Tampere University

Aysun Yozgyur

CLUSTERING OF GEARBOXES BASED ON LOAD-TEST VIBRATION MEASUREMENTS

Faculty of Information Technology and Communication Sciences (ITC)
Master of Science thesis
May 2023

Abstract

Aysun Yozgyur: CLUSTERING OF GEARBOXES BASED ON LOAD-TEST VIBRATION MEASUREMENTS
Master of Science thesis
Tampere University
Masters Degree Programme in Computational Big Data Analytics
May 2023

Renewable energy sources such as wind power has gained more importance as the climate change has become inevitable and governments all around the world have taken actions to stop the degradation of the world ecosystem. Therefore, it is crucial to increase the reliability and lower the cost of maintenance for gearboxes used in the wind turbines. The function of gearboxes in the wind turbines is to increase the rotational speed of a low-speed rotor and transfer it to a higher speed electrical generator. Considering their large size and critical position inside the turbine, reliability of each gearbox has to be tested before they are mounted to the wind turbines since their repair is a major issue. This thesis aims at clustering gearboxes from a single product type based on their vibration measurements that are recorded during test runs. The secondary aim is to investigate the relationship between the vibration measurements of a gearbox and its endurance. Focusing on high-speed shaft of the gearboxes, a dataset consisting of 337 gearboxes having a key performance index (KPI) value in 3 particular measurement points within 3 load steps, which generates 9 variables, is used. In order to come up with statistically meaningful clustering results, a special type of k-means clustering named as *KmL3D* and group-based multi-trajectory modeling that is an application of finite mixture modeling is employed on the collected data set. The clustering algorithms are tuned with changing the hyperparameters, where applicable once at a time. The model accuracies have been evaluated according to the Calinski-Harabasz information criterion for clustering with *KmL3D* and the BIC (Bayesian Information Criterion) and AIC (Akaike Information Criterion) for group-based multi-trajectory modeling. Based on these fit statistics, the model with two clusters was the best in *KmL3D* while the model with six clusters and quadratic trajectories was the best according to the BIC and AIC scores, and additional model selection criteria. The results obtained show that the clustering algorithms can be successfully applied to the collected data, meaning that it is possible to extend it to other product families of gearboxes.

Keywords: Clustering, Vibration Measurements, Finite Mixture Modeling, Group-Based Multi-Trajectory Modeling.

The originality of this thesis has been checked using the Turnitin Originality Check service.

Preface

This thesis was funded by Moventas Gears Oy. I would like to thank my supervisors for their support and guidance throughout the thesis process. I also want to express my gratitude to my family and my husband for their support.

Tampere, May 2023

Contents

1 Introduction . . . 1
1.1 The focus of this thesis . . . 3
1.2 The structure of this thesis . . . 4
2 Method . . . 5
2.1 K-means clustering . . . 5
2.1.1 KmL . . . 7
2.1.2 KmL3D . . . 7
2.1.3 Model selection in KmL3D . . . 9
2.2 Finite mixture models . . . 10
2.2.1 Maximum likelihood fitting and Expectation Maximization Algorithm . . . 11
2.2.2 Group-based trajectory modeling . . . 15
2.2.3 Generalization of group-based trajectory modeling to multi-variate data . . . 16
2.2.4 Model selection in group-based multi-trajectory modeling . . . 21
3 Collected Data . . . 23
4 Analysis and Results . . . 27
4.1 KmL3D analysis . . . 27
4.2 KmL3D results . . . 28
4.3 Group-based multi-trajectory modeling analysis . . . 31
4.4 GBMT Modeling results . . . 32
4.4.1 GBMT Modeling results with polynomial degree 1 and using raw data . . . 32
4.4.2 GBMT Modeling results with polynomial degree 2 and using raw data . . . 36
4.4.3 GBMT Modeling results with polynomial degree 1 and centering on the data set . . . 38
4.4.4 GBMT Modeling results with polynomial degree 2 and centering on the data set . . . 41
5 Conclusion . . . 44
References . . . 49
APPENDIX A. The trajectories obtained for the GBMT model with six clusters and quadratic trajectories fit on the raw data . . . 50

List of Figures

3.1 Normalized vibration measurements collected from X coordinate of the individual gearboxes over 3 load steps 24
3.2 Normalized measurements collected from Y coordinate of the individual gearboxes over 3 load steps 25
3.3 Normalized measurements collected from Z coordinate of the individual gearboxes over 3 load steps 26

4.1 Calinski-Harabasz scores obtained for each iterative run 29
4.2 The mean trajectories of the model with two clusters 29
4.3 The mean trajectories of the model with five clusters 30
4.4 Estimated trajectory curves and their 95% confidence intervals for the model with five clusters, polynomial degree = 1, using raw data . 34
4.5 Estimated trajectory curves and their 95% confidence intervals for the model with six clusters, polynomial degree = 1, using raw data . 35
4.6 Estimated trajectory curves and their 95% confidence intervals for the model with six clusters, polynomial degree = 2, using raw data . 37
4.7 Estimated trajectory curves and their 95% confidence intervals for the model with eight clusters, polynomial degree = 1, centering . . . 40
4.8 Estimated trajectory curves and their 95% confidence intervals for the model with nine clusters, polynomial degree = 2, centering 42

1 Trajectories obtained in the first cluster of the best model 50
2 Trajectories obtained in the second cluster of the best model 51
3 Trajectories obtained in the third cluster of the best model 52
4 Trajectories obtained in the fourth cluster of the best model 53
5 Trajectories obtained in the fifth cluster of the best model 54
6 Trajectories obtained in the sixth cluster of the best model 55

List of Tables

3.1 Number of the gearboxes and the state of the gearboxes 26

4.1 Number of gearboxes from each type in each cluster, model with two clusters 30
4.2 Number of gearboxes from each type in each cluster, model with five clusters 31
4.3 AIC and BIC scores of clustering outcomes for polynomial degree 1 and using raw data 33
4.4 Percentage of the gearboxes in each cluster in the model with five clusters, polynomial degree = 1, using raw data 34
4.5 Percentage of the gearboxes in each cluster in the model with six clusters, polynomial degree = 1, using raw data 35
4.6 AIC and BIC scores of clustering outcomes for polynomial degree 2 and using raw data 36
4.7 Percentage of the gearboxes in each cluster in the model with six clusters, polynomial degree = 2, using raw data 37
4.8 Average posterior probability of assignments for the model with six clusters, polynomial degree=2, using raw data 38
4.9 Number of gearboxes from each type in each cluster, model with six clusters, polynomial degree=2, using raw data 38
4.10 AIC and BIC scores of clustering outcomes for polynomial degree 1 and centering 39
4.11 Percentage of the gearboxes in each cluster in the model with eight clusters, polynomial degree = 1, centering 40
4.12 AIC and BIC scores of clustering outcomes for polynomial degree 2 and centering 41
4.13 Percentage of the gearboxes in each cluster in the model with nine clusters, polynomial degree = 2, centering 42
4.14 Average posterior probability of assignments for the model with nine clusters, polynomial degree=2, centering 43
4.15 Number of gearboxes from each type in each cluster, model with six clusters, polynomial degree=2, using raw data 43

1 Introduction

Sustainability is the key for the future of our world. Considering the increase in consumption and the amount of the reserves left, there is a risk of fossil fuels running out by 2060 (Burek 2010). Therefore, renewable energy sources such as solar energy and wind power gained more importance as the climate change has become inevitable, and governments all around the globe take actions to stop the degradation of the world ecosystem. According to Global Wind Report (GWEC) 2019, the global wind energy market grows by 2.7% each year. With the awareness and more investments in the field, this growth is expected to be 4% each year as described in Global Wind Report (GWEC) 2021. In such a fast-growing area, it is essential to increase the reliability of the products and lower the maintenance costs. In order to achieve these, precautions must be considered in subassemblies of the wind turbines such as gearboxes.

Gearboxes, which are commonly used in a variety of mechanical engineering operations, are widely used in the wind turbines taking into account their technical concept, their capability and their lifetime (Hau 2013). Their function in a wind turbine is to increase the rotational speed of a low-speed rotor and transfer it to a higher speed electrical generator. Due to their large size and critical location inside the turbine, the reliability of each gearbox has to be tested before they are mounted to the wind turbines, because their reparation is a major issue in terms of time and high costs. Moreover, gearboxes have the second highest downtime per failure after their electrical systems inside the wind turbine (Spinato et al. 2009). Therefore, it is crucial to pay attention to gearbox technology to avoid unexpected cost increases.

There are variety of research works on the subject of condition monitoring of the gearboxes after their installation on the wind turbines. For example, acoustic emission (AE) is defined as a construction technique for condition monitoring (Sharma and Parey 2017). Vibration analysis is another way to detect the faulty gearboxes (Salameh et al. 2018). It is possible to detect numerous failures and identify the location of the failure in the gearbox using the vibration measurements (Ojala 2013). Vibration analysis is also performed on the gearboxes before their installation to find out possible defects. However, clustering the gearboxes using their vibration measurements to investigate the hidden relationships is the main objective of this research, rather than the fault detection.

Cluster analysis is a widely used approach for different scientific areas such as image analysis, machine learning, information retrieval, and bioinformatics. Many different clustering algorithms have been developed since 1970s, when this topic gained its popularity (Bock 2008). Clustering algorithms are one of the fields under

unsupervised machine learning techniques, where there is no training set or label assigned on the data set under consideration, and in principle clustering methods can be divided into two main groups as non-parametric methods and parametric methods (Fung 2001). While non-parametric approaches use density-based methods and hierarchical algorithms, the parametric approaches focus on solving an optimization problem by attempting to minimize cost functions. The parametric approaches can be further categorized into two groups as generative or probability-based models and reconstructive models (Ahmed, Barua, and Begum 2020). Numerous algorithms have been developed over the years in both parametric and non-parametric approaches and these algorithms are used in many different fields such as psychology, sociology, and biology.

Finite mixture models, a subset of parametric clustering methods, have become more popular in the recent years due to the fact that these algorithms are capable of modeling the data sets that are not normally distributed and they evaluate cluster membership probabilities for each data point (Zhang and Huang 2015). In this study, group-based trajectory modeling, that is an application of finite mixture models, is used in clustering the gearbox data. In the basic approach of group-based trajectory modeling, a single parameter is taken into account in order to assign the data points into groups using polynomial regression. While the basic model considers univariate data sets, generalization of the basic model allows one to use the method for bivariate or multivariate data sets. Even if it is possible to extend the basic model of the group-based trajectory modeling to a joint trajectory model in theory, group-based multi trajectory modeling is preferred to avoid high-dimensional conditional probability tables and complex calculations (D. Nagin et al. 2016).

The collected data set consists of vibration measurements obtained from the vibration sensors that are placed on the high-speed shaft of the gearboxes. Each test run has been performed under three different load steps, and vibration measurements have been recorded in three coordinates: X, Y and Z. As a result, nine-dimensional data is obtained for each gearbox. Such data of repetitive measurements can be analysed as longitudinal data. In addition to group-based trajectory modeling, a special form of k-means clustering algorithm that is a reconstructive clustering method under the parametric approaches, k-means clustering for longitudinal data, has been used in this study. The version of this method that clusters multivariate data is named as KmL3D which follows the famous Expectation Maximization (EM) algorithm. In the expectation step, calculation of each cluster centers is performed and then each observation is allocated to its nearest cluster in the maximization step (Genolini et al. 2013). It is a well-known problem of EM algorithm that it might not always converge to the optimal solution. In order to overcome this problem, the

algorithm is run several times so that the initial cluster centers are changed to find the optimal solution. However, running the algorithm number of times still does not guarantee that the best clustering result, that is the optimal solution, is obtained.

The reason why group-based trajectory modeling is selected to be used as a probability-based parametric clustering method in this research is that group-based trajectory modeling has a different assumption for conditional independence of current outcomes from past outcomes than the most of other finite mixture models applied on longitudinal data. In other words, most of the finite mixture models utilized on longitudinal data assume that the current observations are independent from the past observations. However, group-based trajectory modeling has such an assumption only on latent trajectory group level but not conditionally independent on the population level because they depend on a latent variable that is the trajectory group membership, making the assumption of conditional independence more reliable. As the basic model of group-based trajectory modeling deals with univariate data, the generalization of the basic model is utilized in order to cluster the multivariate data consisting of all the vibration measurements from the high-speed shafts' X, Y, and Z coordinates. Since the data set under consideration is multivariate in this case, the used approach, the generalization of the basic model, is named as group-based multi-trajectory modeling (Magrini 2022).

In order to select the most accurate clustering result, the group-based multi-trajectory modeling algorithm has been run number of times, each time changing only one hyperparameter, either the number of the clusters or the degree of the polynomials, in the model. Afterwards, the accuracies of each model have been evaluated based on their BIC (Bayesian Information Criterion) and AIC (Akaike Information Criterion). As a consequence, the best results have been obtained by making use of six clusters with the second-degree polynomials as the parameters of the algorithm that is run on the multi-variate data. On the other hand, the results obtained with k-means algorithm arranged to handle longitudinal data, KmL3D, are evaluated based on the BIC and Calinski & Harabasz quality criterion and it has been observed that the best results are obtained with 5 clusters in the reconstructive case.

1.1 The focus of this thesis

The main aim of this study is on clustering the gearboxes based on the vibration measurements recorded during test runs in the factory environment. More specifically, vibration measurements taken from the high-speed shaft of the gearboxes have been used to perform the clustering. The secondary aim is to investigate the relationship between vibration measurements of a gearbox and its endurance. The collected data set is provided by Moventas Oy and it consists of vibration measurements from 337

individual gearboxes of a single product family.

In order to cluster the gearboxes, a special form of k-means clustering, KmL3D, and group-based trajectory modeling that is an application of finite mixture modeling is employed. Data analysis is done with R Programming (R Core Team 2013) using kml3d and gbmt packages, which are developed by (Christophe Genolini, Alacoque, et al. 2015), (Magrini 2021), respectively.

1.2 The structure of this thesis

The methods used for clustering the gearboxes according to the vibration measurements is explained detailly in the Chapter 2. The collected data is introduced in Chapter 3 and the application of the clustering algorithms on the collected data set and the results obtained are described in the Chapter 4. The conclusion and recommendations on future work are provided in the Chapter 5.

2 Method

There are variety of clustering algorithms. As introduced in the Chapter 1, clustering algorithms can be classified to two main groups as parametric approaches and non-parametric approaches, based on the way these algorithms work. As indicated in their names, parametric approaches make the assumption that population formed by subgroups have distinct distributions with their own parameters such as mean and variance. However, non-parametric approaches are conducted on distance-based metrics, and they do not presume about the distributions of the subgroups. On the other hand, a common feature both types of clustering algorithms have is that their purpose is to break down the data set into a number of disjoint clusters. To achieve this goal, an initial number of clusters is needed for initiating both k-means clustering algorithm and group-based trajectory modeling, that are the examples of parametric methods utilized in this thesis, respectively. Details of both algorithms are explained in this chapter.

A famous and basic reconstructive parametric clustering algorithm, k-means clustering, which has many different implementation options, is described in Section 2.1. Sections 2.1.1 and 2.1.2 describe the application of k-means on longitudinal data with single-trajectory and longitudinal data with joint-trajectories, respectively. Also, as stated in Chapter 1, another method used in this research is group-based trajectory modeling, that is a specialized form of finite mixture models (D. Nagin 2014) and explained in detail in Section 2.2.

2.1 K-means clustering

K-means is a widely used popular clustering algorithm across several fields of science due to its various advantages such as higher computational speed compared to the other algorithms and being applicable to continuous data. The aims of the algorithm are to obtain data points that are similar and can be put into the same clusters and to make sure that different clusters are further apart. K-means, being a type of hill-climbing algorithms, works in a way that a partition is repeatedly changed by moving each object to the group which has the closest mean and then calculating the group mean afterwards. Assuming some specifications, it can be shown that this algorithm handles clustering by minimizing trace(W), the trace of within-clusters covariance matrix, when the measure of the similarity is determined as a pre-defined distance calculation method such as Euclidean distance, Chebychev distance, Manhattan distance and Minkowski distance (Singh, Yadav, and Rana 2013) according to the data specific needs. These algorithms, including the computation of the centroids of

all clusters and iterative cluster reassignment of points and recomputation of cluster means are generally referred to as k-means clustering algorithms (Everitt 2011). Also, belonging to the Expectation Maximization class of clustering algorithms, k-means obtains optimal partition by alternating between two stages which are expectation and maximization. These steps can be summarized as determining the centre of each cluster and assigning the observation to its nearest cluster, respectively (Christophe Genolini and Falissard 2011).

k-means algorithm that uses Euclidean distance as the measure of similarity is explained below (Singh, Yadav, and Rana 2013):

Consider $X = \{x_1, x_2, ..., x_n\}$ as the data set and $V = \{v_1, v_2, ..., v_c\}$ as the set of centres.

1. Randomly select cluster centers from the data points for the given number c of clusters.

2. Obtain the Euclidean distance between each data point and cluster centres as presented in the equation 2.1 below where x_i is a data point, x_j is a randomly selected cluster centre, and m is the dimension:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2}, \tag{2.1}$$
$$k = 1, \ldots, m$$
$$j = 1, \ldots, c$$

3. Assign the data point to the cluster whose centre's distance to the data point has the minimum Euclidean distance among all the cluster centres.

4. Calculate the new cluster centres as shown in the equation 2.2 where v_i is a cluster centre:

$$v_i = \left(\frac{1}{c_i}\right) \sum_{j=1}^{c_i} x_j \tag{2.2}$$

   Where the number of data points in the i^{th} cluster is denoted as 'c_i' and x_j is the data point assigned to the i^{th} cluster.

5. Recalculate the distance between each data point and cluster centres that are newly obtained.

6. Stop if there is no reassignment of data points. Else, repeat from step 3 to step 5.

2.1.1 KmL

Since the data set used in this thesis work consists of repeated measurements from the same individuals with three different load steps, it requires to employ an implementation of the k-means algorithm which can deal with longitudinal data. KmL is an R package that enables applying the k-means clustering algorithm to analyse the longitudinal data sets. For instance, KmL offers different options to handle the problem of missing values that can be very common among studies with longitudinal data. However, the collected data set for this research does not have any missing data points and dealing with missing data is kept out of scope of this work.

Another advantage KmL provides is that this implementation of the k-means overcomes the common concern of the predefined number of clusters issue concerning k-means algorithm by running the algorithm several times by changing the initial number of seeds (clusters) and then selecting the best suitable number of clusters according to variety of quality criteria such as Calinski & Harabasz criterion, Ray & Turi criterion and Davies & Bouldin criterion. Also, in order to overcome the well-known weakness of the k-means algorithm, which is converging to a local maximum and not providing the best possible partition concerning the homogeneity of the hill-climbing algorithms, there are several starting conditions and run options that the users can set.

2.1.2 KmL3D

KmL3D is an R software which provides an implementation of k-means algorithm that can cluster several variable trajectories, that are also known as joint trajectories, jointly (Genolini et al. 2013) and it can be considered as a generalization of KmL algorithm that is explained in the Section 2.1.2. While KmL can be used to cluster single variable trajectories, KmL3D can deal with joint variable trajectories (Christophe Genolini, Alacoque, et al. 2015).

Having X, Y, and Z dimensions along with the measurements obtained by repeated load steps direct this research to employ this specific form of the K-means algorithm which is explained below (Christophe Genolini, Alacoque, et al. 2015):

1. Initialization of parameters
2. Loop: i=1
    2.1. Loop: k=1 First of K
        2.1.1. Seeds generation using various starting conditions
        2.1.2. K-means with k clusters
        2.1.3. Calculation of various quality criteria

2.1.4. partition storage

2.1.5. k = Next in K

2.1.6. Until: All k in K have been dealt with

2.1.7. Repeat from step 2.1.1 until the condition in step 2.1.6 is met

2.2. i = i+1

3. Until: i > number of runs

4. Repeat from step 2 until the condition in step 3 is met

Assuming that M outcomes that are the X, Y, Z vibration measurements in this study, $Y_{..A}, Y_{..B}, \ldots, Y_{..M}$ are measured at T different times that are the three different load steps from n experimental units that are the individual gearboxes. y_{ijA} denotes the value of $Y_{..A}$ at the j^{th} time instance for the i^{th} experimental unit. Then, $y_{i.A} = (y_{i1A}, y_{i2A}, \ldots, y_{iTA})$ is known as a single trajectory, and several single trajectories are called joint trajectories which are denoted as $y_{i..} = (y_{i.A}, y_{i.B}, \ldots, y_{i.M})^T$. To sum up, $y_{i..}$ is a matrix whose rows are the single variable trajectories and it is presented in the equation 2.3 below:

$$y_{i..T\times M} = \begin{pmatrix} y_{i1A} & y_{i2A} & \cdots & y_{iTA} \\ y_{i1B} & y_{i2B} & \cdots & y_{iTB} \\ \cdots & \cdots & \cdots & \cdots \\ y_{i1M} & y_{i2M} & \cdots & y_{iTM} \end{pmatrix} \tag{2.3}$$

For a fixed j, $y_{ij.} = (y_{ijA}, y_{ijB}, \ldots, y_{ijM})^T$ represents the state of the i^{th} experimental unit at the j^{th} time instance. Therefore, it must be noted that the experimental unit's state at time j is the j^{th} column of $y_{i..}$ which is given in the equation 2.3 above (Genolini et al. 2013).

In order to utilize K-means clustering algorithm, the distance between the joint trajectories of two experimental units, $y_{1..}$ and $y_{2..}$ must be found out. That distance is basically the distance between two matrices, and there are various methods that could be applied to calculate it. Two of those methods are as the followings (Genolini et al. 2013):

1. The distance between rows of the two matrices can be found and then a function to combine these distances can be used.

2. The distance between columns of the two matrices can be identified and then a function can be used to find the resulting distance value.

That function to combine the row-wise or column-wise obtained vector of distances can be simply the norm of the distance vector. Then, using the first method,

the vector of Euclidean distances between $y_{1..}$ and $y_{2..}$ is found as given in the equation 2.4 below:

$$\begin{pmatrix} d_1(y_{1..}, y_{2..}) \\ d_2(y_{1..}, y_{2..}) \\ \vdots \\ d_M(y_{1..}, y_{2..}) \end{pmatrix} = \begin{pmatrix} \sqrt{\sum_{t=1}^{T}(y_{1tA} - y_{2tA})^2} \\ \sqrt{\sum_{t=1}^{T}(y_{1tB} - y_{2tB})^2} \\ \vdots \\ \sqrt{\sum_{t=1}^{T}(y_{1tM} - y_{2tM})^2} \end{pmatrix} \tag{2.4}$$

The final distance between $y_{1..}$ and $y_{2..}$, $Dist(y_{1..}, y_{2..})$, is found by taking the norm of the vector $[d_1(y_{1..}, y_{2..}), d_2(y_{1..}, y_{2..}), \ldots, d_M(y_{1..}, y_{2..})]$ as provided in the equation 2.5 below:

$$Dist(y_{1..}, y_{2..}) = ||d_1(y_{1..}, y_{2..}), d_2(y_{1..}, y_{2..}), \ldots, d_M(y_{1..}, y_{2..})|| \tag{2.5}$$

A common problem in utilizing the k-means clustering is that one or more outcomes in the data-set are dominating the distance values due to different unit scales which might cause inaccurate clustering results. Normalization is a method that is widely used in order to overcome this problem. In KmL3D package, the option of normalization is offered, with a difference with the classic normalization in which the normalization is performed for entirety of each variable-trajectory, not for each variable-trajectory at each time. Denoting the mean and standard deviation of all $y_{ij}^{(A)}$ for each i and j with $\overline{Y^{(A)}}$ and $sd^{(A)}$, respectively, the normalisation is done as shown in the equation 2.6:

$$y_{ij}^{\prime(A)} = \frac{(y_{ij}^{A} - \overline{Y^{(A)}})}{sd^{(A)}} \tag{2.6}$$

The normalized joint trajectory of the i^{th} experimental unit, $y_{i..}^{'}$, is found by normalizing its single trajectories, $y_{i.}^{\prime(A)}$ one by one (Genolini et al. 2013).

2.1.3 Model selection in KmL3D

In order to apply the k-means algorithm on the collected data set, there are many options available for selecting the initial configurations. Among the seven available choices offered by $KmL3D$ R package (Christophe Genolini, Alacoque, et al. 2015), the $kmeans--$ is the chosen one due to its advantages against the remaining six initialization configurations such as guaranteeing the cluster centers being distant from one another and having low time cost. The logic followed in this configuration is as follows:

1. Choose a center c_0 from the data points randomly and uniformly.

2. Calculate the distance between each data point, x and c_0, $D(x)$.

3. Select a new center, c_1, randomly using a weighted probability distribution in proportion to $D(x)^2$.

4. Delete c_0 from the centers' list.

5. Start a loop for distance calculation and cluster center selection:

    5.1. Calculate the distance between data point x and the closest center that has already been selected, $D(x)$, for each point.

    5.2. Choose the new center c_i from the data points at random, using a weighted probability distribution to select a point x with probability in proportion to $D(x)^2$.

    5.3. Repeat from 5.1. to 5.2. until k number of centers are selected.

In this thesis work, the used distance metric is Euclidean distance. In order to prevent the algorithm from converging to local maxima, the algorithm is run several times. Because of the fact that the starting point changes, the convergence might be to a different maximum. In the end, it is expected that the global maximum, that is the best partition, is reached. However, it must be noted that there is still no certainty for the partition optimality (Christophe Genolini, Alacoque, et al. 2015).

In this research, Calinski & Harabasz quality criterion has been used in order to evaluate the model adequacy, and it is formulated in the equation 2.7 below :

$$C(k) = \frac{trace(B)}{trace(W)} \left(\frac{n-k}{k-1} \right) \tag{2.7}$$

Where trace(B) is trace of the between-clusters covariance matrix B, trace(W) is trace of within-clusters covariance matrix, n is the number of trajectories and k is the number of clusters (Christophe Genolini and Falissard 2011).

2.2 Finite mixture models

Unlike the non-parametric approaches, parametric clustering methods generally try to minimize a cost function by utilizing an optimization problem (Fung 2001). To do so, parametric clustering algorithms need more detailed knowledge and assumptions about the data distribution (Ahmed, Barua, and Begum 2020). Finite mixture modeling, an example of parametric clustering, has gained popularity in the recent years.

The reasons why finite mixture modeling has become to be widely used are that it can deal with longitudinal data and enables the removal of the homogeneity assumption on the data. In other words, it does not assume that the observations

collected are from the same distribution and it also considers the fact that the sample in hand can be associated to mixture of populations which are not determined (Zhang and Huang 2015).

To elaborate, a finite mixture model is a statistical model of the mixture of populations from which the data is collected, and it makes use of the finite mixture densities that are the multivariate density functions of each different subpopulations that are also known as clusters (Everitt 2011).

In order to write down the basic form of finite mixture model, let $Y_1, Y_2, Y_3, ..., Y_n$ denote a random sample of size n, where Y_j is a p dimensional random vector with the probability density function $f(y_j)$. In other words, Y_j has the random variables of p measurements of the j^{th} recording of some attributes on the topic of interest. To be consistent with (McLachlan 2000), a random vector is represented by a lower-case letter while the entire sample is denoted with upper-case letter. For instance, an observed random sample of the entire sample $Y = (Y_1^T, ..., Y_n^T)^T$ is denoted with $y = (y_1^T, ..., y_n^T)^T$. Then, the basic form of a finite mixture model with g number of components is given in the equation 2.8 where $f_i(y_i)$ denotes the density function corresponding to the i^{th} mixture component:

$$f(y_j) = \sum_{i=1}^{g} \pi_i f_i(y_j) \tag{2.8}$$

Where π_i is the i^{th} nonnegative mixing proportion, or weight, and it shall be noted that the summation of π_1 to π_g is equal to one, as stated in the equations 2.9 and 2.10, respectively.

$$0 \leq \pi_i \leq 1, \qquad i = 1, ...g \tag{2.9}$$

$$\sum_{i=1}^{g} \pi_i = 1 \tag{2.10}$$

2.2.1 Maximum likelihood fitting and Expectation Maximization Algorithm

Since the introduction of the expectation maximization (EM) algorithm (Dempster, Laird, and Rubin 1977), maximum likelihood (ML) has gained importance and has been the most used approach for fitting the mixture distributions (McLachlan 2000). To summarize, the focus of maximum likelihood estimation is to find an estimate, $\hat{\Psi}$ for the parameter vector Ψ in a postulated density $f(y_j; \Psi)$ for the random vector Y_j linked to the j^{th} recording on the topic of interest so that the likelihood function is maximized for the estimated parameter vector, $\hat{\Psi}$, that is

provided in regular situations by an appropriate solution of the likelihood equation. A common approach for obtaining the estimated parameter vector, $\hat{\Psi}$, maximizing the likelihood function is to find the zero-point of the derivative of the likelihood function with respect to the parameter vector, Ψ, as provided in the equation 2.11:

$$\frac{\partial L(\Psi)}{\partial \Psi} = 0 \tag{2.11}$$

Or alternatively the log-likelihood equation 2.12:

$$\frac{\partial log L(\Psi)}{\partial \Psi} = 0 \tag{2.12}$$

Note that $L(\Psi)$ is the likelihood function for Ψ and it requires the assumption of mutual independence between the vectors $y_1, ..., y_n$, and it is interpreted as the following equation 2.13:

$$L(\Psi) = \prod_{j=1}^{n} f(y_j; \Psi) \tag{2.13}$$

Within a mixture model consisting of g components, one can write the probability density function for the random vector Y_j in parametric form as the following equation 2.14:

$$f(y_j; \Psi) = \sum_{i=1}^{g} \pi_i f_i(y_j; \theta_i) \tag{2.14}$$

Where the unknown parameter vector is represented with θ_i in the postulated form of the i^{th} component density in the mixture. Then, the vector ξ of all the parameters in $\theta_1, ..., \theta_g$ can be defined and finally, Ψ denotes the parameter vector of all the unknown parameters of the mixture model as given in the following equation 2.15, considering the fact that one of the mixing proportions is redundant as they sum to one (McLachlan 2000):

$$\Psi = (\pi_1, ..., \pi_{g-1}, \xi^T)^T \tag{2.15}$$

It is very common in most of mixture models that the solution of the likelihood equation has more than one root that are resulting in the local maxima. In general, finding the root that is resulting with the global maximum of the likelihood function is extremely hard and time consuming. On the other hand, utilization of EM algorithm allows one to obtain the solutions of the likelihood equation that are associated with the local maxima values of the likelihood function (McLachlan 2000) However, EM algorithm does not allow to find all the local maxima, rather it converges to one of them.

In the formulation of EM algorithm, the collected data, $y = (y_1^T, ..., y_n^T)^T$ is

considered as incomplete because of the fact that group indicator vectors $z_1, z_2, ..., z_n$ are not realized. Here, y_j is a component of the mixture model presented in the equation 2.8 and z_j is the g dimensional group indication vector and it is noted as $z_{ij} = (z_j)_i$ equal to 1, if y_j comes from the i^{th} component of the mixture ($i = 1, ..., g$ and $j = 1, ..., n$), and 0 if not. Then, the complete data, y_c, can be formulated as stated below in the equation 2.16:

$$y_c = (y^T, z^T)^T \tag{2.16}$$

Where

$$z = (z_1^T, ..., z_n^T)^T \tag{2.17}$$

The group indication vectors $z_1, ..., z_n$ are to be found values of the random vectors $Z_1, ..., Z_n$ that are assumed to be distributed unconditionally according to the multinomial distribution as described in the equation 2.18 below:

$$(Z_1, ..., Z_n) \overset{\text{i.i.d.}}{\sim} Mult_g(1, \pi) \tag{2.18}$$

Where $\pi = (\pi_1, \ldots, \pi_g)^T$ (McLachlan 2000).

With the help of the assumption above, the appropriate distribution for the incomplete data vector, Y, is implied by the distribution of the complete data vector Y_c. Then, one can write the complete data log likelihood for Ψ, $logL_c(\Psi)$, as stated in the equation 2.19:

$$logL_c(\Psi) = \sum_{i=1}^{g} \sum_{j=1}^{n} z_{ij} \{log\pi_i + logf_i(y_j; \theta_i)\} \tag{2.19}$$

EM algorithm is an iterative computation where each iteration contains two steps that are named as expectation step, or E-step, that goes before the maximization step, or M-step (Dempster, Laird, and Rubin 1977) and it is used considering z_{ij} as the missing data, whose distribution is computed in E-step and then the M-step optimizes the expected value of the complete-data log-likelihood over that distribution, which takes the conditional expectation of $logL_c(\Psi)$, $Q(\Psi; \Psi^{(t)})$, log-likelihood of the complete data, given y, the observed data, that uses the current of Ψ as written in the equation 2.20 below (McLachlan 2000):

$$Q(\Psi; \Psi^{(t)}) = E_{\Psi^{(t)}}\{logL_c(\Psi)|y\} \tag{2.20}$$

Where $\Psi^{(t)}$ is the value of Ψ after the t^{th} EM iteration. One can initialize the algorithm by setting $\Psi^{(0)}$ as the initial value for Ψ. As a result, the equation 2.20 can be rewritten for the first iteration as the equation 2.21:

$$Q(\Psi;\Psi^{(0)}) = E_{\Psi^{(0)}}\{logL_c(\Psi)|y\} \tag{2.21}$$

Then, on the t^{th} iteration, the posterior probability of j^{th} sample belonging to the i^{th} component of the mixture, can be obtained with the following equation 2.22:

$$E_{\Psi^{(t)}}(Z_{ij}|y) = pr_{\Psi^{(t)}}\{Z_{ij}=1|y\} = \tau_i(y_j;\Psi^{(t)}) \tag{2.22}$$

Where $\tau_i(y_j;\Psi^{(t)})$ represents the posterior probability that y_j, j^{th} member of the sample with the observed value, belongs to the mixture's i^{th} component and it is given in the equation 2.23 below:

$$\tau_i(y_j;\Psi^{(t)}) = \frac{\pi_i^{(t)} f_i(y_j;\theta_i^{(t)})}{\sum_{h=1}^{g}\pi_h^{(t)} f_h(y_j;\theta_h^{(t)})} \tag{2.23}$$

For $i = 1, ..., g$ and $j = 1, ..., n$

Finally, one can write the conditional expectation of complete data log-likelihood for Ψ given y by using the equations 2.18, 2.19 and 2.20 as shown in the equation 2.24 below.

$$Q(\Psi;\Psi^{(t)}) = \sum_{i=1}^{g}\sum_{j=1}^{n}\tau_i(y_j;\Psi^{(t)})\{log\pi_i + logf_i(y_j;\theta_i)\} \tag{2.24}$$

In the M-step global maximization of $Q(\Psi;\Psi^{(t)})$ with respect to Ψ over the parameter space, Ω, is required in $(t+1)^{th}$ iteration to provide the updated estimate, $\Psi^{(t+1)}$.

It should be noted that the updated weights, $\pi_i^{(t+1)}$, are updated independently from the updated estimate for the vector of unknown parameters,$\xi^{(t+1)}$.

If z_{ij} were not missing, the complete data maximum likelihood estimation of π_i would easily be obtained by using the following equation 2.25:

$$\hat{\pi}_i = \sum_{j=1}^{n}\frac{z_{ij}}{n} \qquad\qquad (i = 1, ..., g) \tag{2.25}$$

As z_{ij} are missing from the data, the new estimate of π_i is obtained by replacing each z_{ij} with their current conditional expectation $\tau_i(y_j;\Psi^{(t)})$ in the complete data log likelihood at the E-step (McLachlan 2000) as stated in the equation 2.26 below:

$$\pi_i^{t+1} = \sum_{j=1}^{n}\frac{\tau_i(y_j;\Psi^{(t)})}{n} \qquad\qquad (i = 1, ..., g) \tag{2.26}$$

It is important to observe from the equation 2.26 above that each estimate for π_i at the $(t+1)^{th}$ iteration is calculated by making use of currently obtained posterior probability of each sample y_j belonging to i^{th} component of the mixture model.

In the M-step of $(t+1)^{th}$ the equation 2.24 can be used to find $\xi^{(t+1)}$ by solving its partial derivative with respect to ξ and equating to zero as done in the equation 2.27

$$\sum_{i=1}^{g}\sum_{j=1}^{n}\frac{\tau_i(y_j;\Psi^{(t)})\partial\{log\pi_i + logf_i(y_j;\theta_i)\}}{\partial\xi} = 0 \tag{2.27}$$

It is shown in (Dempster, Laird, and Rubin 1977) that each likelihood value that is obtained in the $(t+1)^{th}$ iteration is larger than or equal to the value obtained in the t^{th} iteration ($L(\Psi^{(t+1)}) \geq L(\Psi^{(t)})$) and the algorithm is run iteratively until the convergence is achieved, i.e., until the difference between two consecutive likelihood values, $L(\Psi^{(t+1)}) - L(\Psi^{(t)})$, becomes negligible (McLachlan 2000).

2.2.2 Group-based trajectory modeling

One of the special applications of finite mixture models is group-based trajectory modeling that focuses on clustering the experimental units within the data under study. The estimated parameters of the model are obtained as a result of maximum likelihood estimation where constructing the likelihood function depends on the research data (D. S. Nagin 2005).

As stated earlier, the data used in group-based trajectory modeling shall be in the form of longitudinal sequence of measurements in i^{th} experimental unit over T periods, that can be denoted as $Y_i = \{y_{i1}, y_{i2}, ..., y_{iT}\}$ with y_{it} being the behaviour of an experimental unit and $P(Y_i)$ being the unconditional probability of observing the i^{th} experimental unit's longitudinal data of behaviour, Y_i. The main purpose is to estimate a parameter set, Ω, , that maximizes the probability of Y_i independent from the form of the probability distribution it has. These estimated parameters play the role in defining the group membership probabilities and trajectory shapes that are characterized by polynomial functions (D. S. Nagin 2005).

The group-based trajectory modeling method assumes that a finite set of distinct polynomials can encapsulate the individual differences in the trajectories. Each set represents a trajectory group which is denoted with the index j from now on. Since the group membership is not observed, it is not possible to sort the experimental units by their group membership and estimate their trajectories with readily available software packages. Hence, it is a necessity to estimate the probability of a member of the population, that is randomly chosen, belonging to j^{th} group, π_j. As a result, J conditional likelihood functions are aggregated to construct the likelihood function, $P^j(Y_i)$, in order to form the unconditional probability of Y_i, the data, that is given by the equation 2.28 below (D. S. Nagin 2005):

$$P(Y_i) = \sum_{j}^{J} \pi_j P^j(Y_i) \tag{2.28}$$

Since $P(Y_i)$ is the summation of a finite number of different groups which are forming the population in the equation 2.28, it is in the form of a "finite mixture model". It is assumed that the sequential realizations of the units in Y_i, y_{it} are conditionally independent over the T repetitive measurements for a given j. Given membership in group j, denoting the probability distribution function of y_{it}, the sequential realizations of the units in Y_i with $p^j(y_{it})$ at the t^{th} measurement, the conditional likelihood function can be written down as given in the equation 2.29 below (D. S. Nagin 2005):

$$P^j(Y_i) = \prod_{t}^{T} p^j(y_{it}) \tag{2.29}$$

Finally, the likelihood function for the entire sample with size N, L, is given in the equation 2.30 below, as a product of the likelihood functions of each experimental unit, that is provided in the equation 2.28 above:

$$L = \prod_{i}^{N} P(Y_i) \tag{2.30}$$

As mentioned earlier, the specific form of the likelihood function that is given in the equation 2.30 above changes depending on each data type (D. S. Nagin 2005).

2.2.3 Generalization of group-based trajectory modeling to multi-variate data

In most of the cases, the data collected from the experimental units included in the research is multivariate. Thus, generalization of group-based trajectory modeling for univariate data is required in order to make use of the provided information as a whole. If the generalization is done by following the idea of using "joint-trajectory" modeling, the heterogeneity in the linkage between the trajectories could be highlighted. However, a disadvantage of such phenomenon is that it requires tables of conditional probabilities and the number of the tables needed is magnified as the number of variables in the data increases. For instance, two tables are required for bivariate data but three tables are necessary for a data with three outcomes at a time and a data with four outcomes is in need of six tables to construct the model. In order to avoid using many conditional probability tables and such complex calculations, multi-trajectory modeling is designed by defining a trajectory group as trajectories for multiple outcomes (D. Nagin et al. 2016).

In order to be convenient with the notation established in Section 2.2.2, let Y_i denote the i^{th} experimental unit's observed measurements and the vector T_i represent the time for each measurement taken from the i^{th} experimental unit. Hence, $P(Y_i|T_i)$ represents the distribution of outcome trajectories. Then, the likelihood of i^{th} experimental unit conditional on the J group number can be written as in the equation 2.31 below.

$$P(Y_i|T_i) = \sum_{j=1}^{J} \pi_j \cdot P(Y_i|T_i, j; \beta_j) \tag{2.31}$$

In the equation 2.31 above, π_j represents the membership probability of the i^{th} experimental unit in group j and the vector of parameters specifying the group-specific trajectory's shape is denoted with β_j. In general, the shape of the trajectory is modelled by a polynomial function of time.

Again, with the same assumption of conditional independence at level of the latent trajectory group for sequential realizations of y_{it}, that are the elements of Y_i, over the T repetitive measurements for a given j, one can write the conditional distribution of Y_i given membership in j^{th} group with index β_j can be written as provided in the equation 2.32 below (D. Nagin et al. 2016):

$$P(Y_i|T_i, j; \beta_j) = \prod_{t=i}^{T} p(y_{it}|T_{it}, j; \beta_j) \tag{2.32}$$

The notation p(.) is utilized to denote the conditional distribution of elements in Y_i, specifically pertaining to the j^{th} group membership and the measurement acquired from i^{th} experimental unit at time t.

In the framework of multi-trajectory model, each J trajectory group is distinguished by a set of trajectories that collectively depict multiple outcomes. The random vector Y_i^k signifies the sequential measurements for the k^{th} outcome obtained from the i^{th} experimental unit. Furthermore, $P_k(Y_i^k|T_i, j; \beta_j^k)$ denotes the distribution of Y_i^k conditioned on group j and parameter vector β_j^k that is unknown. It is important to note that the number of outcomes are represented with K. The assumption made is that the joint distribution of $Y_i^1, Y_i^2, \ldots, Y_i^K$, denoted as $P_j(Y_i^1, Y_i^2, \ldots, Y_i^K)$, can be factorized into the product of individual distributions for each outcome Y_i^k. This can be expressed as $P_j(Y_i^1, Y_i^2, \ldots, Y_i^K) = P_{1j}(Y_i^1)P_{2j}(Y_i^2)\ldots P_{Kj}(Y_i^K)$. Essentially, this assumption implies that the variables Y_i^k are distributed independently, conditioned on membership in group j. Finally, the likelihood for each experimental unit conditional on J group number can be written as given in the equation 2.33 below and it can be seen from the second part of the equation 2.33 expresses the assumption of conditional independence (D. Nagin et al. 2016):

$$P(Y_i^1, Y_i^2, \ldots, Y_i^K|T_i) = \sum_{j=1}^{J} \pi_j \left[\prod_{k=1}^{K} P_k(Y_i^k|T_i, j; \beta_j^k) \right] with$$
$$P_k(Y_i^k|T_i, j; \beta_j^k)] = \prod_{t=1}^{T^k} p_k(y_{it}^k, j; \beta_j^k) \tag{2.33}$$

As stated earlier, $i = 1, \ldots, n$ represents the i^{th} experimental unit, $t = 1, \ldots, T$ the time, and $k = 1, \ldots, K$ the k^{th} outcome in the sequential measurement from the the i^{th} experimental unit. In addition, y_{it} are the elements in Y_i. Then, the collected data can be denoted as given in the equation 2.34 below (Magrini 2022):

$$Y_i = \begin{pmatrix} y_{i1}^T \\ y_{i2}^T \\ \cdots \\ y_{it}^T \\ \cdots \\ y_{iT}^T \end{pmatrix} = \begin{pmatrix} y_{i1}^1 & \cdots & y_{i1}^k & \cdots & y_{i1}^K \\ y_{i2}^1 & \cdots & y_{i2}^k & \cdots & y_{i2}^K \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ y_{it}^1 & \cdots & y_{it}^k & \cdots & y_{it}^K \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ y_{iT}^1 & \cdots & y_{iT}^k & \cdots & y_{iT}^K \end{pmatrix} \tag{2.34}$$

Moreover, $j = 1, \ldots, J$ denote the groups and G_i represents a latent variable that takes the value j, if the i^{th} experimental unit belongs to j^{th} group. Different than the model that (D. Nagin et al. 2016) developed, in the group-based multi-trajectory (GBMT) model, that is developed in (Magrini 2022), with polynomial degree of d, it is assumed that the multivariate outcome observations at time t on i^{th} experimental unit is produced by one multivariate normal (MVN) polynomial regression for each group $j = 1, \ldots, J$, that removes the assumption of conditional independence as shown in the equation 2.35 below (Magrini 2022):

$$y_{it}|G_i = j \sim MVN(\mu_{jt}, \Sigma_j)$$
$$\mu_{jt} = B_j^T(1, t, t^2, \ldots, t^d)^T \tag{2.35}$$

Where B_j is a $(d+1) \times K$ matrix for the regression coefficients in the j^{th} group as can be seen in the equation 2.36, and Σ_j is the covariance matrix of the outcomes in the group j which is a square matrix with K dimensions. Combining all these, the likelihood function of the model can be written as stated in the equation 2.37 below (Magrini 2022):

$$B_j = \begin{pmatrix} \beta_{j0}^1 & \dots & \beta_{j0}^k & \dots & \beta_{j0}^K \\ \beta_{j1}^1 & \dots & \beta_{j1}^k & \dots & \beta_{j1}^K \\ \dots & \dots & \dots & \dots & \dots \\ \beta_{jd}^1 & \dots & \beta_{j0}^k & \dots & \beta_{j0}^K \end{pmatrix} \tag{2.36}$$

$$L(B_1, \dots, B_J, \Sigma_1, \dots, \Sigma_J, \pi_1, \dots, \pi_J) = \prod_{i=1}^{n} \left[\sum_{j=1}^{J} \pi_j \prod_{t=1}^{T} \phi(y_{it}|B_j, \Sigma_J) \right] \tag{2.37}$$

Where $\phi(y_{it}|B_j, \Sigma_J)$ represents the multivariate normal density for y_{it} in j^{th} group and π_j is the probability of membership of the i^{th} experimental unit in group j .

Normalisation and standardization are usually applied in order to find out the groups fully based on the trend and also to allow the comparison between different outcomes, i.e. the indicators or the different attributes of experimental units, that are included in the data. As a result, the collected data is made unitless and the groups are determined homogeneously in trend, not mainly by the outcomes whose levels are dominating the sample variability (Magrini 2022).

Four different normalization methods; centering, standardization, ratio to the mean and logarithmic ratio to the mean are provided below in the equations 2.38, 2.39, 2.40, 2.41, respectively.

$$y_{it}^{k\,*} = y_{it}^k - \bar{y}_i^k \tag{2.38}$$

$$y_{it}^{k\,*} = \frac{y_{it}^k - \bar{y}_i^k}{s_i^k} \tag{2.39}$$

$$y_{it}^{k\,*} = \frac{y_{it}^k}{\bar{y}_i^k} \tag{2.40}$$

$$y_{it}^{k\,*} = log\left(\frac{y_{it}^k}{\bar{y}_i^k}\right) \approx \frac{y_{it}^k - \bar{y}_i^k}{y_i^k} \tag{2.41}$$

Where $y_{it}^{k\,*}$ denotes the the k^{th} outcome obtained at the time t on the i^{th} experimental unit, $\bar{y}_i^k$ denotes the sample mean of the k^{th} outcome obtained at all times on the i^{th} experimental unit, and finally s_i^k denotes the standard deviation of the k^{th} outcome obtained at all times on the i^{th} experimental unit. Note that $y_{it}^{k\,*} = y_{it}^k$ without any normalisation (Magrini 2021). On the other hand, it is inappropriate to standardize the longitudinal data since the outcomes are obtained by taking the measurements repeatedly over time (Heggeseth and Jewell 2018).

The parameters of the likelihood function that is given in the 2.37, π_j, B_j and

Σ_j are obtained by running the EM algorithm that is explained in detail in the Section 2.2.1 until convergence is achieved for each group $j = 1, \ldots, J$. One should remember that the EM algorithm described in the Section 2.2.1 solves a problem of missing data that are started from random values. The posterior probability of each group $j = 1, \ldots, J$ for each experimental unit is calculated in the expectation step as shown in the equation 2.42 where the hat symbol represents a current estimate of a parameter (Magrini 2022):

$$Pr(G_i = j|Y_i) \equiv \pi_{ij} = \frac{\hat{\pi}_j \prod_{t=1}^{t} \phi(y_{it}|\hat{B}_j, \hat{\Sigma}_j)}{\sum_{j=1}^{J} \hat{\pi}_j \prod_{t=1}^{t} \phi(y_{it}|\hat{B}_j, \hat{\Sigma}_j)} \tag{2.42}$$

Maximum likelihood estimate of parameters given posterior probabilities are obtained in the maximization step. Obtaining $\hat{\pi_j}$, the maximum likelihood of prior probability of j^{th} group is found by taking the average of posterior probabilities of of j^{th} group as stated in the equation 2.43 below (Magrini 2022):

$$\hat{\pi}_j = \frac{1}{n} \sum_{i=1}^{n} Pr(G_i = j|Y_i) = \frac{1}{n} \sum_{i=1}^{n} \pi_{ij} \tag{2.43}$$

The maximum likelihood estimation of parameters B_j and Σ_j, $\hat{B}_j$ and $\hat{\Sigma}_j$ respectively, are obtained as given in the equation 2.44 below, through regression for each group $j = 1, \ldots, J$ by making use of the equation 2.35 and weighted least squares estimation method where posterior probabilities of j number of groups, $\hat{\pi}_j$, are used as the weights (Magrini 2022) :

$$\begin{aligned} \hat{B}_j &= \left(\boldsymbol{X^T W X}\right)^{-1} \boldsymbol{X^T W Y} \\ \hat{\mu}_{jt} &= \hat{B}_j^T (1, t, t^2, \ldots, t^d)^T \\ \hat{\Sigma}_j &= \frac{\sum_{i=1}^{n} \pi_{ij} \sum_{t=1}^{T} (y_{it} - \hat{\mu}_j t)(y_{it} - \hat{\mu}_j t)^T}{\sum_{i=1}^{n} \pi_{ij}} \end{aligned} \tag{2.44}$$

With X, W, and Y as defined in the equation 2.45 below:

$$\boldsymbol{X}_{(nT)\times(d+1)} = \begin{pmatrix} \boldsymbol{x_1} \\ \cdots \\ \boldsymbol{x_i} \\ \cdots \\ \boldsymbol{x_n} \end{pmatrix}$$

$$
\begin{aligned}
\boldsymbol{x_i}_{T\times(d+1)} &= \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & 2 & 2^2 & \dots & 2^d \\ \dots & \dots & \dots & \dots & \dots \\ 1 & t & t^2 & \dots & t^d \\ \dots & \dots & \dots & \dots & \dots \\ 1 & T & T^2 & \dots & T^d \end{pmatrix} \\
\boldsymbol{Y}_{nT\times K} &= \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_i \\ \dots \\ Y_n \end{pmatrix} \\
W_{nT\times nT} &= \begin{pmatrix} \mathbf{1}_{T\times T}\pi_{1j} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \mathbf{1}_{T\times T}\pi_{ij} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \mathbf{1}_{T\times T}\pi_{nj} \end{pmatrix}
\end{aligned} \tag{2.45}
$$

As discussed earlier in the end of the section 2.2.1, the EM algorithm is run iteratively until the convergence is achieved. Then, the experimental units in the study are assigned to the group with the highest probability:

$$\hat{G}_i = argmax_l Pr(G_i = l|y_i) \equiv argmax_l \pi_{il} \tag{2.46}$$

Where $\hat{G}_i = argmax_l \pi_{il}$ denotes the assignment of the experimental units to the groups (Magrini 2022).

2.2.4 Model selection in group-based multi-trajectory modeling

As of today, the selection of the number of groups J remains to be an open discussion (D. Nagin et al. 2016, Magrini 2022). However, different types of fit statistics such as Bayesian Information Criterion (BIC), and Akaike Information Criterion (AIC) have been used to understand the goodness of fit. It was indicated in number of studies that fit statistics' performances are highly dependent on the data under study and there is no general conclusion that one criterion is better than the others (Magrini 2022; Nest et al. 2020).

Denoting the likelihood with L, number of independent parameters within the model p, and total number of measurements with $nT = N$, the equation 2.47 below illustrates how BIC (Schwarz 1978) and AIC (Akaike 1974) are calculated,

respectively:

$$BIC = -2logL + plogN$$
$$AIC = -2logL + 2p \tag{2.47}$$

Models with lower valued BIC and AIC as calculated in the equation (2.47) are usually reported to be more reliable (Magrini 2022). In addition for establishment of number of groups, it is also required to select the polynomial degree for each outcome (or indicator) within each group in the framework of group-based multi-trajectory modeling. However, choosing the number of groups is more important than deciding on the polynomial degree of the trajectories (D. S. Nagin 2005; Magrini 2022).

A procedure was developed for selecting the model that fits the data under study the best which consists of the following steps (Koohi et al. 2021):

- Run the GBMT modeling algorithm number of times with different number of groups with polynomial degree one to find out the optimal number of groups to express the heterogeneity in the data,
- Increase the polynomial degree of trajectories to two to allow the curved developmental patterns,
- Assess the goodness of fit based on BIC,
- To compare two models, identify the significant changes by using the Bayes factor, $e^{BIC_1 - BIC_2}$, where BIC_1 and BIC_2 are the BIC values of the two models,
  - Bayes factor $\leq \frac{1}{10}$ or ≥ 10 is assumed to indicate a significant change.
- Select the model with the least BIC value and enough sample size in each distinct group (number of units in the group $\geq 5\%$ of the sample) as the best.
- Evaluate 4 different fit diagnostic criteria to ensure that the model fits the data well as recommended by (D. S. Nagin 2005):
  - Check if average posterior probability of group assignments, $\frac{1}{n}\sum_{i=1}^{n}\hat{\pi}_{ij} \geq 0.7$ for all groups,
  - Check if the odds of correct classification $(OCC_j) \geq 5$ for all groups,
  - Check if the ratio of the number of experimental units in a group to the number of all experimental units is similar to the group membership probabilities that are estimated from the model,
  - Look for narrow confidence intervals for group membership probabilities.

3 Collected Data

There are variety of gearbox models with different mechanical designs that are produced for wind turbines. These gearboxes are tested in the factories after production, before they are installed on the wind turbines at sites. Number of tests have been developed to ensure the quality of each gearbox such as the analysis of oil cleanliness data, sound data, and vibration measurements. In addition, analysis of vibration data is very common practice in condition monitoring (Ojala 2013) and fault diagnostics as it helps to understand the type and severity of the defect (Aherwar 2012).

Vibration measurements are collected by making use of several accelerometers that are mounted on a gearbox during three different load steps. Sensors placed on different predefined places of the machine record the data throughout the process and transform the collected measurement to the frequency domain from the time domain. Location of the sensors on a gearbox might differ depending on the model of the gearbox. Initially, in the data collected for this specific study, there were 12 different sensors that had been placed on a gearbox which recorded the data from 3 different directions of 4 preset strategic points.

The data used in this thesis work were collected from 337 distinct planetary gearboxes of the same model. The 3 sensors on the high speed shaft recorded the data from X, Y and Z directions during the tests. As a result, a data set with dimensions 337×9 were obtained. The reason why only high speed shaft measurements have been taken into account is that the high speed shaft measurements reflect the data characteristics the best, according to the experts' opinion. The vibration measurements that are collected from X, Y and Z coordinates of all gearboxes are visualized as the parallel coordinate plots in the figures 3.1, 3.2 and 3.3, respectively.

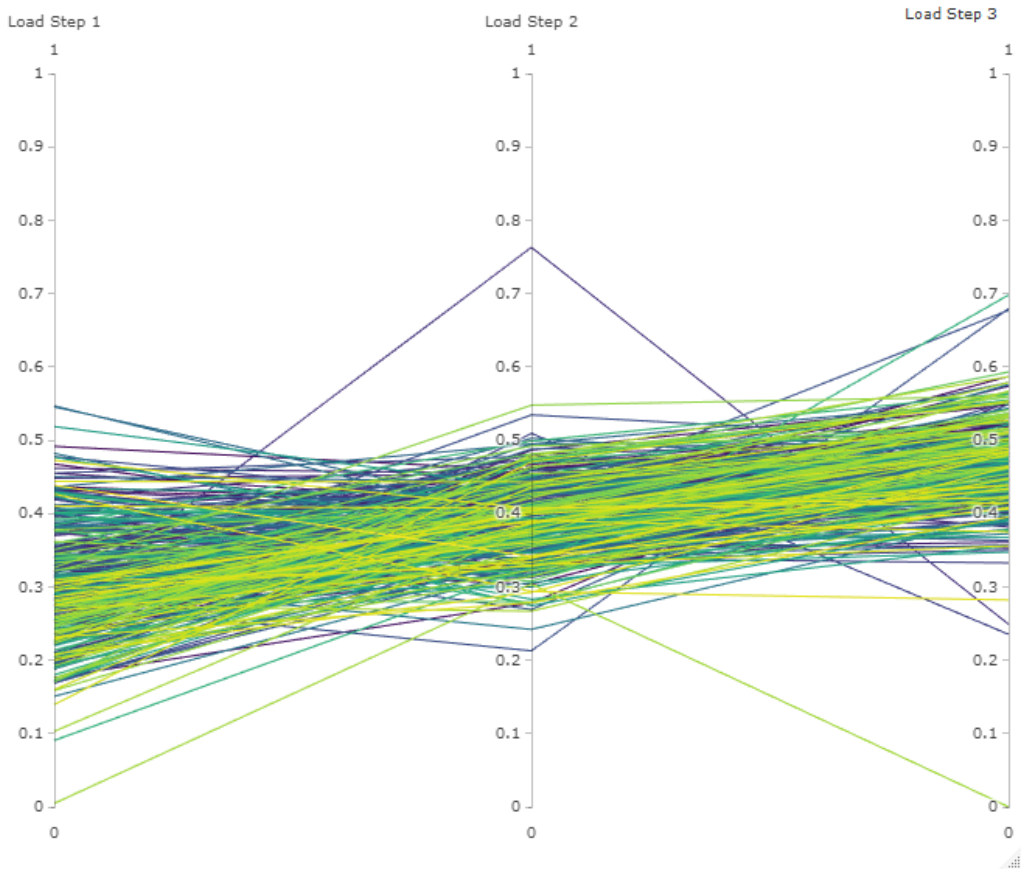

Figure 3.1 *Normalized vibration measurements collected from X coordinate of the individual gearboxes over 3 load steps*

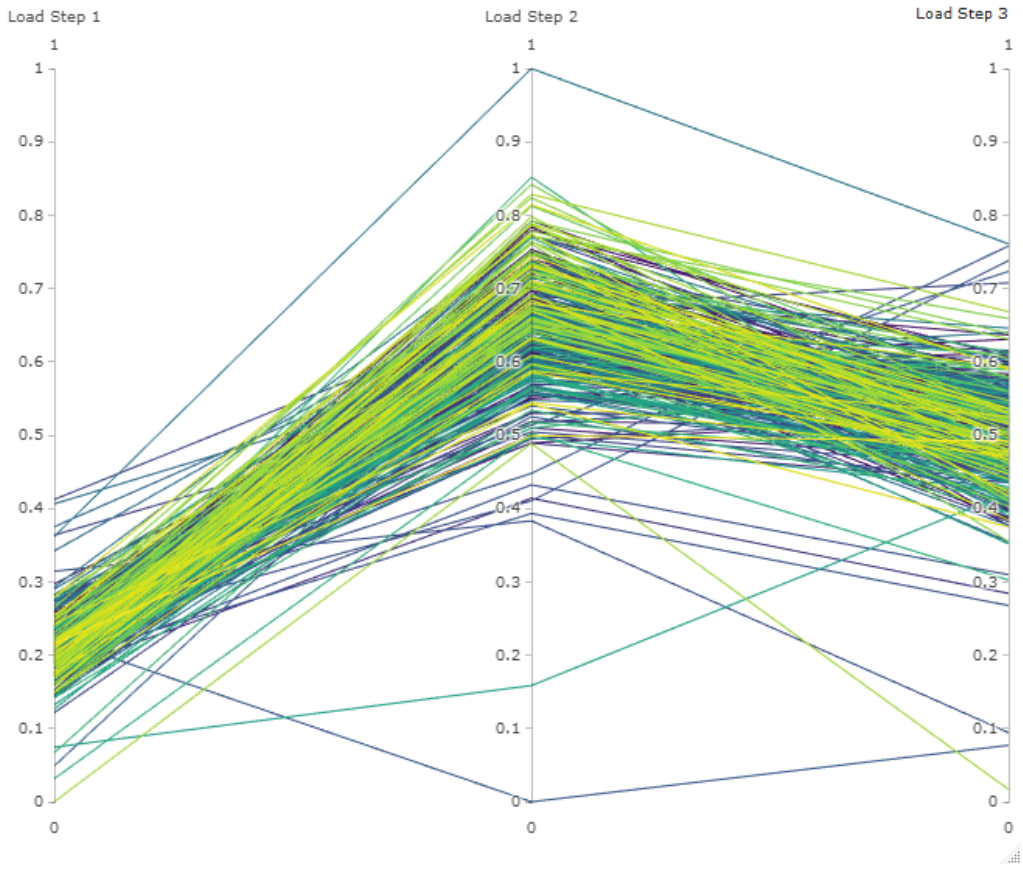

Figure 3.2 *Normalized measurements collected from Y coordinate of the individual gearboxes over 3 load steps*

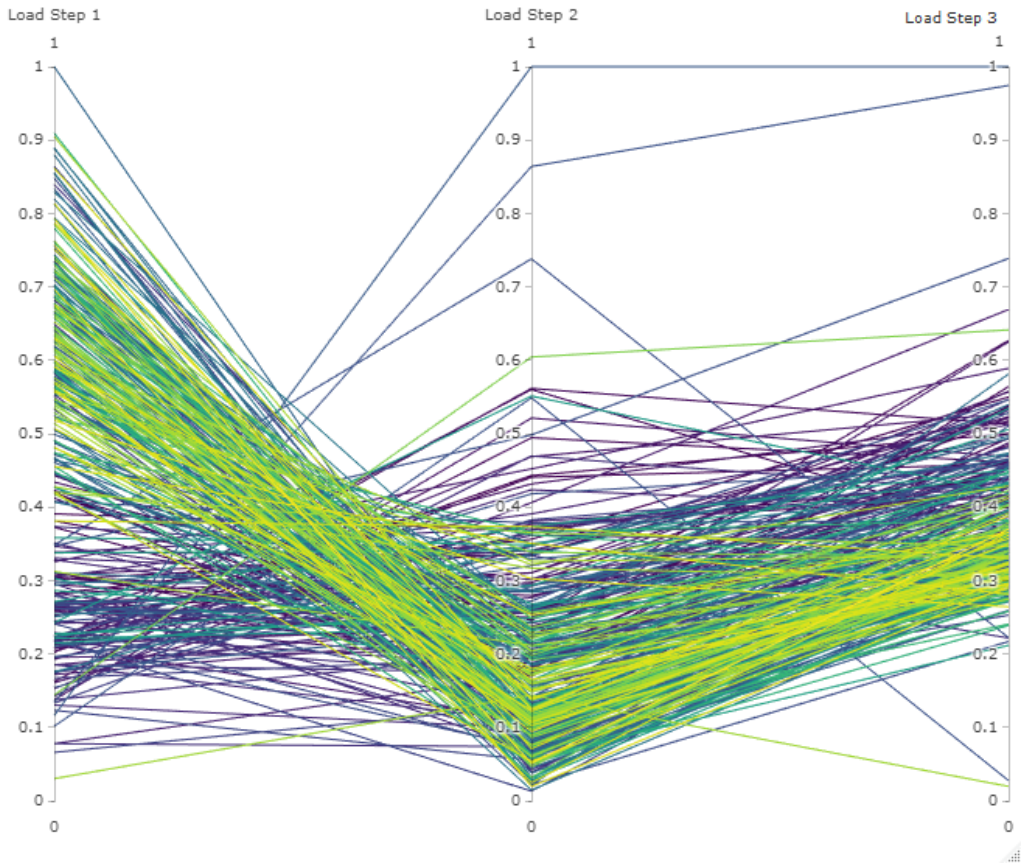

Figure 3.3 *Normalized measurements collected from Z coordinate of the individual gearboxes over 3 load steps*

17 of 337 gearboxes included in this research are known to be suffered from four different issues such as tooth damages and fractures in the course of their operation on the wind turbines. The Table 3.1 shows the number of gearboxes and corresponding states of the gearboxes indicating the related issue from type 1 to type 4. The remaining 320 gearboxes are known to be healthy which is referred as type 5 in the Table 3.1.

Table 3.1 *Number of the gearboxes and the state of the gearboxes*

Number of gearboxes	State
4	type 1
4	type 2
8	type 3
1	type 4
320	type 5

4 Analysis and Results

There has been a natural interest on clustering with five groups throughout this research due to the fact that there were four distinct issues that the gearboxes suffered from and the healthy ones. Therefore, the results obtained with five clusters have been investigated whether they were the best in relation to the criteria used to evaluate the model adequacies or not in both reconstructive and probability based parametric clustering approaches that are used in this thesis work.

As stated in the chapters 1 and 2, a specific form of k-means clustering for multivariate longitudinal data is employed in this research work as the reconstructive parametric clustering algorithm in order to deal with such data set and it is named as KmL3D (Genolini et al. 2013). Since the analysis on the collected data set is performed in R software (R Core Team 2013), an available R package which has the same name with the algorithm, *kml3d*, is used (Christophe Genolini, Alacoque, et al. 2015). On the other hand, the generalization of group-based trajectory modeling, group-based multi-trajectory modeling, that belongs to the finite mixture modeling family of probability based parametric clustering methods is applied on the data set by making use of the R package named as *gbmt* (Magrini 2021) for the same reason that the collected data is consisting of multivariate repetitive measurements.

The detailed description on the application of *kml3d* package on the collected data set is provided in the section 4.1 and the results obtained are presented in the section 4.2. The section 4.3 details how the package *gbmt* is utilized in this thesis work. Finally, the results corresponding to the group-based multi-trajectory modeling are demonstrated in the section 4.4.

4.1 KmL3D analysis

KmL3D clustering is a reconstructive clustering method that is a type of k-means clustering and specifically designed to cluster multivariate longitudinal data. As described in the chapter (3), the measurements were taken from X, Y, and Z coordinates of high-speed shafts of the 337 gearboxes from a single product family under three different load steps. As a result, 337×9 sized data set was obtained which is considered a multivariate longitudinal data. Moreover, a pre-processing for the data is applied in order to be consistent with (Genolini et al. 2013). The vibration measurements obtained for X, Y and Z coordinates are denoted as $Y_{..X}, Y_{..Y}, Y_{..Z}$ which are the variable trajectories. For the i^{th} gearbox at the j^{th} load step ($j = 1, 2, 3$), the value of $Y_{..X}$ is noted as y_{ijX}. Similarly, y_{ijZ} and y_{ijZ} represent the values of $Y_{..Y}$ and $Y_{..Z}$, respectively. To summarize, the measurements collected from the i^{th} gearbox

is a matrix as provided in the equation 4.1 below that has the same structure with 2.3 that is set in the chapter 2:

$$y_{i..} = \begin{pmatrix} y_{i1X} & y_{i2X} & y_{i3X} \\ y_{i1Y} & y_{i2Y} & y_{i3Y} \\ y_{i1Z} & y_{i2Z} & y_{i3Z} \end{pmatrix} \qquad i = 1, \ldots, 337 \tag{4.1}$$

Such a modification on the data allows the usage of the *kml3d* package (Christophe Genolini, Alacoque, et al. 2015) as explained comprehensively in the section 2.1.2 of the chapter 2. Euclidean distance is used in this clustering algorithm as the distance metric as also stated in the section 2.1.2. Since the collected data is composed of vibration measurements from X, Y and Z coordinates of the high-speed shaft of the gearboxes for three different load steps, there is not any scaling method is used to process the data before proceeding with the utilization of the clustering algorithm. In other words, the KmL3D clustering is applied directly to the raw data.

4.2 KmL3D results

KmL3D clustering algorithm has been run on the collected data for pre-selected cluster numbers from two to ten. Since the Expectation-Maximization (EM) algorithm is used, it is run 20 times for each model iteratively in order to ensure that the algorithm is not converged to a local maximum, which is a well-known problem of EM algorithm. Eventually, it is expected that the algorithm converges to the global maximum which is the optimal solution for clustering, or at least to a good local maximum.

In order to evaluate the model adequacy, Calinski-Harabasz quality criterion is used that is explained in detail in the Chapter 2 and formulated in the equation 2.7. The corresponding Calinski-Harabasz scores are visualized in the figure 4.1 below in which each partition is shown with its corresponding cluster number (Christophe Genolini, Alacoque, et al. 2015).

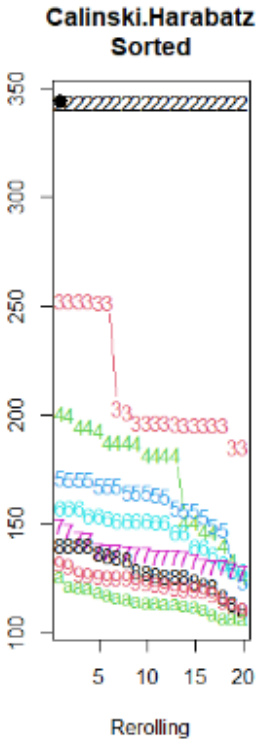

Figure 4.1 *Calinski-Harabasz scores obtained for each iterative run*

It can be seen from the 4.1 that the highest scores are obtained for the model with two clusters which is also indicated with a black dot on top-left corner of the same figure 4.1. The mean trajectories of the clusters for the model with two cluster that is shown to be the best model are illustrated in the figure 4.2 below.

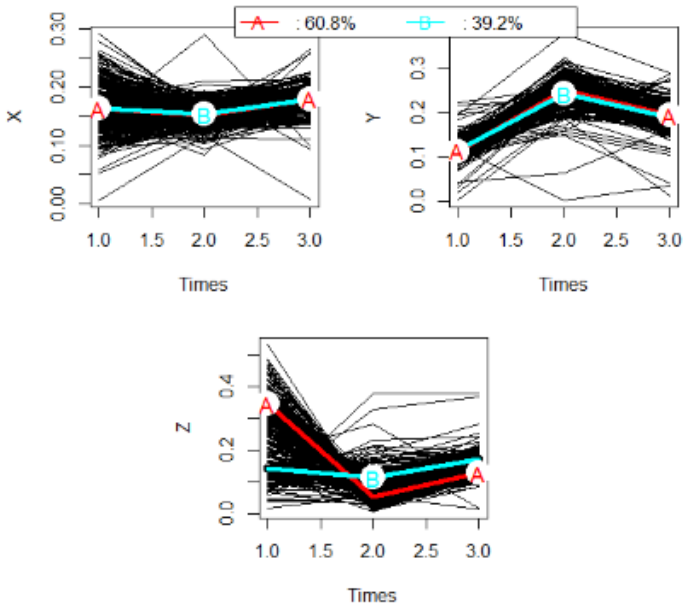

Figure 4.2 *The mean trajectories of the model with two clusters*

The trajectories that are shown in black in the figure 4.2 represent the single-variable trajectories that are parts of joint-trajectories Christophe Genolini, Alacoque, et al. 2015.

The number of gearboxes that belongs to type 1 to type 4 as described in the chapter 3 in each cluster for the model with two clusters is provided in the Table 4.1 below:

Table 4.1 *Number of gearboxes from each type in each cluster, model with two clusters*

Number of gearboxes from type 1	Number of gearboxes from type 2	Number of gearboxes from type 3	Number of gearboxes from type 4	Cluster number
4	2	4	1	1
0	2	4	0	2

It can be realized from the Table 4.1 that the first cluster included 11 of the 17 gearboxes which is approximately 65% of the gearboxes suffered from various issues during their operation, which is a close percentage that the first cluster has involved 60.8% of the sample. Similarly, almost 35% of the 17 gearboxes belong to the second cluster which has 39.2% of the sample.

The model with five clusters has also been investigated due to the attention for this specific case as a result of the collected data set. As can be seen from the figure 4.1 that the Calinski-Harabasz scores are not the best but also not the worst. The mean trajectories of the clusters for the model with five cluster are presented in the figure 4.3 below:

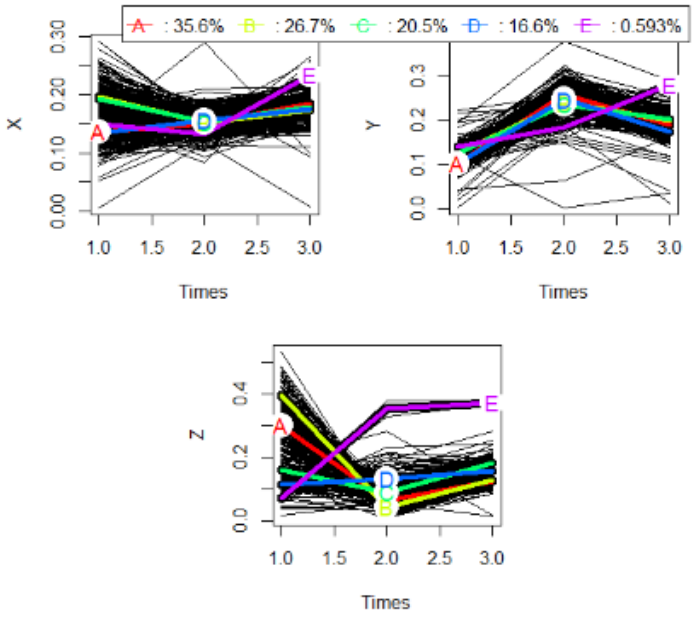

Figure 4.3 *The mean trajectories of the model with five clusters*

The number of the 17 gearboxes belonging to type 1 to type 4 as stated in the

chapter 3 in each cluster for the model with five clusters is depicted in the Table 4.2 below:

Table 4.2 *Number of gearboxes from each type in each cluster, model with five clusters*

Number of gearboxes from type 1	Number of gearboxes from type 2	Number of gearboxes from type 3	Number of gearboxes from type 4	Cluster number
4	2	4	1	1
0	2	1	0	2
0	0	2	0	3
0	0	1	0	4
0	0	0	0	5

Similar to the model with two cluster, 11 of the 17 gearboxes of interest are in the largest cluster. However, the largest cluster includes 35.6% of the sample. The rest of the clusters include 3, 2, 1 of the remaining six gearboxes, respectively. The fifth cluster contained only two gearboxes which are from type 5 as described in the chapter 2.

4.3 Group-based multi-trajectory modeling analysis

The probability based parametric clustering method that is used in this thesis work is group-based multi-trajectory (GBMT) modeling which is depicted in detail in the chapter 2. The data size is 337×9, as described in the section 4.1 and the R package *gbmt* that is used in implementation of group-based multi-trajectory modeling requires data pre-processing similar to the case in the usage of *kml3d* for k-means clustering for longitudinal multivariate data. Again, the vibration measurements of X, Y, and Z coordinates of the high-speed shafts of all the gearboxes are denoted with, $Y_{..X}, Y_{..Y}, Y_{..Z}$, respectively. Moreover, the vibration measurements from X, Y, and Z coordinates for the i^{th} gearbox at the j^{th} load step are represented with $y_{ijX}, y_{ijY}, y_{ijZ}$, correspondingly. Eventually, the nine measurements, $Y_{i..}$, are collected for the i^{th} gearboxes put in the form as shown in the equation 4.2 below:

$$Y_{i..} = \begin{pmatrix} y_{i1X} & y_{i1Y} & y_{i1Z} \\ y_{i2X} & y_{i2Y} & y_{i2Z} \\ y_{i3X} & y_{i3Y} & y_{i3Z} \end{pmatrix} \tag{4.2}$$

It must be noted that the equation 4.2 follows the structure that is presented in the equation 2.34 in the chapter 2 as needed by the R package *gbmt* (Magrini 2021) that is used in this research.

Group-based multi-trajectory modeling has been done with different polynomial degrees for the trajectories, on the raw data and on the transformed data by making use of one of the scaling methods that are discussed in the chapter 2. The package

gbmt also makes it available for backward elimination of the non-significant coefficients to adapt the most suitable polynomial degree which does not exceed the pre-set value for each outcome in the collected data set by pruning option. The grup-based multi-trajectory modeling in this research is employed by enabling the pruning option in the *gbmt* package (Magrini 2021). Furthermore, the clustering has been performed 20 times for each configuration in order to ensure that the algorithm is not converged to one of the local maxima and to find out the global maximum value of the log-likelihood function since the EM algorithm has been utilized in the group-based multi-trajectory modeling. As the initial values change in each run, it is expected that the EM algorithm converges to the global maximum value eventually, or at least to a good local maximum.

4.4 GBMT Modeling results

Group-based multi-trajectory modeling has been run on the raw data and the transformed data by the scaling method used on the raw data by adjusting the polynomial degree of the trajectories. Two different polynomial degrees which are 1 and 2 are set for the trajectories and the algorithm is run for the raw data and the pre-processed data where centering has been applied on each measurement as described in the chapter 2.

There are two reasons why the remaining scaling methods are not used. First, the collected data set consists of repeated measurements for different load steps and such data sets are considered as longitudinal data. It is stated that normalizing such longitudinal data is inappropriate and not necessary (Heggeseth and Jewell 2018).Therefore, the standardization that was described in the chapter 2 is not used. Secondly, the collected data set does not need to be made unitless because all the outcomes are vibration measurements which are on the same scale. Hence, the scaling methods ratio to the mean and log ratio to the mean which are also explained in the chapter 2 are not utilized throughout this research.

By making use of the pre-processed data and the raw data, and two different polynomial degrees, four different configurations have been generated. For each configuration, the group-based multi-trajectory modeling clustering algorithm has been run with different pre-selected cluster numbers varying from two to ten.

4.4.1 GBMT Modeling results with polynomial degree 1 and using raw data

In the case of no scaling method used, the group-based multi-trajectory modeling algorithm has been directly applied to the raw data without any data manipulation on the data set. The clustering algorithm has been run for pre-selected cluster

numbers starting from two to ten. The AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) scores are provided in the Table 4.3 below:

Table 4.3 *AIC and BIC scores of clustering outcomes for polynomial degree 1 and using raw data*

Number of clusters	AIC score	BIC score
2	-9801.969	-9679.0023
3	-9995.853	-9808.943
4	-10141.585	-9890.73
5	-10214.471	-9899.675
6	-10278.666	-9899.926
7	-10317.174	-9874.491
8	-10331.857	-9825.231
9	-10346.660	-9781.010
10	-10346.037	-9711.525

As can be seen from the Table 4.3, the AIC scores kept decreasing for the increasing cluster numbers. On the other hand, the lowest BIC score was obtained for cluster number 6 which is only slightly less than the BIC score of clustering number being 5. The Bayes factor is calculated and found to be far less than 10 and close to 1 as given in the equation 4.3. Therefore, it can be claimed that there is a weak evidence for model with six clusters is better than the model with five clusters (D. S. Nagin 2005).

$$B_{56} = e^{(-9899.675+9899.926)} = 1.2853 \tag{4.3}$$

In order to select the number of clusters within a model, it is also important to keep the model complexity as less as possible (D. S. Nagin 2005; Den Teuling, Pauws, and Heuvel 2020). Considering this, and the fit statistics that are illustrated in the Table 4.3, it is decided to select the model with five clusters and the obtained trajectories for the model with five clusters is shown in the figure 4.4 below.

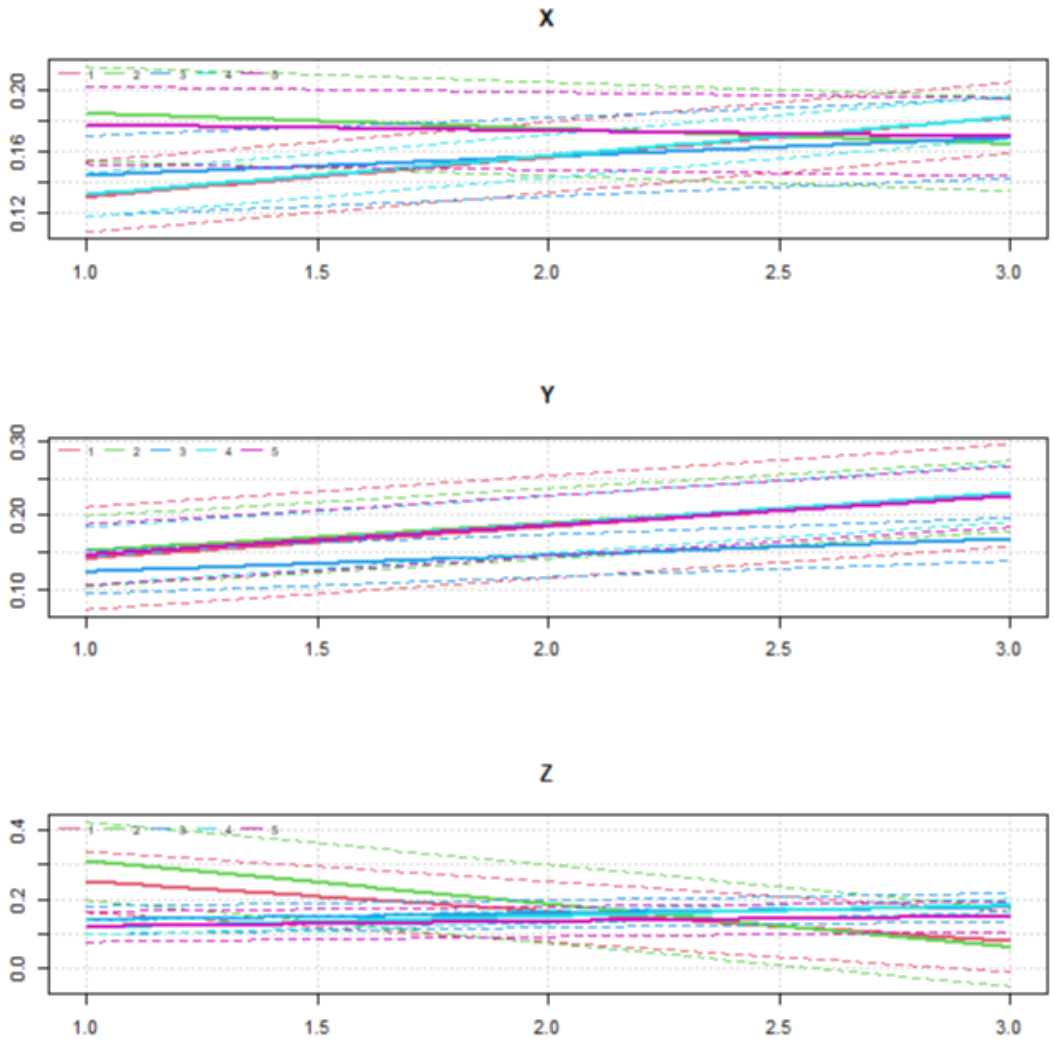

Figure 4.4 *Estimated trajectory curves and their 95% confidence intervals for the model with five clusters, polynomial degree = 1, using raw data*

The percentage of the gearboxes in each cluster are provided in the Table 4.4 below.

Table 4.4 *Percentage of the gearboxes in each cluster in the model with five clusters, polynomial degree = 1, using raw data*

Percentage	Cluster number
34%	1
30%	2
3%	3
10%	4
23%	5

As introduced in the end of the Chapter 2, it is required that each cluster in the model shall include at least 5% of the sample size where this model fails. Therefore,

it is needed to check the model with six clusters for which the trajectories obtained are illustrated in the figure 4.5 below.

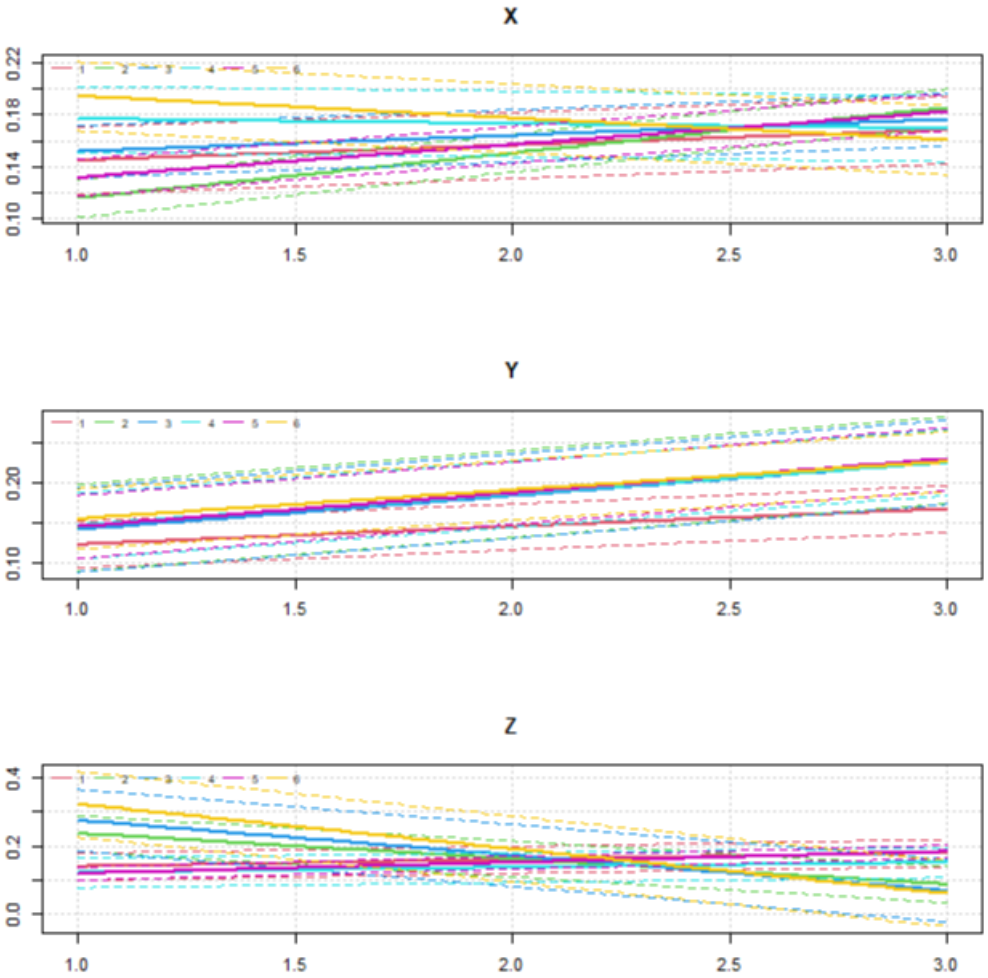

Figure 4.5 *Estimated trajectory curves and their 95% confidence intervals for the model with six clusters, polynomial degree = 1, using raw data*

The percentage of the gearboxes in each cluster are provided in the Table 4.5 below.

Table 4.5 *Percentage of the gearboxes in each cluster in the model with six clusters, polynomial degree = 1, using raw data*

Percentage	Cluster number
3.5%	1
16.5%	2
27%	3
23%	4
10%	5
20%	6

As can be seen in the Table 4.5, this model also fails since the percentage of the gearboxes in the first cluster is less than 5%. Therefore, it could be concluded that this configuration with polynomial degree one and using raw data does not yield for a sensible clustering.

4.4.2 GBMT Modeling results with polynomial degree 2 and using raw data

In this case, group-based multi-trajectory modeling algorithm is applied to the data without any scaling performed and polynomial degree of the trajectories is set to two. The pre-selected cluster numbers are varied from two to ten, and the AIC and BIC scores for each run have been provided in the Table 4.6 below:

Table 4.6 *AIC and BIC scores of clustering outcomes for polynomial degree 2 and using raw data*

Number of clusters	AIC score	BIC score
2	-12015.53	-11863.06
3	-12508.37	-12282.11
4	-12821.94	-12526.82
5	-12981.14	-12607.32
6	-13064.81	-12612.29
7	-13132.64	-12601.42
8	-13186.52	-12576.60
9	-13228.36	-12544.660
10	-13259.65	-12487.42

Comparing the Table 4.3 presenting the results in the previous section to the Table 4.6, one can see that there is a strong improvement in the AIC scores and BIC scores for the configuration with polynomial degree two and no scaling on the data. Similar to the case with polynomial degree for the trajectories being one and no scaling applied to the data, the AIC score decreases as the number of clusters in the model with polynomial degree two and no scaling increases. The lowest BIC score is achieved for the model with six clusters, as can be seen from the Table 4.6. Since there have been a special interest for the model with five clusters, the Bayes factor is calculated as given in the equation 4.4 in order to investigate if there is a strong evidence that the model with six clusters is better than the model with five clusters.

$$B_{56} = e^{(-12607.32+12612.29)} = 144.03 \tag{4.4}$$

Since the Bayes factor is greater than 10, it can be claimed that there is a strong evidence that the model with six clusters is better than the one with five clusters.

The trajectories for the model with six clusters are shown in the figure 4.6 below.

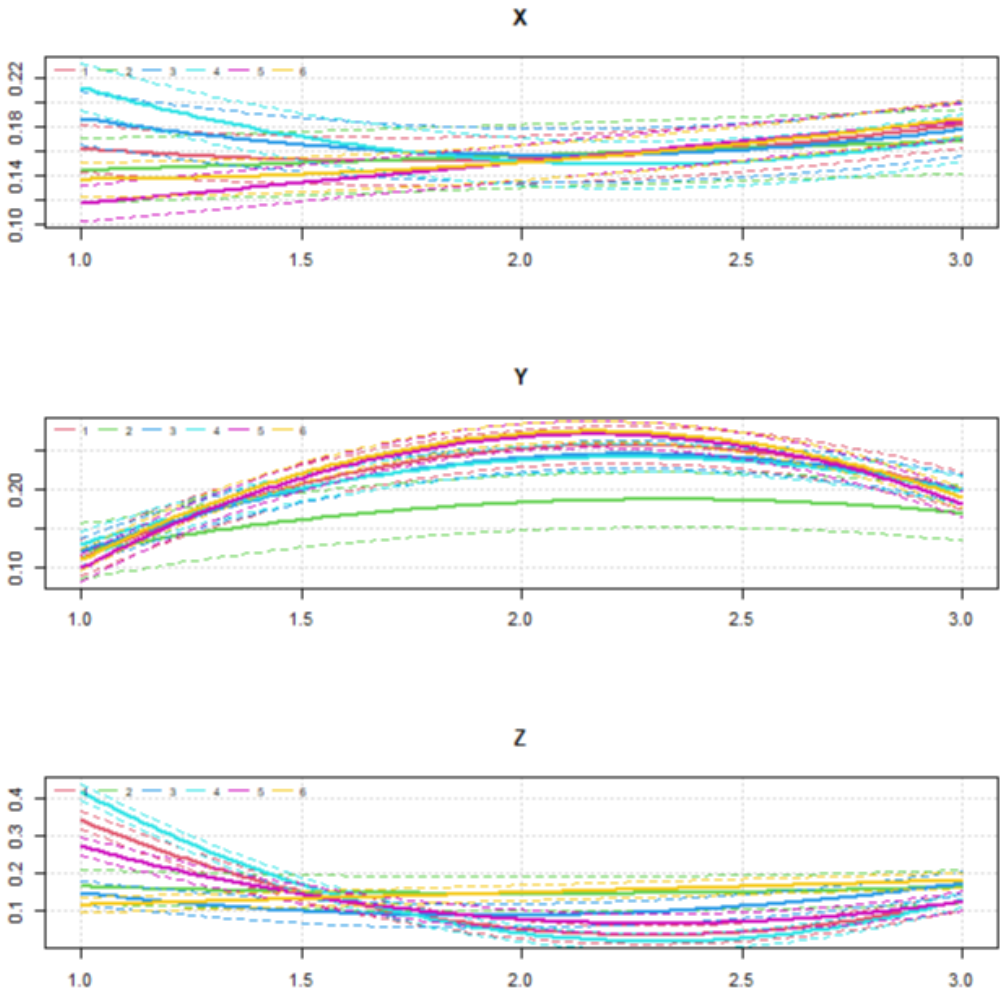

Figure 4.6 *Estimated trajectory curves and their 95% confidence intervals for the model with six clusters, polynomial degree = 2, using raw data*

The percentage of the gearboxes in each cluster are shown in the Table 4.7 below.

Table 4.7 *Percentage of the gearboxes in each cluster in the model with six clusters, polynomial degree = 2, using raw data*

Percentage	Cluster number
29%	1
5%	2
22%	3
15%	4
18%	5
11%	6

It can be seen from the Table 4.7 that each cluster contains at least 5% of the

sample. The average posterior probability of assignments for the model (APPA) are provided in the Table 4.8 below.

Table 4.8 *Average posterior probability of assignments for the model with six clusters, polynomial degree=2, using raw data*

APPA	Cluster number
0.922	1
0.998	2
0.936	3
0.937	4
0.955	5
0.948	6

It can be observed from the Table 4.8 that all the average posterior probability of assignments are larger than 0.7, which is introduced as another criterion for evaluation of model adequacy in the end of the Chapter 2.

The number of the gearboxes that had issues during their operation according to their type of issue in each cluster is shown in the Table 4.9 below:

Table 4.9 *Number of gearboxes from each type in each cluster, model with six clusters, polynomial degree=2, using raw data*

Number of gearboxes from type 1	Number of gearboxes from type 2	Number of gearboxes from type 3	Number of gearboxes from type 4	Cluster number
1	2	3	1	1
0	0	0	0	2
0	0	2	0	3
0	1	0	0	4
3	1	3	0	5
0	0	0	0	6

It must be observed from the Table 4.9 that the second and the sixth clusters have no gearbox suffering from any issue. Therefore, it can be concluded that this configuration with polynomial degree for trajectories being two and usage of the raw data for the model with six clusters yields consistent results.

4.4.3 GBMT Modeling results with polynomial degree 1 and centering on the data set

Before proceeding with the application of the group-based multi-trajectory modeling clustering algorithm on the data set, a pre-processing on the data set has been performed in this configuration which is centering. In other words, sample mean is subtracted from each measurement, as described in the equation 2.38 in the Chapter 2. Afterwards, the clustering algorithm is run by setting the polynomial degree for

the trajectories as one for cluster numbers from two to ten. The AIC and BIC scores of each model are provided in the Table 4.10 below:

Table 4.10 *AIC and BIC scores of clustering outcomes for polynomial degree 1 and centering*

Number of clusters	AIC score	BIC score
2	-10832.82	-10709.86
3	-11189.35	-11002.44
4	-11379.98	-11134.05
5	-11558.44	-11243.64
6	-11688.34	-11309.60
7	-11782.03	-11339.35
8	-11866.79	-11365.08
9	-11913.01	-11347.36
10	-11943.99	-11309.48

It can be seen from the Table 4.10 that the AIC score declines as the number of clusters in the models increases which is similar to the previous configurations discussed before. Unlike the AIC score, the minimum value for the BIC score is obtained for the model with eight clusters.

There have been an attention on the adequacy for the models with five clusters, as discussed in the beginning of the chapter 4. The Bayes factor calculation between the model with five clusters and the model having 8 clusters is given in the equation 4.5 below, which is found to be much larger than ten that concludes that there is a strong evidence that the model with eight clusters is better than the model with five clusters.

$$B_{58} = e^{(-11243.64+11365.08)} \approx 5.5 \times 10^{52} \tag{4.5}$$

The trajectories for the model with eight clusters are shown in the figure 4.7 and the percentage of the gearboxes in each cluster are given in the Table 4.11.

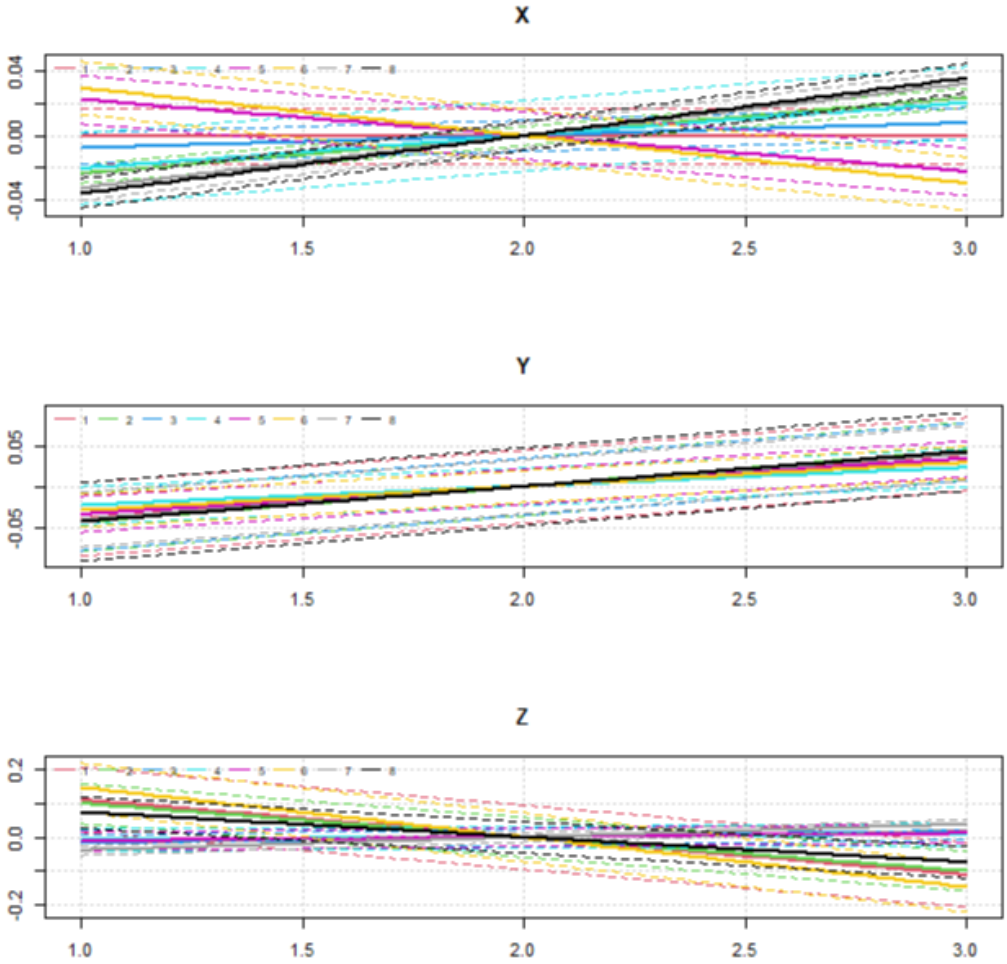

Figure 4.7 *Estimated trajectory curves and their 95% confidence intervals for the model with eight clusters, polynomial degree = 1, centering*

Table 4.11 *Percentage of the gearboxes in each cluster in the model with eight clusters, polynomial degree = 1, centering*

Percentage	Cluster number
26%	1
13%	2
16%	3
4%	4
9%	5
10%	6
7%	7
15%	8

The fourth cluster in this model has only 4% of the sample which fails to be a consistent clustering according to (D. S. Nagin 2005). Therefore, it is not necessary to investigate the average posterior probabilities of the assignments and gearbox distribution into the clusters.

4.4.4 GBMT Modeling results with polynomial degree 2 and centering on the data set

Similar to the previous section, the data set has been centered by subtracting the sample mean from each measurement as described in the equation 2.38 in the Chapter 2. Then, the group-based multi-trajectory modeling clustering algorithm has been applied nine times on the manipulated data set for cluster numbers increasing from two to ten in each model. The AIC and BIC scores obtained for each model are shown in the Table 4.12 below:

Table 4.12 *AIC and BIC scores of clustering outcomes for polynomial degree 2 and centering*

Number of clusters	AIC score	BIC score
2	-13647.33	-13494.85
3	-14095.32	-13864.14
4	-14531.39	-14226.43
5	-14793.48	-14409.83
6	-14956.66	-14499.22
7	-15097.93	-14561.79
8	-15241.43	-14626.59
9	-15343.92	-14660.22
10	-15383.09	-14620.69

The minimum BIC score in this configuration where the polynomial degree of the trajectories are set to two and centering is applied on the data set is obtained for the model with nine clusters as can be seen from the Table 4.12 above where the AIC score monotonically decreases for increasing number of clusters in the models. In order to justify that the model with nine cluster would yield more accurate results than the model with five clusters, the Bayes factor has been calculated as shown in the equation 4.6 below:

$$B_{59} = e^{(-14409.83+14660.22)} \approx 5.53 \times 10^{108} \tag{4.6}$$

Since the Bayes factor is much larger than ten, it is shown that there is a strong evidence for the model with nine clusters being better than the model with five clusters. The trajectories obtained for the model with nine clusters are illustrated in the figure 4.8 below.

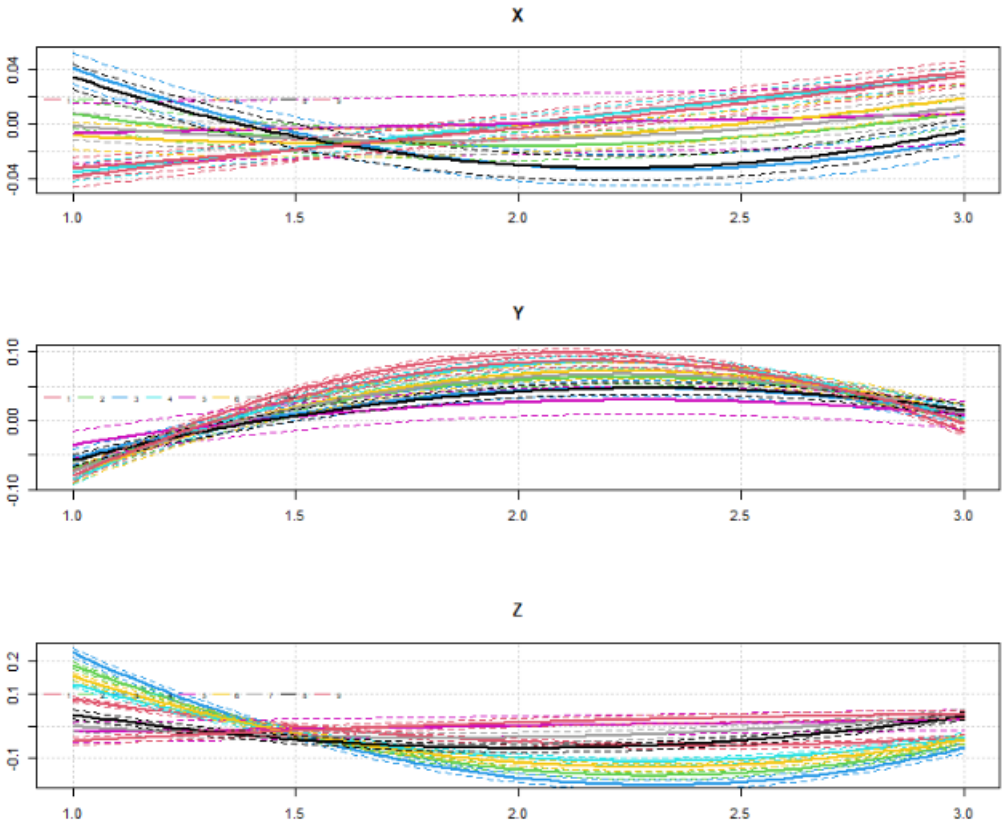

Figure 4.8 *Estimated trajectory curves and their 95% confidence intervals for the model with nine clusters, polynomial degree = 2, centering*

The percentages of the gearboxes in each cluster of the model with nine clusters are demonstrated in the Table 4.13 below:

Table 4.13 *Percentage of the gearboxes in each cluster in the model with nine clusters, polynomial degree = 2, centering*

Percentage	Cluster number
6%	1
15%	2
12%	3
11%	4
5%	5
18%	6
18%	7
8%	8
7%	9

Each cluster contains at least 5% of the sample as can be seen from the Table 4.13. The average posterior probabilities of assignments for each cluster are provided in the table 4.14.

Table 4.14 *Average posterior probability of assignments for the model with nine clusters, polynomial degree=2, centering*

APPA	Cluster number
0.972	1
0.963	2
0.996	3
0.960	4
0.997	5
0.939	6
0.982	7
0.946	8
0.996	9

It can be seen from the Table 4.14 that all the values of the average posterior probability of assignment are larger than 0.7, which was set as one of the criteria for model selection in the end of the Chapter 2.

The number of the gearboxes that had issues during their operation according to their type of issue in each cluster is shown in the Table 4.15 below:

Table 4.15 *Number of gearboxes from each type in each cluster, model with six clusters, polynomial degree=2, using raw data*

Number of gearboxes from type 1	Number of gearboxes from type 2	Number of gearboxes from type 3	Number of gearboxes from type 4	Cluster number
0	1	2	0	1
0	1	1	0	2
0	1	0	0	3
2	0	1	0	4
0	0	0	0	5
2	1	2	1	6
0	0	1	0	7
0	0	1	0	8
0	0	0	0	9

The fifth and the ninth clusters in this model do not have any gearboxes that issues during their operation as can bee seen from the Table 4.15. In the light of the results obtained, it can be concluded that this model would yield meaningful clusters.

5 Conclusion

The main aim in this thesis was to cluster the gearboxes using their vibration measurements. The vibration measurements from X, Y and Z dimensions of 337 gearboxes from a single product family were recorded over three distinct load steps. As a result, a data set sized 337 $\times$ 9 were obtained, which is considered a longitudinal data set since the vibration measurements has been taken repeatedly.

The first clustering algorithm that is employed on the data set in this research is a special form of k-means clustering from reconstructive parametric clustering algorithms named as $KmL3D$ which can be used on multivariate longitudinal data. The algorithm has been directly run 20 times on the raw data for different models in which cluster numbers are changed from two to ten. The reason why the clustering has been done 20 times is to prevent it from converging to a poor local maximum since it uses the Expectation-Maximization algorithm. Eventually, it is expected that the global maximum which is the optimal solution for each model is achieved, or at least a good local maximum. The distance metric used for utilization of $KmL3D$ was Euclidean distance for all nine models. Calinski-Harabasz information criterion is used to check the adequacy for each model. According to the Calinski-Harabasz scores, the model with two clusters was the best. The clusters obtained in this model had gearboxes that had issues during their operation approximately proportional to their size. For instance, 11 of 17 gearboxes having issues ($\sim$65%) are involved in the cluster containing 60.8% of the sample.

There have been a special interest on the results obtained for the models with five clusters. The results obtained by making use of *KmL3D* algorithm on the data set, it was shown that for the model with five clusters, the first cluster involved the majority of the gearboxes that have had issues but also it contained 35.6% of the entire sample. Besides, the fifth cluster had no such gearbox of interest but the size of this cluster is comparably small.

The second clustering algorithm that is applied on the data set in this thesis is a generalization of group-based trajectory modeling which is a unique form of finite mixture models from probability based parametric clustering algorithms and named as group-based multi-trajectory (GBMT) modeling. The clustering has been performed both on the raw data and centered data in which within-individual means are subtracted from the outcomes obtained in the data set and polynomial degrees of the trajectories are set as either one or two. As a result, four different configurations were obtained. Since this is a parametric approach, maximum likelihood estimation with Expectation-Maximization algorithm is utilized in order for estimating the parameters of the likelihood function. Similar to $KmL3D$, the algorithm is run 20

times for the models with different cluster numbers that vary from two to ten to prevent the remedy the problem of EM algorithm converging to one of local maxima.

For the evaluation of the results obtained with GBMT modeling, the AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) were used. In all four configurations, the AIC score continuously decreased as the number of clusters in the models increase. Therefore, no conclusion were made based on AIC scores. On the other hand, there was a minimum BIC score obtained in all the configurations. The results show that the models with quadratic trajectories have had lower AIC and BIC scores in clustering with both of the raw-data and centered data than the models having linear trajectories that are failed to include at least 5% of the sample in each cluster.

There were two clustering results that were able to pass the criteria set for the model selection which are the followings:

- The model with six clusters and second degree polynomials for the trajectories which is fit to the raw data,
- The model with nine clusters and second degree polynomials for the trajectories which is fit to the centered data.

Both of the models have had the least BIC scores in their respective configurations with significant evidence over the models with second least BIC score that is identified by calculating the Bayes factor. Moreover, each cluster in both of the models contains at least 5% of the sample, and average posterior probabilities were larger than 0.7 as pointed in the section 4.4 which is set as another criterion for model selection. Since there are two models that have been able to pass the model selection criteria, the best one is selected to be the one with less clusters. In other words, the model with six clusters and quadratic trajectories that is fit to the raw data has been chosen to be the best model since one of the aims in clustering problems is to explain the data with as less clusters as possible. It must be observed that the sixth cluster in the selected model as the best has 11% of the sample, 37 gearboxes which had no issue along with the second cluster which has 17 gearboxes, 5% of the sample. The trajectories obtained for the model with six clusters and quadratic trajectories are presented in the Appendix A.

To conclude, this research employed two different types of clustering methods on a data set of vibration measurements obtained from gearboxes used in wind turbines which has become essential for production of green energy over the recent years. It cannot be observed that one clustering method performed better than the other, but meaningful results have been obtained in each of them.

In order to further improve the results obtained in this research, a data set consisting of more vibration measurements from the gearboxes might be helpful.

Also, quality criteria different than or in addition to the vibration measurements can be included in the study. If the data set includes more gearboxes that have had issues during their usage in the wind turbines, then more reliable inference can be made on the gearbox characteristics and hidden relationships between them.

References

(GWEC), Global Wind Energy Council (Apr. 2019). *Global Wind Report 2018*. URL: https://gwec.net/wp-content/uploads/2019/04/GWEC-Global-Wind-Report-2018.pdf.

— (Mar. 2021). *Global Wind Report 2021*. URL: https://gwec.net/wp-content/uploads/2021/03/GWEC-Global-Wind-Report-2021.pdf.

Aherwar, A (2012). "An investigation on gearbox fault detection using vibration analysis techniques: A review". eng. In: *Australian journal of mechanical engineering* 10.2, pp. 169–183. ISSN: 1448-4846.

Ahmed, Mobyen Uddin, Shaibal Barua, and Shahina Begum (2020). "Artificial Intelligence, Machine Learning and Reasoning in Health InformaticsAn Overview". eng. In: *Intelligent systems reference library*. Vol. 192. Intelligent Systems Reference Library. Cham: Springer International Publishing, pp. 171–192. ISBN: 3030549313.

Akaike, H (1974). "A new look at the statistical model identification". eng. In: *IEEE transactions on automatic control* 19.6, pp. 716–723. ISSN: 0018-9286.

Bock, Hans-Hermann (Jan. 2008). "Origins and extensions of the k-means algorithm in cluster analysis". In: *Journal Électronique dHistoire des Probabilités et de la Statistique [electronic only]* 4.

Burek, S. (Oct. 2010). "When will fossil fuels finally run out and what is the technical potential for renewable energy resources?" In: *International Journal of COMADEM* 13, pp. 22–27.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". eng. In: *Journal of the Royal Statistical Society. Series B, Methodological* 39.1, pp. 1–38. ISSN: 0035-9246.

Den Teuling, N. G. P., S. C. Pauws, and E. R. van den Heuvel (2020). "A comparison of methods for clustering longitudinal data with slowly changing trends". eng. In: *Communications in statistics. Simulation and computation* ahead-of-print.ahead-of-print, pp. 1–28. ISSN: 0361-0918.

Everitt, Brian (2011). *Cluster analysis*. eng. 5th ed. Wiley series in probability and statistics. Chichester, West Sussex, U.K: Wiley. ISBN: 0-470-97844-9.

Fung, Glenn (Jan. 2001). "A Comprehensive Overview of Basic Clustering Algorithms". In: *Citeseer*.

Genolini, C, J.B Pingault, T Driss, S Côté, R.E Tremblay, F Vitaro, C Arnaud, and B Falissard (2013). "KmL3D: A non-parametric algorithm for clustering joint trajectories". eng. In: *Computer methods and programs in biomedicine* 109.1, pp. 104–111. ISSN: 0169-2607.

Genolini, Christophe, Xavier Alacoque, Mariane Sentenac, and Catherine Arnaud (2015). "Kml and kml3d: R packages to cluster longitudinal data". eng. In: *Journal of statistical software* 65.4, pp. 1–34. ISSN: 1548-7660.

Genolini, Christophe and Bruno Falissard (June 2011). "KmL: A package to cluster longitudinal data". In: *Computer methods and programs in biomedicine* 104, e112–21. DOI: 10.1016/j.cmpb.2011.05.008.

Hau, Erich (2013). *Wind Turbines: Fundamentals, Technologies, Application, Economics*. eng. 3rd ed. 2013. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 9783642271502.

Heggeseth, Brianna C. and Nicholas P. Jewell (2018). "HOW GAUSSIAN MIXTURE MODELS MIGHT MISS DETECTING FACTORS THAT IMPACT GROWTH PATTERNS". eng. In: *The annals of applied statistics* 12.1, pp. 222–245. ISSN: 1932-6157.

Koohi, Fatemeh, Davood Khalili, Mohammad Ali Mansournia, Farzad Hadaegh, and Hamid Soori (2021). "Multi-trajectories of lipid indices with incident cardiovascular disease, heart failure, and all-cause mortality: 23ăyears follow-up of two US cohort studies". eng. In: *Journal of translational medicine* 19.1, pp. 1–286. ISSN: 1479-5876.

Magrini, Alessandro (2021). *gbmt: Group-Based Multivariate Trajectory Modeling*. R package version 0.1.2. URL: https://CRAN.R-project.org/package=gbmt.

— (Mar. 2022). "Assessment of agricultural sustainability in European Union countries: A group-based multivariate trajectory approach". In: *AStA Advances in Statistical Analysis*. DOI: 10.1007/s10182-022-00437-9.

McLachlan, Geoffrey J. (2000). *Finite mixture models*. eng. Wiley series in probability and statistics, Applied probability and statistics section. New York (N.Y.): Wiley. ISBN: 0-471-00626-2.

Nagin, Daniel (Nov. 2014). "Group-Based Modeling: An Overview". In: *Annals of nutrition & metabolism* 65, pp. 205–10. DOI: 10.1159/000360229.

Nagin, Daniel, Bobby Jones, Valeria Lima Passos, and Richard Tremblay (Oct. 2016). "Group-based multi-trajectory modeling". In: *Statistical Methods in Medical Research* 27. DOI: 10.1177/0962280216673085.

Nagin, Daniel S. (2005). *Group-based modeling of development*. eng. Cambridge, Mass: Harvard University Press. ISBN: 0-674-04131-3.

Nest, Gavin van der, Valeria Lima Passos, Math Candel, and Gerard Breukelen (Mar. 2020). "An overview of mixture modelling for latent evolutions in longitudinal data: Modelling approaches, fit statistics and software". In: *Advances in Life Course Research* 43, p. 100323. DOI: 10.1016/j.alcr.2019.100323.

Ojala, Mikko-Ilari (2013). “Testing, commissioning and additional development of neural network application for condition monitoring of wind turbine drive train”. MA thesis. Finland: University of Jyväskylä.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Salameh, Jack P., Sebastien Cauet, Erik Etien, Anas Sakout, and Laurent Rambault (2018). “Gearbox condition monitoring in wind turbines: A review”. eng. In: *Mechanical systems and signal processing* 111, pp. 251–264. ISSN: 0888-3270.

Schwarz, G (1978). “Estimating the Dimension of a Model”. eng. In: *The Annals of statistics* 6.2, pp. 461–464. ISSN: 0090-5364.

Sharma, Ram Bihari and Anand Parey (2017). “Condition Monitoring of Gearbox using Experimental Investigation of Acoustic Emission Technique”. eng. In: *PLASTICITY AND IMPACT MECHANICS*. Procedia Engineering 173, pp. 1575–1579. ISSN: 1877-7058.

Singh, Archana, Avantika Yadav, and Ajay Rana (Apr. 2013). “K-means with Three different Distance Metrics”. In: *International Journal of Computer Applications* 67, pp. 13–17. DOI: 10.5120/11430-6785.

Spinato, F, P.J Tavner, G.J.W Bussel van, and E Koutoulakos (2009). “Reliability of wind turbine subassemblies”. eng. In: *IET renewable power generation* 3.4, pp. 387–401. ISSN: 1752-1416.

Zhang, Hanze and Yangxin Huang (Mar. 2015). “Finite Mixture Models and Their Applications: A Review”. In: *Austin Biometrics and Biostatistics* 2, p. 1013.

APPENDIX A. The trajectories obtained for the GBMT model with six clusters and quadratic trajectories fit on the raw data

The trajectories of the best model obtained for group-based multi-trajectory modeling which is the model with quadratic trajectories and six clusters are illustrated in this appendix. Therefore, the trajectories of the vibration measurements from X, Y and Z coordinates of the gearboxes in each group are shown below in the figures 1, 2, 3, 4, 5, 6 starting from the group one to the group six, respectively. It must be noted that the effect of pruning in order for allowing the backward elimination of the non-significant terms is clearly visible in all the figures where some of the trajectories are first degree polynomials.

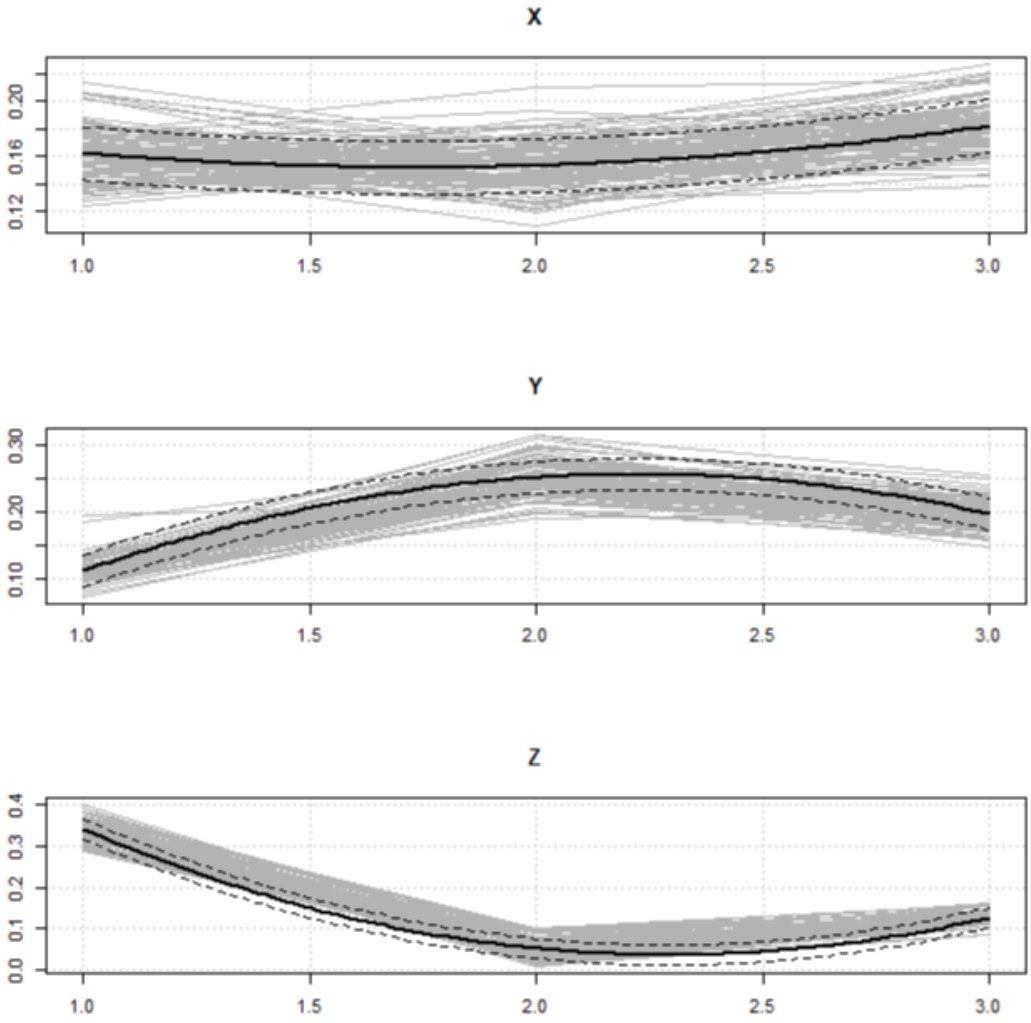

Figure 1 *Trajectories obtained in the first cluster of the best model*

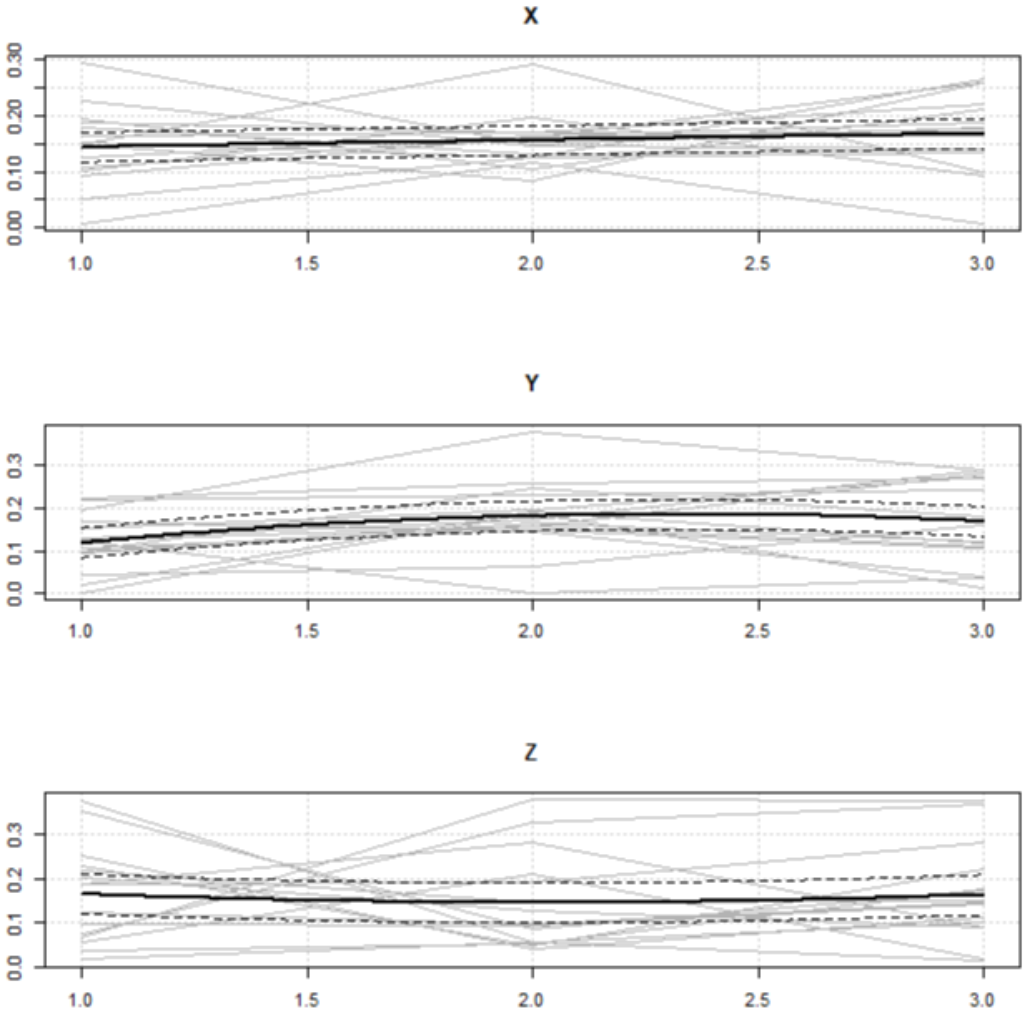

Figure 2 *Trajectories obtained in the second cluster of the best model*

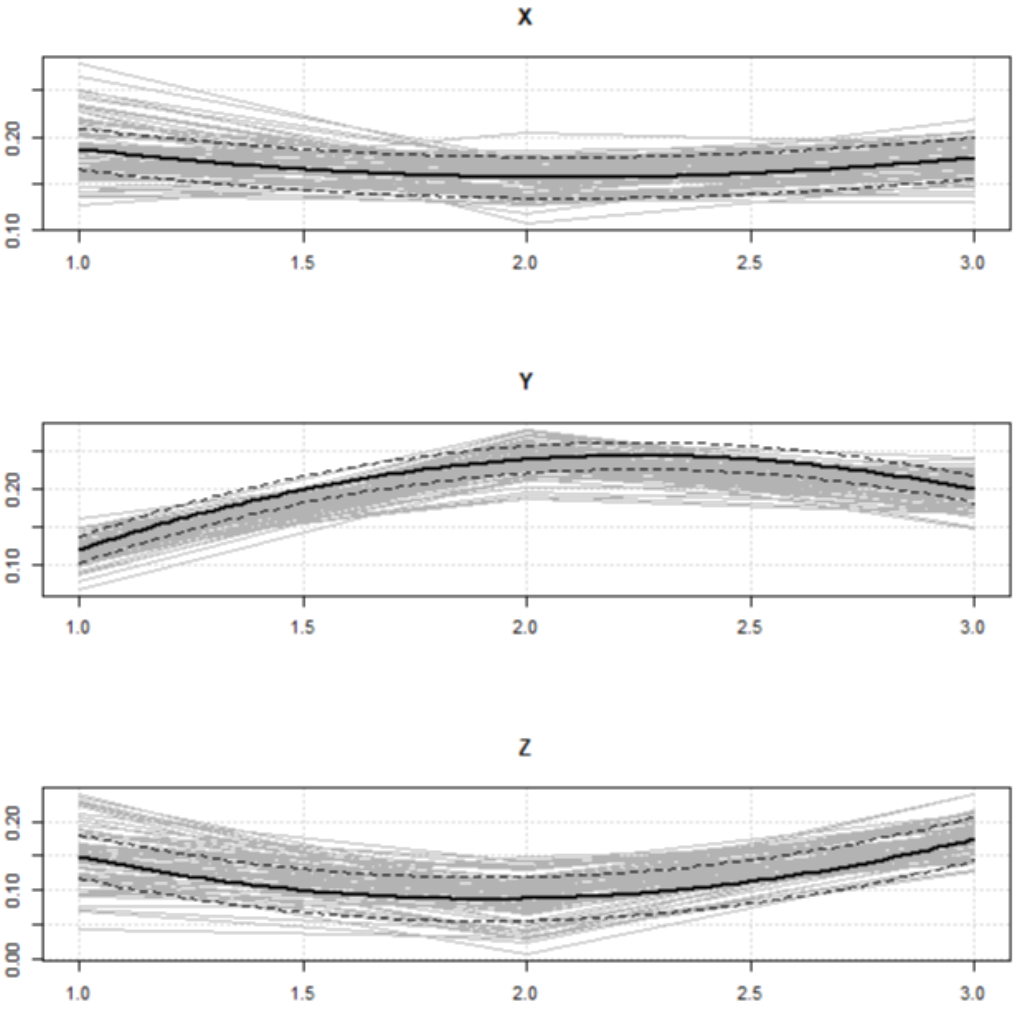

Figure 3 *Trajectories obtained in the third cluster of the best model*

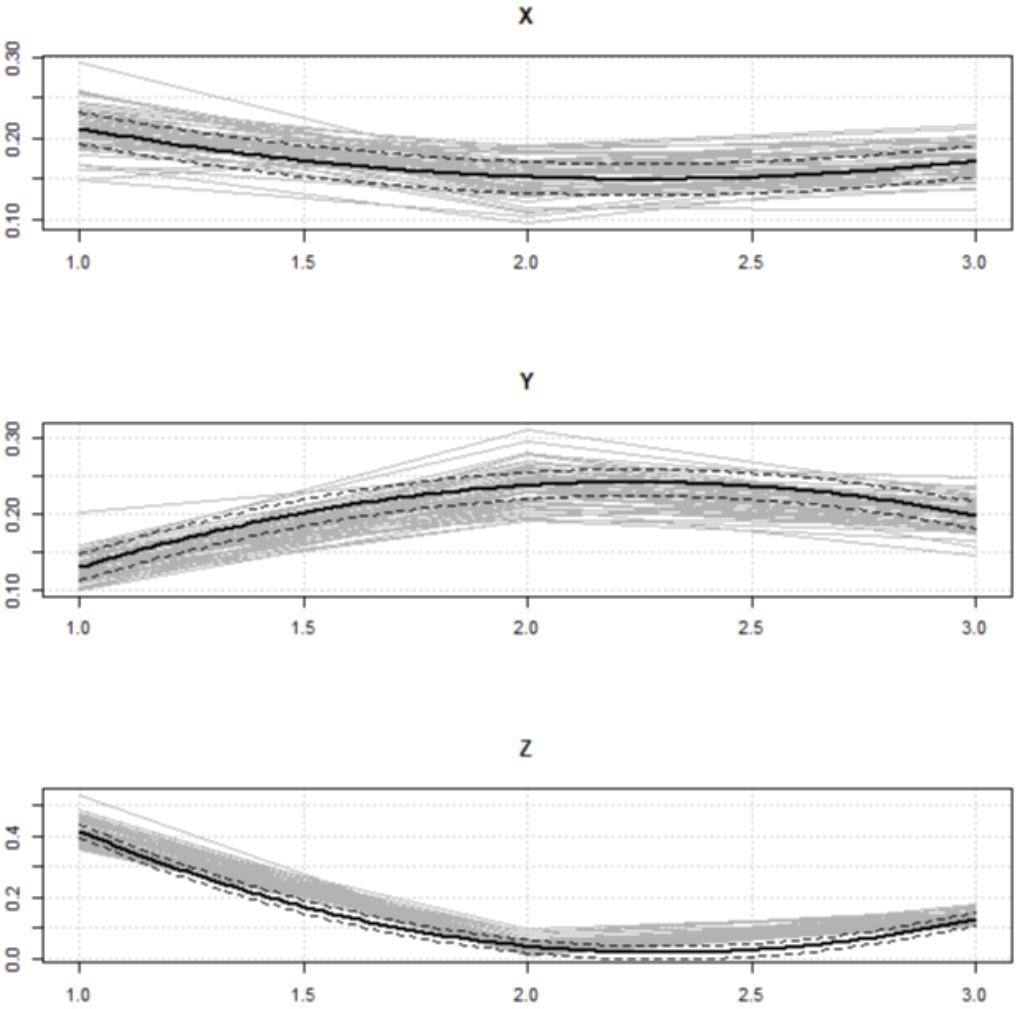

Figure 4 *Trajectories obtained in the fourth cluster of the best model*

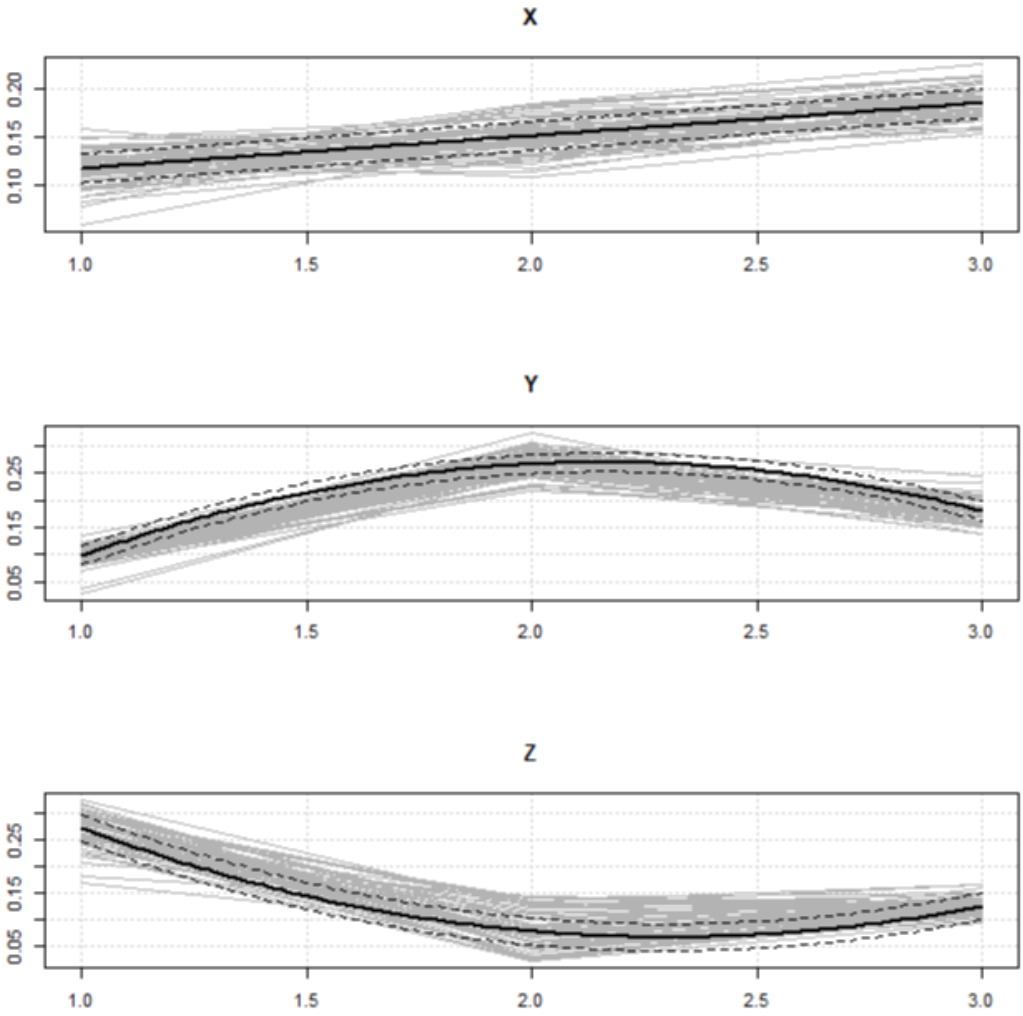

Figure 5 *Trajectories obtained in the fifth cluster of the best model*

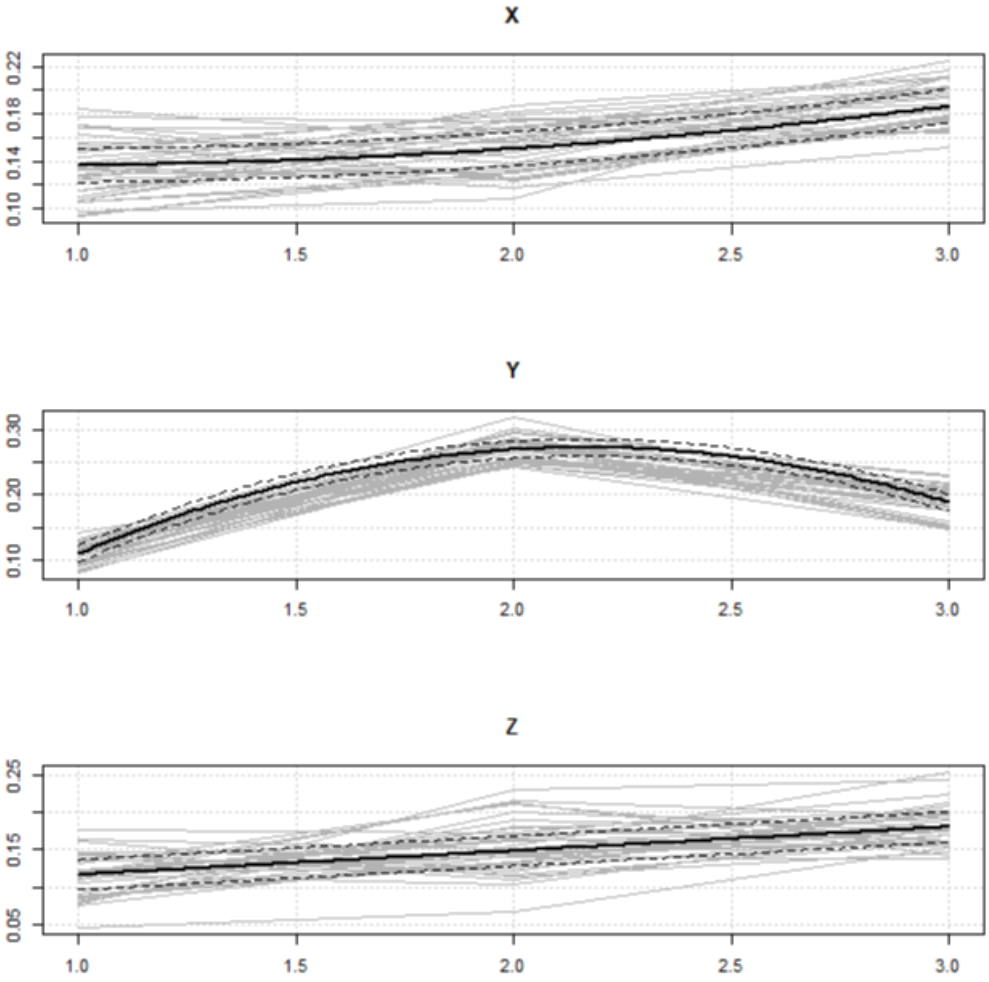

Figure 6 *Trajectories obtained in the sixth cluster of the best model*