AN UNSUPERVISED CLUSTERING METHODOLOGY BY MEANS OF AN IMPROVED DBSCAN ALGORITHM FOR OPERATIONAL CONDITIONS CLASSIFICATION IN A STRUCTURE

JUAN CARLOS PERAFAN LÓPEZ

UNIVERSIDAD PONTIFICIA BOLIVARIANA
ESCUELA DE INGENIERÍAS
MAESTRÍA EN INGENIERÍA
MEDELLÍN
2018

AN UNSUPERVISED CLUSTERING METHODOLOGY BY MEANS OF AN IMPROVED DBSCAN ALGORITHM FOR OPERATIONAL CONDITIONS CLASSIFICATION IN A STRUCTURE

JUAN CARLOS PERAFAN LÓPEZ

Trabajo de grado para optar al título de Magíster en Ingeniería

Director
Julián Sierra Pérez
Doctor of Philosophy

UNIVERSIDAD PONTIFICIA BOLIVARIANA
ESCUELA DE INGENIERÍAS
MAESTRÍA EN INGENIERÍA
MEDELLÍN
2018

DECLARACIÓN DE ORIGINALIDAD

"Declaro que esta tesis (o trabajo de grado) no ha sido presentada para optar a un título, ya sea en igual forma o con variaciones, en ésta o cualquier otra universidad". Art. 82 Régimen Discente de Formación Avanzada, Universidad Pontificia Bolivariana.

Firma Autor:

To My Dear Mom

ACKNOWLEDGEMENTS

I would like to express my gratitude to my family, specially my parents, my MSc. research work director PhD. Julián Sierra, and to the *Universidad Pontificia Bolivariana* for the continuous support.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Structural Health Monitoring (SHM) is highly relevant nowadays, not only for aerospace maintenance, but for a large number of newly engineering applications. Pattern recognition has become an important part of SHM for signal processing and anomalies or damage detection, assuring structural integrity. New methods are created day by day and more researchers and engineers feel the interest to generate techniques which can make SHM become a more compacted, sophisticated and automatized system, eliminating human factors and intrinsic errors. This work evaluates the computational complexity and accuracy of a novel methodology of unsupervised clustering called FA+GA-DBSCAN which employs a combination of machine learning techniques including factor analysis for dimensionality reduction and a density clustering algorithm called DBSCAN enhanced with a genetic algorithm. In order to automatically detect a variety of structural behaviors using the novel methodology an experiment with a beam in cantilever under dynamic loads was taken in consideration.

Keywords:

Clustering, operational conditions, pattern recognition, SHM, unsupervised learning.

RESUMEN

El monitoreo de salud estructural (SHM) es altamente relevante ahora, no solo en el mantenimiento aeroespacial, sino también en un sinnúmero de nuevas aplicaciones ingenieriles. El reconocimiento de patrones se ha convertido en una parte importante en el SHM para el procesamiento de señales y detección de anomalías o daños que aseguran una integridad estructural. Día tras día son creados nuevos métodos en los que ingenieros e investigadores sienten el interés de generar nuevas tecnologías que puedan convertir al SHM en un sistema más compacto, sofisticado y automatizado, eliminando los factores humanos y errores intrínsecos. Este trabajo evalúa la precisión y el costo computacional de una metodología novedosa de *clustering* no supervisado que ha sido llamada FA+GA-DBSCAN la cual emplea una combinación de técnicas de aprendizaje automático incluyendo análisis factorial para reducción dimensional y un algoritmo de agrupamiento por densidad llamado DBSCAN mejorado con un algoritmo genético. Para detectar automáticamente una variedad de condiciones operacionales utilizando esta metodología novedosa un experimento con una viga en voladizo bajo cargas dinámicas fue tenido en consideración.

Palabras Clave:

Aprendizaje no supervisado, *clustering*, condiciones operacionales, reconocimiento de patrones, SHM.

1 INTRODUCTION

Historically, classification has been one of the most important methods for the modern human being to keep things in order. Classification, in a primitive way, is the simple action to group objects in a specific order [1]; initially, classification was used by humans to survive, they made different classifications between similar species of animals depending on their benefits or their dangerousness and the same could have happened also with poisonous and edible plants.

After the ancient and rudimentary classification models created by Greek and Roman thinkers, the concept of taxonomy was introduced and became an important field of scientific research since Swedish Carl Linnaeus' era (1707-1778), who described, classified and gave a scientific name to a great number of living things [2]. Simpson [1] gave a general concept of taxonomy as the "theoretical study of classification, including its foundation, principles, procedures and rules", being this one of the first steps in the construction of pattern recognition as an area of study.

Currently, pattern recognition has become the science behind classification and it has several applications in many science fields. While technology improves every day, the amount of information in relation with it increases; pattern recognition is a part of machine intelligence devoted to decision making, as a result of a need for handling information in an automatic way [3]. Machine intelligence is also known as machine learning and its goal is to make computers capable of learning from information, understanding and processing it through algorithms [4].

Bishop states that pattern recognition is "a field that is concerned with automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions such as classifying the data into different categories" [5]. Classification has facilitated the development of technology as it has helped to understand new phenomena looking for features that describe them [6], those features can be a starting point to find other similar characteristics over new information based

on similarity or dissimilarity between features and this new information [7].

Pattern recognition is divided into two major groups: *supervised pattern recognition* classifies objects in previously known classes, determining to which of these classes new data belongs [8]. In other words, supervised learning can generate a set of rules using classifiers to generalize new information [9]. The type, extension of damage and the remaining life of the structure system can be determined using supervised learning attached to analytical models [10]. *Unsupervised pattern recognition*, unlike supervised pattern recognition, is used when labeled data are not available. Unsupervised pattern recognition is also called clustering [11] and its goal is to search for groups in the data universe that discriminate data in a particular group. Clustering methods can work as a part of a supervised method, looking for the representative classes [12].

Pattern recognition algorithms have the ability to detect related damage characteristics using obtained features. Damage in structures can be considered as a physical change in the system which can affect it negatively on its regular performance [13], besides it can put lives, operations and money at risk. The need of having global information about structural health has had an impact in the way maintenance is done. SHM, which by now is one of the most notable methods of maintenance and damage prevention in many structures, consists on an implementation of a system or strategy for damage detection.

Aerospace, civil, and mechanical engineering are the most concerned fields in SHM [14] since it helps to reduce the risk due to the human factor, reducing the redundancy methods, time and costs without affecting safety [15]. SHM-related algorithms usually fall in one of three categories depending on the availability of information about damaged and pristine structure: *group classification* when the algorithm has the ability to classify pristine and damaged feature characteristics, *analysis of outliers* when there are not available data to make a comparison and *regression analysis* when data are correlated with particular types of damage including information about their extension and location [13].

In this work, a methodology for operational conditions classification in a structure under dynamic loads in which the relationships between strain signals sensed by means of Fiber Bragg Gratings FBGs and operational loads that are unknown is presented. An unsupervised clustering methodology which combines a variety of machine learning techniques including factor analysis for dimensionality reduction and a genetic algorithm

around of a density clustering algorithm called Density Based on Spatial Clustering of Application with Noise DBSCAN for operational condition classification is presented. The computational complexity and accuracy include a comparative study with a similar methodology based on a Local Density-based Simultaneous Two-Level Self-Organizing Maps DS2L-SOM clustering developed by Sierra [16].

1.1 APPROACH OF THE PROBLEM

The continuous measuring of strain in a structure using fiber optic technologies is a promising technology [17]. Besides, it is one of the methods used for SHM in which clustering techniques are involved for novelty detection or anomaly detection methods [18, 19, 20, 21], however, the clustering techniques used for SHM have received less attention in the technical literature [10]. The novelty or anomaly detection methods are specially used in structures in which, under their normal operational conditions, a variety of dynamic loads can be present simultaneously.

In general, the mechanical behavior conditions associated to an aircraft's structure are presented in other kind of systems such as offshore oil platforms, bridges, ships, wind turbines, among other. In such system the operational conditions may vary due to thermal variation, vibration and acoustic environments, changing mass (fuel consumption), aerodynamic forces due to atmospheric variations, among others [10].

Aircraft structures, are usually divided in major groups: wings, fuselage, tail units and control surface; the configuration into these groups may vary depending on the aircraft's final application. Aircraft structures may support two different major loads: ground loads related to movement and transportation on the ground, and air loads which are the loads presented due the flight maneuvers.

Besides, the forces induced in the aircraft can be divided into *surface forces* related to the forces applied in the surface of the structure (aerodynamic forces) and *body forces* produced by the interaction between the structure and the gravitational and inertia effects. Therefore, all aerodynamic loads are the result of the pressure distribution across the surface produced by flights maneuvers and external conditions.

The loads in the aircraft's principal structures (e.g. the wing), can be defined as a result of direct loads, bending, shear, and torsion in addition to pressure loads [22]. Since the number of loads presented in the wing is large, a constant monitoring in this type of structures is decisive for its proper performance. Unfortunately, some of the aircraft's principal structures such as the wing, have portions which are difficult to reach, thus, a lightweight and non sparkling sensing system is a good option for monitoring their health condition. However, not every pattern recognition technique may work adequately, as a part of a monitoring system, considering the aircraft's structures natural response due to changes in loads.

Even if the structure and material physical behaviors are well understood, modeling the structure's behavior under a variety of loads could be an extensively time-consuming process. In some cases, due to the complexity of the structural behaviors, the relationships between normal operational conditions and strain signals are unknown and it becomes necessary to identify such relationships in order to discriminate pristine strain signals from damage strain signals. The use of an unsupervised clustering technique in order to look for such relationships may be an accurate methodology following the concept of novelty detection presented by Farrar and Worden [10].

To determine such relationships, a simplified structure which represents a system's main structure such an aircraft's spar beam was taken in consideration in a experiment carried out by Sierra [16]. A general representation of the experimental setup is presented in Figure 1. The experiment setup consists in a hollow rectangular aluminum beam with a 20 mm for 40 mm cross section, 1 mm thickness (See Figure 2) and a cantilever length of 1200 mm. The region where the beam was fixed to the testing workbench using C clamps, was filled with a wooden core to prevent plastic deformation.

The beam was submitted to dynamic bending loads using the same amplitude at the beam's tip, at a frequency around 4 Hz. The maximum amplitude at the beam's tip was around 14mm. Thus, it was expected loads around 4 3Kg and strains around 700 $\mu\varepsilon$. The bending loads were applied in the opposite end of the clamping zone, using a rod connected from the beam's longitudinal axis to an electric motor. Experiments included 13 variations of $2\,^\circ$ in the pitch angle in order to generate data between $-8\,^\circ$ to $16\,^\circ$.

The electric motor is not considered as a part of the experiment, and it does not rotate in conjunction with the beam. The intention of changing the pitch angle is to simulate

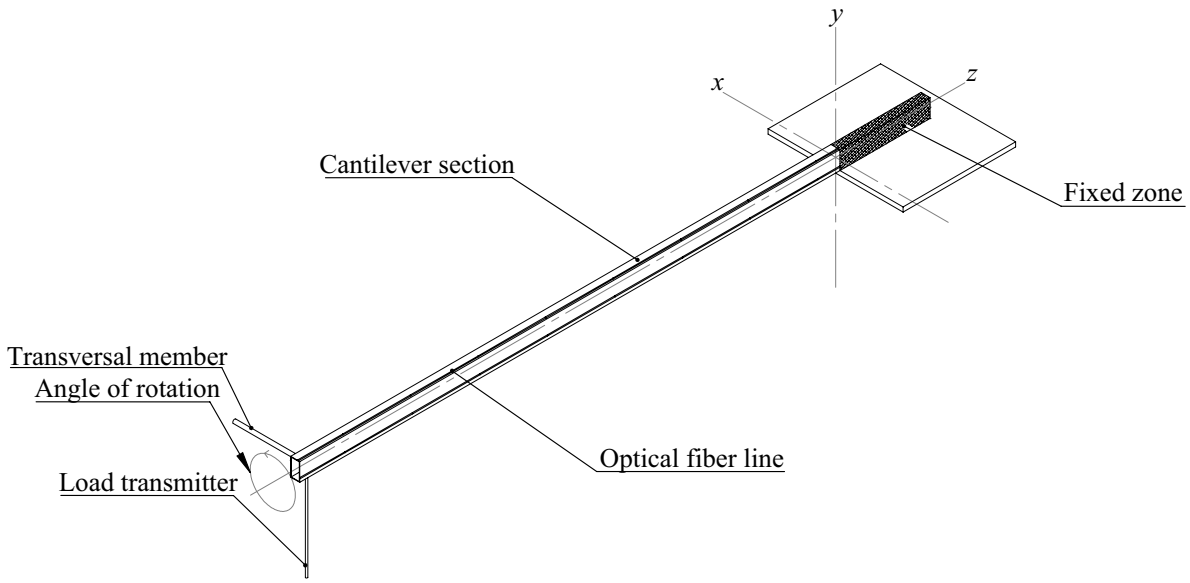


Figure 1. Experimental setup.

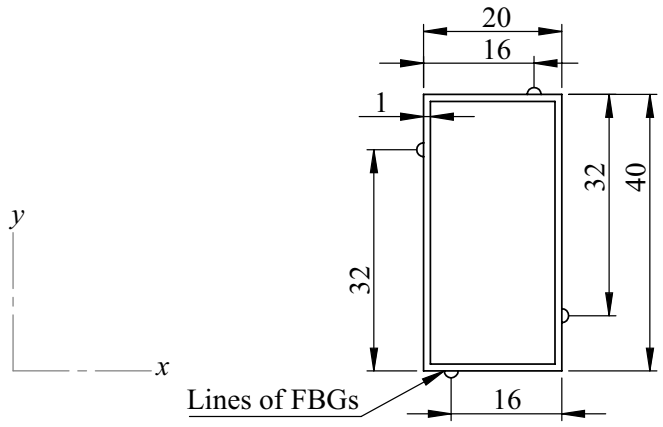


Figure 2. Cross section (all measures are given in mm).

an aerospace structure (e.g. wing's spar beam or a wind turbine blade's main structure where a set of variable pitch conditions can be presented). A pitch angle indicator was placed in a side of the beam to have control on the pitch shift in every required experiment. Some of the variations are presented in Figure3

Although, the technique carried out by Sierra, DS2L-SOM, was capable to determine such 13 variations into 12 different clusters in the aluminum beam's experiment with high accuracy, the computational complexity was a drawback since it is desired in the

near future to implement a damage detection system in an Unmanned Aerial Vehicle UAV.

Thus, it is considered appropriate that a new methodology designed to avoid these drawbacks is needed. Hence, in the current work it was intended to evaluate off-line, if there is an unsupervised clustering methodology with the ability of classifying unknown operational conditions using strain signals with low computational complexity and reliable accuracy which could be part of a future SHM methodology for damage identification.

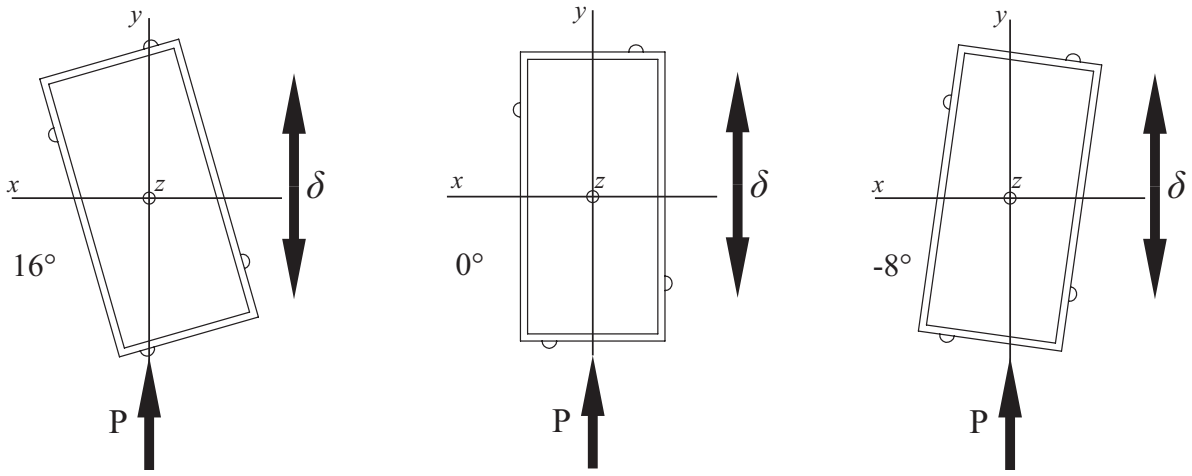


Figure 3. Representation of $-8\,^{\circ}$, $0\,^{\circ}$ and $16\,^{\circ}$ pitch angle variations.

1.2 OBJECTIVES

1.2.1 Main Objective

To develop an unsupervised clustering methodology by means of an improved DBSCAN algorithm in order to classify operational conditions from strain signals measurements.

1.2.2 Specific Objectives

- To apply preprocessing techniques in which data cleansing and scaling are included in order to improve the quality of characteristics related with the system's physics (features).
- To process strain signals by means of a dimensionality reduction technique in a suitable form that allows a following clustering by using a density based algorithm.
- To determine an unsupervised density-based algorithm for which previous knowledge of the number of clusters is not required.
- To implement the methodology of the unsupervised clustering algorithm in a programming language.
- To determine the accuracy and computational complexity of the proposed clustering methodology in comparison with a tested technique based on a Local Density-based Simultaneous Two-Level Self-Organizing Maps (DS2L-SOM) clustering.
- To test the proposed clustering methodology in a real case with strain signals obtained from an instrumented Unnamed Aerial Vehicle wing's beam flying under regular operational conditions.

1.3 Outline

The proposed methodology for automatic operational conditions classification in a structure is developed considering a experiment developed by Sierra [16]. Section 2 deals with the state of art of different approaches to dimensionality reduction methods however, its major focus is related with solutions for SHM problems using unsupervised learning techniques. Section 3 deals with an outline of proposed axioms and paradigms presented in the use of SHM and pattern recognition. Section 4 presents briefly the reception and the preliminary processing of experimental data. Section 5 deals with the dimensionality reduction of the experimental data; a technique called factor analysis FA was selected for this task. A Machine learning technique selected for clustering the dimensionality reduced data from the experiment called DBSCAN is presented in Section 6. Here the DBSCAN algorithm is improved by means of a genetic algorithm. Section 7 introduce the pseudocodes of the proposed methodology. Section 8 presents the the precision and computational complexity of the proposed clustering methodology

in comparison with a proven classification technique called DS2L-SOM. Section 9 deals with an approach of the methodology in a real case scenario, considering an acquisition system flying prototype. Finally, conclusions and additional comments are presented.

2 STATE OF THE ART

Several approaches have been made during the last decade to detect operational conditions in structures or structure's monitoring and prevent the damage growth; those approaches are included in one of five general subjects reviewed by Farrar and Worden [23]: SHM, condition monitoring [24], non destructive evaluation [25], statistical process control [26] and damage prognosis [27]. However, some subjects are common for most of the available proposed methodologies such as the use of statistical and mathematical tools, the handling and physical interpretation of a large amount of information, and SHM is not the exception.

2.1 DIMENSIONALITY REDUCTION METHODS

In addition to engineering, there have been parallel developments on statistics and mathematics, specially to establish reduction techniques in fields such as astronomy, biology, remote sensing, economics, psychology and consume transactions, due to the large amount of data involved in actual problems. Besides, there is an obstacle on handling big data problems with traditional statistical methods [28]. The data dimensionality reduction process is also known as *data compression*. The implementation of a data reduction technique is inherent to an SHM methodology due to the large amount of data [10]. D.L. Donoho [29] states that high dimensional datasets needed new theoretical backgrounds and methodologies to handle large quantities of information.

Nowadays, several reduction techniques have been developed in order to describe big data problems in a lower dimensional representation. Among the most widely used methods for dimensionality reduction rely the techniques based on orthogonal projections due to their simple geometric interpretation in a lower dimensional space. Those techniques are known as linear dimensionality reduction methods with orthogonal ma-

trix constrains [30]. According to K. Fodor [28], the most widely used dimensionality reduction techniques within the group of linear dimensionality reduction techniques are principal component analysis or PCA [31] and factor analysis, FA [32]. Besides, other developments in dimensionality reduction techniques have been explored, like: projection pursuit [33], independent component analysis, [34] and random projections [35]. Among other non-linear methods have been explored lately such as non linear PCA [36], principal curves [37], multidimensional scaling [38], topologically continuous maps (self-organized maps) or SOM [39], neural networks or NN [40], and genetic algorithms for dimensionality reduction [41].

The selection between factor analysis and principal component analysis is not simple. The computational issues and the performance of the algorithms facing a specific dataset leads to an analysis of which of those algorithms is more suitable. FA and PCA aim to the same goal: to describe a large number of observed variables into a compacted and reduced number of new representative ones.

W. Velicer and D. Jackson [42] established an extensive comparative study between PCA and FA. The two methods differ in the fact that in FA a reduction of the variance or diagonal elements of the covariance matrix is involved, otherwise, PCA does not allow operation on the diagonal elements, thus, the covariance or the elements out of the diagonal will not be affected.

The algebraic variations between techniques lead to different results, however, it is common to see that if the same number of components or factors have been selected the results may be similar. If a dataset analyzed by the two methods produce different results, it is precise to redefine the number of factors selected. It has been found that PCA's computational complexity is slightly lower than FA's, although, if the problem is well defined the convergence of FA is faster.

Otherwise, FA is aimed to manage latent or unobserved variables unlike PCA which is intended to describe manifested or observed variables. W. Velicer and D. Jackson also affirmed that a drawback for FA is that one of its techniques to fit a model works adding a random component to the linear weighted result by an iterative process, this technique is known as Maximum Likelihood factor analysis and sometimes may lead to convergence problems.

PCA and FA are partly exploratory and confirmatory analysis, nevertheless, PCA is a more fashioned technique, in consequence it has been frequently used in a variety of fields of study. PCA has been widely applied in the framework of SHM [43]. Ni et al. [44] presented a PCA methodology for dimensionality reduction of frequency responses for damage detection. Mujica et al. [45] explored statistical indexes based on PCA for a damage detection. de Latour [46] performed a damage classification and estimation using PCA for dimensionality reduction with acceleration time series. Sierra et al. [47] presented a methodology for SHM based on the Self Organized Maps and PCA statistical tools. Katsikeros et al. [48] developed a methodology for damage detection in an aerospace lap-joint structure using strain signals. Rao et al. [49] created a comparison study with strain signals detection in a variety of signal sensors using PCA. Magalhães and Caetano [50] developed an SHM study using an on-line modal analysis in a bridge located in Portugal. A database was created and a dynamic regression complemented with PCA was performed for damage identification.

Nevertheless, FA is a technique that can not be rejected at all, since it has some derived useful features when the retained factors describe a specific number of the original variance. The common factors or factor scores may become a helpful tool to detect clusters or outliers if they work together with a pattern recognition algorithm which may also compensate the overall complexity cost of the clustering methodology if it is selected adequately. FA has not been extensively used; overall, FA has been used in the SHM framework for removing environmental effects such as temperature or humidity in an SHM system [51, 52, 53].

2.2 UNSUPERVISED PATTERN RECOGNITION IN SHM

As it was discussed in Section 1, the algorithms for handling a large amount of information such as supervised pattern recognition and unsupervised pattern recognition may be part of an SHM methodology. In this study case, the attention was focused on SHM-related applications with unsupervised pattern recognition used for novelty detections, such as the existence of damage in a structure and in some cases the location of it [10].

2.2.1 Unsupervised artificial neural networks

Unsupervised artificial neural network ANN for SHM works in a similar way as supervised ANN, but there is not a priori knowledge about structure damage [17], usually studies named as unsupervised NN, are developed with trained supervised NN following to an unsupervised clustering algorithm method.

Dervilis et al. [54] presented a group of machine learning techniques including multilayer perceptron. A method based on auto-associative networks using radial basis function RBF in which it can be generated a statistical representation of training data with the help of pristine information was presented. Afterwards, unknown information was compared with previous undamaged information for damage detection.

Wen et al. [55] presented a methodology with the use of modal parameters on a structure introducing a feature representation called damage localization feature DLF and a learning model based on ANN and an unsupervised fuzzy algorithm, called unsupervised fuzzy neural networks UFN. The study was made on a five story frame building damaged, also the influences on measured noise and incomplete modal data were evaluated.

2.2.2 Fuzzy clustering algorithms

Unlike probabilistic algorithms in the fuzzy clustering algorithms a feature can be simultaneously part of more than one cluster. A representative point is used in case of compact clusters, with dissimilarity measured between two points, depending on it, the resulting algorithm can be fuzzy C-means or fuzzy k-means algorithm [3]. Unsupervised fuzzy clustering is based on the optimization of a fuzzy objective function where it is not necessary to specify the number of clusters, usually those kind of cluster methods combine Fuzzy-c means and K-nearest neighbors [56], leading different cluster results [57].

The article presented by Baraldi et al. [58] illustrates a real industrial case of fault diagnosis on a steam turbine of a nuclear plant as a particular requirement of Electricite de France (EDF) who had the need of unlabeled transient conditions identification by operational or faulty circumstances based on 148 shut down transients. It was used

pattern recognition for undefined signals behavior evolution classification, the general aim in this publication was to detect similar vectors with different behaviors.

2.2.3 Gustafson-Kessel algorithms or GK algorithms

Following the previous topic Gustafson-Kessel algorithms or GK algorithms are based in the fuzzy methodology, GK algorithm was made by D. Gustafson and W. Kessel [59] with the intention to have visualization and differentiation between classes. In GK algorithm is made a natural metric generalization through the use of a fuzzy covariance matrix instead hard covariance matrix, arguing that a crisp membership is not quiet realistic because pattern vectors will have characteristics of several classes, furthermore a set of memberships were assigned to a pattern vector.

Dinh et al. [60] presented an article in which acoustic emission AE data from a Carbon Fiber Reinforced Polymers CFRP composite were used for unsupervised pattern recognition (natural clustering) using GK clustering. The work presented by Dinh et al. was followed by some others studies such as the paper presented by Doan et al. [61] which presented a methodology for fatigue detection based on acoustic emission AE and piezoelectric sensors on CFRP composite element (composite split disks), this component presented a high noise data into a massive group of data becoming more difficult damage detection. GK algorithm and Quasi-Static QS test for low data obtainment were used for clustering.

Feature selection and clustering optimization was made by reducing the value the DB index proposed by Davies and Bouldin [62], then, the set of denoised features S were partitioned using a GK cluster algorithm (having in consideration the DB index) selecting a feature f by iteration to a current set of features S. The Mahalanobis distance defined by the GK algorithm was used for a distance estimation between a signal and a cluster.

Ramasso et al. [63] proposed another approximation of early detection of damage in reinforced carbon fiber matrix using unsupervised clustering, with the use of AE time series signals. The method presented a clustering optimization procedure using probabilistic formulation instead distance measures, clusters were obtained automatically by multiple subsets of features, a variety of clustering methodologies were taken in

consideration in a technique called consensus clustering, but GK algorithm had special attention for data scattering, with this technique sensitivity of kinetics and onsets of damage detection was high.

2.2.4 Self-organized maps SOM

Self-organized maps or SOM were proposed by Kohonen [39], it is based in a topologically ordered mapping from signals in a a neural network by regression, those neurons have the ability of developing into specific decoders or detectors of their respective signal in a meaningful order. Moreover, a feature coordinate system in a network is defined, those mappings are known as self organized maps; the map of features can be used as a part of a pattern recognition methodology.

Sierra et al. [47] developed a methodology for SHM using a U-Matrix for clustering visualization based on SOM and PCA's statistical tools Q and T^2 indexes [45], for automatic damage detection. The heuristic procedure was made by placing FBGs in a variety of structures in which different kind of damages were induced then data were from pristine and damaged structures were acquired. A baseline was created after PCA was applied to data, moreover an automatic clustering methodology DS2L-SOM was performed to process the information and determining the changes on load conditions.

A non-parametric algorithm based on vibration damage detection was developed by Avci and Abdeljaber [64], and presents a novelty SOM technique. A structural damage quantification following the previous article was presented by Abdeljaber et al. [65] in which damage detection SOM algorithm was implemented on a Phase II Experimental Benchmark Problem of SHM data, validated in a finite element method FEM experiment.

Five different cases were considered for creating artificial damage. In order to generate an SOM topology, accelerometer signals were measured under random excitation for damaged and undamaged conditions, two different groups of accelerometers were taken in consideration. To training the algorithm features from accelerometers response were created, then a baseline was made following the original algorithm. The structure was equipped with 15 accelerometers. A numerical demonstration was made with five damage cases and 10 damage indexes were computed. The algorithm shown a high

sensitivity to damaged structures numerically simulated.

2.2.5 Genetic algorithms GA

Genetic algorithms are inspired in natural selection, operators are applied to a population of solutions, *offspring* are compared with its previous population according to a specific function [3]. GA can be used for fitting quantitative models and for parameters selection for optimizing the performance of a system [66]. A deeper explanation of the GA algorithms is presented in Section 6.4. In SHM, GA are also known as bio-inspired damage detection methods.

A newly unsupervised method based on GA is presented by Silva et al. [67] called Genetic Algorithm for Decision Boundary Analysis GADBA developed for damage detection in bridges; a dataset which belongs to the Z-24 Bridge in Switzerland and Tamar Bridge in England in presence of operational and environmental influences were taken in consideration for the experiment. It was carried out an algorithm for clustering optimization by a concentric hypersphere algorithm, as well, it was developed a technique for clusters characterization.

A paper presented by Betti et al. [68] associates ANN and GA for structural damage identification. The methodology was applied on a three-story steel frame instrumented with 12 accelerometers, modal singularities were taken into consideration. Loads were applied to the structure, an ANN was trained by four specific signals spectra looking for structure eigenvalues from their modes and frequencies. Afterwards, data output were used to build update a finite element model. Besides, a GA algorithm was tested to recognize damaged areas. Results showed that, the associations of ANN and GA were effective for damage identification.

Meruane et al. [69] developed a methodology called hybrid real coded genetic algorithm or Hybrid-RCGA, to detect structural damage in aerospace frame structures, in a single and multiple damage scenarios. Different objective functions were studied, as frequencies, modal analysis and strain energy, then, it was selected the one that had better convergence. A damage penalization was convenient to determine a false alarm detection problem caused by noise, finally, results were compared with conventional methods.

2.2.6 Algorithms based on cost function optimization (K-means)

Algorithms based on cost function optimization (K-means) are one of the most common unsupervised algorithm, K-means clustering algorithm was introduced by Hartigan and Wong [70]. K-means divides a number of points in a specific dimension into a number of clusters using the (K-initial cluster centers) at the beginning.

A damage sensing system presented by Diez et al. [71] employed a novelty unsupervised algorithm applied in the Sydney harbor bridge (Australia) for clustering a subgroup of joints with similar behavior and detected failures or anomalies by vibration data. It was not previous information about those anomalies, thus an unsupervised method was performed. Tri-axial accelerometers were placed on the bridge, then, there were collected structure's vibration caused for the passing vehicles.

The methodology followed a feature extraction and outliers removal. Signal processing were performed by a Fast Fourier transform FFT for representing the spectra in a new sequence. Finally a behavior characterization was carried out by a K-means clustering. Once the FFT was applied to the signals, K-means was performed with an Euclidean metric detecting damage patterns represented by small clusters, which were used afterwards to classify unknown data.

Santos et al. [72] established a methodology to classify different structural conditions in a bridge's cable through an on-line concept without data references. Neural networks were used to estimate the actual structural condition, then, a clustering methodology were applied using K-means with Gowda-Diday dissimilarity measures. Then, the methodology was used with numerical and experimental data, and real-time detection capabilities were reached by ANN and a K-means algorithm. The methodology was successful and highly sensitive to damage detection allowing the algorithm to detect small stiffness reductions.

Al-Jumaili et al. [73] performed a study of transient acoustic emission AE signals from a carbon fiber laminate buckling test by the implementation of an unsupervised clustering technique, besides, a hierarchical clustering was used to group distinctive features. Data reduction was performed using a combination of PCA and Fuzzy C-means, then, K-means was used for unsupervised pattern recognition. A clustering quality criterion was taken in consideration looking for an optimal number of clusters.

The methodology presented meaningful results for damage mechanisms detection and AE signals classification.

3 SHM USING PATTERN RECOGNITION

A change in the original conditions in a structure due to external or internal forces may derive in damage, this involves mechanical relationships that can be determined since the very first steps of the structure's fabrication process. As time goes by, the complexity of aerospace structures and the necessity of constant monitoring of structural systems have impacted directly the way maintenance is carried out.

Non destructive testing NDT methods have been the cutting edge of damage detection in the last few decades but some of those methods are time consuming and imply high costs. SHM can be determined as the natural evolution of the NDT methods since the SHM process can help in the optimization of structures, less weight, and improving the economic income by reducing the maintenance expenditures [74]. Moreover, over the last 20 years, authors have carried out the SHM process to a state of maturation in which a group of axioms have emerged. Worden and Farrar [75] have developed the axioms which are presented as follows:

Axiom I. All materials have inherent flaws or defects.
Axiom II. The assessment of damage requires a comparison between two system states.
Axiom III. Identifying the existence and location of damage can be done in an unsupervised learning mode, but identifying the type of damage present and the damage severity can generally only be done in a supervised learning mode.
Axiom IVa. Sensors cannot measure damage. Feature extraction through signal processing and statistical classification is necessary to convert sensor data into damage information.
Axiom IVb. Without intelligent feature extraction, the more sensitive a measurement is to damage, the more sensitive it is to changing operational and environmental conditions.
Axiom V. The length and time scales associated with damage initiation and evolution dictate the required properties of the SHM sensing system.

Axiom VI. There is a trade-off between the sensitivity to damage of an algorithm and its noise rejection capability.

Axiom VII. The size of damage that can be detected from changes in system dynamics is inversely proportional to the frequency range of excitation.

The integrity of the structure can be determined through a specific array of sensors incorporated directly in the structure which can register a series of physical or mechanical magnitudes (damage-sensitive features) such as strain, acceleration, vibration, etc. via on-line (real time) or off-line. Some SHM techniques like the strain-based or vibration-based damage detection often imply the use of pattern recognition methods, specially unsupervised pattern recognition techniques because most of the time there is no information available of the damaged system [76].

The paradigms established for SHM and damage detection proposed by farrar et al. [13] were followed in the present work as a guideline to construct the general methodology. Roughly Farrar et al. argue that it is necessary to generate a reliable baseline of the pristine structure to obtain a robust damage detection system. Thus, if the behavior of the structure under a combination of loads in which the structure remains pristine are unknown, unsupervised clustering is a good way to generate such baseline. Although, some actions have to be taken in consideration before achieving a robust damage detection methodology:

3.1 SYSTEM EVALUATION

The operational evaluation as it can be seen in Figure 4, delimits the monitoring outlook from the way that it will be done and how it will be accomplished. The system evaluation has to respond to important questions that will shape the specific focus of the SHM process in a specific application such as what the life, safety or economic justifications are, how the damage is investigated, what the operational and environmental conditions in which the system has to be monitored are, and what the limitations in acquiring data under a specific operational conditions are.

Figure 4. System evaluation.

3.2 ACQUISITION METHODS

This part of the methodology is based on a strategy that has to be designed taking into account the system evaluation, the main objective here is to determine sensitive damage-inferring sensors. The number and location of sensors is determined here. Besides, data acquisition, hardware for transmission and storage, which depends on the sampling frequency are selected. Sometimes the data acquisition is performed continuously, (e.g. if a crack propagation is a concern). An example of an acquisition methods is presented in Figure 5.

However, the amount of data will increase significantly, thus, a signal selection boundary needs to be determined to preserve just the most relevant information. The constant monitoring in the change of the strain field in a structure under specific load conditions is based on this paradigm. Since the data is collected under a variety of external circumstances, it is helpful to scale the data to support the comparison of data measured at similar times in a specific environmental scenario.

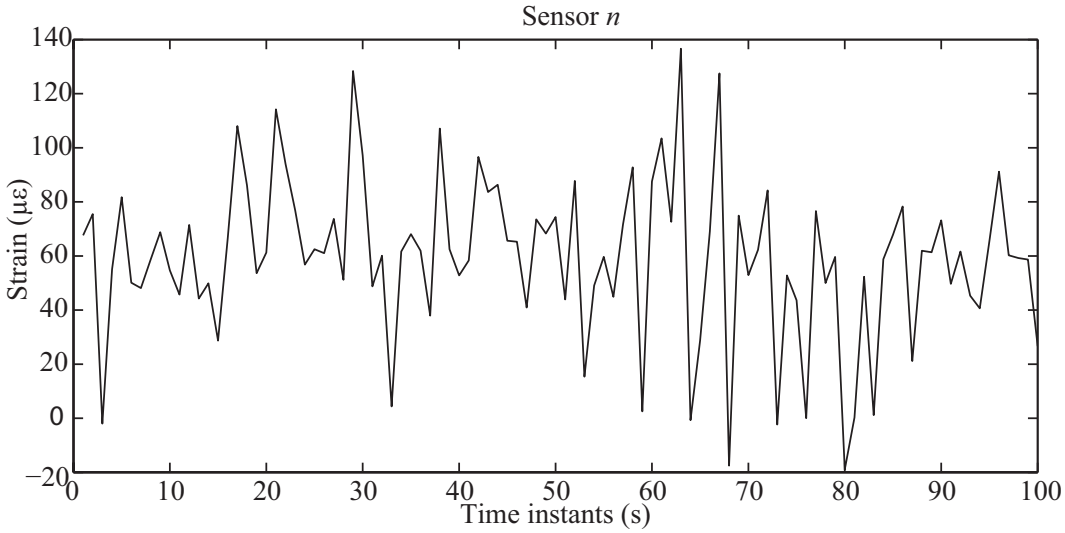


Figure 5. Time-domain strain signals sensing by an FBG sensor as an acquisition method.

3.3 SENSITIVE FEATURES

Every quantity measured using an acquisition method that may indicate the presence or not of damage can be named as a feature. Those obtained features are the elements needed to perform a further pattern recognition process for a subsequent discrimination between an operational condition and a damage manifestation, an example is shown in Figure 6. The machine learning algorithm may lead to a better and faster damage detection if there are selected appropriate features. *Feature extraction* and *feature selection* are really important elements here.

The first one refers to a transformation of the acquired data into a new representation which can clarify correlations that are originally hidden including damage; on the other hand, the selection of features relies in the fact that some features are better than others depending on the system nature. The ideal feature is the one that is really sensitive to the presence of damage in the structure but is less or non altered when it is affected by environmental or sub-operational conditions. A condensation of the data is necessary if there is a handling of large datasets to retain just the most significant changes in the system operation.

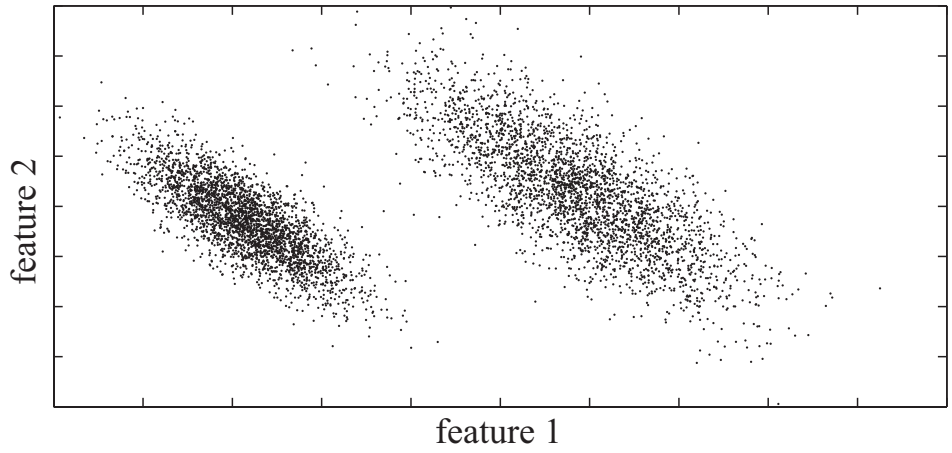


Figure 6. Sensitive features.

3.4 STATISTICAL MODEL

Here the algorithms capable to manage those discovered features are treated with the aim to identify damages in the system, as it is illustrated in Figure 7. This is the part that has had the least attention and is still in development, specially the field concerned on damage detection, when there is no damage information available *unsupervised learning*. However, developments related to detection of outliers have been constructed. In contrast, when damage information is available the information can be treated through an algorithm belonging to the *supervised learning* field. Finally the performance of the statistical model needs to be quantified in order to determine the performance and sensitivity of the Machine learning algorithm and the selected features.

3.5 STRAIN FIELD PATTERN RECOGNITION

The damage occurrence in a structure can change the global stiffness, hence, the strain field may vary. The strain field can be inferred by means of the measuring of strain signals using fiber optic sensors such as FBGs, those elements are described with more detail in Section 4.1. Depending on the system evaluation, specific physical principles may lead the phenomena presented in the system.

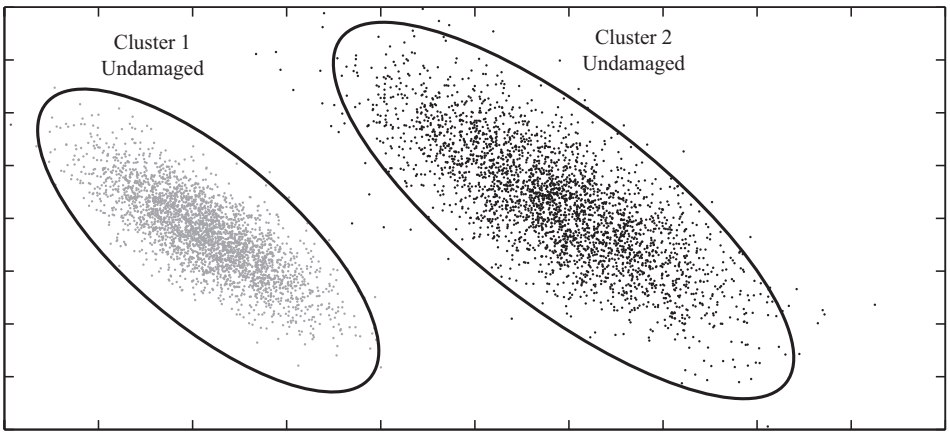


Figure 7. Statistical model.

For the validation process in this methodology, a beam in cantilever which varies in a set of specific angles, was submitted to a cyclic flexural load and, due to the moment of inertia's variation in accordance to the specific angle its stiffness will change. However, those physical variables simultaneously presented will preserve relationships and trends. This redundancy, permits the handling of data represented by means of new virtual variables.

In this specific case, as it was determined by Sierra et al. [77] the strain field had a direct correlation with the damage occurrence, besides, changes in load conditions where linear relationships are slightly conserved. In a plot of strain vs. strain between two different sensors, a nonlinear relationship was evident as it can be seen in Figure 8.

However, the information can be adjusted linearly reducing the computational cost of the overall process, nevertheless, loss of information may have been presented in some cases; those relationships become a hint in a further data processing, for example, data reduction in which the use of a linear method would lead to convenient results. However, when damage is presented in the structure, the strain field may lead to more marked nonlinear results, thus, more complex algorithms may be necessary to ensure reliable information about the structure's health.

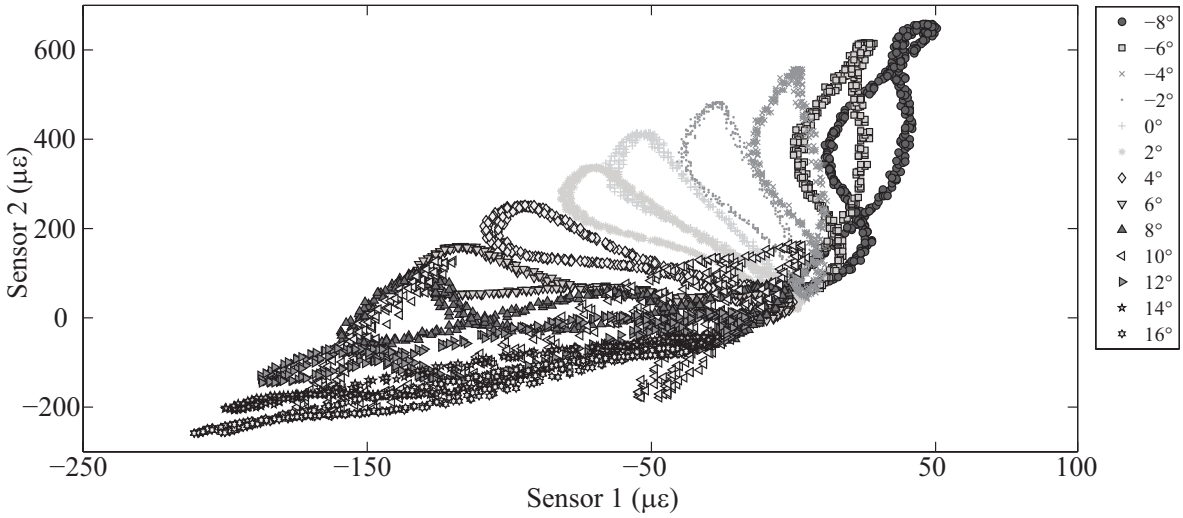


Figure 8. Strain field variation.

3.6 DAMAGE DETECTION

Following Rytter [78] a damage detection system has to achieve five fundamental and hierarchical levels, the more levels are reached the better the capability of the algorithm:

1. **Detection:** the algorithm indicates the presence of damage.
2. **Localization:** the method submits information about the location of the damage.
3. **Type:** the method is able to determine the kind of damage presented in the structure.
4. **Extent:** the method can determine the severity of the damage.
5. **Prediction:** the method is capable to establish the remaining lifetime of the system.

The further idea which is not covered in this methodology is the identification of damage using strain signals starting from a comparative study. This could be performed based on the benchmarks created by the algorithm from the pristine structure and new unknown information where damage may be present. An example of how it may work is presented in Figure 9. The challenge relies on how pristine conditions are identified when loads do not have a specific magnitude (as it may happen in an aircraft flying under regular conditions), where loads are unexpected and not clearly defined.

Here lies the importance of a structure's baseline under regular load conditions. With the purpose of designing the present methodology, information about strain signals from the experiment designed by Sierra [16] (Politécnico de Madrid, Spain, 2014) presented in Section 1.1 is taken into consideration.

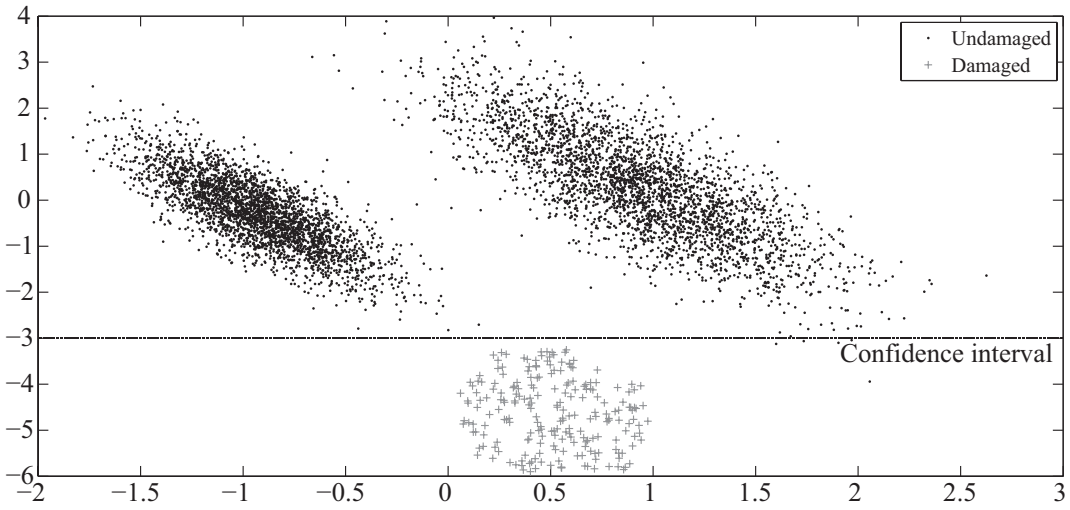


Figure 9. Damage detection.

4 DATA ACQUISITION AND PRELIMINARY PROCESSING

4.1 FUNDAMENTALS PRINCIPLES OF SENSORS

When a load is applied to a structure, perturbations are presented, sometimes manifested as damage, which can change the strain field, making the global stiffness vary. Consequently, the strain field will always vary in the time-line, changing the magnitude of the signals measured in a specific period of time.

A way to measure those strain changes in a structure is through the use of optical fiber sensors called Fiber Bragg Gratings FBGs strain gages. Basically, when the light passes through an FBG, a proportional part of the Bragg's wavelength is reflected, as it is illustrated in Figure 10.

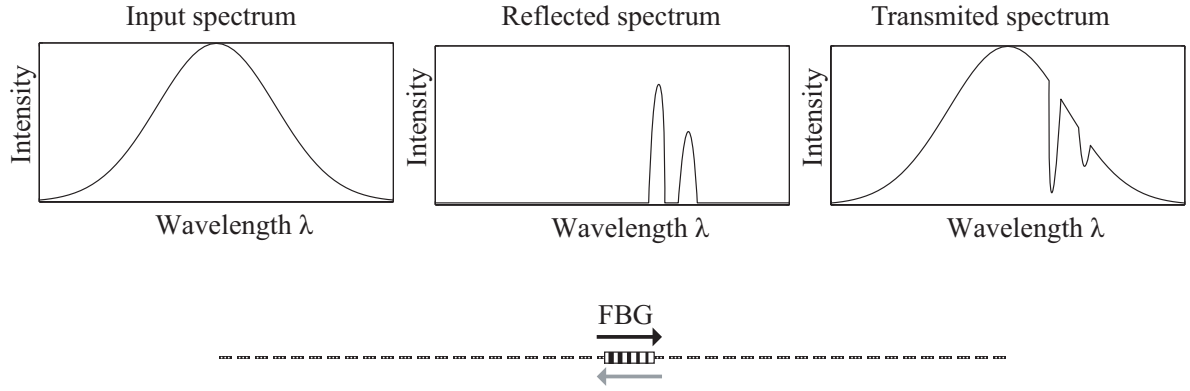


Figure 10. Bragg's wavelength reflection.

The reflected wavelength varies proportionally with the deformation of the fiber optic sensors which are bonded to or embedded into the structure. The Bragg's wavelength is given by the expression developed by Kersey et al. [79]:

$$\lambda_B = 2n_i\Lambda, \tag{1}$$

where n_i is the effective refractive index of the fiber's core and Λ is the grating pitch.

Furthermore, the relation between physical phenomena such as strain and/or temperature and the Bragg's wavelength is given by the following equation:

$$\frac{\Delta\lambda_B}{\lambda_B} = (1 - \rho_\alpha)\Delta\varepsilon + (1 + \xi)\Delta T, \tag{2}$$

where the fiber photo-elastic coefficient is represented by ρ_α and ξ represents the thermo-optic coefficient of the fiber. The values for $\Delta\varepsilon$ and ΔT were obtained experimentally using the following expressions [80]:

$$\Delta\varepsilon = (803.9 \pm 5.6)\frac{\mu\varepsilon}{\text{nm}}(\Delta\lambda) \rightarrow k_\varepsilon = (0.7991 \pm 0.0055)\mu\varepsilon^{-1}, \tag{3}$$

$$\Delta T = (101.9 \pm 1.2)\frac{\text{K}}{\text{nm}}(\Delta\lambda) \rightarrow k_T = (6.334 \pm 0.0074) \times 10^{-6}\text{K}^{-1}, \tag{4}$$

where $k_\varepsilon = 1 - \rho_\alpha$ and $k_T = 1 + \xi$.

The location and characteristics of the FBGs have to be pragmatically done, considering the zones where high levels of stain are expected. The location criteria depends widely on the level of understanding of the structure's mechanical and physical behaviors. FBGs are small and not invasive, they can be embedded using resins such as epoxy. FBGs can be placed on the surface of the structure during or after the fabrication of the piece with insignificant physical changes. Unlike other strain sensors such as strain gauges that may be more precise and cheap, Kreuzer [81] proposes considerable advantages of the FBGs:

- FBGs are lightweight and small sized.
- As it was mentioned above, FBGs and composite materials are quite easily "mergeable", therefore, FBGs can be an active part of a structure turning them into smart structures considering that modern structures like some aerospace structures are made of with composite materials.
- They have a high range of measurement, more than $10\,000\,\mu\varepsilon$ in some cases, thus this kind of sensors are appropriate for structures that exhibit high strains.
- They have extended lifespan, besides they do not have electromagnetic interference and corrosive vulnerability.

- FBGs are electrically passive, they can be placed in high voltage zones or flammable areas, consequently it is possible to attach FBGs in critical systems such as aircrafts, wind turbines, nuclear reactors, etc. Besides, there are specific FBGs that have been developed to operate under high temperature environments. On the other hand, they can also operate under low temperature environments due to their low thermal conductivity.

Moreover, FBGs sensors present questionable disadvantages such as temperature dependence, thus, a temperature sensor is always required in order to perform a thermal compensation to estimate strain measurements. FBGs exhibit high sensitivity to lateral forces and in some cases the high stiffness of the FBGs may cause parallel forces, and finally, nowadays the interrogators are yet considered expensive.

4.2 DATA ACQUISITION

Raw datasets of strain signals can be stored using an optical sensing interrogator connected to the FBGs [81]. Commonly strain signals are measured in micro strain $\mu\varepsilon$ units. If the strain acquisition is made in a structure without any manufacturing defect and under regular load conditions, those signals belonging to the structure will represent the pristine condition of the system. Otherwise, The scan frequency is limited by the acquisition module capability. A Basic FBGs reflective signal acquisition scheme is presented in Figure 11.

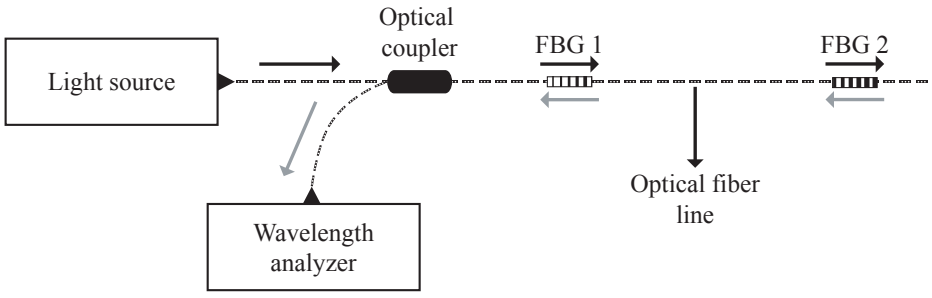


Figure 11. Basic FBGs reflective signal acquisition scheme.

In the experiment taken into consideration for the methodology, four optical lines were placed on the surface of the aluminum beam, each one having eight FBGs placed in parallel with the beam's longitudinal axis (x axis), the location of the FBGs are presented in Figure 12. Each FBG had a periodic modulation length of 2 mm and wavelengths between 1510 nm and 1590 nm.

A total of 32 FBGs were used in the experiment, one of them was mechanically isolated in order to measure temperature with the aim to perform a thermal compensation. The first group of FBGs were placed 50 mm from the clamping ending, the rest of the groups were placed with a spacing of 150 mm starting from the first group. A detailed view of the FBGs distribution is presented in Figure 12. Each experiment was placed under dynamic loads for a period of time of 410 s.

The acquisition system consisted in a four-channel Micron Optics SM130 optical fiber interrogator and strains were measured at a sampling rate of 100 Hz. Each position was tested twice, in order to generate validation data. The size of raw data acquired in the experiments is presented in Table 1.

Table 1. Aluminum beam's acquired data.

Name	**Pitch angle °**	**Experiment trials**	**Number of sensors**
BL_0	-8	41000	32
BL_2	-6	41002	32
BL_4	-4	41000	32
BL_6	-2	41000	32
BL_8	0	41000	32
BL_10	2	41000	32
BL_12	4	40998	32
BL_14	6	41002	32
BL_16	8	41001	32
BL_18	10	41001	32
BL_20	12	41000	32
BL_22	14	41001	32
BL_24	16	41001	32

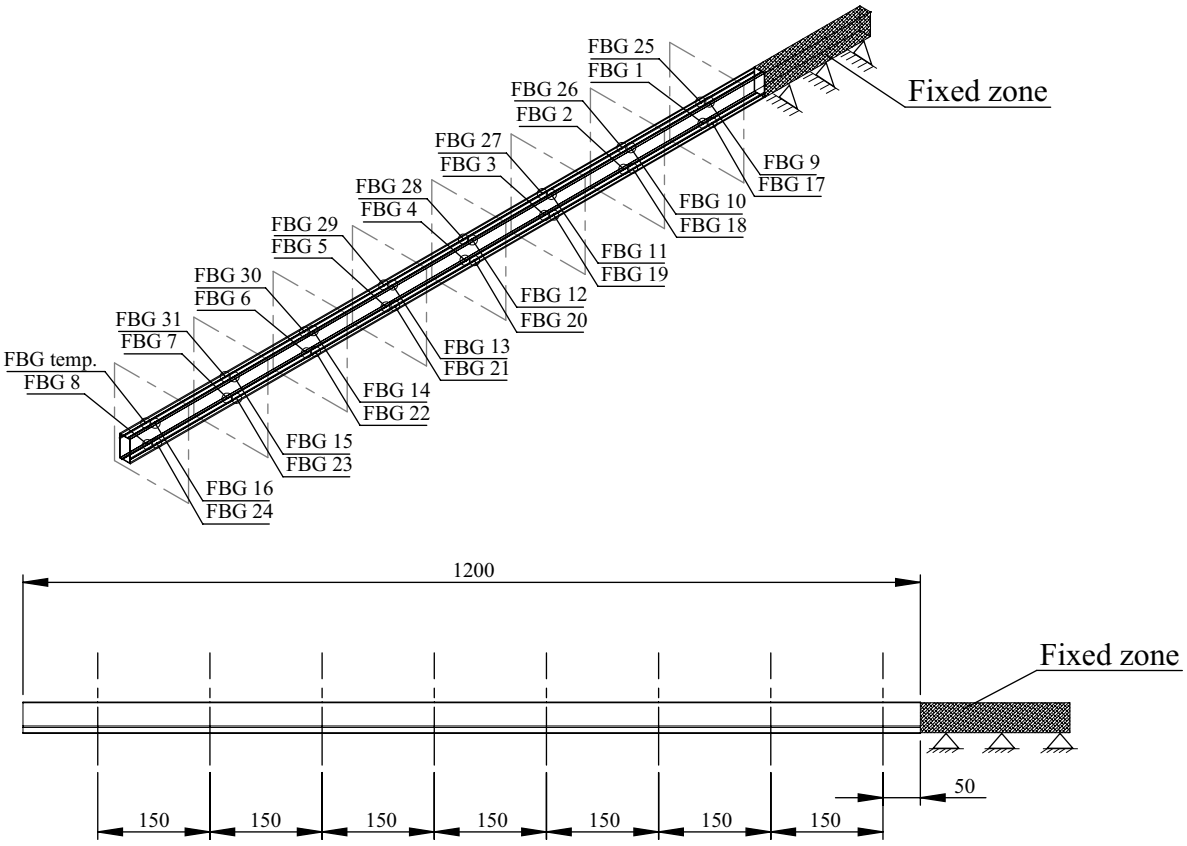


Figure 12. Sensors distribution (all measures are given in mm).

4.3 DATA PRELIMINARY PROCESSING

It is convenient to perform a preliminary processing of the strain signals with the aim of enhancing the clustering performance, removing the less significant or redundant information. The selection of representative signal characteristics has to be made in accordance with the mechanical behavior of the structure. Thus, not every signal recorded is meaningful for clustering.

In experiments during long periods of time under cyclic loads, large quantity of information is stored and as a result, undesired information is usually recorded. This undesired information is commonly known as noise and because its high frequency nature it will occupy a large amount of space which may decrease the classification performance. The quantity of noise is subjective and it depends on the instrument accuracy, capability, limitations and especially on the nature of the experiment. Failure in the experimental

model as a debonding of a sensor line or wrong manipulation may have repercussions in the signal acquisition; those flaws unrelated with the structure's behavior are also known as outliers. With the intention of refining the model by detrending, averaging, smoothing, denoising or filtering the signals, outliers can be removed from the data set before the processing or even after it [82].

The focus of the preliminary processing may vary in accordance with the nature of the experiment. As it was found experimentally by Sierra [16], for a suitable selection of significant strain signals, the Signal to Noise Ratio SNR over 50 may improve the overall classification process, therefore, for the used Micron Optics interrogator, which has a white noise range around 1 $\mu\varepsilon$, magnitudes over 50 $\mu\varepsilon$ are desired.

Moreover, the physical phenomena that governs the experiment structural, beam deflection, can be simplified to a state of tension/compression in the opposite faces parallel to the principal axis, as it is presented in Figure 13. The selection of discretized maximum and minimum signal peaks and peaks vicinity points that represent the state of maximum tension/compression loads avoid outliers and noise.

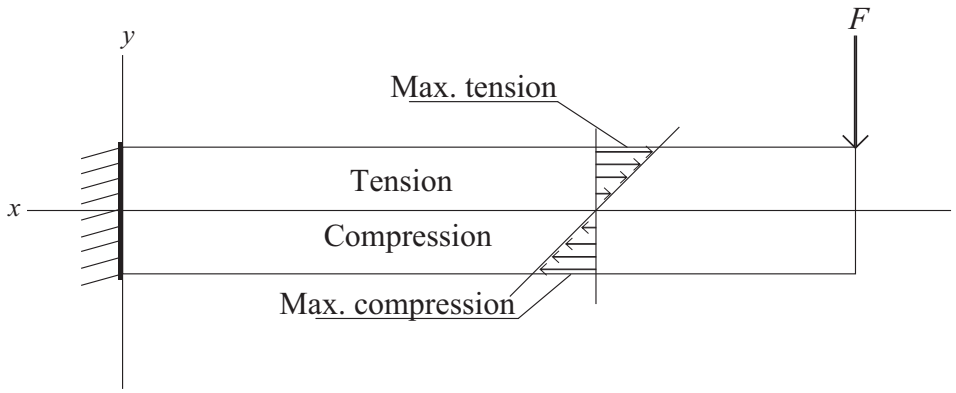


Figure 13. Structural beam deflection.

Therefore, the number of time trials required was reduced applying this process commonly considered as cleansing. As it is graphically exemplified in Figure 14, specific strain information from the sensor 1 was selected for a subsequent signal handling as a part of the outliers removing.

Furthermore, different magnitudes and scales may be determined by each sensor, since different circumstances e.g. temperature or humidity can be presented and vary during

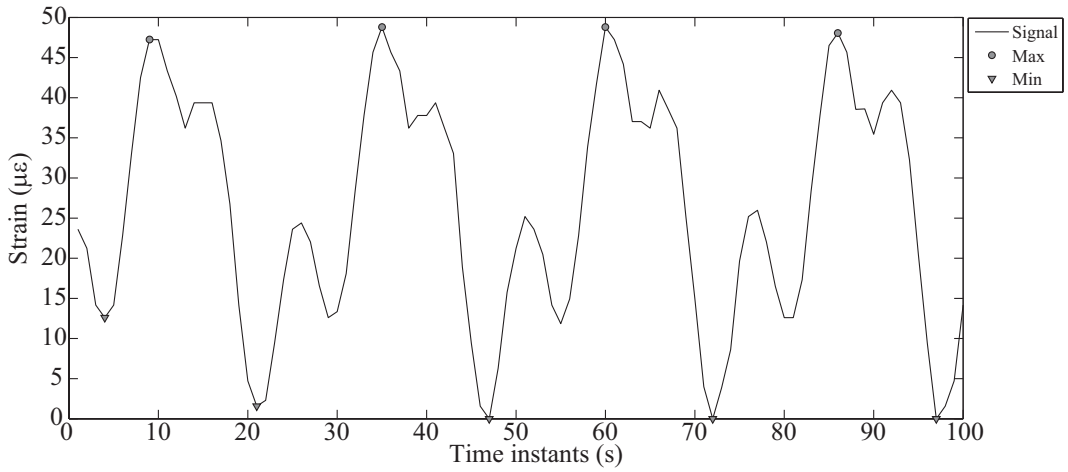


Figure 14. Signal cleansing.

the recording time. The action of scaling the data allows to avoid differences that are not relevant in the classification process; using the auto-scaling process each variable is re-scaled to have zero mean and unity variance, this procedure is carried out in the dimensionality reduction process, Section 5, however, technically it is part of the preliminary processing. Following the notation of Mujica et al. [45], each sensor vector v_{ij} is redefined as:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \tag{5}$$

$$\sigma_{vj}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \mu_{vj})^2, \tag{6}$$

$$\overline{x}_{ij} = \frac{x_{ij} - \mu_{vj}}{\sigma_{vj}}, \tag{7}$$

where μ_{vj} represents the mean and σ_{vj}^2 the variance of sensor j measurements v_j, thus, $\overline{x}_{ij}$ is the rescaled sample. Although the rescaled sample symbol is considered just like x_{ij} for simplicity. The resultant matrices' magnitudes are presented in Section 5.2, Table 3, before the dimensionality reduction process of the experiment. Following the previous action, a normality test is presented before and after the auto scaling was performed and, as it was expected, it has variance one (see Figure 15) and mean zero

(see Figure 16). Thus, the variance error is constant (0.0069), therefore, it can be considered that the dataset presents homogeneity of variance.

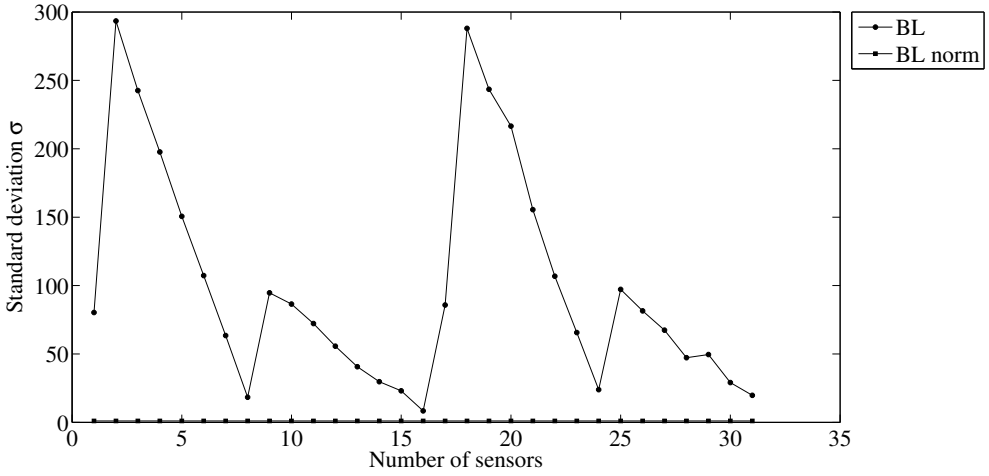


Figure 15. Baseline and Baseline normalized standard deviation.

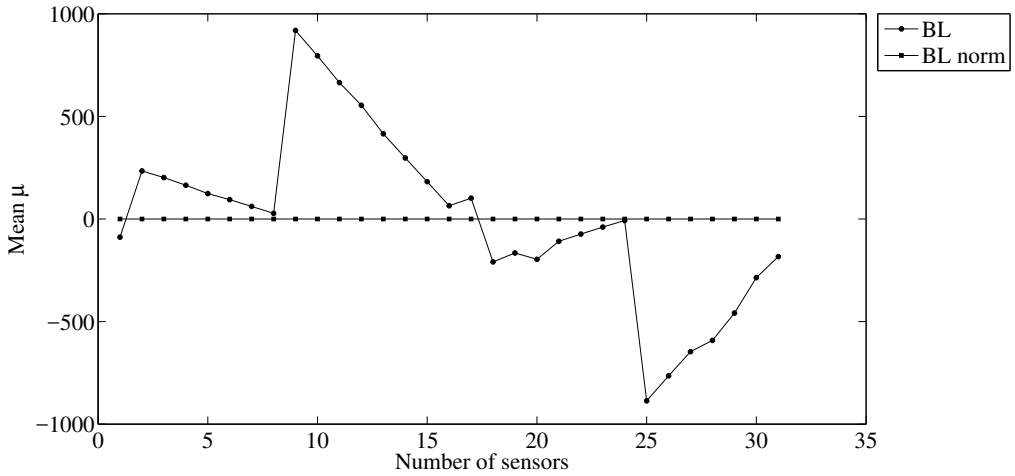


Figure 16. Baseline and Baseline normalized mean.

Further, after the auto-scaling process, a correlation between each pair of sensors, is presented using the covariance matrix of the dataset in Table 2. Some sensors must have to be intercorrelated to determine a unique contribution to a factor in the factor analysis process.

Table 2. Covariance matrix of the Baseline matrix after the auto-scaling process.

	BL 1	BL 2	BL 3	BL 4	BL 5	BL 6	BL 7	BL 8	BL 9	BL 10	BL 11	BL 12	BL 13	BL 14	BL 15	BL 16	BL 17	BL 18	BL 19	BL 20	BL 21	BL 22	BL 23	BL 24	BL 25	BL 26	BL 27	BL 28	BL 29	BL 30	BL 31
BL 1	1	0,98	0,98	0,98	0,99	0,98	0,98	0,96	-0,75	-0,77	-0,76	-0,7	-0,67	-0,69	-0,82	-0,76	-0,99	-0,99	-0,99	-0,98	-0,99	-0,99	-0,98	-0,97	0,76	0,74	0,73	0,38	0,79	0,71	0,74
BL 2	0,98	1	0,99	0,99	0,99	0,99	0,99	0,98	-0,65	-0,68	-0,66	-0,6	-0,56	-0,58	-0,74	-0,69	-0,99	-0,99	-0,99	-0,99	-0,99	-0,99	-0,99	-0,98	0,67	0,64	0,63	0,25	0,7	0,6	0,64
BL 3	0,98	0,99	1	0,99	0,99	0,99	0,99	0,98	-0,65	-0,68	-0,67	-0,6	-0,56	-0,59	-0,74	-0,69	-0,99	-0,99	-0,99	-0,99	-0,99	-0,99	-0,99	-0,98	0,67	0,65	0,64	0,25	0,7	0,6	0,64
BL 4	0,98	0,99	0,99	1	0,99	0,99	0,99	0,98	-0,65	-0,68	-0,67	-0,6	-0,56	-0,59	-0,74	-0,69	-0,99	-0,99	-0,99	-0,99	-0,99	-0,99	-0,99	-0,98	0,67	0,65	0,64	0,25	0,7	0,61	0,64
BL 5	0,99	0,99	0,99	0,99	1	0,99	0,99	0,98	-0,65	-0,68	-0,67	-0,6	-0,57	-0,59	-0,74	-0,7	-0,99	-0,99	-0,99	-0,99	-0,99	-0,99	-0,99	-0,98	0,67	0,65	0,64	0,25	0,71	0,61	0,64
BL 6	0,98	0,99	0,99	0,99	0,99	1	0,99	0,98	-0,66	-0,68	-0,67	-0,6	-0,57	-0,59	-0,74	-0,69	-0,99	-0,99	-0,99	-0,99	-0,99	-0,99	-0,99	-0,98	0,67	0,65	0,64	0,26	0,71	0,61	0,65
BL 7	0,98	0,99	0,99	0,99	0,99	0,99	1	0,98	-0,65	-0,68	-0,67	-0,6	-0,57	-0,59	-0,74	-0,68	-0,99	-0,99	-0,99	-0,99	-0,99	-0,99	-0,99	-0,98	0,67	0,65	0,64	0,25	0,7	0,61	0,65
BL 8	0,96	0,98	0,98	0,98	0,98	0,98	0,98	1	-0,61	-0,64	-0,63	-0,56	-0,52	-0,54	-0,69	-0,6	-0,96	-0,98	-0,98	-0,98	-0,97	-0,97	-0,96	-0,94	0,64	0,61	0,6	0,22	0,67	0,58	0,62
BL 9	-0,75	-0,65	-0,65	-0,65	-0,65	-0,66	-0,65	-0,61	1	0,99	0,99	0,99	0,99	0,99	0,98	0,92	0,74	0,65	0,66	0,65	0,66	0,67	0,67	0,67	-0,99	-0,99	-0,99	-0,89	-0,99	-0,99	-0,99
BL 10	-0,77	-0,68	-0,68	-0,68	-0,68	-0,68	-0,68	-0,64	0,99	1	0,99	0,99	0,98	0,99	0,98	0,93	0,77	0,68	0,69	0,67	0,69	0,69	0,7	0,7	-0,99	-0,99	-0,99	-0,87	-0,99	-0,99	-0,99
BL 11	-0,76	-0,66	-0,67	-0,67	-0,67	-0,67	-0,67	-0,63	0,99	0,99	1	0,99	0,99	0,99	0,98	0,92	0,76	0,67	0,67	0,66	0,68	0,68	0,69	0,69	-0,99	-0,99	-0,99	-0,88	-0,99	-0,99	-0,99
BL 12	-0,7	-0,6	-0,6	-0,6	-0,6	-0,6	-0,6	-0,56	0,99	0,99	0,99	1	0,99	0,99	0,97	0,91	0,7	0,6	0,61	0,6	0,61	0,62	0,62	0,62	-0,99	-0,99	-0,99	-0,92	-0,98	-0,99	-0,98
BL 13	-0,67	-0,56	-0,56	-0,56	-0,57	-0,57	-0,57	-0,52	0,99	0,98	0,99	0,99	1	0,99	0,96	0,91	0,67	0,56	0,57	0,56	0,57	0,58	0,59	0,59	-0,98	-0,99	-0,99	-0,93	-0,98	-0,99	-0,98
BL 14	-0,69	-0,58	-0,59	-0,59	-0,59	-0,59	-0,59	-0,54	0,99	0,99	0,99	0,99	0,99	1	0,97	0,92	0,69	0,59	0,59	0,58	0,6	0,61	0,61	0,62	-0,99	-0,99	-0,99	-0,92	-0,98	-0,99	-0,98
BL 15	-0,82	-0,74	-0,74	-0,74	-0,74	-0,74	-0,74	-0,69	0,98	0,98	0,98	0,97	0,96	0,97	1	0,96	0,82	0,74	0,75	0,74	0,75	0,76	0,76	0,77	-0,98	-0,98	-0,97	-0,82	-0,99	-0,96	-0,96
BL 16	-0,76	-0,69	-0,69	-0,69	-0,7	-0,69	-0,68	-0,6	0,92	0,93	0,92	0,91	0,91	0,92	0,96	1	0,77	0,7	0,7	0,69	0,71	0,71	0,72	0,76	-0,92	-0,91	-0,91	-0,77	-0,93	-0,89	-0,88
BL 17	-0,99	-0,99	-0,99	-0,99	-0,99	-0,99	-0,99	-0,96	0,74	0,77	0,76	0,7	0,67	0,69	0,82	0,77	1	0,99	0,99	0,99	0,99	0,99	0,99	0,98	-0,76	-0,74	-0,73	-0,37	-0,79	-0,7	-0,73
BL 18	-0,99	-0,99	-0,99	-0,99	-0,99	-0,99	-0,99	-0,98	0,65	0,68	0,67	0,6	0,56	0,59	0,74	0,7	0,99	1	0,99	0,99	0,99	0,99	0,99	0,98	-0,67	-0,65	-0,64	-0,25	-0,7	-0,61	-0,64
BL 19	-0,99	-0,99	-0,99	-0,99	-0,99	-0,99	-0,99	-0,98	0,66	0,69	0,67	0,61	0,57	0,59	0,75	0,7	0,99	0,99	1	0,99	0,99	0,99	0,99	0,98	-0,68	-0,65	-0,64	-0,26	-0,71	-0,61	-0,65
BL 20	-0,98	-0,99	-0,99	-0,99	-0,99	-0,99	-0,99	-0,98	0,65	0,67	0,66	0,6	0,56	0,58	0,74	0,69	0,99	0,99	0,99	1	0,99	0,99	0,99	0,98	-0,67	-0,64	-0,63	-0,24	-0,7	-0,6	-0,63
BL 21	-0,99	-0,99	-0,99	-0,99	-0,99	-0,99	-0,99	-0,97	0,66	0,69	0,68	0,61	0,57	0,6	0,75	0,71	0,99	0,99	0,99	0,99	1	0,99	0,99	0,99	-0,68	-0,66	-0,65	-0,26	-0,71	-0,61	-0,65
BL 22	-0,99	-0,99	-0,99	-0,99	-0,99	-0,99	-0,99	-0,97	0,67	0,69	0,68	0,62	0,58	0,61	0,76	0,71	0,99	0,99	0,99	0,99	0,99	1	0,99	0,99	-0,68	-0,66	-0,65	-0,27	-0,72	-0,62	-0,65
BL 23	-0,98	-0,99	-0,99	-0,99	-0,99	-0,99	-0,99	-0,96	0,67	0,7	0,69	0,62	0,59	0,61	0,76	0,72	0,99	0,99	0,99	0,99	0,99	0,99	1	0,99	-0,69	-0,66	-0,65	-0,27	-0,72	-0,62	-0,65
BL 24	-0,97	-0,98	-0,98	-0,98	-0,98	-0,98	-0,98	-0,94	0,67	0,7	0,69	0,62	0,59	0,62	0,77	0,76	0,98	0,98	0,98	0,98	0,99	0,99	0,99	1	-0,69	-0,67	-0,66	-0,28	-0,72	-0,62	-0,65
BL 25	0,76	0,67	0,67	0,67	0,67	0,67	0,67	0,64	-0,99	-0,99	-0,99	-0,99	-0,98	-0,99	-0,98	-0,92	-0,76	-0,67	-0,68	-0,67	-0,68	-0,68	-0,69	-0,69	1	0,99	0,99	0,88	0,99	0,99	0,99
BL 26	0,74	0,64	0,65	0,65	0,65	0,65	0,65	0,61	-0,99	-0,99	-0,99	-0,99	-0,99	-0,99	-0,98	-0,91	-0,74	-0,65	-0,65	-0,64	-0,66	-0,66	-0,66	-0,67	0,99	1	0,99	0,89	0,99	0,99	0,99
BL 27	0,73	0,63	0,64	0,64	0,64	0,64	0,64	0,6	-0,99	-0,99	-0,99	-0,99	-0,99	-0,99	-0,97	-0,91	-0,73	-0,64	-0,64	-0,63	-0,65	-0,65	-0,65	-0,66	0,99	0,99	1	0,9	0,99	0,99	0,99
BL 28	0,38	0,25	0,25	0,25	0,25	0,26	0,25	0,22	-0,89	-0,87	-0,88	-0,92	-0,93	-0,92	-0,82	-0,77	-0,37	-0,25	-0,26	-0,24	-0,26	-0,27	-0,27	-0,28	0,88	0,89	0,9	1	0,86	0,92	0,89
BL 29	0,79	0,7	0,7	0,7	0,71	0,71	0,7	0,67	-0,99	-0,99	-0,99	-0,98	-0,98	-0,98	-0,99	-0,93	-0,79	-0,7	-0,71	-0,7	-0,71	-0,72	-0,72	-0,72	0,99	0,99	0,99	0,86	1	0,98	0,98
BL 30	0,71	0,6	0,6	0,61	0,61	0,61	0,61	0,58	-0,99	-0,99	-0,99	-0,99	-0,99	-0,99	-0,96	-0,89	-0,7	-0,61	-0,61	-0,6	-0,61	-0,62	-0,62	-0,62	0,99	0,99	0,99	0,92	0,98	1	0,99
BL 31	0,74	0,64	0,64	0,64	0,64	0,65	0,65	0,62	-0,99	-0,99	-0,99	-0,98	-0,98	-0,98	-0,96	-0,88	-0,73	-0,64	-0,65	-0,63	-0,65	-0,65	-0,65	-0,65	0,99	0,99	0,99	0,89	0,98	0,99	1

5 DIMENSIONALITY REDUCTION TECHNIQUE

5.1 THEORETICAL BACKGROUND

As it was mentioned before, through a dimensionality reduction method, it is possible to describe a large quantity of the original dataset and project the obtained information in a new dimensional space. The technique called factor analysis similar to PCA was selected to perform the experiment's dimensionality reduction. Like PCA, FA is also a linear method, and was applied initially in the field of psychology [28]. The main goal of FA is to describe how the original x variables in a dataset depend on a small number of variables k, with $k < x$, capable to describe a large part of the observed model [83], except for an error term due to the linear adjust.

Following the notation of Jolliffe [84], FA can represent the original variables $x_1, x_2, \ldots, x_p$ as a linear combination of hypothetical variables called common factors $f_1, f_2, \ldots f_m$, factor loadings Λ_{jk}, where $j = 1, 2, \ldots, p$; $k = 1, 2, \ldots, m$ which represent the correlation of each original variable with a common factor and specific factors or errors e_j Therefore, FA model can be illustrated as:

$$\begin{aligned} x_1 &= \lambda_{11} f_1 + \lambda_{12} f_2 + \ldots + \Lambda_{1m} f_m + e_1 \\ x_2 &= \lambda_{21} f_1 + \lambda_{22} f_2 + \ldots + \Lambda_{2m} f_m + e_2 \\ &\vdots \\ x_p &= \lambda_{p1} f_1 + \lambda_{p2} f_2 + \ldots + \Lambda_{pm} f_m + e_p, \end{aligned} \tag{8}$$

then, the general equation form of FA is represented as follows,

$$X = \Lambda f + e. \tag{9}$$

Unlike the standard regression model, there will be different estimation techniques in FA taking into account that at the beginning Λ and f are unknown. Therefore, there

would not exist one best or a unique solution. After mathematical handling shown widely in the literature as in [28, 83, 84, 32, 85], FA model can be represented in terms of covariances following the next equation:

$$\Sigma = \Lambda\Lambda' + \Psi, \tag{10}$$

where Σ is the data covariance matrix, and Ψ the covariance of the specific factors. Assuming T as an orthogonal matrix, Λ and Ψ are initially calculated using the equation $\Lambda^* = \Lambda T$ which after an algebraic handling result in $\Lambda\Lambda'$. A unique initial solution can be found, placing some restrictions over Λ. Therefore multiplying the orthogonal matrix T to the initial solution, other solutions can be determined when Λ is rotated until a particular "best" solution is found.

Varimax [86], Quartimax [87] and Promax [88] are common rotation matrices used by mostly all popular computer packages [84]. Using multivariate normality of f and e, more precise values of Ψ and Λ can be estimated using the maximum likelihood by an iterative process as it was explained by Lawley [32].

Selecting a number of factors desired to rotate will reduce the dimension of the original dataset. In consequence, there is a number of eigenvalues which describe the information retained, related to the selected number of factors. There are several rules to determining the amount of factors necessary to describe a reliable quantity of the original information. The most common method is called "eigenvalues greater than one" rule [89], the main goal is based on retaining the factors with eigenvalue greater than one, those factors may perform the best description of the original samples.

Other rules also included in PCA, are focused on keeping the factors in which the cumulative variance describes more than the 80% of the original variance [89]. In addition, there is also a graphic method called the "scree plot" in which the factors before the breaking point have to be retained, as it can be seen in Figure 17. Those rules are going to be verified using a statistical software program.

The aim with the present methodology is to group the common factors that keep a relation between variables, since further on it was necessary to cluster the common factors using the density based algorithm which works in a two-dimensional space, it was necessary to retain two common factors. Moreover, it was possible to graphically represent the formed clusters.

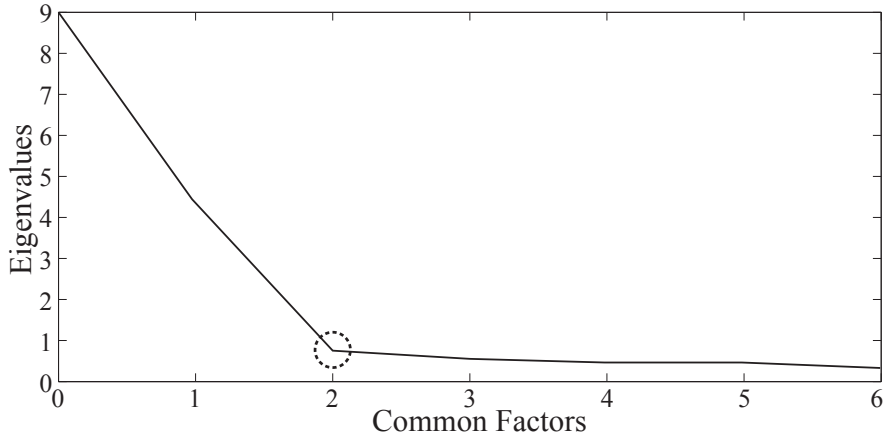


Figure 17. Scree plot example.

It was expected that variables related in between due to the preservation of the difference in magnitude among signals in a lower space, and in consequence, a large quantity of the original information could be described. Therefore, those signals represented in a new dimensional space by the common factors should fall in a same cluster.

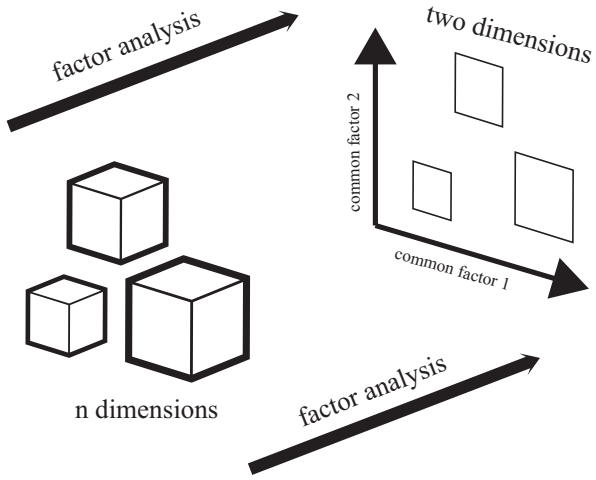


Figure 18. Dimensionality reduction.

Following the Mujica et al. [45] notation, the original data set D is a $D_{m\times n}$ matrix, with m time instants or experimental trials and n variables (sensors). Before the dimensionality reduction, the rows (experimental trials) must have been randomly combined. When FA is applied to the experimental matrix, the dataset can be represented in a lower dimension of $D_{m\times 2}$ elements, selecting the first two common factors. Each row

of the new matrix will result in a point of (x, y) coordinates that can be clustered and represented graphically, see Figure 18.

5.2 CASE OF STUDY

Each pitch angle was represented by the most significant data after the preliminary processing described in Section 4. Every pitch angle matrix was concatenated in a biggest matrix and was considered as the baseline matrix $D_{m\times n}$. Each row in the baseline matrix (aluminum beam's dataset) $D_{m\times n}$ was randomly arranged with the aim to recreate a real case scenario in which unknown load magnitudes appear arbitrarily. In Table 3 the pertinent size of each pitch angle's matrix after the preliminary processing is shown.

Table 3. Aluminum beam dataset.

Name	**Pitch angle °**	**Relevant trials**	**Number of sensors**
BL_0	-8	1609	31
BL_2	-6	1609	31
BL_4	-4	1609	31
BL_6	-2	1609	31
BL_8	0	1609	31
BL_10	2	1610	31
BL_12	4	1609	31
BL_14	6	1611	31
BL_16	8	1611	31
BL_18	10	1612	31
BL_20	12	1612	31
BL_22	14	1612	31
BL_24	16	1613	31
TOTAL		**20935**	**31**

As a result, a matrix with 20935 of the most representative experimental trials belonging to each pitch angle and 31 sensors was created; the size of the matrix was determined

as $D_{20935\times31}$. The 31 sensors were assumed as 31 dimensions as it was explained in Section 5.1. Hence, the beam's dataset was treated as a high dimensionality problem, thus, the importance of using a precise dimensionality reduction technique. As it was mentioned above, the dimensionality reduction technique was performed to project those 31 dimensions in a lower space.

The first two rotation matrices Varimax, (see Figure 19) and Quartimax (see Figure 20) are orthogonal and they seemed to be similar in terms of point group's spatial location. It was evident that in the two-dimensional space the dimensionality reduction algorithm projects well-separated specific group of points. Most of those groups appeared to have a tendency to the second common factor; the plot of the factor loadings Λ_{jk} presented in Subsection 5.1 represented with their specific number would indicate a hidden relationship of each sensor to a common factor.

Although the rotation matrices results are similar, the Promax projection which is an oblique method, was selected for the dimensionality reduction given that it performed a most balanced projection of the original information with two common factors retained. In consequence, it could be easier to interpret. Besides, it was evident that performing this projection resulted in well-separated and condensed groups that could contain the information of a particular pitch angle, as it is represented in the Promax rotation matrix Figure 21.

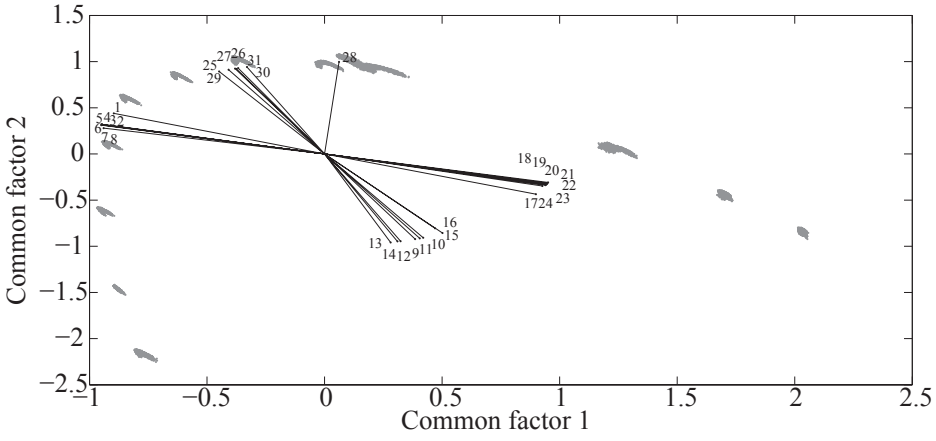


Figure 19. Varimax rotation matrices.

12 well-discretized groups emerged when the common factors were scattered in all of the rotation matrices, as it can be seen in Figures 19, 20 and 21. However, the shape of

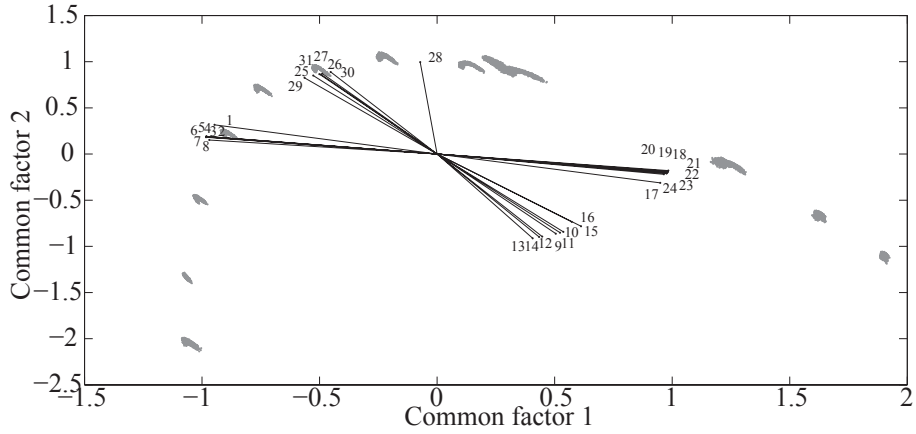


Figure 20. Quartimax rotation matrices.

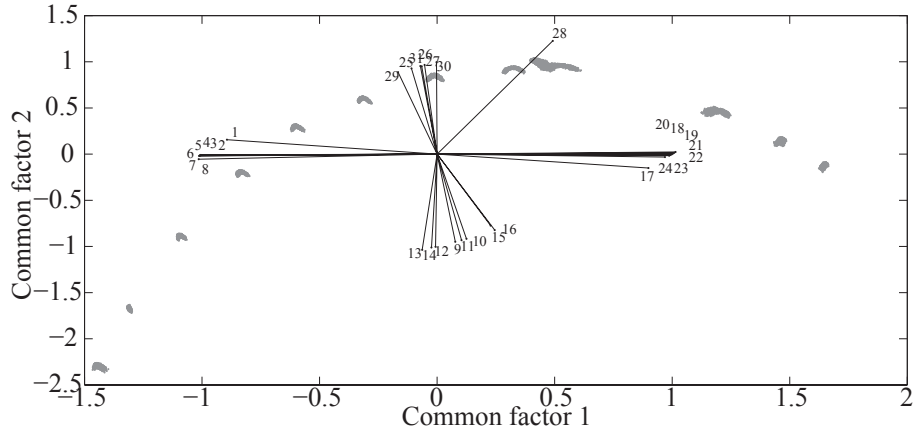


Figure 21. Promax rotation matrices.

the group of points established by the Varimax rotation, Figure 22 and the Quartimax rotation, Figure 23, seemed to be elongated and not as condensed as the one shown by the Promax rotation.

In contrast to the first two rotation matrices, the Promax rotation seemed to carry out a most condensed projection of the beam's strain signals in a lower dimension, as it can be confirmed in Figure 24. The groups generated by this rotation were well-distributed in the spatial location and also had a more likely spherical shaped form. This could result in an easier pattern recognition clustering, having in mind that, the clustering algorithm used in this methodology works in a two-dimensional space under special

conditions discussed in detail in Section 6.

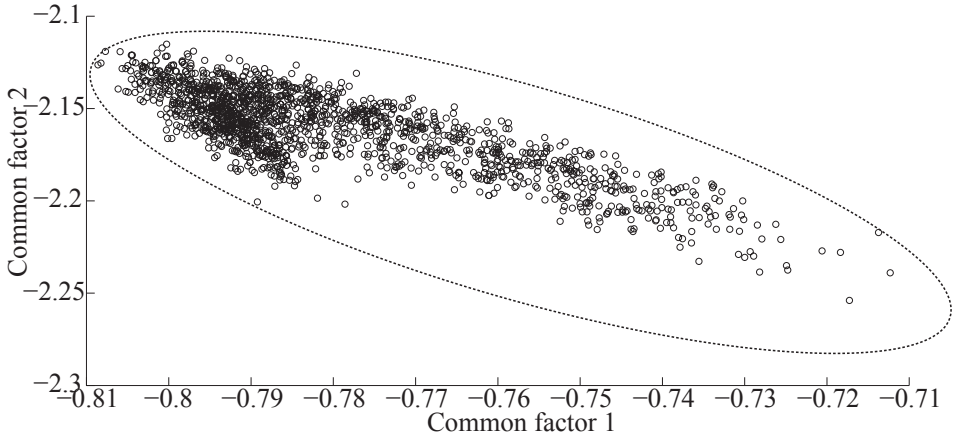


Figure 22. Varimax group shape.

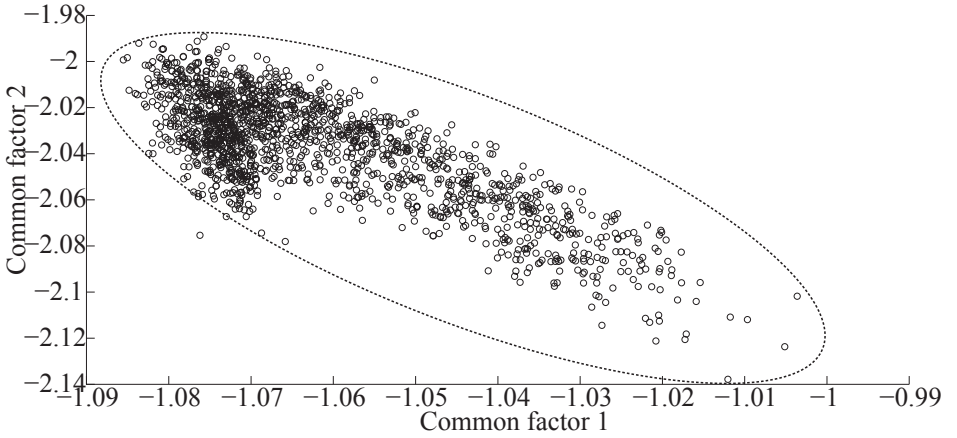


Figure 23. Quartimax group shape.

To guarantee that the reduced matrix $D_{20935\times 2}$ adequately represents the original information some rules had to be achieved. Those rules were mentioned before in Subsection 5.1 and included an analysis of the variance and eigenvalues of each common factor of the dimensionality reduced information $D_{20935\times 2}$. The original information may be graphically represented with just two common factors in a two dimensional space depending on the percentage of retained available information.

As it is shown in Table 4 it was just necessary to retain two common factors to describe the 99.234 % of the original information. Moreover, the first two common factors had

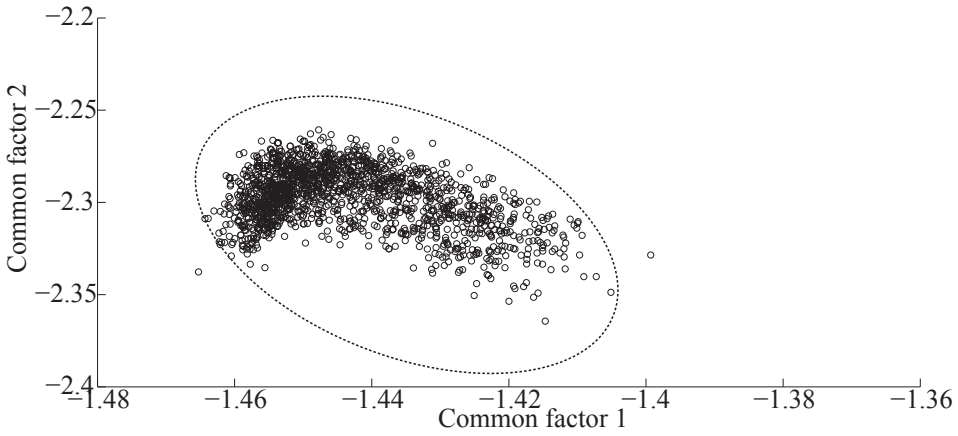


Figure 24. Promax group shape.

a eigenvalue greater than 1, the common factor 1 had an eigenvalue of 25.277 and the common factor 2 an eigenvalue of 5.485. Therefore, the rules established to determine the performance of the dimensionality reduction technique were notably reached, and a large quantity of the original information was able to be handled with high reliability in a two-dimensional space.

Furthermore, the scree plot, (see Figure 25), clarified the FA performance detailed in Table 4, illustrating the capacity of FA to retain large amounts of information. After the second factor a breaking point was easily recognizable in the scree plot and then the eigenvalues had an asymptotic behavior close to zero.

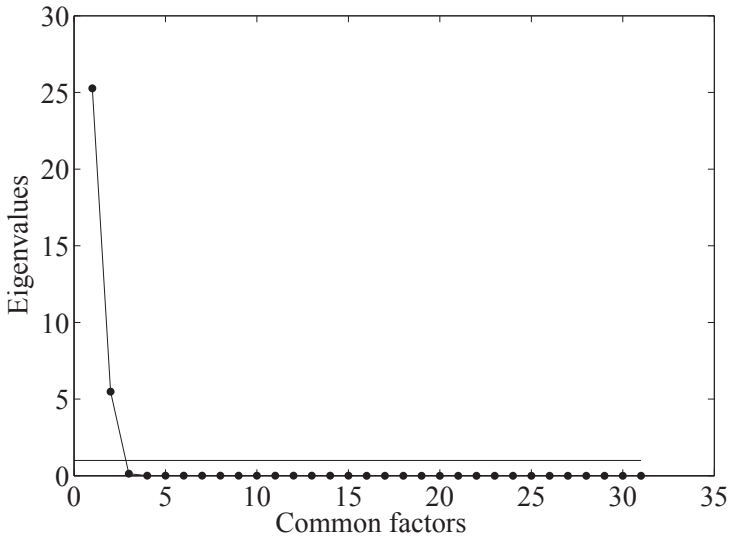


Figure 25. Beam dataset's scree plot.

Table 4. Eigenvalues greater than one.

Common factor	**Eigenvalue**	**Percent of variance**	**Cumulative percentage**
1	25.2775	81.54	81.540
2	**5.4852**	**17.694**	**99.234**
3	0.2065	0.666	99.900
4	0.0126	0.041	99.941
5	0.0060	0.019	99.960
6	0.0054	0.017	99.978
7	0.0017	0.006	99.983
8	0.0008	0.003	99.986
9	0.0008	0.003	99.989
10	0.0007	0.002	99.991
11	0.0006	0.002	99.993
12	0.0005	0.001	99.994
13	0.0003	0.001	99.995
14	0.0003	0.001	99.996
15	0.0002	0.001	99.997
16	0.0002	0.001	99.998
17	0.0002	0.001	99.998
18	0.0001	0	99.998
19	0.0001	0	99.999
20	0.0001	0	99.999
21	0.0001	0	99.999
22	0.0001	0	99.999
23	0	0	100
24	0	0	100
25	0	0	100
26	0	0	100
27	0	0	100
28	0	0	100
29	0	0	100
30	0	0	100
31	0	0	100

6 AUTOMATIC CLUSTERING EMPLOYING THE DBSCAN ALGORITHM

6.1 DBSCAN ALGORITHM

The DBSCAN clustering algorithm, is an unsupervised algorithm created by Ester et al. [90]. The selection of the DBSCAN algorithm was based on its ability to determine clusters without predefined class labels. DBSCAN often offers a superior performance than known techniques such as *k-means* or hierarchical algorithms given that it is not necessary to determine the number of desired groups previously. DBSCAN configuration parameters are minimal since it was designed to perform an unsupervised clustering as an exploratory search for features in large datasets; besides, it was developed to keep a low computational cost with a time complexity of $O(n \log n)$ and $O(n^2)$ in the worst case [91].

DBSCAN works detecting a common density of points in a two-dimensional euclidean space (in this case the dimensionality reduced baseline $D_{20935\times 2}$). It is considered that such density will be greater with points related in between than in surrounded zones out of the generated cluster. DBSCAN selects a random point and measures the distance between the random point and a next point, and so on successively with the other points.

The algorithm correlates the points belonging to the dataset depending on the initial parameters configuration *Eps* and *MinPts*. The value of *Eps* and *MinPts* have to be designated as input parameters. The *Eps* input parameter manages a circle radius in which a specific density is desired, and the *MinPts* parameter determines a minimum number of points desired in the circle.

Several distance measurements have been considered for the association of points, depending on their similarity or dissimilarity (being one the opposite of the other mea-

sure). Even though, like several clustering algorithms including DBSCAN work using the Euclidean distance proximity measure, there are different measurement alternatives, here are presented some of the most common proximity measurements between two points which can be used in similar applications:

- **Minkowski distance:** the Minkowski distance can be assumed as a generalization of the Manhattan and Euclidean distance,

$$dist_{x,y} = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} \tag{11}$$

  where, x_i, y_i belong to a vector X and are coordinates of x, y with $p > 0$.
- **City block or Manhattan distance:** the Manhattan norm is defined as:

$$dist_{x,y} = \sum_{i=1}^{n} |x_i - y_i| \tag{12}$$

  where, x_i, y_i belong to a vector X and are coordinates of x, y with $p = 1$.
- **Euclidean distance:** the squared Euclidean distance can be represented as:

$$dist_{x,y} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{13}$$

  where, x_i, y_i belong to a vector X and are coordinates of x, y with $p > 2$.
- **Mahalanobis distance:** let define C as the covariance matrix of the variables analyzed, the Mahalanobis distance was defined to measure differences among mean vectors:

$$dist^2_{x,y} = (x_i - y_i) C^{-1} (x_i - y_i)' \tag{14}$$

  where, Where x_i, y_i belong to a vector X and are coordinates of x, y.
- **Cosine similarity:** the cosine similarity has a close relation with the inner product and is defined as:

$$cosdist_{x,y} = \frac{x_i^T y_i}{\|x_i\| \|y_i\|} \tag{15}$$

  where, x_i, y_i belong to a vector X and are coordinates of x, y.

Following the original notation created by Ester et al. [90], DBSCAN algorithm follows six basics rules for clustering a dataset D:

- **Eps-neigborhood of a point** p, denoted as $N_{eps}(p)$, is defined by $N_{eps}(p) = \{q \in D | dist(p, q) \leq Eps\}$.

- A point p is **directly density-reachable** from a point q if
  1. $p \in N_{Eps(q)}$ and
  2. $|N_{Eps(q)}| \leq MinPts$ *'Core point condition'*.

  In this circumstance the core point condition is symmetric, (see Figure 26).
- A point p is **density-reachable** from a point q if there is a chain of points $p_1, \ldots, p_n, p_1 = q, p_n = p$ such that p_{i+1} is directly density-reachable from p_i, (see Figure 27).
- A point p is **density-connected** to a point q with regard to *Eps* and *MinPts* if exists a point o such that p and q are density-reachable from o with regard to *Eps* and *MinPts*, (see Figure 28).
- A **Cluster** is a non-empty subgroup of D with regard to *Eps* and *MinPts* in which the following conditions are satisfied,
  - $\forall p, q$ if $p \in C$ and q is density-reachable from p, then $q \in C$,
  - (2) $\forall p, q \in C : p$ is density connected with q.
- **Noise** is defined as a set of points in D that do not belong to any cluster C_i, i.e. noise $= \{p \in D | \forall i : p \notin C_i\}$.

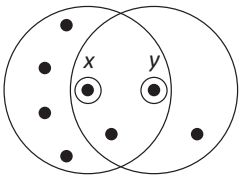


Figure 26. Directly density-reachable.

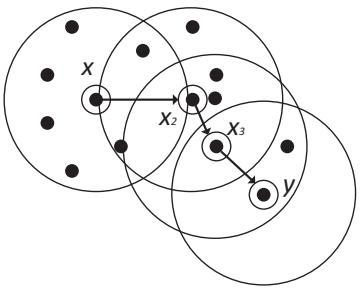


Figure 27. Density-reachable.

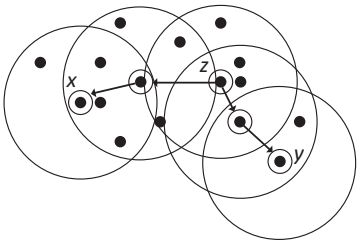


Figure 28. Density-connected.

6.2 SELECTION OF THE INPUT PARAMETERS *Eps* AND *MinPts*

The biggest drawback in the DBSCAN algorithm is the selection of an appropriate set of input parameters *Eps* and *MinPts* in a particular density of points. For a good performance of the DBSCAN algorithm, input parameters have to be adjusted properly with regards to a specific dataset D; in this special case the $D_{20935\times2}$ dataset with a Promax matrix rotation, (see Figure 29).

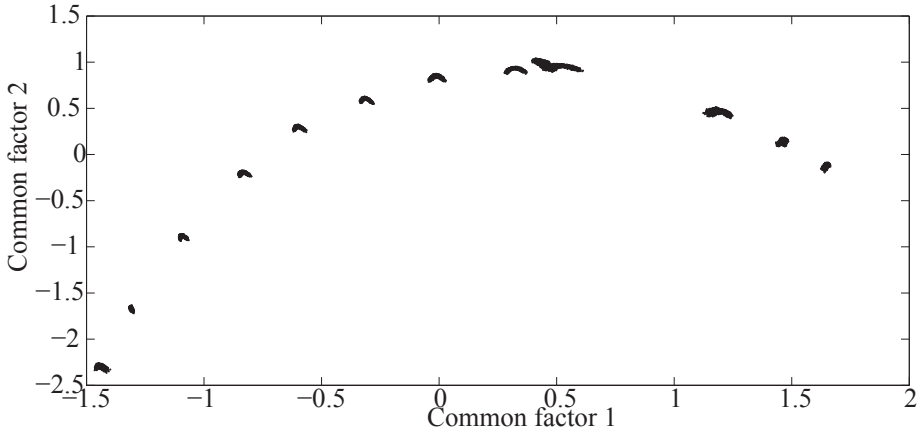


Figure 29. Aluminum beam's dataset graph.

Thus, it was necessary to set up a strategy with the aim to select the DBSCAN's initial parameters depending on the scatter and the space location of the points. The selection of the initial input parameters *Eps* and *MinPts* deliberately, becomes the classification algorithm less "automatized"; besides the performance and the accuracy of clustering can substantially decrease. Automatizing the selection of the input parameters may avoid the seeking of values by trial-and-error.

Afterwards it was generated a simple dataset to show the influence on the variation of the parameters *Eps* and *MinPts*. The two-dimensional dataset had two defined condensed shape clusters, as it can be seen in Figure 30, each group contains 50 points. There was a notable separation between the two clusters, thus, the groups are easily recognizable. The objective with this example was to illustrate the impact of the initial input parameters, varying one parameter while the other was fixed.

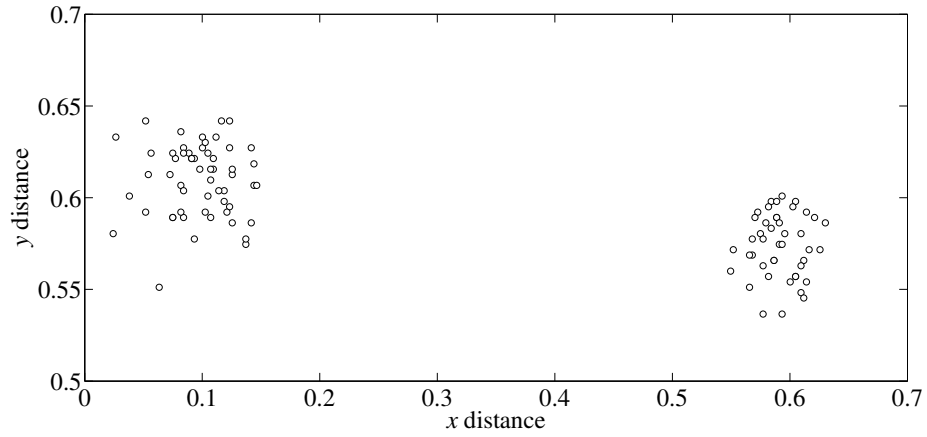


Figure 30. Two clusters artificial dataset.

As it can be seen in Figure 30 the points are spread in a space with a horizontal scale variation between 0 and 0.7 units and a vertical scale variation between 0.52 and 0.7 units. Therefore, a clue of an initial size for the *Eps* parameter could be around the scale of those values. As it was mentioned above, the parameter *MinPts* determines a minimum number of points into the circle formed by *Eps*, thus, in large datasets this parameter may not have the same impact in the final clustering than the *Eps* parameter.

Each clustering figure presented bellow will indicate every formed group in a different gray scale and a *CL* label; if noise is detected it will be represented with an asterisk, besides, the selected *Eps* parameter is also represented graphically with a circle.

6.2.1 Eps variation

The value of *Eps* will vary starting in 1 until 0.01 and the value of *MinPts* is fixed in 10. In the first test belonging to Figure 31, one cluster was generated, the diameter of the *Eps* was too big, thus the input parameters were not well tuned. In the second test

belonging to Figure 32, two clusters were generated, the diameter of *Eps* is convenient and and there was no presence of noise. In the third test belonging to Figure 33, there were no clusters generated, and in the other side the presence of noise was massive.

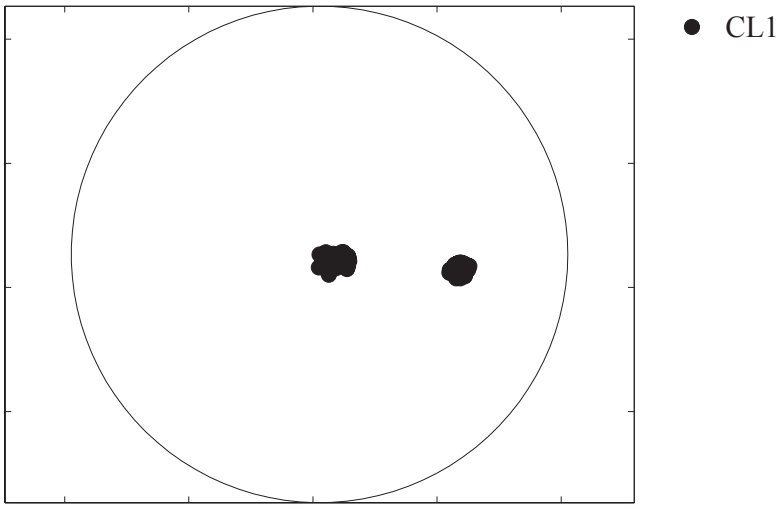


Figure 31. *Eps*=1 *MinPts*=10

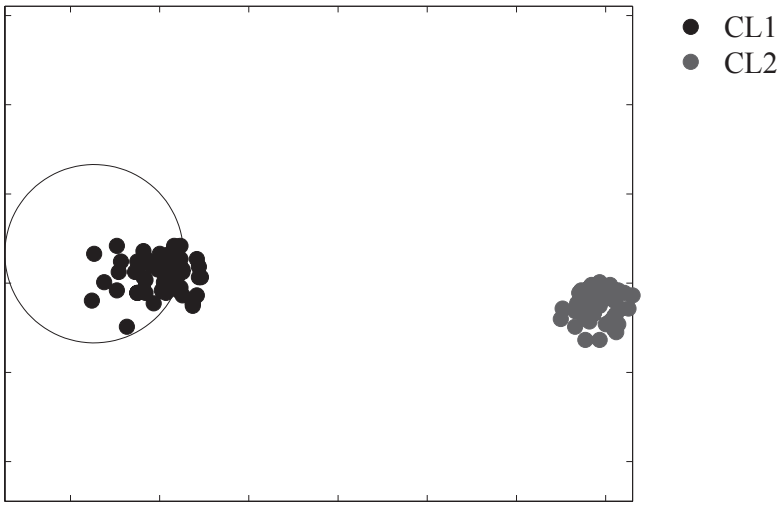


Figure 32. *Eps*=0.1 *MinPts*=10

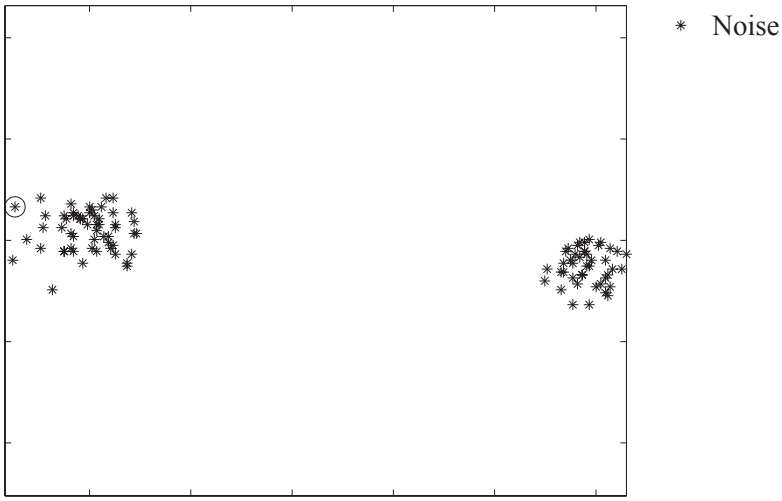


Figure 33. *Eps*=0.01 *MinPts*=10

6.2.2 MinPts variation

Taking into consideration that the maximum number of points in a cluster is 50, it was convenient to have *MinPts* tested using descendant values between 20 and 1; the parameter *Eps* was fixed in 0.1. In the first test belonging to Figure 34, two clusters were generated, the parameter *MinPts* defined worked well and there was no presence of noise. Figure 32, test 2 in this case, is the natural descendant variation in this Section with *Eps*=0.1 and *MinPts*=10 as initial parameters. Two clusters were generated, the parameter *MinPts* defined worked well and and there was no presence of noise.

In the third test belonging to Figure 35, two clusters were generated, the parameter *MinPts* defined worked well and there was no presence of noise. In the fourth test belonging to Figure 36, two clusters were generated, the parameter *MinPts* defined worked well and there was no presence of noise.

As it is demonstrated, the output information provided by the DBSCAN algorithm seemed to be more sensitive to a change in the results due to the parameter *Eps*, hence it was evident that a strategy to automatically determine the parameter *Eps* was needed. Thus, Section 6.4 is dedicated to develop a solution for to this drawback.

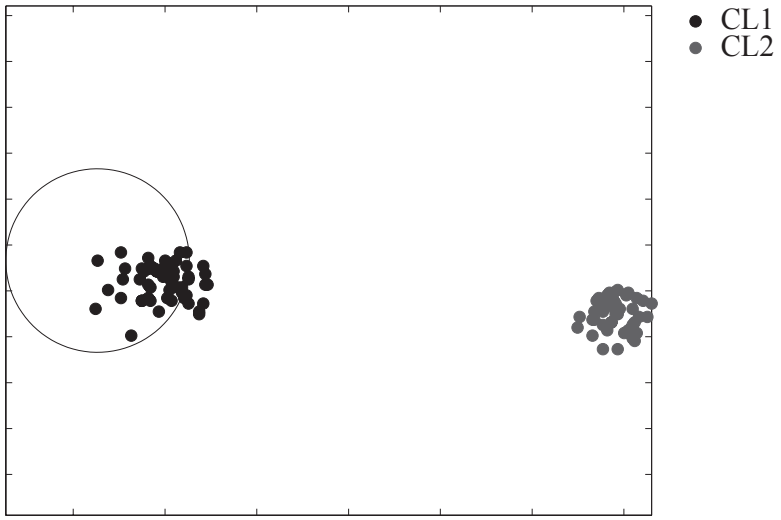


Figure 34. *Eps*=0.1 *MinPts*=20

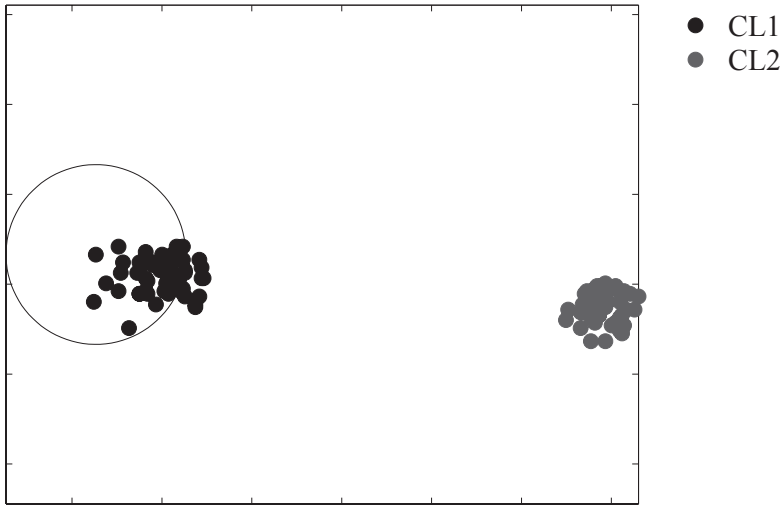


Figure 35. *Eps*=0.1 *MinPts*=5

6.3 DEFINITION OF *MinPts*

A development carried out by Gaonkar et al. [92] is one of those efforts to improve the automatism of the DBSCAN algorithm. In this article a technique to automatically determine the parameter *Eps* was evaluated. However, the authors proposed a brief

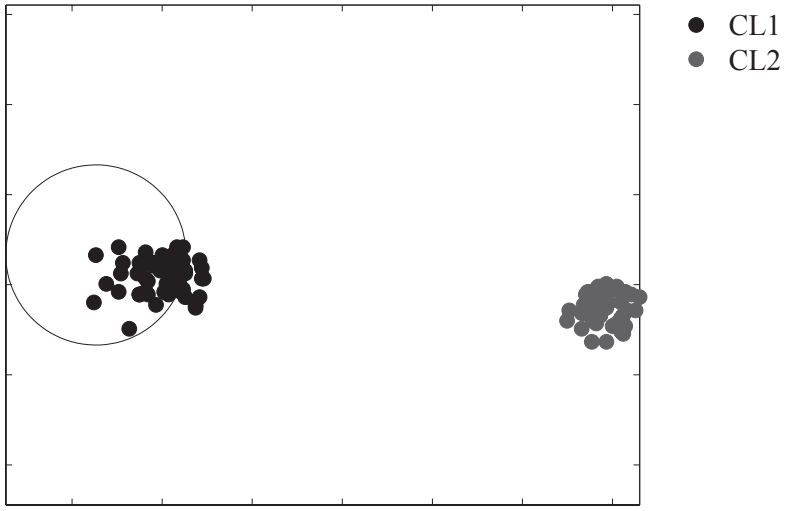


Figure 36. *Eps*=0.1 *MinPts*=1

way to determine the parameter *MinPts*; nevertheless, in the present methodology the determination of the parameter *MinPts* was slightly altered having in mind that some changes delivered better result.

The machine learning function *nearest neighbor*, permits the discovering of distances using the Euclidean distance between a selected point and the leftover points in a specific dataset D; this function also may defines the densities contained in the specific dataset D. Therefore, those values may become possible *Eps* values. Those values are also treated further in Section 6.5 as a part of the methodology to chose the best *Eps* value. The selection of *MinPts* is represented as follows:

$$MinPts = \sum_{i=1}^{n} d_i, \tag{16}$$

where d_i is the i-th value of each density measured in the neighborhood of a point content in a specific dataset D with n number of experiments.

6.4 DEFINITION OF *Eps* USING A GENETIC ALGORITHM

Due to the simplicity in terms of computational cost, computer implementation, and automatism of the DBSCAN algorithm, scientists have made efforts with new techniques to include new characteristics to the DBSCAN algorithm with the aim of a better overall performance.

A brief state of the art with regard to the selection of initial parameters for DBSCAN is presented bellow. An automatic selection of the parameter *Eps* developed by Gaonkar and Sawant [92] using the *k-dist* graph to determine different densities into a specific dataset D. Kumar and Reddy [93] developed a methodology based on Group methods called G-DBSCAN, accelerating the neighbor searching in a specific dataset D. Patwary et al. [94] presented an methodology called PDSDBSCAN using graph algorithms concepts. Furthermore, a tree-based approach was implemented to generate the clusters.

Xiong et al. [95] presented a DBSCAN modification called DBSCAN-DLP based on density levels partitioning. An algorithm based on DBSCAN called I-DBSCAN was developed by Zhou et al. [96] this paper presented a methodology to determine the initial parameters *Eps* and *MinPts* analyzing statistics properties from a density distribution matrix called $DIST_{n\times n}$. Some other works related to DBSCAN include the ST-DBSCAN[97], DSets-DBSCAN [98], C-DBSCAN [99], BDE-DBSCAN [100], E-DBSCAN [101] and PACA-DBSCAN [102].

In the present methodology, a GA was implemented to enhance the performance of DBSCAN by automatic definition of the parameter *Eps* for a specific dataset D. GA are techniques of optimization which are inspired by natural selection. GA simulates the natural evolution allowing a population of a particular number of individuals to evolve in a specific form under a variety of conditions or rules to a state that maximizes a fitness function [103]. In other words, the main goal is to reach the specific solution through a fitness value.

Following the article published by Srinivas and Patnaik [104], the most common actions to perform a GA involve:

- *The encoding mechanism*, which is the base of a GA, represents the information

to be optimized in a data string, the encoding type in which the data string is defined depends on the nature of the problem. Some of the most common encoding methods involve the use of integer numbers or binary arrays.

- *The fitness function* is the function to be optimized, each chromosome which represents strings of information (e.g. strings of binary numbers representing possible solutions to a problem) has to be analyzed using the fitness function with the aim to preserve the arrays with better results.
- *The selection method* emulates the nature's mechanism of survival, this technique may vary depending on the problem to be solved or optimized, the most common alternatives for a selection method include a tournament selection, a rank-based selection, elitist strategies, steady-state selection and a proportionate selection scheme in which a roulette wheel selection scheme may be included.
- *The crossover process* consists in the selection of random fraction of elements belonging to a chromosome and perform an exchange of those elements among selected chromosomes.
- Finally after crossover, some chromosomes can be exposed to a *mutation*. The mutation consists in changing a fraction of elements belonging to a randomly selected chromosome with new information randomly selected from a specific population.

The parameter *Eps* has a large influence over the performance of the DBSCAN algorithm since it has a direct connection with the typical density of the dataset. Determining an adequate *Eps* value for a specific dataset could significantly increase the automation process, accuracy and reduce processing time. In this methodology construction DBSCAN was automatized by a GA based on the Lin et al. [105] scheme.

The initial population of the genetic algorithm was determined using the *nearest neighbor* function, which determines the densities associated to a particular dataset D, measuring the distances between a selected point and the leftover points. Those distances can be defined as the most common radii associated to a specific dataset D, therefore the initial population was made up with 50 radii between the average radius r_{avg} and the maximum radius r_{max}.

Each chromosome is made up of genes, which contains a section of information in particular. In this case, the chromosomes of the initial population had two different genes, one gene was composed by a point p_i from the common factors dataset $D_{m\times 2}$ with x, y coordinates and the other one had a radius r_i found using the *nearest neighbor*

function. Hence, the chromosome could be a combination of real numbers as it is represented bellow:

Table 5. GA chromosome.

x	y	radius
25.8035	6.3917	0.5184

The *fitness function* was formed by three principal components:

coverage ratio CR:

$$CR = \frac{|S_{p1,r1} \cup S_{p2,r2} \ldots \cup S_{pn,rn}|}{|D|}, \tag{17}$$

sum of density SD:

$$SD = \sum_{i=1}^{n} \frac{|S_{pi,ri}|}{|r_i^2|}, \tag{18}$$

and *duplicate ratio DR:*

$$DR = \frac{\sum_{i=1}^{n} |S_{pi,ri}|}{|S_{p1,r1} \cup S_{p2,r2} \ldots \cup S_{pn,rn}|}, \tag{19}$$

then, the *fitness function* was determined by the following general equation:

$$F = \frac{CR \times SD}{DR}, \tag{20}$$

where each $S_{pi,ri}$ represent a set of points with center pi radius ri.

There were selected radii with maximum values including their belonging points, avoiding redundant points, this is known as the tournament method. In this methodology, half of the initial chromosomes was replaced after each sample was evaluated with the *fitness function*.

Every time that it was needed a random procedure in the crossover and mutation procedures was performed using the Mersenne Twister algorithm created by Matsumoto

and Nishimura [106] which generates uniform pseudo-random numbers, this algorithm has a computational cost on $O(P^n)$, where P is the degree of the polynomial.

Besides, for better comprehension, replaced positions can be identified in tables by the bold text. The crossover operation was applied selecting randomly cross positions between points and radii.

The crossover process was performed as follows:

Table 6. GA before crossover.

x	y	radius
25.8035	6.3917	0.5184
16.4488	10.5420	0.2024
12.5300	10	0.1022

The *crossover* process was made by fixing the point coordinates and randomly varying radii positions, the fitter chromosomes *offsprings*, could result as follows:

Table 7. GA after crossover.

x	y	radius
25.8035	6.3917	**0.2024**
16.4488	10.5420	**0.1022**
12.5300	10	**0.5184**

The mutation process was carried out replacing radii values from some selected chromosomes with new ones in random positions, the new radii values could be larger than the maximum value of the initial population but not less than the lower value. In the same way, points with center p_i were replaced with new values in random positions.

An example of the mutation process is indicated in Table **??** and Table 9. In the mutation not every chromosome was modified, some of them just changed the radius or the point p_i belonging to the initial population.

A new family of chromosomes emerged after the previous processes related to the genetic algorithm and were evaluated over and over again using the fitness function with the aim to improve the information inside each chromosome until a solution converged in

Table 8. GA before mutation.

x	y	radius
25.8035	6.3917	0.2024
16.4488	10.5420	0.1022
12.5300	10	0.5184

Table 9. GA after mutation.

x	y	radius
25.8035	6.3917	0.2024
16.4488	10.5420	**0.1529**
8.5240	**16.3920**	0.5184

a specific value. In this case, 50 iterations of the sequential process were necessary to guarantee a reliable convergence.

Finally, the best radii selected by the GA could be used as the *Eps* desired in a specific dataset D. Moreover, with the use of the automatically selected *Eps* the DBSCAN algorithm should perform an accurate and faster clustering performance.

6.5 EXPERIMENTAL EVALUATION

To validate the genetic DBSCAN algorithm, seven different datasets were used; one of them was presented in Subsection 6.2, Figure 30, the rest of them are free access artificial datasets, which are specially designed with the aim of test pattern recognition algorithms under development. Datasets are available on-line at [107].

The overall precision was evaluated using the expressions proposed by Fawcett [108]:

$$TP = \frac{positives\ correctly\ classified}{total\ positives}, \tag{21}$$

$$FP = \frac{negatives\ correctly\ classified}{total\ positives}, \tag{22}$$

where TP are the true positives detected points, and FP the false positive detected points, thus the precision is calculated as follows:

$$precision = \frac{TP}{TP + FP} \tag{23}$$

6.5.1 Two clusters

The *two clusters* dataset, (see Figure 37), was created in the development of this methodology. It has 100 points and two classes marked with a number into the dashed boxes for user guidance.

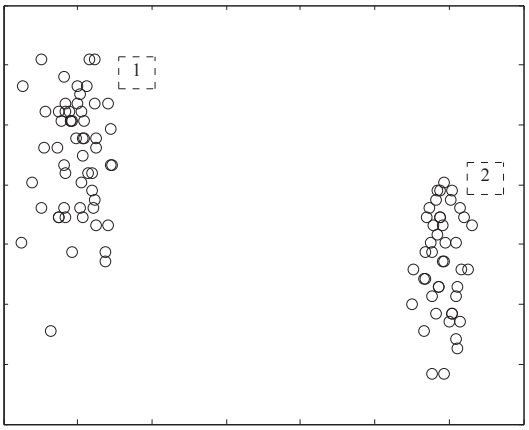


Figure 37. Two clusters.

As a result two clusters were discovered as it can be seen in Figure 38, the initial parameters *Eps*=0.0392 and *MinPts*=6.485 were defined automatically. The elapsed time was 0.236 s.

Figure 38. Two clusters DBSCAN clustering results.

6.5.2 Aggregation

The *Aggregation* dataset (see Figure 39) was developed by Gionis et al. [109]. It has 788 points and seven classes marked with a number into the dashed boxes for user guidance.

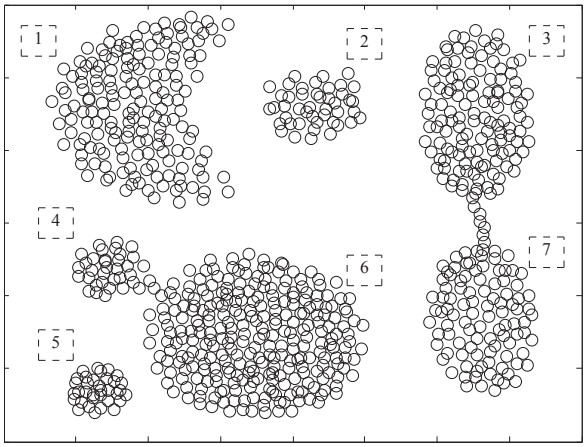


Figure 39. Aggregation.

As a result eight clusters were discovered as it can be seen in Figure 40, the initial parameters *Eps*=1.130 and *MinPts*=5.498 were defined automatically. The elapsed

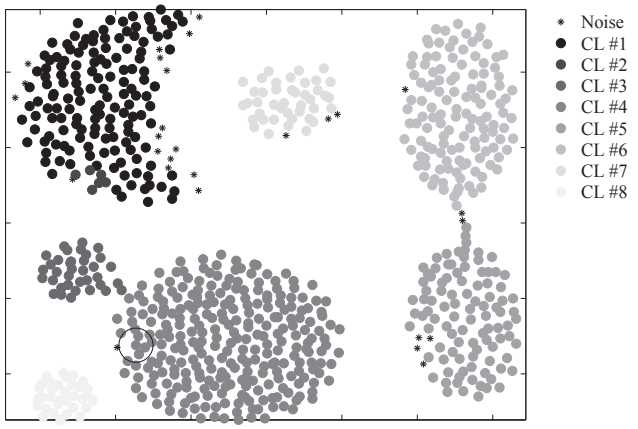


Figure 40. Aggregation DBSCAN clustering results.

time was 0.337 s.

6.5.3 Dim 32 and Dim 64

The high dimensionality datasets (DimSets) *Dim 32* and *Dim 64* were synthetically generated in the framework of the work development by Fränti et al. [110], the *Dim 32* dataset has 32 dimensions, 1024 points and 16 classes, the *Dim 64* dataset has 64 dimensions, 1024 points and 16 classes.

The dimensionality reduction process was carried out using the FA algorithm proposed by the present methodology, where the first two common factors generated shown condensed groups belonging to a probable cluster which are easily recognizable. 10 common factors were necessary to represent the 90.1 % of the *Dim 32* information and 15 common factors were necessary to represent the 99.9 % of the *Dim 64* information. Anyhow, just the first two common factors are graphically represented in Figures 41 and 42 in order to be classified using the DBSCAN algorithm.

As a result 14 clusters were discovered in the Dim 32 dataset as it can be seen in Figure 43, the initial parameters *Eps*=0.112 and *MinPts*=10.055 were defined automatically. The elapsed time was 0.487 s. 16 clusters were discovered in the Dim 64 dataset as it can be seen in Figure 44, the initial parameters *Eps*=0.077 and *MinPts*=4.670 were

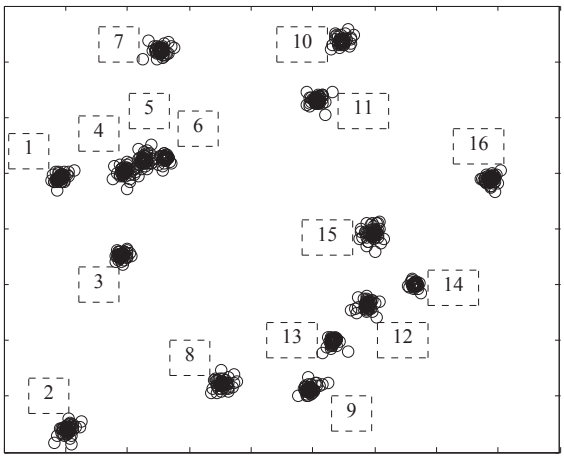


Figure 41. Dim 32.

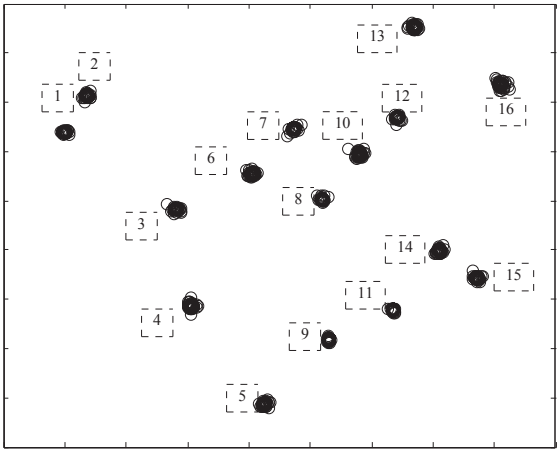


Figure 42. Dim 64.

defined automatically. The elapsed time was 0.544 s.

6.5.4 Flame

The *Flame* dataset, (see Figure 45), was developed by Fu and Medico [111]. It has 240 points and two classes marked with a number into the dashed boxes for user guidance.

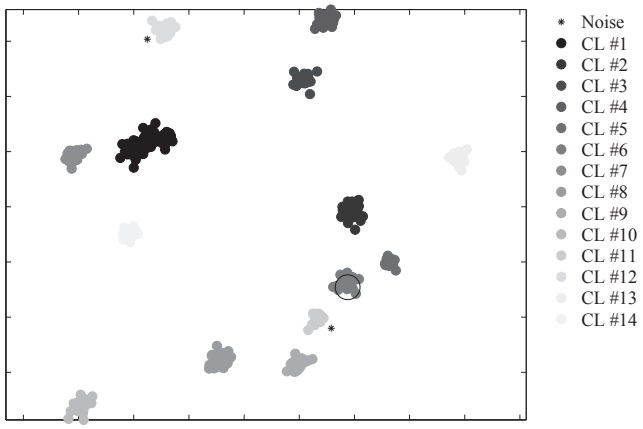


Figure 43. Dim 32 DBSCAN clustering results.

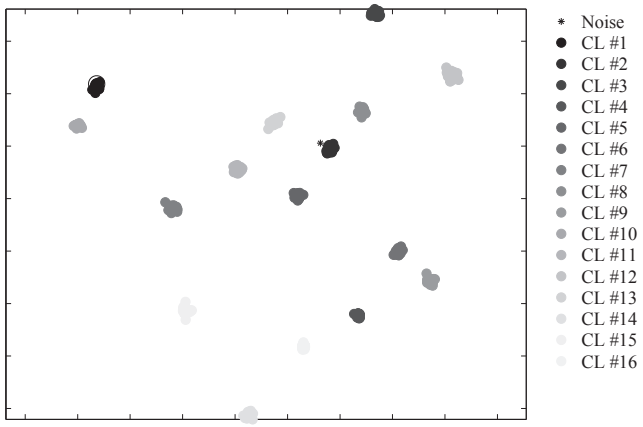


Figure 44. Dim 64 DBSCAN clustering results.

As a result one cluster was discovered as it can be seen in Figure 46, the initial parameters *Eps*=1.244 and *MinPts*=5.871 were defined automatically. The elapsed time was 0.204 s.

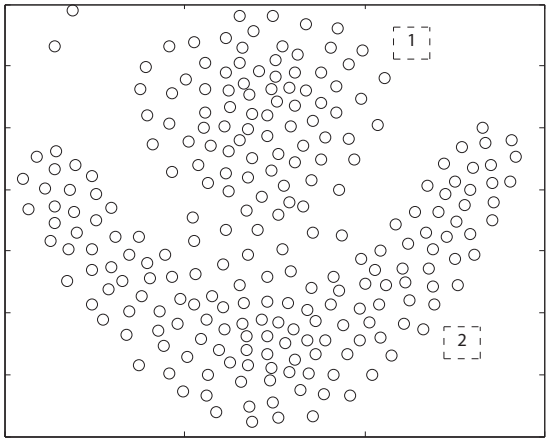


Figure 45. Flame.

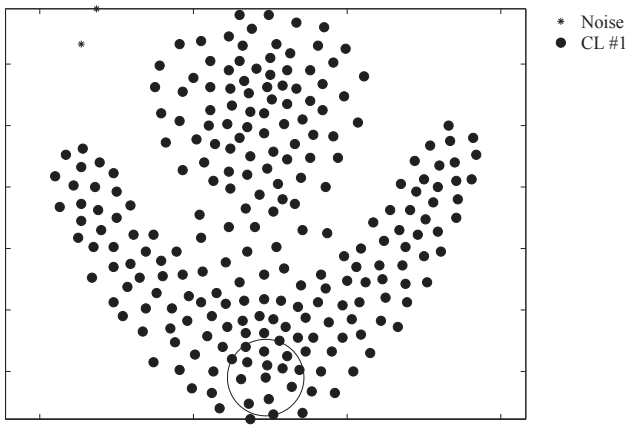


Figure 46. Flame DBSCAN clustering results.

6.5.5 r15

The *r15* dataset,(see Figure 47), was developed by Veenman [112]. It has 600 points and and 15 classes marked with a number into the dashed boxes for user guidance.

As a result eight clusters were discovered as it can be seen in Figure 46, the initial parameters *Eps*=0.617 and *MinPts*=1.043 were defined automatically. The elapsed time was 0.314 s. The algorithm was not capable of discrete points among the core

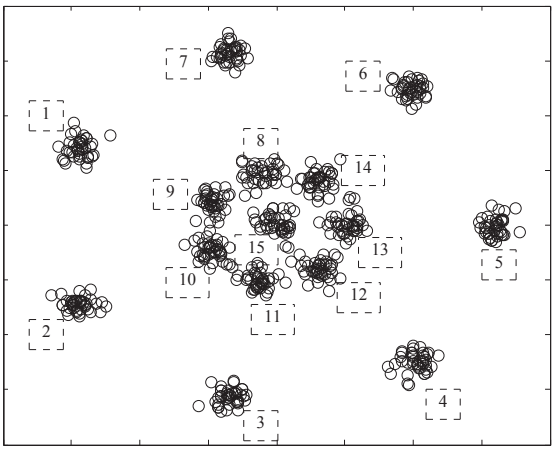


Figure 47. r15.

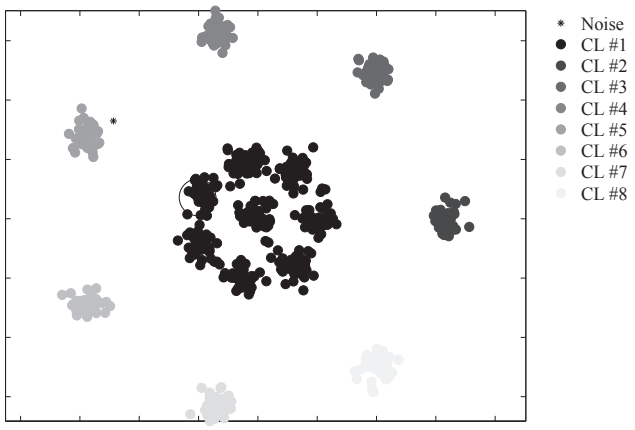


Figure 48. r15 DBSCAN clustering results.

groups, it recognized them as a one big cluster due to a characteristic of the DBSCAN algorithm named as density reachable presented in Section 6.

6.5.6 Jaine

The *Jaine* dataset, (see Figure 49), was developed by Jain and Law [113], it has 373 points and two classes marked with a number into the dashed boxes for user guidance.

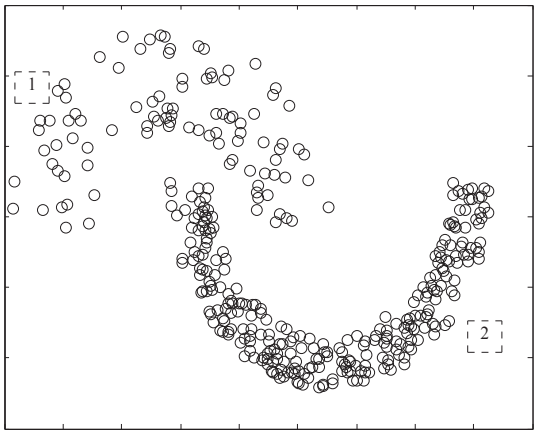


Figure 49. Jain.

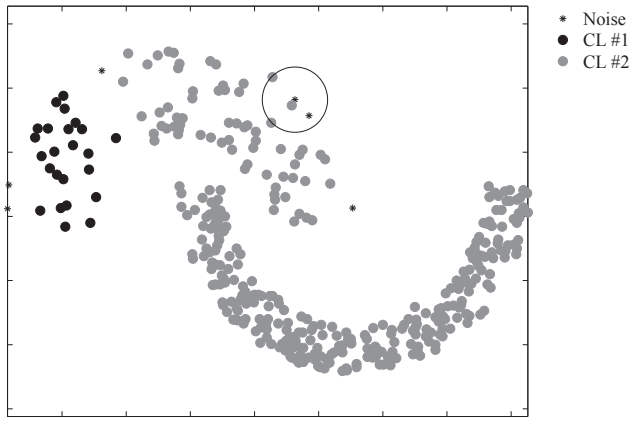


Figure 50. Jain DBSCAN clustering results.

As a result two clusters were discovered as it can be seen in Figure 50, the initial parameters *Eps*=2.544 and *MinPts*=5.228 were defined automatically. The elapsed time was 0.288 s.

Well condensed datasets such as Dim 32, Dim64, and Aggregation reveal a fine outcome from the automatic processing of the algorithm with a superior level of precision; otherwise, the Flame, R15 and Jain datasets due their fuzzy nature presented

a drawback for the algorithm, where the algorithm's precision decreases considerably. Clustering results are condensed in Table 10 including the overall precision.

Table 10. Artificial datasets.

Dataset	**Number of points**	**Features**	**Clusters** C	**Precision** %
Two clusters	100	2	2	99
Aggregation	788	7	8	95.304
Dim 32	1024	16	14	87.304
Dim 64	1024	16	16	99.902
Flame	240	2	1	34.165
R15	600	15	8	53.167
Jain	373	2	2	80.428

7 IMPLEMENTATION IN A COMPUTER PROGRAMMING

The implementation carried out for the unsupervised classification methodology included the preliminary processing, processing and post-processing. It was performed using Matlab R2014a numerical programing software for Windows 7 in an Intel Core i7, 2.2 GHz processor, 6 GB of RAM and 500 GB hard drive PC. This automatized procedure is part of the SHM methodology called *statistical model development*. There were taken in consideration a combination of pre-established Matlab functions and the use of scripts based on the unsupervised clustering theory.

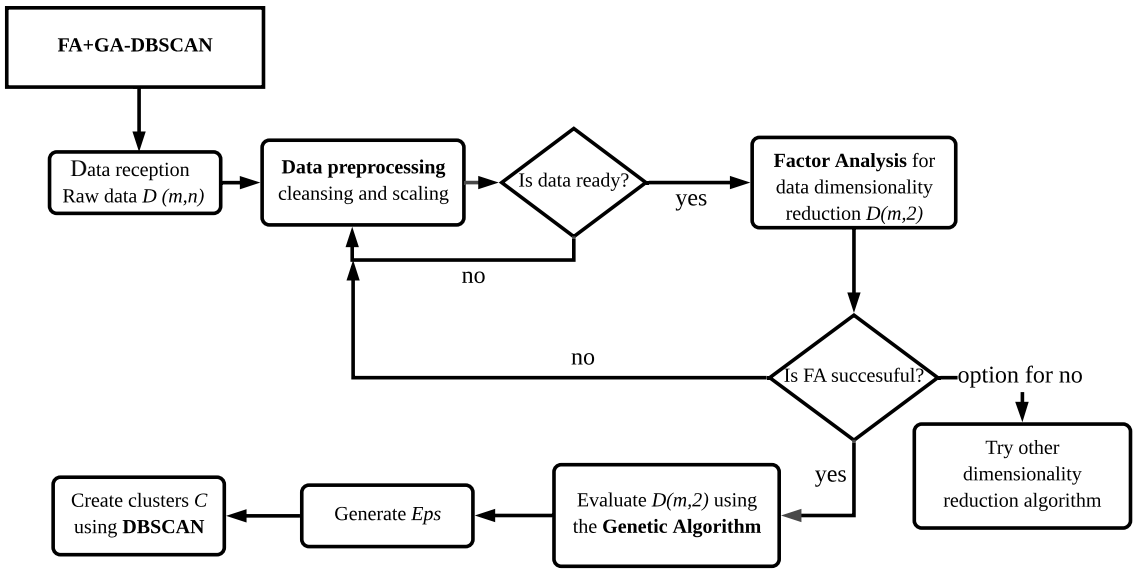


Figure 51. General algorithm flow chart.

The strain matrices were preliminary processed using techniques which included removing undesired information (cleansing), reduction techniques such FA and signal filters which allowed to keep just important features of each signal, thus, just the most representative information of a structural behavior in particular was presented. For the processing

step, the scripts belonging to the DBSCAN and the GA for the *Eps* parameter determination were implemented in a programming language.

After data were clustered into a particular group, post-processing actions were performed to determine the processed signals belonging accuracy. Each row of a specific dataset $D_{n\times 2}$ were labeled within their related cluster, (e.g. if a group of points belongs to the cluster 1, they are labeled with a number one and so on with the other generated groups). with the use of the software's tools it was possible to create figures desired for a better comprehension of the clustering performance.

Following a general pseudocode of the overall methodology is presented:

Data: Raw data reception $RawD_{m,n}$, with m rows (time instants) and n columns (number of sensors), probably includes "unclean" data
Result: $D_{m,n}$
`/* Concatenate` D_i `matrices if it is necessary, with i-th number of experiments to be analyzed in a set of` D `matrices */`
`/* Cleansing */`
for $i \leftarrow 1$ **to** m **do**
  Remove n rows with undefined values such as infinite values or invalid characters
end
`/* Evaluate the signal to noise ratio SNR */`
if $SNR <= 50$ **then**
  Delete the rows with SNR under 50 (probably noise);
end
`/* Scaling of data (if it is needed) */`
for $i \leftarrow 1$ **to** m **do**
  Calculate the mean μ Calculate the variance among sensors σ^2 Evaluate $\overline{D}_{m,n} = (D_{m,n} - \mu)/\sigma$
end
Set $\overline{D}_{m,n} = D_{m,n}$ for simplicity;

Algorithm 1: Data reception and cleansing.

Data: $D_{m,n}$, with m rows (time instants) and n columns (number of sensors)

Result: $D_{m,2}$, dimensionality reduced matrix

`/* Determine the number of factors */`

if $Eigenvalues => 1$ **then**

  Select the number of factors with eigenvalues greater than one $(number\,of\,factors < n)$;

else

  Select the number of factors desired by user criteria;

end

`/* Choose the ideal rotation matrix: varimax, promax, quartimax or none */`

`/* Select the number of common factors desired */`

while $D_{m,n} \neq \lambda \times f + e$ **do**

  Determine the maximum likelihood estimate factor loadings λ ;

  Generate the common factor predictions f;

  Define the specific variances or specific factors e;

  Rotate factors until the equality is achieved;

end

`/* plot the scree plot */`

`/* plot the common factors coefficients */`

Algorithm 2: Factor Analysis.

Data: $D_{m,2}$
Result: *Eps*
`/* Nearest neighbor densities */`
for $i \leftarrow 1$ **to** m **do**
  Choose a random point $p_{i,j}$ belonging to the dataset $D_{m,2}$;
end
Make a measure of nearest neighbor densities from $p_{i,j}$.;
foreach *density measure* d_i **do** determine the *mean density* and the *maximum density* ;
`/* Create initial population */`
for *mean density* **to** *max density* **do**
  Generate randomly a initial population;
end
while *number of iterations desired* **do**
  `/* Call the fitness function` ff `*/`
  $ff = coverage\,ratio/(sum\,of\,density \times duplicate\,ratio)$;
  `/* Eval. the initial population with the` ff `*/`
  Preserve the better population $\rightarrow$ parents;
  `/* Crossover */`
  for $i \leftarrow 1$ **to** *number of parents* **do**
    Vary a desired number of density measures in random positions;
  end
  `/* Mutation */`
  for $i \leftarrow 1$ **to** *number of parents* **do**
    replace a desired number of new random density measures in random positions;
  end
  `/* Eval. the new population with the` ff `*/`
end
Select the minimum density value d_i from the new population d_{min};
Set $d_{min} = Eps$;

Algorithm 3: Genetic Algorithm.

Data: $D_{m,2}$ and *Eps*
Result: A N number of clusters C_N and noise
/* Determine the value of *MinPts* */
$Minpts = \sum d_i$;
/* Let X_{un} a set of unvisited points from $D_{m,2}$ */
Set $C = 0$;
while $X_{un} \neq \emptyset$ **do**
    Randomly select a point $p_{i,j} \in X_{un}$;
    if $p_{i,j}$ *is a noncore point* **then**
        Mark $p_{i,j}$ as noise;
        $X_{un} = X_{un} - p_{i,j}$;
    else
        $N = N + 1$;
        Determine all density reachable points from $p_{i,j}$;
        Assign $p_{i,j}$ and previous points to a cluster C_N;
        $X_{un} = X_{un} - C_N$;
    end
    Points marked as noise are also assigned to a special cluster C_N;
end
/* Tag each cluster generated C_N with a natural number, tag noise with 0 */
/* plot the clusters with a specific color */

Algorithm 4: DBSCAN algorithm.

8 RESULTS, COMPUTATIONAL COMPLEXITY AND PRECISION

After confirming the methodology notable performance with condensed datasets, the clustering methodology could be tested with an acceptable performance using the experimental dataset (the dimensionality reduced aluminum beam's dataset $D_{20935\times 2}$). The parameter Eps=0.011 was defined automatically by the genetic algorithm presented in Section 6.4 and the parameter $MinPts$=13.934 was found using the corresponding equation presented in Section 6.3.

The algorithm was able to find 12 clusters, from 13 original classes belonging to the pitch angles. A total of 20932 points were clustered into 12 different clusters and 3 points were identified as noise, each formed cluster can be identified in a specific scale of gray, Figure 52.

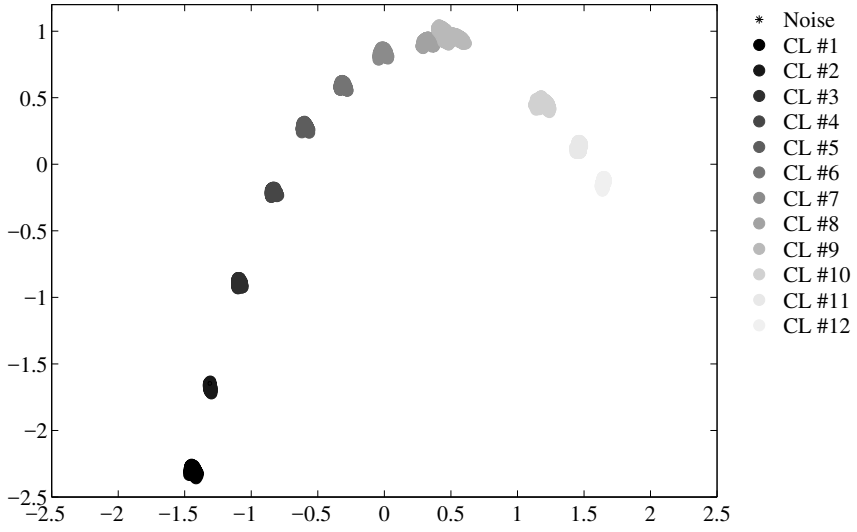


Figure 52. Beam's dataset DBSCAN clustering results.

In the specific case of the aluminum beam, the load was always normal to the cross

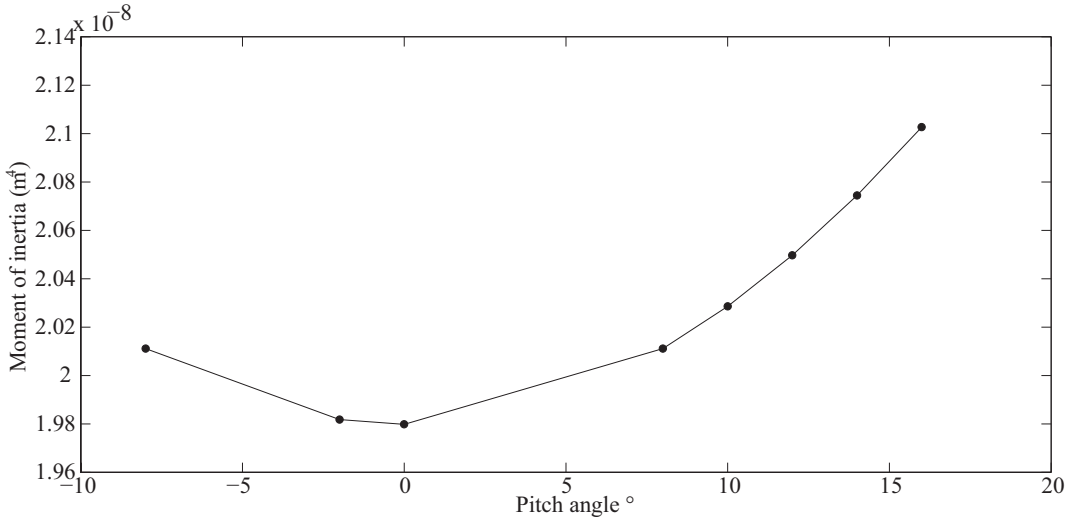
section plane, parallel to the y axis. As it is proposed by the beam theory under elastic deformations, when the pitch angle changed under specific loads, the moment of inertia varied, and because the deflection is inversely proportional to the moment of inertia the aluminum's beam strain field could have changed. The maximum deflection of the beam can be determined using the following equation:

$$\delta_{max} = \frac{PL^3}{3EI}, \tag{24}$$

where P is a specific load, L the distance measured between a reference fixed point and a specific load, E the Young's modulus, and I moment of inertia of the cross-sectional area. The variation of the moment of inertia related to some pitch angles is presented in Figure 53.

Furthermore, the beam's stiffness can be considered as the combination of the Young's modulus and the moment of inertia. In some pitch angles the stiffness was relatively low given that there was a significant change in the moment of inertia in addition to a relatively low Young's modulus. Therefore, due to the previous conditions, a group of well-defined clusters was easily detected by the algorithm.

Figure 53. Variation of the moment of inertia.



8.1 FA+GA-DBSCAN and DS2L-SOM performance comparison

In order to summarize the overall methodology presented in this work, it was named as FA+GA-DBSCAN. The intention of this Section was to contrast the performance and computational complexity taking into account the dataset from the aluminum's beam. The comparison was performed using the results obtained with the FA+GA-DBSCAN and a well proven methodology named DS2L-SOM proposed by Sierra [16].

This comparison allowed to have a benchmark for the performance of the FA+GA-DBSCAN considering that the algorithm presented by Sierra derived reliable results. This comparison was carried out using two common applications for algorithm classification assessment such as the confusion matrix and the ROC curves presented in the following Subsections 8.1.1 and 8.1.2.

The comparison was performed using Matlab R2014a numerical programing software for Windows 10 in an Intel Core i7, 2.6 GHz processor, 16 GB of RAM and 1 TB hard drive PC. The algorithms were analyzed running the same dataset. The aluminum beam's dataset $D_{20935\times 2}$, which has 13 known groups, was the dataset selected for the performance evaluation. Each algorithm was run for six trials in order to create trends on time and precision and avoid compilation anomalies.

As a result, both algorithms seemed to perform a stable and fast classification, deriving on similar results, thus the overall precision as it is discussed in the following Sections 8.1.1 and 8.1.2. However, the computational cost had a notable difference between algorithms. The running time of the overall process for both algorithms is presented in Figure 54.

It was evident that the FA+GA-DBSCAN achieved the best, classifying more than 20000 signal projections in a quarter of a second. As it can be identified in Figure 54, the average computational cost of the DS2L-SOM was 31.459 s almost twice in comparison with the FA+GA-DBSCAN which had an average processing time of 16.576 s.

This remarkable computational time was achieved taken into account that the DBSCAN algorithm performance remains under the $O(n \log n)$. Nevertheless, the selection of accurate initial parameters, in which the function *nearest neighbor* was involved, may have reduced the computational complexity of the overall process to $O(n + nd)$, where

d is the maximum number of distance computations related to the function *nearest neighbor* as it was stated by Kumar and Reddy [93] in a similar approach.

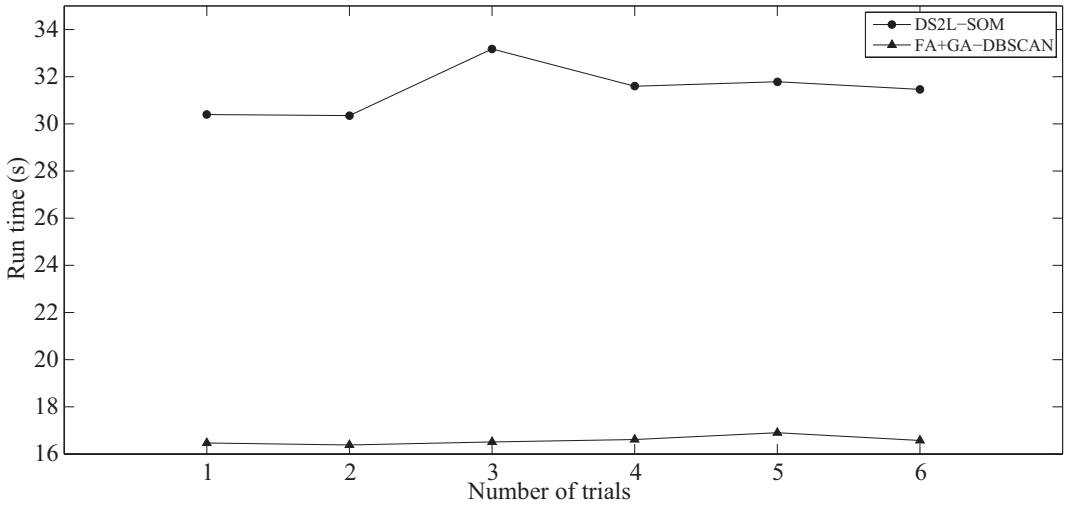


Figure 54. FA+GA-DBSCAN and DS2L-SOM time complexity.

8.1.1 Confusion matrix

The main objective of the confusion matrix is to determine a correlation between the clusters found by the classification algorithm and the original groups. The precision of the FA+GA-DBSCAN and th DS2L-SOM was determined with the use of a confusion matrix. The correlation of each original signal with a clustered point was defined into it.

The confusion matrix for the FA+GA-DBSCAN algorithm is represented in Table 11 and for the DS2L-SOM in Table 12. The overall performance of the FA+GA-DBSCAN classifier was remarkable, however the performance of the DS2L-SOM was also notable, both with precisions over 90 %. The precision of the FA+GA-DBSCAN was slightly poorest than the DS2L-SOM, the first one had an overall precision of 92.285 % and the last one a precision of 92.291 %, both of them were calculated with the equation 23 presented in Section 6.5.

The cause of the decreased algorithms precision is related to the cluster number 9 (CL9), which contains signal information about two specific classes (pitch angles) 8° and 10°, this phenomenon appeared to be similar in both algorithms. These two original pitch

angles were not as mechanically different as expected, since the inferred strain field could have been similar. The classifiers were not able to determine a difference between classes, in this case their strain magnitudes were particularly almost equal.

Table 11. FA+GA-DBSCAN confusion matrix.

		Clusters found													
Baseline	**Pitch angle**	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**C9**	**C10**	**C11**	**C12**	**Noise**	**True positives**
BL_0	**-8**	**1609**	0	0	0	0	0	0	0	0	0	0	0	0	1609
BL_2	**-6**	0	**1609**	0	0	0	0	0	0	0	0	0	0	0	1609
BL_4	**-4**	0	0	**1609**	0	0	0	0	0	0	0	0	0	0	1609
BL_6	**-2**	0	0	0	**1609**	0	0	0	0	0	0	0	0	0	1609
BL_8	**0**	0	0	0	0	**1609**	0	0	0	0	0	0	0	0	1609
BL_10	**2**	0	0	0	0	0	**1610**	0	0	0	0	0	0	0	1610
BL_12	**4**	0	0	0	0	0	0	**1609**	0	0	0	0	0	0	1609
BL_14	**6**	0	0	0	0	0	0	0	**1610**	0	0	0	0	1	1610
BL_16	**8**	0	0	0	0	0	0	0	0	**1611**	0	0	0	0	1611
BL_18	**10**	0	0	0	0	0	0	0	0	1612	0	0	0	0	0
BL_20	**12**	0	0	0	0	0	0	0	0	0	**1611**	0	0	1	1611
BL_22	**14**	0	0	0	0	0	0	0	0	0	0	**1612**	0	0	1612
BL_24	**16**	0	0	0	0	0	0	0	0	0	0	0	**1612**	1	1612
Total		**1609**	**1609**	**1609**	**1609**	**1609**	**1610**	**1609**	**1610**	**3223**	**1611**	**1612**	**1612**	**3**	**20935**

Table 12. DS2L-SOM confusion matrix.

		Clusters found													
Baseline	**Pitch angle**	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**C9**	**C10**	**C11**	**C12**	**Noise**	**True positives**
BL_0	**-8**	**1609**	0	0	0	0	0	0	0	0	0	0	0	0	1609
BL_2	**-6**	0	**1609**	0	0	0	0	0	0	0	0	0	0	0	1609
BL_4	**-4**	0	0	**1609**	0	0	0	0	0	0	0	0	0	0	1609
BL_6	**-2**	0	0	0	**1609**	0	0	0	0	0	0	0	0	0	1609
BL_8	**0**	0	0	0	0	**1609**	0	0	0	0	0	0	0	0	1609
BL_10	**2**	0	0	0	0	0	**1610**	0	0	0	0	0	0	0	1610
BL_12	**4**	0	0	0	0	0	0	**1609**	0	0	0	0	0	0	1609
BL_14	**6**	0	0	0	0	0	0	0	**1611**	0	0	0	0	0	1611
BL_16	**8**	0	0	0	0	0	0	0	0	**1611**	0	0	0	0	1611
BL_18	**10**	0	0	0	0	0	0	0	0	1612	0	0	0	0	0
BL_20	**12**	0	0	0	0	0	0	0	0	0	**1612**	0	0	0	1612
BL_22	**14**	0	0	0	0	0	0	0	0	0	0	**1610**	0	2	1610
BL_24	**16**	0	0	0	0	0	0	0	0	0	0	0	**1613**	0	1613
Total		**1609**	**1609**	**1609**	**1609**	**1609**	**1610**	**1609**	**1611**	**3223**	**1612**	**1610**	**1613**	**2**	**20935**

8.1.2 Receiving Operating Curves ROC

Recently, the ROC graphs have had an increase on their application to determine the classifiers' performance due to the lack of proper and simple metrics [108]. Mainly, the ROC graphs can determine how many data points are clustered as *true positives*, if the sample is positive and it is clustered as positive and *false positives*, if the sample is positive and it is clustered as negative. Moreover, the classification capability of the

algorithm can be determined visualizing the Area Under the Curve AUC in the ROC graph.

To understand the capability of the classifier, it is just necessary to represent graphically the true positives vs. the false positives. As it was found by Fawcett [108], the classification algorithms achieve their highest accuracy if the AUC remains around 70 %. For the DS2L-SOM the AUC index was 70.829 %, and the FA+GA-DBSCAN AUC index was 74.982 %. The curves ROC of both methodologies are presented graphically in Figure 55 for the DBSCAN methodology and Figure 56 for the DS2L-SOM technique.

In this particular case the AUC of the DS2L-SOM and the FA+GA-DBSCAN algorithm remains around 70 %, however, the AUC FA+GA-DBSCAN is slightly superior. Hence, the FA+GA-DBSCAN algorithm was not over-trained, and the results obtained so far by the methodology were reliable.

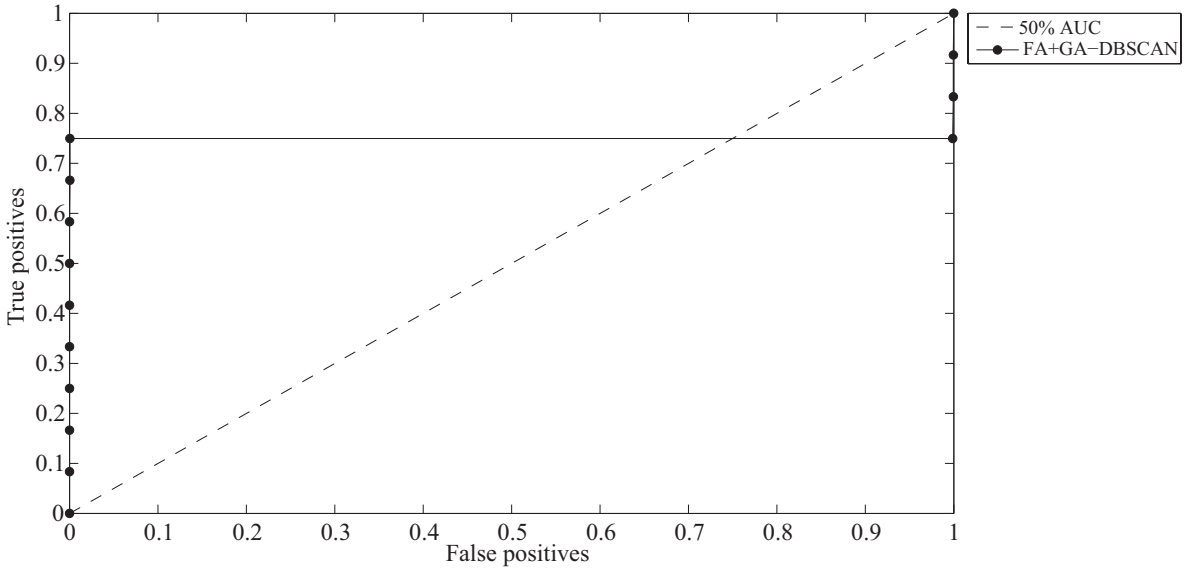


Figure 55. FA+GA-DBSCAN ROC curve.

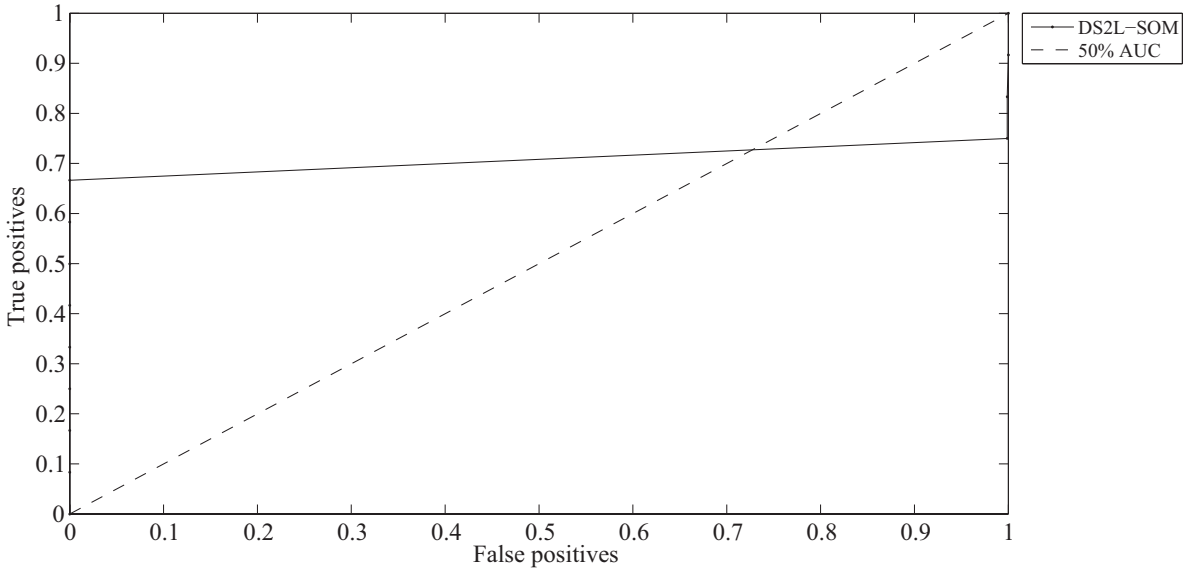


Figure 56. DS2L-SOM ROC curve.

9 TEST OF THE FA+GA-DBSCAN IN A REAL CASE

At this point the methodology was able to be tested in a real case with unknown structural operational conditions. An electric powered UAV designed by Sierra et al. [114], with a rectangular beam made of CFRP skin and a wooden core which is part of the wing main structure was taken into consideration. The wing's beam was instrumented with FBGs sensors to gather strain signals datasets. The UAV have a wingspan of 4 m and a take off weight TOW of 13 kg. A general overview of the prototype for strain signal acquisitions is presented in Figure 57.

Figure 57. Strain signals acquisition prototype.

The aim with the instrumented wing was to validate the performance of damage identification methodologies in a real case. Hence, the necessity to create techniques such as FA+GA-DBSCAN, to generate natural clusters as an exploratory development based on the obtainment of strain signals related to regular operational conditions.

The idea with this UAV was to develop methodologies capable to automatically detect damage. This UAV was considered as a prototype for strain signal acquisition and it belongs to a project sponsored by the *Universidad Pontificia Bolivariana UPB* under

the science and technological production plan *UPB Innova* and the UPB's *School of Aeronautics* with name *"Desarrollo de un sistema remoto de adquisición y transmisión de medidas de deformación en una aeronave con el fin de inferir la integridad de la estructura"*.

The UAV had an acquisition system consisting of a 6 kHz one channel miniaturized Ibsen Photonics interrogator. Further, the prototype wing's beam was instrumented with 20 FBGs with a wavelength range between 1525 nm and 1570 nm, with the aim to perform a thermal compensation, an external temperature sensor was carried inside the PixHawk flight controller hardware's speed sensor located close to the aircraft wing main beam.

Five FBGs were embedded on the top face of the beam, the same number of FBGs were embedded on the bottom face of the beam; the intention of these configurations was to obtain tension/compression strain measurements. Moreover, five sensors were embedded in a $-45°$ configuration on the beam's left face; additionally, other five FBGs were embedded in a $45°$ on the right face; the purpose of these arrangements was to obtain torsion strain measurements. The general arrangement of FBGs is presented in Figure 58.

9.1 Preliminary processing and dimensionality reduction

A preliminary processing was performed in the same way as the one explained in the Section 4. Different actuations were performed by the aircraft's pilot on land in a semi-controlled manner. The remote pilot executes a flight control movement, e.g. a right turn slip, the flight control will performs the actuation under a set of velocity and angle orientation parameters.

A dataset of $D_{195431\times20}$ was gathered from a regular operation flight. Besides, a smaller dataset of $D_{77883\times20}$ was extracted for data validation. For this experimental case, there were selected 15000 representative experiment trials from the $D_{195431\times20}$ matrix, hence, the input dataset was a $D_{15000\times20}$ matrix.

Successively, the dimensionality reduction was carried out using the same methodology

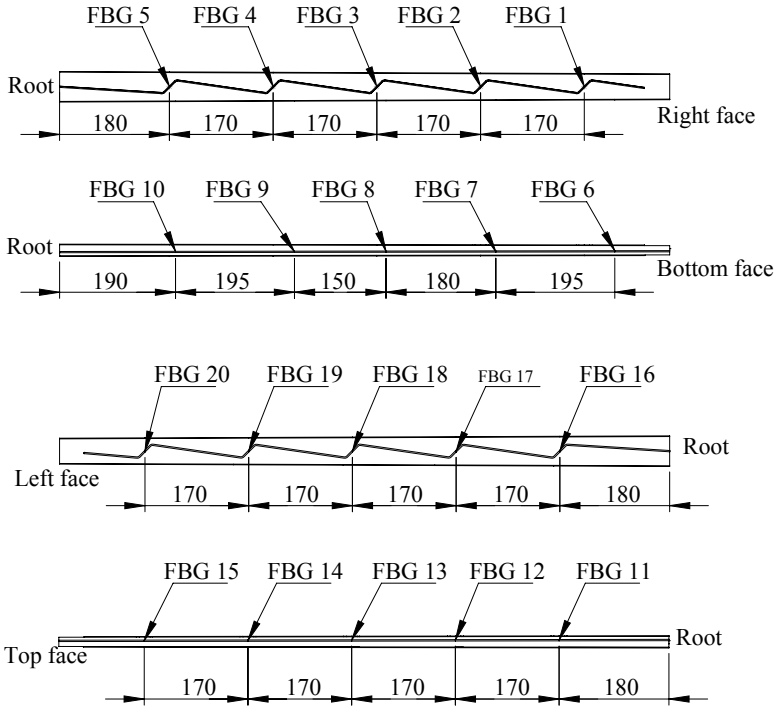


Figure 58. Instrumented beam (all measures are given in mm).

than the one explained in Section 5. The rotation matrices Varimax, Quartimax and Promax, were considered for this experiment. Varimax and Promax rotation matrices projected strain signal information in more spread points but into condensed groups, that could be considered as clusters by the DBSCAN, however, the nature of the projections were fuzzy.

The aircraft's beam beneath regular operational conditions, was submitted under a diversity of maneuvers, including pitch angles from $-15°$ to $15°$ and roll angles from $-30°$ to $30°$. The air-stream which passes through the wing, induces a variation on upper and lower pressures due to the chamber in the airfoil profile, which derives in a resultant force. Nevertheless, in regular flight conditions, the resultant force will act in the center of pressure of the body, which accordingly to Matthews [115] is a point in the section where there is no pitching moment and where the aerodynamic forces will act.

The lift force is defined as the force totally perpendicular to the relative wind, yet, the

drag force, is perpendicular to it. Consequently, if the UAV is flying, the resultant force between the lift and drag forces is an upward force. Moreover, in a regular flight, the direction of the resultant force will be nearly normal, inclined backward on the airfoil section. Hence, the direction of the resultant force will also be nearly normal to the wing's beam.

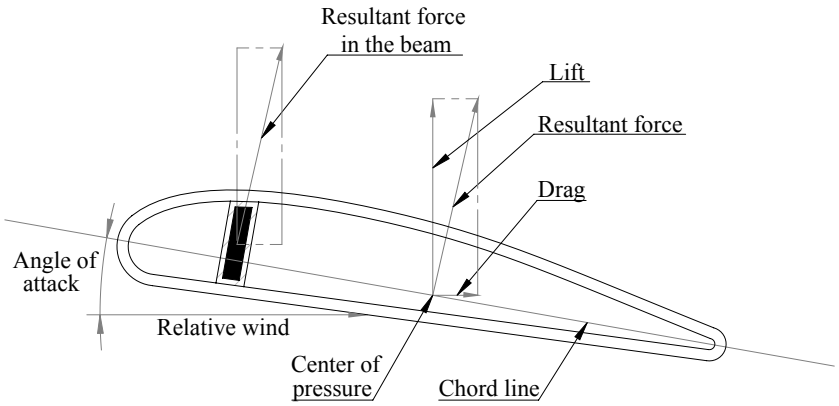


Figure 59. Wing beam section's resultant force.

Although, there were induced torsion loads or load variations since the aircraft is submitted under environmental conditions. Those variations could have been considered insignificant, reason why it can be assumed that the beam experimented almost the same load conditions during the entire flight. For clarity Figure 59 is presented. In addition, the stiffness of the CFRP is considerably higher compared with the aluminum beam, hence, the explanation of why the experimental dataset projected into a two dimensional space had a fuzzy nature where it was not simple to determine point group patterns related to operational conditions.

However, the Varimax rotation was selected since the formed groups seemed to have a more rounded, condensed shapes and besides, it is the most common rotation matrix. Figures 60 and 61 represent the Varimax and Promax rotation successively. The Quartimax rotation preserved condensed groups, however, there could have been a loss of information since there was not a pronounced segregation among formed groups as it can be seen in Figure 62. Further, clustering the Quartimax rotated factors may derive in a over grouping of operational conditions.

After the covariances study was performed, the ideal number of common factors to

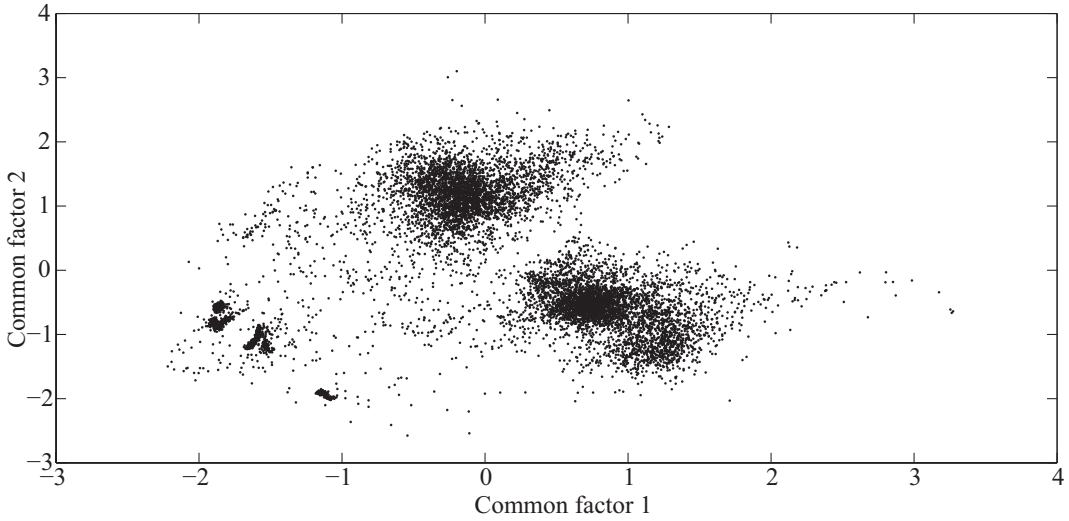


Figure 60. Varimax rotation Prototype signals.

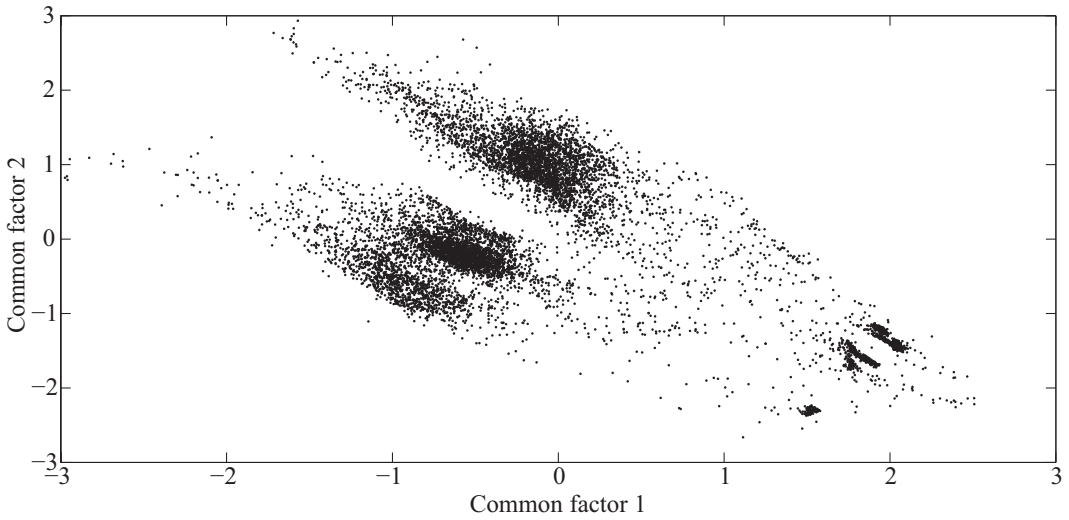


Figure 61. Promax rotation Prototype signals.

retain was three. Those three factors describe the 95.144 % of the original information. However, the DBSCAN algorithm works in a two-dimensional space, thus, there was a percentage of information which was lost in this procedure. The information retained for the first two common factors was 88.09 %. Although the rule of the eigenvalues greater than one was not fully respected, the information retained was close to the 90 %.

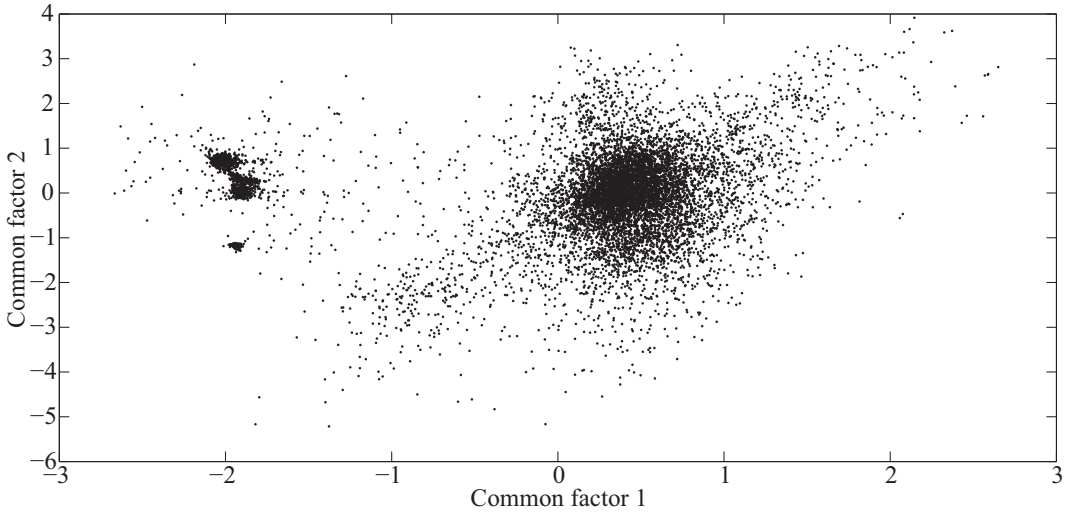


Figure 62. Quartimax rotation Prototype signals.

9.2 Data processing and results

With the use of the GA-DBSCAN algorithm, 1576 clusters were found and the DBSCAN initial parameters were automatically defined $Eps = 0.011436$ and $MinPts = 1.1437$. It was clear that at the beginning, the data clustering, Figure 63 could be confused because of the large quantity of clusters generated. However, not every discovered cluster was meaningful, since the size of a large quantity of clusters was small.

Thus, it was necessary to have a discrimination of the larger clusters which had a high probability of relationship with an operational condition. Finally, 4342 points were marked as noise by the clustering algorithm; it is important to highlight this action performed the DBSCAN algorithm, since the algorithm helped in the cleansing selecting unrelated and spread signals as noise.

10 large clusters were selected because of their significant relationship among strain signals, and are represented graphically in Figure 64. Those clusters may have represented the relevant strain field information related to operational conditions of the aircraft, and may be used as part of a baseline for a further analysis using a damage detection algorithm, employing the concepts explained in Section 3.6. The resultant clusters are presented below in Table 13 and represented graphically in Figure 64. The overall processing time was 308.399 s.

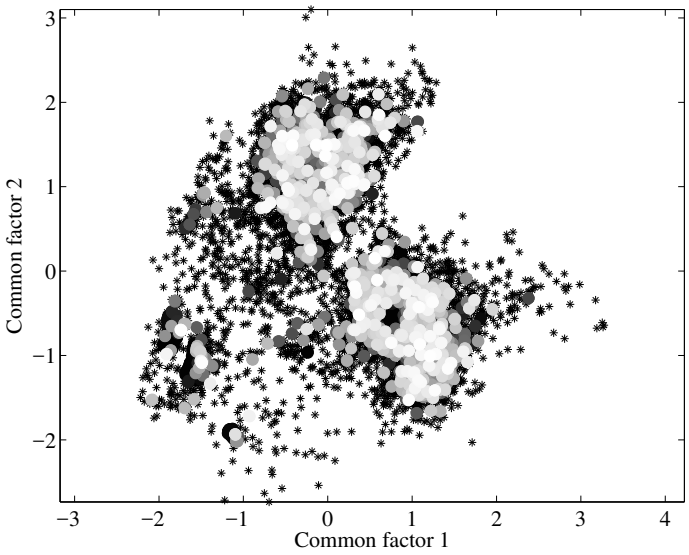


Figure 63. FA+GA-DBSCAN Prototype signal clustering.

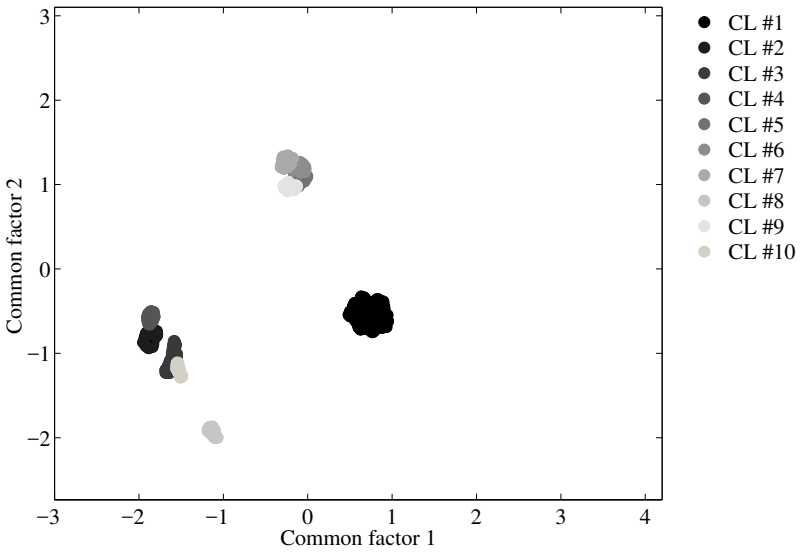


Figure 64. FA+GA-DBSCAN Prototype resultant clusters.

Table 13. Resultant clusters

Cluster Name	**Experiment trials**	**Number of sensors**
C11	1635	20
C3	1085	20
C20	580	20
C73	307	20
C76	204	20
C12	171	20
C19	146	20
C38	139	20
C25	114	20
C61	103	20

CONCLUSIONS

Damage detection techniques under the SHM paradigm, require a handling of large quantity of information in which machine learning methodologies are involved, where a synergy of multidisciplinary processes and novel unsupervised and supervised learning algorithms are needed. The implementation of damage detection techniques in real world problems, leads to execute strategies to create or modify vanguard AI algorithms understanding the accepted SHM fundamental axioms proposed by Worden and Farrar [75].

This work tried out a new methodology proposed for automatic operational condition identification in the framework of the SHM. To complete the proposed methodology, an experiment carried out by Sierra [16] was considered. The experiment consisted in an aluminum beam instrumented with FBGs placed in cantilever under dynamic loads in 13 different operational conditions. A variety of algorithms belonging to the field of Machine learning were taken into account. A system evaluation and a preliminary process were performed to remove outliers.

The use of an accurate preliminary process technique in relation to the physical phenomena enhanced the clustering performance, since it generated clearer groups and remove unnecessary information. A dimensionality reduction technique called factor analysis was explored to perform a dimensionality reduction of the beam's baseline data. The beam's data was projected in a two-dimensional space for a further clustering process. The FA technique seemed to have a solid ability to match specific loads with strain signals in a randomized environment. FA can be considered as an alternative for PCA in the dimensionality reduction process, in which its properties can be explored for handling information in multidimensional datasets.

An unsupervised classifier DBSCAN was employed for grouping the reduced information from the FA process. The Machine learning function *nearest neighbor* was meaningful in the initial parameters *MinPts* and *Eps* determination. Further, the DBSCAN al-

gorithm was automatized with a genetic algorithm looking for determining the initial parameter *Eps*. The handling of a genetic algorithm improved the DBSCAN capabilities substantially; due to the automation of the DBSCAN with a GA, it was possible to infer a variety of structural behaviors presented in an experimental procedure with an aluminum beam given the effect of different discrete loads.

A total of 12 different operational conditions were identified in an unsupervised way. The experimental results lay bare that due to the action of the genetic algorithm the selection of an automatized *MinPts* lead to specific clusters that could have been optimum in size, rejecting specific signals as outliers since DBSCAN is also capable of detecting noise. Under this terms, the algorithm had an overall precision of 92.285 %. The precision of the algorithm in an exploratory clustering was remarkable. In the same way, a comparison made with a well-proven methodology, left clear that the the computational complexity of the FA+GA-DBSCAN was extraordinary, reducing almost in a half the processing time.

Moreover, as it was presented above, FA+GA-DBSCAN algorithm had more sensitivity for clustering detection in condensed datasets, and a loss of accuracy was identified when fuzzy datasets were handled. Hence, it is suggested to perform a dataset clustering in such way that the datasets remain in the most condensed form possible. In rigid structures, the methodology may have decreased in the ability to detect different structural behaviors due to the lack of a considerable difference in the strain field among specific loads.

Further, the stiffness of a structure in cantilever submitted under dynamic loads was an underlying factor in the methodology for the dimensionality reduction and strain signal data classification processes. Two factors had a direct incidence in the structure's deflection, the cross section's moment of inertia and the material stiffness. Although the cross section's moment of inertia varies by operational conditions, if the material stiffens is relatively high, that change in the moment of inertia will be insignificant in proportion to the overall stiffness.

An evidence of the previous state, was observed when data were projected using FA. When the aluminum beam's dataset was projected, clear and defined groups emerged; Otherwise, when the dataset belonging to a more rigid structure such as the UAV dataset, derived in more fuzzy and scattered projection of points.

Finally, some commentaries are presented:

- Non-linear dimensionality reduction methods may perform better projections of dataset, however, nowadays there are some inconveniences doing this, such as the increasing of the computational complexity, or an over-fitting of the generated features, which may increase the number of false positives.
- A density based clustering algorithm in a three-dimensional space, may improve the clustering process, considering that, a projection of data in a three-dimensional space may preserve a greater percentage of the original information.
- Other way to enhance the performance of the density based algorithm, could be, developing a strategy to clustering depending on local densities, some works about such kind of methodology have been explored before.
- It could be a good idea to implement a fuzzy classification algorithm, in a semi-supervised learning methodology, which would improve the quality of the classification, however the precision of the generated baseline may decrease.
- Considering that the presented methodology presents a relatively low computational complexity, it could be implemented on-line in a system, such as the UAV. It could woks into a damage detection scheme, discriminating damages or anomalies.
- More sophisticated machine learning techniques are being developed at present, they could work together with simple, precise and sometimes free access classifiers, to upgrade their capabilities. Such as in the case of the GA-DBSCAN algorithm.

BIBLIOGRAPHY

[1] G. G. Simpson, *Principles of animal taxonomy.* Columbia University Press, 1961, no. 20.

[2] A. Bueno, "Carl von linn. la pasin por la sistemtica," *Ars medica. Revista de humanidades*, pp. 199–214, 2007.

[3] S. Theodoridis and K. Koutroumbas, *Pattern Recognition Fourth Edition.* Elsevier, 2009.

[4] T. Mitchell, *Machine Learning.* McGraw-Hill, 1997.

[5] C. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2009.

[6] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on neural networks*, vol. 16, no. 3, pp. 645–678, 2005.

[7] R. Xu and D. C. Wunsch, *Clustering.* NJ: Wiley, 2009.

[8] B. Vandeginste, D. Massart, L. Buydens, P. Jong, S. De; Lewi, and J. Smeyers-Verbeke, *Data Handling in Science and Technology.* Elsevier Science, 1998.

[9] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," *Informatica*, 2007.

[10] C. R. Farrar and K. Worden, *Structural health monitoring: a machine learning perspective.* John Wiley & Sons, 2013.

[11] J. Tohka, "Sgn-2506: Introduction to pattern recognition," *Tampere University of Technology, Department of Signal Processing*, 2013.

[12] A. Webb, *Statistical Pattern Recognition, Second Edition.* Wiley, 2002.

[13] C. R. Farrar, S. W. Doebling, and D. A. Nix, "Vibration–based structural damage identification," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 359, no. 1778, pp. 131–149, 2001.

[14] H. Sohn, C. R. Farrar, F. M. Hemez, D. D. Shunk, D. W. Stinemates, B. R. Nadler, and J. J. Czarnecki, "A review of structural health monitoring literature: 1996–2001," *Los Alamos National Laboratory*, 2003.

[15] R. P. Rulli, C. G. G. Bueno, F. Dotta, and P. A. da Silva, "Damage detection systems for commercial aviation," in *Dynamics of Smart Systems and Structures*. Springer, 2016, pp. 329–342.

[16] J. S. Pérez, "Smart aeronautical structures: development and experimental validation of a structural health monitoring system for damage detection," Ph.D. dissertation, Aeronauticos, 2014. [Online]. Available: http://oa.upm.es/30438/

[17] M. Martinez-Luengo, A. Kolios, and L. Wang, "Structural health monitoring of offshore wind turbines: A review through the statistical pattern recognition paradigm," *Renewable and Sustainable Energy Reviews*, vol. 64, pp. 91–105, 2016.

[18] C. M. Bishop, "Novelty detection and neural network validation," *IEE Proceedings-Vision, Image and Signal processing*, vol. 141, no. 4, pp. 217–222, 1994.

[19] K. Worden, "Structural fault detection using a novelty measure," *Journal of Sound and vibration*, vol. 201, no. 1, pp. 85–101, 1997.

[20] M. Markou and S. Singh, "Novelty detection: a reviewpart 1: statistical approaches," *Signal processing*, vol. 83, no. 12, pp. 2481–2497, 2003.

[21] Markou and Singh, "Novelty detection: a reviewpart 2: neural network based approaches," *Signal processing*, vol. 83, no. 12, pp. 2499–2521, 2003.

[22] T. H. G. Megson, *Introduction to Aircraft Structural Analysis*. Butterworth-Heinemann, 2013.

[23] C. R. Farrar and K. Worden, "An introduction to structural health monitoring,"

Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, vol. 365, no. 1851, pp. 303–315, 2007.

[24] D. E. Bently and T. Hatch'Charles, "Fundamentals of rotating machinery diagnostics," *Mechanical Engineering-CIME*, vol. 125, no. 12, pp. 53–54, 2003.

[25] P. J. Shull, *Nondestructive evaluation: theory, techniques, and applications*. CRC press, 2016.

[26] D. C. Montgomery, *Introduction to statistical quality control*. John Wiley & Sons, 2007.

[27] C. R. Farrar and N. A. Lieven, "Damage prognosis: the future of structural health monitoring," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 365, no. 1851, pp. 623–632, 2007.

[28] I. K. Fodor, "A survey of dimension reduction techniques," Lawrence Livermore National Lab., CA (US), Tech. Rep., 2002.

[29] D. L. Donoho *et al.*, "High-dimensional data analysis: The curses and blessings of dimensionality," *AMS Math Challenges Lecture*, vol. 1, p. 32, 2000.

[30] J. P. Cunningham and Z. Ghahramani, "Linear dimensionality reduction: survey, insights, and generalizations." *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 2859–2900, 2015.

[31] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[32] D. N. Lawley and A. E. Maxwell, "Factor analysis as a statistical method," *Journal of the Royal Statistical Society. Series D (The Statistician)*, vol. 12, no. 3, pp. 209–229, 1962.

[33] J. H. Friedman and W. Stuetzle, "Projection pursuit regression," *Journal of the American statistical Association*, vol. 76, no. 376, pp. 817–823, 1981.

[34] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*. John

Wiley & Sons, 2004, vol. 46.

[35] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: applications to image and text data," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001, pp. 245–250.

[36] Y. Mori, M. Kuroda, and N. Makino, "Nonlinear principal component analysis," in *Nonlinear Principal Component Analysis and Its Applications*. Springer, 2016, pp. 7–20.

[37] T. Hastie and W. Stuetzle, "Principal curves," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 502–516, 1989.

[38] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, 1964.

[39] T. Kohonen, *Self-Organizing Maps Third Edition*. Springer, 2001.

[40] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[41] M. L. Raymer, W. F. Punch, E. D. Goodman, L. A. Kuhn, and A. K. Jain, "Dimensionality reduction using genetic algorithms," *IEEE transactions on evolutionary computation*, vol. 4, no. 2, pp. 164–171, 2000.

[42] W. F. Velicer and D. N. Jackson, "Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure," *Multivariate behavioral research*, vol. 25, no. 1, pp. 1–28, 1990.

[43] L. E. Mujica, J. Vehí, M. Ruiz, M. Verleysen, W. Staszewski, and K. Worden, "Multivariate statistics process control for dimensionality reduction in structural assessment," *Mechanical Systems and Signal Processing*, vol. 22, no. 1, pp. 155–171, 2008.

[44] Y. Ni, X. Zhou, and J. Ko, "Experimental investigation of seismic damage identification using pca-compressed frequency response functions and neural networks," *Journal of sound and vibration*, vol. 290, no. 1, pp. 242–263, 2006.

[45] L. Mujica, J. Rodellar, A. Fernandez, and A. Güemes, "Q-statistic and t2-statistic pca-based measures for damage assessment in structures," *Structural Health Monitoring*, vol. 10, no. 5, pp. 539–553, 2011.

[46] O. R. de Lautour and P. Omenzetter, "Damage classification and estimation in experimental structures using time series analysis and pattern recognition," *Mechanical Systems and Signal Processing*, vol. 24, no. 5, pp. 1556–1569, 2010.

[47] J. Sierra-Perez, M. A. Torres-Arredondo, G. Cabanes, A. Güeme, and L. E. Mujica, "Structural health monitoring by means of strain field pattern recognition on the basis of pca and automatic clustering techniques based on som** the research included in this document was partially supported by the," *IFAC-PapersOnLine*, vol. 48, no. 28, pp. 987–992, 2015.

[48] C. E. Katsikeros and G. Labeas, "Development and validation of a strain-based structural health monitoring system," *Mechanical Systems and Signal Processing*, vol. 23, no. 2, pp. 372–383, 2009.

[49] M. B. Rao, M. Bhat, C. Murthy, K. V. Madhav, and S. Asokan, "Structural health monitoring (shm) using strain gauges, pvdf film and fiber bragg grating (fbg) sensors: A comparative study," in *Proc. National Seminar on Non-Destructive Evaluation*, 2006, pp. 7–9.

[50] F. Magalhães, A. Cunha, and E. Caetano, "Vibration based structural health monitoring of an arch bridge: from automated oma to damage detection," *Mechanical Systems and Signal Processing*, vol. 28, pp. 212–228, 2012.

[51] A. Deraemaeker, A. Preumont, and J. Kullaa, "Modeling and removal of environmental effects for vibration based shm using spatial filtering and factor analysis," *Proceedings of IMAC XXIV*, 2006.

[52] J. Kullaa, "Vibration-based structural health monitoring under variable environmental or operational conditions," in *New trends in vibration based structural health monitoring*. Springer, 2010, pp. 107–181.

[53] A. Deraemaeker, E. Reynders, G. De Roeck, and J. Kullaa, "Vibration-based structural health monitoring using output-only measurements under changing

environment," *Mechanical systems and signal processing*, vol. 22, no. 1, pp. 34–56, 2008.

[54] N. Dervilis, M. Choi, S. Taylor, R. Barthorpe, G. Park, C. Farrar, and K. Worden, "On damage diagnosis for a wind turbine blade using pattern recognition," *Journal of sound and vibration*, vol. 333, no. 6, pp. 1833–1850, 2014.

[55] C. Wen, S. Hung, C. Huang, and J. Jan, "Unsupervised fuzzy neural networks for damage detection of structures," *Structural Control and Health Monitoring*, vol. 14, no. 1, pp. 144–161, 2007.

[56] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy k-nearest neighbor algorithm," *IEEE transactions on systems, man, and cybernetics*, no. 4, pp. 580–585, 1985.

[57] N. Zahid, O. Abouelala, M. Limouri, and A. Essaid, "Unsupervised fuzzy clustering," *Pattern Recognition Letters*, vol. 20, no. 2, pp. 123–129, 1999.

[58] P. Baraldi, F. Di Maio, M. Rigamonti, E. Zio, and R. Seraoui, "Clustering for unsupervised fault diagnosis in nuclear turbine shut down transients," *Mechanical Systems and Signal Processing*, vol. 58, pp. 160–178, 2015.

[59] D. E. Gustafson and W. C. Kessel, "Fuzzy clustering with a fuzzy covariance matrix," in *Decision and Control including the 17th Symposium on Adaptive Processes, 1978 IEEE Conference on*. IEEE, 1979, pp. 761–766.

[60] D. Dinh, E. Ramasso, V. Placet, L. Boubakar, and N. Zerhouni, "Application of an unsupervised pattern recognition approach for ae data originating from fatigue tests on cfrp," in *31st Conference of the European Working Group on Acoustic Emission (EWGAE)*, 2014.

[61] D. D. Doan, E. Ramasso, V. Placet, S. Zhang, L. Boubakar, and N. Zerhouni, "An unsupervised pattern recognition approach for ae data originating from fatigue tests on polymer–composite materials," *Mechanical Systems and Signal Processing*, vol. 64, pp. 465–478, 2015.

[62] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 1979.

[63] E. Ramasso, V. Placet, and M. L. Boubakar, "Unsupervised consensus clustering of acoustic emission time-series for robust damage sequence estimation in composites," *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 12, pp. 3297–3307, 2015.

[64] O. Avci and O. Abdeljaber, "Self-organizing maps for structural damage detection: a novel unsupervised vibration-based algorithm," *Journal of Performance of Constructed Facilities*, vol. 30, no. 3, p. 04015043, 2015.

[65] O. Abdeljaber, O. Avci, N. T. Do, M. Gul, O. Celik, and F. N. Catbas, "Quantification of structural damage with self-organizing maps," in *Structural Health Monitoring, Damage Detection & Mechatronics, Volume 7*. Springer, 2016, pp. 47–57.

[66] L. Chambers, *The Practical Handbook of Genetic Algorithms Aplications, Second Edition*. CHAPMAN & HALL/CRC, 2001.

[67] M. Silva, A. Santos, E. Figueiredo, R. Santos, C. Sales, and J. C. Costa, "A novel unsupervised approach based on a genetic algorithm for structural damage detection in bridges," *Engineering Applications of Artificial Intelligence*, vol. 52, pp. 168–180, 2016.

[68] M. Betti, L. Facchini, and P. Biagini, "Damage detection on a three-storey steel frame using artificial neural networks and genetic algorithms," *Meccanica*, vol. 50, no. 3, pp. 875–886, 2015.

[69] V. Meruane and W. Heylen, "An hybrid real genetic algorithm to detect structural damage using modal properties," *Mechanical Systems and Signal Processing*, vol. 25, no. 5, pp. 1559–1573, 2011.

[70] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[71] A. Diez, N. L. D. Khoa, M. M. Alamdari, Y. Wang, F. Chen, and P. Runcie, "A clustering approach for structural health monitoring on bridges," *Journal of Civil Structural Health Monitoring*, vol. 6, no. 3, pp. 429–445, 2016.

[72] J. P. Santos, C. Crémona, L. Calado, P. Silveira, and A. D. Orcesi, "On-line unsupervised detection of early damage," *Structural Control and Health Monitoring*, 2015.

[73] S. K. Al-Jumaili, K. M. Holford, M. J. Eaton, J. P. McCrory, M. R. Pearson, and R. Pullin, "Classification of acoustic emission data from buckling test of carbon fibre panel using unsupervised clustering techniques," *Structural Health Monitoring*, vol. 14, no. 3, pp. 241–251, 2015.

[74] H. Speckmann and R. Henrich, "Structural health monitoring (shm)–overview on technologies under development," in *Proc. of the World Conference on NDT, Montreal-Canada*, 2004.

[75] K. Worden, C. R. Farrar, G. Manson, and G. Park, "The fundamental axioms of structural health monitoring," in *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 463, no. 2082. The Royal Society, 2007, pp. 1639–1664.

[76] C. R. Farrar and H. Sohn, "Pattern recognition for structural health monitoring," in *Workshop on Mitigation of Earthquake Disaster by Advanced Technologies*, 2000.

[77] J. Sierra-Pérez, M. A. Torres-Arredondo, G. Cabanes, A. Güemes, L. E. Mujica, and C.-P. Fritzen, "Damage detection in metallic beams from dynamic strain measurements under different load cases by using automatic clustering and pattern recognition techniques," in *EWSHM-7th European workshop on structural health monitoring*, 2014.

[78] A. Rytter, "Vibrational based inspection of civil engineering structures," Ph.D. dissertation, Dept. of Building Technology and Structural Engineering, Aalborg University, 1993.

[79] A. D. Kersey, M. A. Davis, H. J. Patrick, M. LeBlanc, K. Koo, C. Askins, M. Putnam, and E. J. Friebele, "Fiber grating sensors," *Journal of lightwave technology*, vol. 15, no. 8, pp. 1442–1463, 1997.

[80] J. Sierra-Pérez, A. Güemes, and L. E. Mujica, "Damage detection by using fbgs

and strain field pattern recognition techniques," *Smart materials and structures*, vol. 22, no. 2, p. 025011, 2012.

[81] M. Kreuzer, "Strain measurement with fiber bragg grating sensors," *HBM, Darmstadt, S2338-1.0 e*, 2006.

[82] W. J. Staszewski and K. Worden, *Signal Processing for Damage Detection*. John Wiley & Sons, Ltd, 2009, ch. 21. [Online]. Available: http://dx.doi.org/10.1002/9780470061626.shm042

[83] W. Härdle and L. Simar, *Applied multivariate statistical analysis*. Springer Science & Business Media, 2007.

[84] I. T. Jolliffe, "Principal component analysis and factor analysis," *Principal component analysis*, pp. 150–166, 2002.

[85] S. S. Sharma, *Applied multivariate techiques*. John Wiley & Sons,, 1996.

[86] H. F. Kaiser, "The varimax criterion for analytic rotation in factor analysis," *Psychometrika*, vol. 23, no. 3, pp. 187–200, 1958.

[87] J. O. Neuhaus and C. Wrigley, "The quartimax method," *British Journal of Mathematical and Statistical Psychology*, vol. 7, no. 2, pp. 81–91, 1954.

[88] A. E. Hendrickson and P. O. White, "Promax: A quick method for rotation to oblique simple structure," *British Journal of Mathematical and Statistical Psychology*, vol. 17, no. 1, pp. 65–70, 1964.

[89] H. F. Kaiser, "The application of electronic computers to factor analysis," *Educational and psychological measurement*, vol. 20, no. 1, pp. 141–151, 1960.

[90] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[91] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining. 1st*. Boston: Pearson Addison Wesley. xxi, 2005.

[92] M. N. Gaonkar and K. Sawant, "Autoepsdbscan: Dbscan with eps automatic

for large dataset," *International Journal on Advanced Computer Theory and Engineering*, vol. 2, no. 2, pp. 11–16, 2013.

[93] K. M. Kumar and A. R. M. Reddy, "A fast dbscan clustering algorithm by accelerating neighbor searching using groups method," *Pattern Recognition*, vol. 58, pp. 39–48, 2016.

[94] M. A. Patwary, D. Palsetia, A. Agrawal, W.-k. Liao, F. Manne, and A. Choudhary, "A new scalable parallel dbscan algorithm using the disjoint-set data structure," in *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*. IEEE Computer Society Press, 2012, p. 62.

[95] Z. Xiong, R. Chen, Y. Zhang, and X. Zhang, "Multi-density dbscan algorithm based on density levels partitioning," *JOURNAL OF INFORMATION &COMPUTATIONAL SCIENCE*, vol. 9, no. 10, pp. 2739–2749, 2012.

[96] H. Zhou, P. Wang, and H. Li, "Research on adaptive parameters determination in dbscan algorithm," *JOURNAL OF INFORMATION &COMPUTATIONAL SCIENCE*, vol. 9, no. 7, pp. 1967–1973, 2012.

[97] D. Birant and A. Kut, "St-dbscan: An algorithm for clustering spatial–temporal data," *Data & Knowledge Engineering*, vol. 60, no. 1, pp. 208–221, 2007.

[98] J. Hou, H. Gao, and X. Li, "Dsets-dbscan: A parameter-free clustering algorithm," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3182–3193, 2016.

[99] C. Ruiz, M. Spiliopoulou, and E. Menasalvas, "C-dbscan: Density-based clustering with constraints," in *International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing*. Springer, 2007, pp. 216–223.

[100] A. Karami and R. Johansson, "Choosing dbscan parameters automatically using differential evolution," *International Journal of Computer Applications*, vol. 91, no. 7, 2014.

[101] P. Sharma and Y. Rathi, "Efficient density-based clustering using automatic parameter detection," in *Proceedings of the International Congress on Information*

and Communication Technology. Springer, 2016, pp. 433–441.

[102] G. Chaudhari Chaitali, "Optimizing clustering technique based on partitioning dbscan and ant clustering algorithm," *International Journal of Engineering and Advanced Technology (IJEAT) ISSN*, pp. 2249–8958, 2012.

[103] R. L. Haupt and S. E. Haupt, *Practical genetic algorithms*. John Wiley & Sons, 2004.

[104] M. Srinivas and L. M. Patnaik, "Genetic algorithms: A survey," *computer*, vol. 27, no. 6, pp. 17–26, 1994.

[105] C.-Y. Lin, C.-C. Chang, and C.-C. Lin, "A new density-based scheme for clustering based on genetic algorithm," *Fundamenta Informaticae*, vol. 68, no. 4, pp. 315–331, 2005.

[106] M. Matsumoto and T. Nishimura, "Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator," *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, vol. 8, no. 1, pp. 3–30, 1998.

[107] P. F. et al, "Clustering basic benchmark," 2015, accessed 2017-09-01. [Online]. Available: http://cs.uef.fi/sipu/datasets/

[108] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[109] A. Gionis, H. Mannila, and P. Tsaparas, "Clustering aggregation," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, p. 4, 2007.

[110] P. Fränti, O. Virmajoki, and V. Hautamäki, "Fast agglomerative clustering using a k-nearest neighbor graph," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1875–1881, 2006.

[111] L. Fu and E. Medico, "Flame, a novel fuzzy clustering method for the analysis of dna microarray data," *BMC bioinformatics*, vol. 8, no. 1, p. 3, 2007.

[112] C. J. Veenman, M. J. T. Reinders, and E. Backer, "A maximum variance cluster algorithm," *IEEE Transactions on pattern analysis and machine intelligence*,

vol. 24, no. 9, pp. 1273–1280, 2002.

[113] A. K. Jain and M. H. Law, "Data clustering: A users dilemma," in *International conference on pattern recognition and machine intelligence*. Springer, 2005, pp. 1–10.

[114] A. Carvajal-Castrillón, J. Alvarez-Montoya, J. Niño-Navia, L. Betancur-Agudelo, F. Amaya-Fernández, and J. Sierra-Pérez, "Structural health monitoring on an unmanned aerial vehicle wing's beam based on fiber bragg gratings and pattern recognition techniques," *Procedia Structural Integrity*, vol. 5, pp. 729–736, 2017.

[115] C. Matthews, *Aeronautical engineer's data book*. Elsevier, 2001.