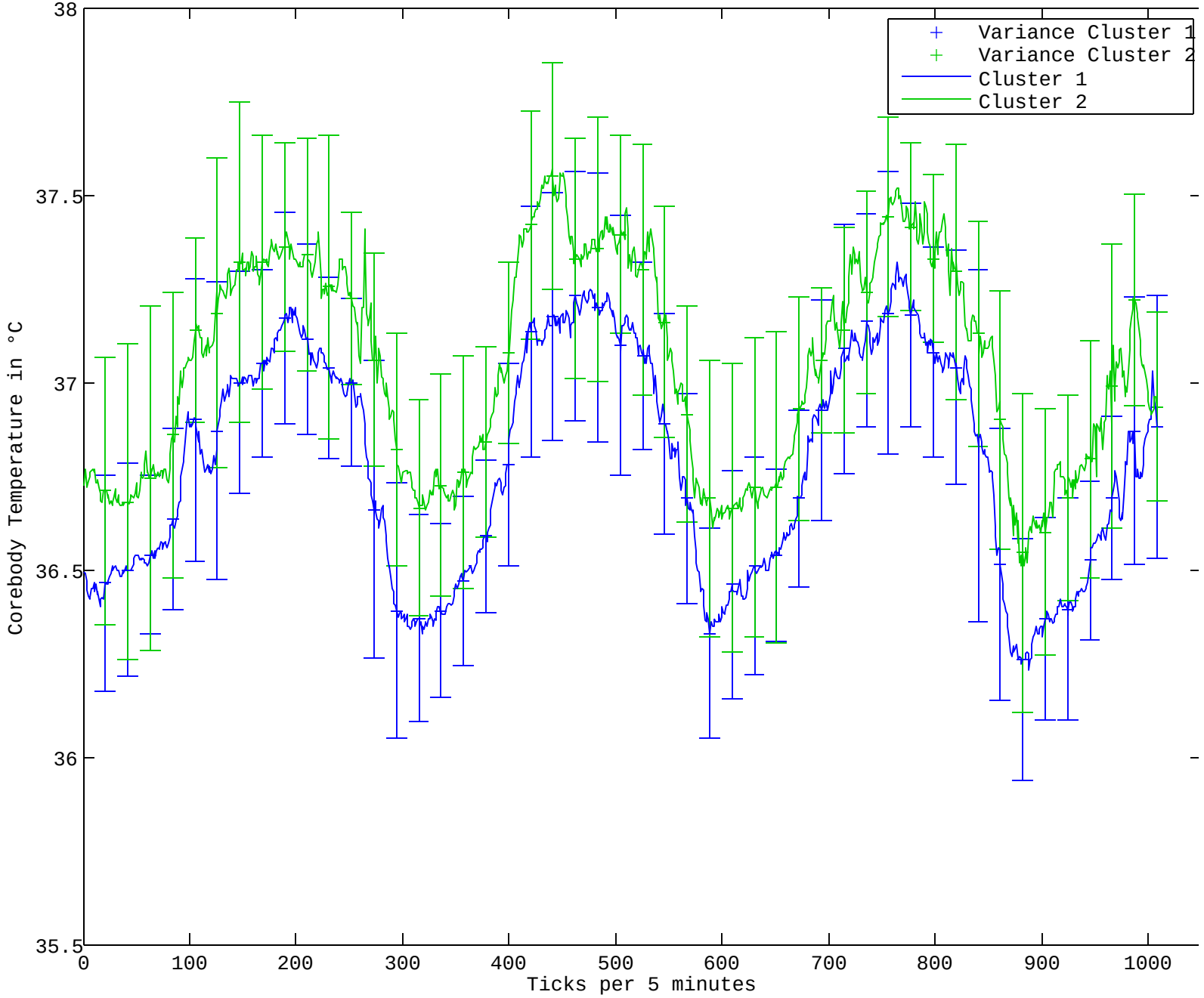



Figure 18: Clustering Results with Euclidean Distance

### 6.1.2 Dynamic Time Warping

The are the clusters which were formed with the Dynamic Time Warping and our missing value's adjusted K-means algorithm, are displayed in Figure 19 on the next page.

The black and blue lines displayed in the graphs represent the mean line of all the participants which where assigned to that respective cluster. It can be seen that these means differ in the sense that cluster 1 in consistently higher than cluster 2, however it can also been see that the standard deviation spread shows that this difference is not strongly significant.

Table 10: Confusion Matrix for K-medoids with Dynamic Time Warping.

| | Longevity | Normal |
|---|---|---|
| $c1$ | 52 | 44 |
| $c2$ | 18 | 18 |

In relation to Longevity Cluster 1 has 96 of the 132 participants in which 52 are longevity and 44 are control . And Cluster 2 has 36 of the 132 participants in which 18 are longevity and 18 are control. So again this means that not a strong relation has been found between these clusters and longevity.

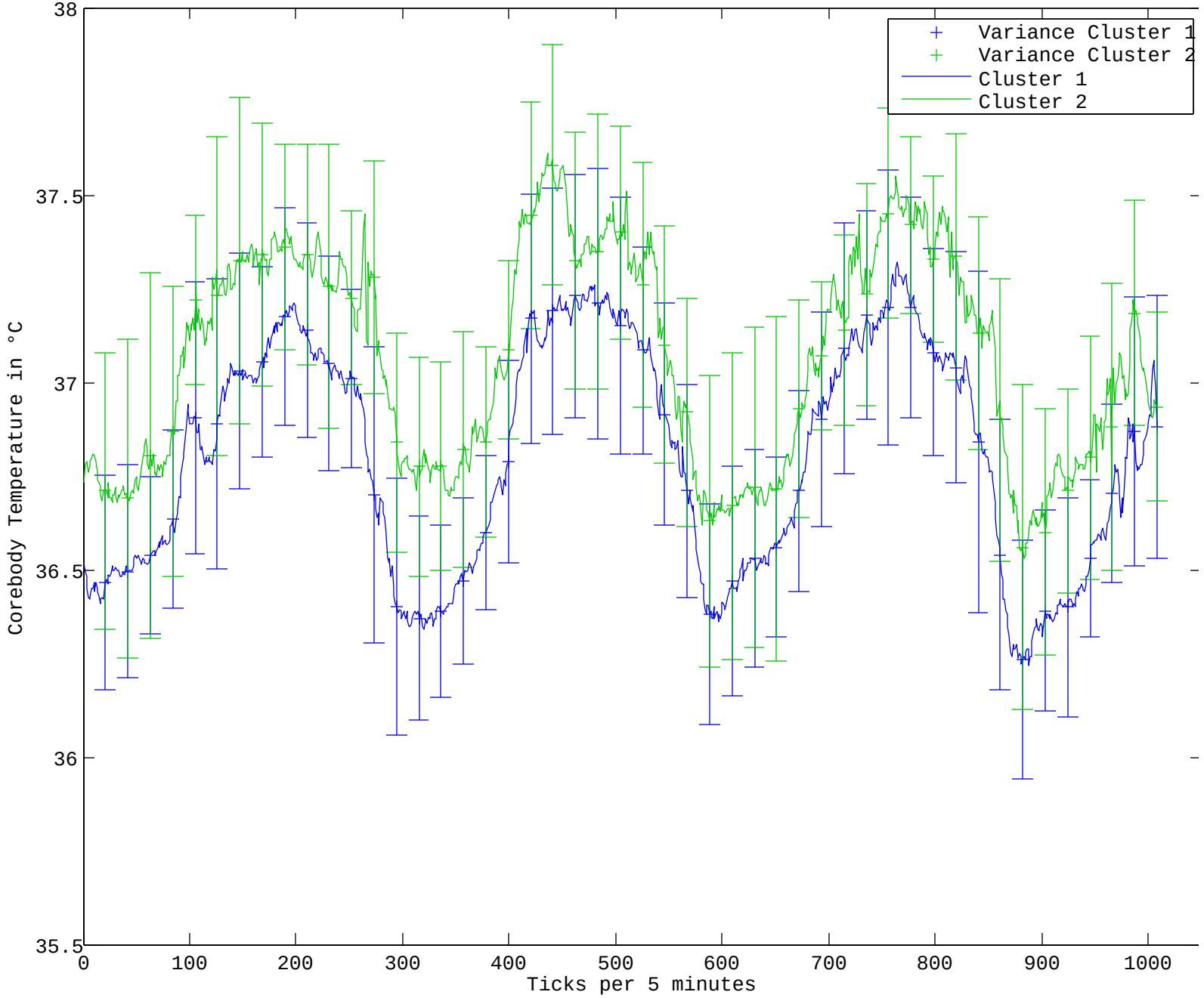



Figure 19: Clustering Results with Dynamic Time Warping

# 7 Conclusion

Already it was know from Figure 8 that the means between off spring from long living people and the partners of the offspring was not very significantly different.

It was reasoned that maybe not every off spring inherited the longevity gene or that there was some other hidden pattern in the data, which both could have been found by time series cluster analyses.

Most of the literature on whole time series clustering assumes that missing value's in time series can be handled with pre processing techniques. However in the case of the Core Body Temperature, it was thought that a missing value ignoring approach might be a better approach. Techniques such as interpolation would only bias and effect the similarity of the data more. This is because it was thought that a similarity pattern occurs on a 24 hour basis. If 1 day of 24 hours of the 84 hours of data is missing there are still 2 and half days left to seek similarity or dissimilarity with in the clustering process.

From the experiments on synthetic data it was shown that indeed a missing value ignoring approach to whole time series clustering can work. But it also gives a better performance compared to interpolation of missing value's. If we assume that a distinctive pattern happens happens every 12 hours.

Yet the clusters found with our missing value approach on the Core Body Temperature could not be explained by longevity or any other reason.

Thus it strongly believed that there just isn't any distinction in Core Body Temperature between the off spring and the partners of the off spring which served as a control group.

# 8 Further Work

There are more variables than the Core Body Temperature collected during the SwitchBox Study, namely: heart rate, glucose,MRI and many more.

All these other variables encounter the same missing value problems, and if we assume they a distinctive pattern every 12 hours, they could benefit from the same missing value ignoring approach as in this thesis work.

A multivariate analysis approach is also strongly suggested, i.e. maybe there is connection between all the SwitchBox variables which could explain longevity.

Such as authors of [3] do, they take a interesting temporal multivariate approach for electronic health record data in the classification of thrombocytopenia in patients , which strongly suggested to also be done for the SwitchBox variables.

In short they suggested the abstraction of time series into nominal trend values: Increasing and Decreasing; and into nominal trend values: low, normal, high.

Then they find temporal patterns by using temporal logic with sliding a window of a certain size across the multiple time series, which finds interactive pattern such as "the administration of heparin precedes a decreasing trend in platelet counts".

But because this type of temporal pattern mining usually return large number of patterns, of which most are irrelevant for classification, they suggest a Minimal Predictive Temporal Pattern framework which filters out non-predictive and spurious temporal patterns.

# References


[1] Aach J, Church G (2001) Aligning gene expression time series with time warping algorithms. Bioinformatics 17:495508.

[2] Aggarwal, Charu C., Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional space. Springer Berlin Heidelberg, 2001.

[3] Batal, Iyad, et al. "A pattern mining approach for classifying multivariate temporal data." Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on. IEEE, 2011.

[4] Bar-Joseph Z, Gerber G, Gifford D, Jaakkola T, Simon I (2002) A new approach to analyzing gene expression time series data. In: Proceedings of the 6th annual international conference on research in computational molecular biology, pp 3948.

[5] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In KDD Workshop, 1994.

[6] BROCKWELL,P.ANDDAVIS, R. 2009. Time Series: Theory and Methods.Springer.

[7] Berkhin, Pavel. "A survey of clustering data mining techniques." Grouping multidimensional data. Springer Berlin Heidelberg, 2006. 25-71.

[8] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York and London,1987

[9] Box, George; Jenkins, Gwilym (1970). Time series analysis: Forecasting and control, San Francisco: Holden-Day

[10] Chakrabarti, K., Keogh, E., Pazzani, M., Mehrotra, S. Locally adaptive dimensionality reduction for indexing large time series databases. ACM Transactions on Database Systems. Volume 27, Issue 2, (June 2002). pp 188-228.

[11] Chung, H. M., & Gary, P. (1999). Special section: Data mining. Journal of Management Information Systems, 16(1), 1116.

[12] DING,H.,TRAJCEVSKI,G.,SCHEUERMANN,P.,WANG,X.,ANDKEOGH, E. 2008. Querying and mining of time series data: Experimental comparison of representations and distance measures.Proc. VLDB Endowm. 1,2, 15421552.pok

[13] Esling, Philippe, and Carlos Agon. "Time-series data mining." ACM Computing Surveys (CSUR) 45.1 (2012): 12.

[14] Estivill-Castro, V. (2002). "Why so many clustering algorithms". ACM SIGKDD Explorations Newsletter 4: 65. doi:10.1145/568574.56857.

[15] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density- based algorithm for discovering clusters in large spatial databases, Proceedings of the 1996 International Conference on Knowledge Discovery and Data Mining (KDD96), Portland, OR, 1996, pp. 226231.

[16] Effros, Michelle, and Diego Dugatkin. "Multiresolution vector quantization." Information Theory, IEEE Transactions on 50.12 (2004): 3130-3145

[17] Fu, Tak-chung. "A review on time series data mining." Engineering Applications of Artificial Intelligence 24.1 (2011): 164-181

[18] Geurts, P. Pattern extraction for time series classification. Proceedings of Principles of Data Mining and Knowledge Discovery, 5th European Conference; 2001 Sep 3-5; Freiburg, Germany, pp 115-127.

[19] Grzymala-Busse, Jerzy W., and Ming Hu. "A comparison of several approaches to missing attribute values in data mining." Rough sets and current trends in computing. Springer Berlin Heidelberg, 2001.

[20] X. Golay, S. Kollias, G. Stoll, D. Meier, A. Valavanis, P. Boesiger, A new correlation-based fuzzy logic clustering algorithm for fMRI, Mag. Resonance Med. 40 (1998) 249260.

[21] S. Guha, R. Rastogi, K. Shim, CURE: an efficient clustering algorithm for large databases, Proceedings of the 1998 ACM- SIGMOD International Conference on Management of Data, Seattle, WA, June 1998, pp. 7384.

[22] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, 2012 pp. 443540

[23] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, 2012 pp. 8889

[24] Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." Journal of the Royal Statistical Society. Series C (Applied Statistics) 28.1 (1979): 100-108.

[25] Harvey, Andrew C., and Richard G. Pierse. "Estimating missing observations in economic time series." Journal of the American Statistical Association 79.385 (1984): 125-131.

[26] HUANG,Y.ANDYU, P. 1999. Adaptive query processing for time-series data. In Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 282286

[27] Jain, Anil K., and Richard C. Dubes. Algorithms for clustering data. Prentice-Hall, Inc., 1988.

[28] Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn. "Data clustering: a review." ACM computing surveys (CSUR) 31.3 (1999): 264-323.

[29] Jiang, Yi, Tuo Lan, and LiHua Wu. "A Comparison Study of Missing Value Processing Methods in Time Series Data Mining." Computational Intelligence and Software Engineering, 2009. CiSE 2009. International Conference on. IEEE, 2009.

[30] G.Karypis, E.-H. Han, V. Kumar, Chameleon: hierarchical clustering using dynamic modeling, Computer August (1999), 68-75

[31] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Wiley, New York, 1990.

[32] KEOGH,E.ANDPAZZANI, M. 1998. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. InProceedings of the 4th International Conference of Knowledge Discovery and Data Mining. AAAI Press, 239241.

[33] Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S. Dimensionality reduction for fast similarity search in large time series databases. Knowledge and Information Systems 2001; 3: 263-286.

[34] Keogh, E., Lonardi, S., Chiu, W. Finding Surprising Patterns in a Time Series Database In Linear Time and Space. In the 8 th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2002 Jul 23 26; Edmonton, Alberta, Canada, pp 550-556

[35] Keogh, Eamonn, and Shruti Kasetty. "On the need for time series data mining benchmarks: a survey and empirical demonstration." Data Mining and knowledge discovery 7.4 (2003): 349-371.

[36] Keogh, Eamonn, Stefano Lonardi, and Chotirat Ann Ratanamahatana. "Towards parameter-free data mining." Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.

[37] Keogh, Eamonn, and Chotirat Ann Ratanamahatana. "Exact indexing of dynamic time warping." Knowledge and information systems 7.3 (2005): 358-386.

[38] R. Krishnapuram, A. Joshi, O. Nasraoui, L. Yi, Low- complexity fuzzy relational clustering algorithms for web mining, IEEE Trans. Fuzzy Systems 9 (4) (2001) 595607

[39] Kruskall JB, Liberman M (1983) The symmetric time warping algorithm: from continuous to discrete. In: Time warps, string edits and macromolecules. Addison

[40] Lakshminarayan, Kamakshi, Steven A. Harp, and Tariq Samad. "Imputation of missing data in industrial databases." Applied Intelligence 11.3 (1999): 259-275.

[41] Warren Liao, T. "Clustering of time series dataa survey." Pattern Recognition 38.11 (2005): 1857-1874.

[42] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: L.M. LeCam, J. Neyman (Eds.), Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281297.

[43] E.A. Maharaj, Clusters of time series, J. Classification 17 (2000) 297314.

[44] Murty, M. Narasimha, and Anil K. Jain. "Knowledge-based clustering scheme for collection management and retrieval of library books." Pattern recognition 28.7 (1995): 949-963.

[45] Moon, Todd K. "The expectation-maximization algorithm." Signal processing magazine, IEEE 13.6 (1996): 47-60.

[46] Naoki Saito. Local feature extraction and its application using a library of bases. PhD thesis, Yale University, December 1994.

[47] Paterson, Mike, and Vlado Dank. Longest common subsequences. Springer Berlin Heidelberg, 1994.

[48] Rabiner L, Juang B (1993) Fundamentals of speech recognition. Prentice, Englewood Cliffs, NJ

[49] Rani, Sangeeta, and Geeta Sikka. "Recent Techniques of Clustering of Time Series Data: A Survey." International Journal of Computer Applications 52 (2012).

[50] Ratanamahatana, Chotirat Ann, et al. "Mining time series data." Data Mining and Knowledge Discovery Handbook. Springer US, 2010. 1049-1077.

[51] Schmill M, Oates T, Cohen P (1999) Learned models for continuous planning. In: 7th international workshop on articial intelligence and statistics

[52] C.T. Shaw, G.P. King, Using cluster analysis to classify time series, Physica D 58 (1992) 288298.

[53] Schoenmaker, M, *Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study.*. European journal of human genetics : EJHG 14, 79-84, doi:10.1038/sj.ejhg.5201508 (2006).

[54] Stefanos Manganaris. Supervised Classification with Temporal Data. PhD thesis, Computer Science Department, School of Engineering, Vanderbilt University, December 1997.

[55] Stijntjes, M. *Familial longevity is marked by better cognitive performance at middle age* the Leiden Longevity Study. PloS one 8, e57962, doi:10.1371/journal.pone.0057962 (2013).

[56] Moon, Todd K. "The expectation-maximization algorithm." Signal processing magazine, IEEE 13.6 (1996): 47-60.

[57] W. Wang, J. Yang, R. Muntz, R., STING: a statistical information grid approach to spatial data mining, Proceedings of the 1997 International Conference on Very Large Data Base (VLDB97), Athens, Greek, 1997, pp. 186195.

[58] Waleed Kadous. Temporal Classi cation: Extending the Classi cation Paradigm to Multivariate Time Series. PHD Thesis. The University of New South Wales, October 2002.

[59] T. Zhang, R. Ramakrishnan, M. Livny, BIRCH: an efficient data clustering method for very large databases, Proceedings of the 1996 ACM-SIGMOD International Conference on Management of Data, Montreal, Canada, June 1996, pp