
**University Institute of Lisbon**

Department of Information Science and Technology


# Big Data Analytics Applied to Sensor Data of Engineering Structures: Predictive Methods


Filipe Galvão Chambel Caçador


A Dissertation presented in partial fulfillment of the Requirements for the Degree of

**Master in Computer Science**


**Supervisor**

José Eduardo de Mendonça Tomás Barateiro, PhD

LNEC

**Co-Supervisor**

Elsa Alexandra Cabral da Rocha Cardoso, Assistant Professor, PhD

ISCTE-IUL


October 2017

# *Resumo*

Modelos preditivos são instrumentos fundamentais para a análise da segurança de barragens. São importantes para obter conclusões acerca da segurança estrutural destas. Os dados utilizados nos modelos preditivos, são obtidos através de sensores que se encontram embutidos nas estruturas. Apesar dos algoritmos preditivos serem ferramentas poderosas para a análise e previsão, outras técnicas de Machine Learning e modelos estatísticos, como as redes neuronais, têm sido desenvolvidas e utilizadas nestas áreas ao longo dos anos. Devido às diferentes formas que a monitorização destas estruturas é feita, o foco está em melhorar os métodos existentes, através de uma análise comparativa. Este trabalho tem como finalidade o desenvolvimento de uma metodologia que compare os diferentes algoritmos preditivos, como a Multiple Linear Regression, a Ridge Regression, a Principal Component Regression e as Redes Neuronais, bem como a aplicação de diferentes técnicas de separação de dados. Esta metodologia será aplicada a um caso de estudo, com a finalidade de determinar qual ou quais as combinações de variáveis que obtêm o melhor desempenho na previsão do seu comportamento.

**Palavras-chave:** Análise Preditiva, Aprendizagem Automática, Data Mining, Análise de Big Data, Análise Estatística, Monitorização de Barragens.

# *Abstract*

Predictive models are fundamental instruments for providing dam safety analysis. They are important tools to retrieve conclusions about the structural safety of these dams. The data for these predictive models is gathered through sensors embedded within these structures. Even though predictive models are powerful tools for analysis and prediction, other machine learning and statistical models, like neural networks, have been developed over the years. Due to the many ways dam safety analyses is performed, the focus is to improve the existing methods by comparing them with each other. This work is focused on developing the methodology that compares different predictive models, like the Multiple Linear Regression Model, the Ridge Regression Model, the Principal Component Regression Model and Neural Networks, as well as comparing different re-sampling techniques for separating the data. This methodology is applied to a case study, with the purpose of finding which combinations of input variables provide the highest accuracy for predicting the behavior of these structures.

**Keywords:** Predictive Analytics, Machine Learning, Data Mining, Big Data, Statistical Analysis, Dam Monitoring.

# *Acknowledgements*

The path to the completion of this dissertation proved not to be as easy as it was thought in the beginning. Its completion is thanks, in great part, to a group of special people who challenged and stuck with me through its development.

I would like to express my deepest appreciation to my Supervisor, Professor José Barateiro whose enthusiasm in this area of studies and feedback provided great insights to the development of this dissertation.

I would also like to express my gratitude to my Co-Supervisor, Professor Elsa Cardoso, whose support and thoughtful feedback provided the development of a more thorough dissertation.

A thank you to my good friend, António Antunes, with whom I shared my college experience with and throughout the development of this dissertation.

To my family a big thank you as well, for bearing with me through the difficult times, late nights and absences while writing this dissertation.

I am also greatfull for the institution LNEC for providing me with the data sources thus allowing me to have a more realistic approach on the development.

A great thank you to the institution ISCTE-IUL in which i took my bachelors and masters, by giving me the tools to further my education as well my personality.

To all a great Thank You.

# Contents

**Bibliography** **75**

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| **LNEC** | **L**aboratório **N**acional de **E**ngenharia **C**ivil |
| **DSR** | **D**esign **S**cience **R**esearch |
| **DSRM** | **D**esign **S**cience **R**esearch **M**ethodology |
| **BI** | **B**usiness **I**ntelligence |
| **IoT** | **I**nternet of **T**hings |
| **DM** | **D**ata **M**ining |
| **ML** | **M**achine **L**earning |
| **KDD** | **K**nowledge **D**iscovery **P**rocess |
| **CRISP-DM** | **C**ross **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining |
| **DM-LC** | **D**ata **M**ining **L**ife **C**ycle |
| **SL** | **S**tatistical **L**earning |
| **PCA** | **P**rincipal **C**omponent **A**nalysis |
| **SVD** | **S**ingular **V**alue **D**ecomposition |
| **PC** | **P**rincipal **C**omponent |
| **LR** | **L**inear **R**egression |
| **MLR** | **M**ultiple **L**inear **R**egression |
| **LS** | **L**east **S**quares |
| **RSS** | **R**esidual **S**um of **S**quares |
| **RR** | **R**idge **R**egression |
| **PCR** | **P**rincipal **C**omponent **R**egression |
| **NN** | **N**eural **N**etwork |
| **MLP** | **M**ulti-**L**ayer **P**erceptron |
| **R** | **C**orrelation **C**oefficient |
| **R2** | **C**oefficient of **D**etermination |
| **R2Adj** | **C**oefficient of **D**etermination **A**djusted |
| **MSE** | **M**ean **S**quared **E**rror |
| **RMSE** | **R**oot **M**ean **S**quared **E**rror |
| **MAE** | **M**ean **A**bsolute **E**rror |
| **HO** | **H**old-**O**ut |

| | |
|---|---|
| **KFCV** | **K**-**F**old **C**ross-**V**alidation |
| **LOOCV** | **L**eave-**O**ne-**O**ut **C**ross-**V**alidation |
| **SWCV** | **S**liding **W**indow **C**ross-**V**alidation |
| **SHM** | **S**tructural **H**ealth **M**onitoring |

# Chapter 1

# Introduction

Engineering structures like bridges, dams and buildings, have become indispensable instruments for human society. These structures ensure and provide a diverse range of benefits, from an economic, social or environmental point of view. Once these structures are built and constantly used they start aging and begin to deteriorate. Due to the constant usage and environmental effects suffered by these structures and the growing vulnerability associated with their aging, there has been an increasing need to assess, manage and monitor the risks associated with them, as well as to provide constant improvements to their safety throughout their entire lifespan, meaning that their structural integrity and maintainability must be guaranteed in order to prevent possible catastrophic events that may occur, either economic, environmental or humanitarian.

With these problems in mind, the goal of structural health monitoring (SHM) of engineering structures consists in determining with high accuracy the location and severity of damages on the structures as soon as they happen. The methods that are currently being used for structural health monitoring can only determine whether there is an existing damage within the structures but not the entirety extent of these damages (Chang, Flatau, & Liu, 2003).

To monitor, assess and evaluate these structures several factors must be taken into consideration to make sure that these structures are functioning as intended and to provide a way of detecting any abnormal behavior that could endanger their safety and the safety of the surrounding areas. These factors are gathered either through manually inspections by the engineering teams or specialists from these areas, or automatically by instrumentation within and around these structures,

mainly through the use of sensors, but it can also be generated based on knowledge from engineering experts. Figure 1.1 exemplifies these situations where the Experimenter can be identified as the visual inspector and the Embedded sensor as the automatic instrument generating the necessary information, both monitoring a structure that is interacting with its surrounding environment.

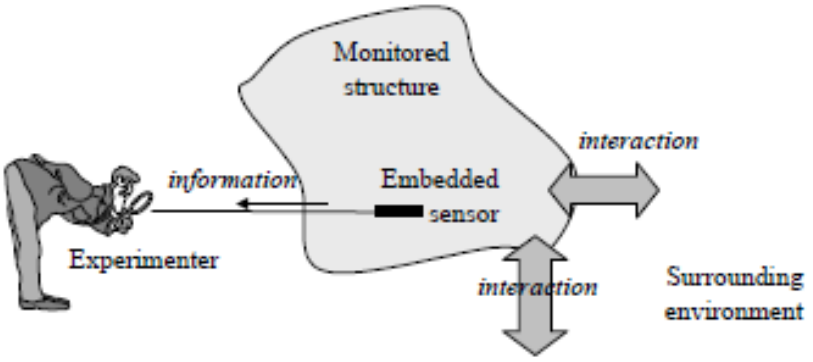


FIGURE 1.1: Structural Health Monitoring, manual inspection and automatic data retrieval (Balageas et al., 2010)

The most important factors that allow the monitoring of these structures are, when applied, the level and temperature of the water, air movement and temperature, the weight and movement of external elements on the structure, structural shifts, the age of the structure, among others. Despite the existence of instruments capable of obtaining most of these factors automatically, there is still the need for visual intervention and inspection to determine the possibility of existing undetected damages or deterioration. The task of getting reliable measurements from the instruments located within the structure is not easy, due to their placement often in hostile environments, where they are not easily reachable or the conditions inside the structure are not favorable for mechanical or electronic tools, like humidity for instance, which can, in time, cause these instruments to malfunction and provide unreliable information.

The emerging ability to acquire data from several different sources creates a new and different paradigm where science is now able to generate knowledge from pattern detection, correlations or dependencies from sources with different properties or representations. In the case of sensors, a great level of potential can be gathered from the data where they represent real events, and can easily assume a great volume of information, leading to a Big Data scenery where this generated information can aid analysts adding more value to businesses and find hidden knowledge that was not previously identified. It is very important to correctly

analyze the data to further rely on the engineering structures either when they are functional and in use or abandoned, and to be able to assess their ability to withstand unlikely events like earthquakes or other environmental causes.

Engineering structures like dams, which will be the focusing structure covered throughout the dissertation, are artificial reservoirs that are able to hold large amounts of water, ensuring a diverse range of benefits, either from an economic or from a social point of view, where their roles are to prevent floods, generate and provide hydroelectric power, to reclaim land that otherwise would be submerged and to provide water supply to several human activities, either for consumption or industrial use. Water, especially fresh water, is relatively scarce and needs to be preserved and so, it is imperative to ensure the successful monitoring of these structures. Thus, this dissertation proposes an approach for monitoring and evaluating dam response behavior based on the analyses of predictive, statistical and machine learning modeling techniques for the assessment of the structural engineering safety and maintainability of this type of structures.

The research methodology followed throughout this dissertation is the Design Science Research Methodology (DSRM) for developing and evaluating the successfulness of the artifacts to solve the identified research problems (Von Alan, March, Park, & Ram, 2004); (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007).

## 1.1 Research Methodology

The Design Science Research Methodology (DSRM) focuses on the importance of creating, developing and evaluating different artifacts to meet and solve the proposed and relevant objectives and problems. According to (Von Alan et al., 2004), Design Science Research (DSR) is a problem-solving process which means that its main objective is the acquisition of knowledge and understandability of the problems and their respective solutions to allow for the development and application of these created artifacts. And so, the authors propose seven DSR guidelines to be followed in order to prove the successfulness of each of the artifacts. The guidelines are presented as follows, including a very brief explanation about each of them:

1. **Design as an Artifact**: DSR must produce successful and viable artifacts that can either be defined as a construct, model, method or instantiation artifacts;

2. **Problem Relevance**: The objective of DSR is to develop solutions capable of solving either technological or relevant business problems;
3. **Design Evaluation**: The utility, quality and efficacy of each of the created artifact must be demonstrated though proper evaluation methods;
4. **Research Contributions**: Effective DSR must provide contributions that are able to be verified in the areas of focus of each of the artifacts;
5. **Research Rigor**: DSR relies on the application of rigorous methods for developing, demonstrating and evaluating the artifacts;
6. **Design as a Search Process**: An effective artifact requires using all available means to reach a desired end under the scope of the environment of the problem;
7. **Communication of Research**: DSR must be presented to both, technological and management oriented audiences.

To (Peffers et al., 2007), the DSRM revolves around six main steps: (1) problem identification and motivation, (2) definition of the objectives, (3) the design and development of a proposal for the solution, (4) demonstration of the use of the developed proposal, (5) evaluation of the proposed artifacts and their results and (6) communication. Depending on the type of investigation, its entry point can vary depending on the problem at hand. In the case of this dissertation, the entry point is a Problem-Centered Initiation because in this case the objectives, as suggested in both Figure 1.2 and Figure 1.3, can only be inferred from first defining the problem and its motivation (Section 1.2).

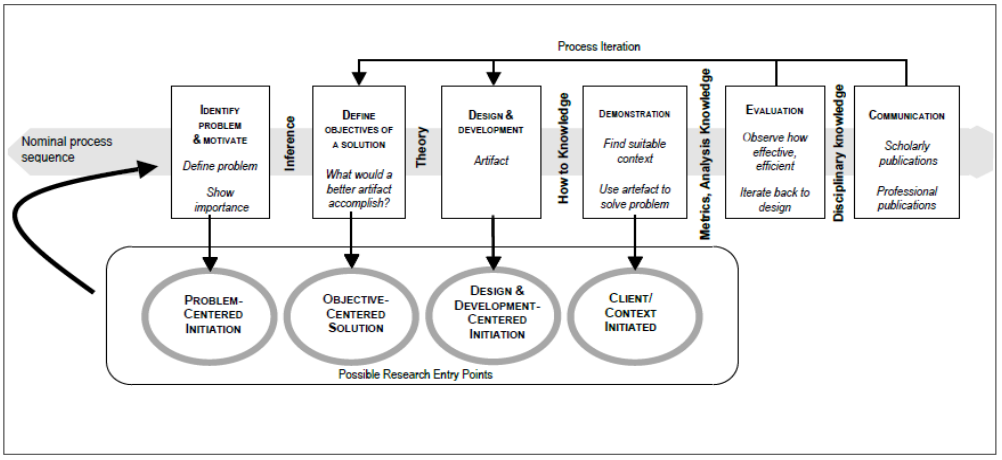


FIGURE 1.2: Design Science Research Methodology (extracted from (Peffers et al., 2007))

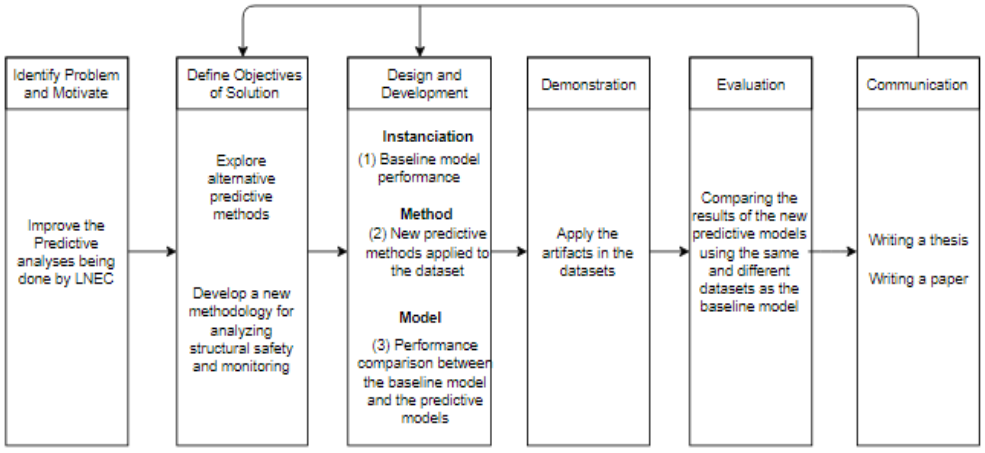


FIGURE 1.3: Adaptation Design Science Research Methodology (extracted from (Peffers et al., 2007))

The artifacts that have been developed in this dissertation are: an instantiation artifact which focuses on analyzing the currently used (Baseline) methods; a method artifact for the identification of other predictive models for monitoring the behavioral responses of the structures; and a model artifact for the comparison of the predictive algorithms identified, against the baseline as well as against each other. The combination of these artifacts and the possibility for some improvements from these models are supposed to lead to advancements on the current body of knowledge for the identified problem.

The artifacts demonstration is done through their application on a case study with data from a real dam in Portugal which has been provided by the GestBarragens software developed and currently being used by Laboratório Nacional de Engenharia Civil (hereafter noted as LNEC). The focus of this demonstration is to identify the best combinations of input variables for the models and the response variables themselves, with the goal of obtaining a generic approach to the problem as well as obtaining a higher accuracy in predicting the behavioral structural responses. The evaluation of the artifacts is done using a framework proposed by (Von Alan et al., 2004) and using the identified model and combination of variables that provide the highest accuracy from the same source of datasets, to prove the validity of each of the artifacts.

## 1.2 Problem Identification and Motivation

This section of the dissertation corresponds to the Problem Identification and Motivation step of the DSRM, which purpose is to define and justify the research problem and what value will be given by applying the proposed solution.

The main problem, shared by several authors in the area, is the limited monitoring and representation currently being applied to the engineering structures that, in time, provide for a loss of information that could endanger the structures culminating in disaster. The application of models to predict the behavior of structural responses has become a standard for SHM. But for the most part, the input variables are chosen without considering the best combination for each of the behavioral responses. On one hand, it is expected that area specialists (i.e. civil engineers) know which variables to choose from to positively affect the response variables, thus generating high accuracy models. But on the other hand, there are other unexpected factors that could impact the structures response, such as the existence of patterns not yet discovered that could prove to be beneficial, thus increasing the predictions, and that could also provide higher insights on how the safety of these structures should be monitored and maintained.

The motivation of this dissertation derives from the necessity of improving the monitoring of the behavioral structural responses, using different predictive model techniques and input variables combinations as well as other differentiating factors that could allow for beneficial improvements to the models. The success of the proposed research could provide a significant positive impact on the analysis and monitoring of generic engineering structures and their behavioral structure response. The case study used to exemplify the application of the different predictive model techniques is a real dam in Portugal with automatic and manual acquisition sensors.

## 1.3 Contributions of the Solution

Considering the definition of the problem and the motivation behind it, as well as the related work, which is comprised by the knowledge of what has been accomplished in the past, the objectives can then be inferred. This dissertation focuses on one main objective: explore alternative predictive and statistical methods as well as machine learning algorithms for structural behavior monitoring and

safety to improve results and their interpretation by the analysts, through comparison between them and those that are currently being used. The application of these techniques has the potential to improve the predictive capabilities of engineering structures and structural problem monitoring and detection, facilitating interventions on these structures. Furthermore, to determine if the objectives of the solution have been correctly defined and successfully resolved, three research questions have been created:

1. Can there be a better alternate method and combination of input variables for improving the predictive accuracy of each of the different structural behavior responses of dams?
2. Can the representation of results be improved to provide new insights and help decision-makers improve their business decisions?
3. Can the application of the methodology developed for demonstrating the created artifacts be applied to other generic engineering structures, and not only for the application on dams?

Throughout the dissertation, several artifacts have been created, focused on resolving the previously identified problems (refer to section 1.2), that are able to improve the body of knowledge already possessed as well as use this knowledge in new ways. And so, from the following artifacts, contributions provided from this dissertation can also be inferred:

- An **instantiation** artifact to determine the baseline model performance from the techniques currently in place that, according to (Mata, 2011), are considered to be good practices for structural behavior prediction and safety on dams. This baseline model performance will then be compared to the other predictive models to determine what models, combination of input variables and necessary parameters, provide a higher accuracy of the results, considering what is currently being done.
- A **method** artifact for applying other predictive statistical models or machine learning algorithms to the datasets and generate new models to improve the basis knowledge and to allow for further comparison of their performance against the baseline model.
- A **model** artifact for comparing the performance of the resulting predictive models with the performance of the baseline model performance.

The demonstration of how the developed artifacts are going to solve the identified problems is done through the development of a methodology (Chapter 3) applied to a real case study (Chapter 3.1) in which, example datasets, each referring to one of the five behavioral response variables considered, are going to be used for determining the performance results of each of the models that are then compared to the performance of the baseline model.

And finally, to evaluate and validate the success as well as the performance of the artifacts, taking into account the identified problems, the model artifact is then applied to different datasets to demonstrate the generalization of the artifact as well as to determine its efficiency and effectiveness.

## 1.4 Document Structure

The remainder of this dissertation is structured as follows:

- **Chapter 2 - Related Work**: In this chapter, it is provided an overview of the existent literature in the area of this research, as well as the related work on predicting the behavior of engineering structures, with a focus on Dams, as introduced in Chapter 1;
- **Chapter 3 - Design and Development**: In this chapter, the objectives of the solution are identified through the development of the artifacts applied on the Demonstration phase (Chapter 4) as well as the case study (Section 3.1);
- **Chapter 4 - Demonstration and Evaluation**: In this chapter, the proposed solution is applied to the datasets and the results are demonstrated and evaluated in order to determine their efficiency and validity on this research;
- **Chapter 5 - Conclusions**: In this chapter, it is concluded the dissertation, the research questions proposed on Chapter 1 are answered and is defined the future work to this research.

# Chapter 2

# Related Work

This chapter of the dissertation covers the theoretical background and work related to the problem and motivation that have been previously identified (Section 1.2), which allows for the definition of the objectives of the solution of the DSRM. The knowledge of the body of work contained throughout this chapter will serve as input for the advancement of the following steps of the research, as expressed in Figure 2.1.

- **Section 2.1**: In this section it is introduced the Big Data Analytics paradigm to motivate the capabilities of Big Data and the value that it gives to businesses;
- **Section 2.2**: In this section it is introduced the notion of Data Mining, in order to introduce the accepted methodologies for creating a valuable Data Mining Project and the different data preparation techniques that go along with it;
- **Section 2.3**: In this section it is introduced the concept of Predictive Modeling in which are presented the different predictive models used throughout the Demonstration and Evaluation phase;
- **Section 2.4**: In this section there will be presented the concept of Models Evaluation where are explained the different evaluation metrics and resampling methods used to evaluate the different predictive models;
- **Section 2.5**: In this section it is presented the current Predictive Modeling techniques for Dam Behavior where the existing work, related to dam behavior monitoring, is shown.

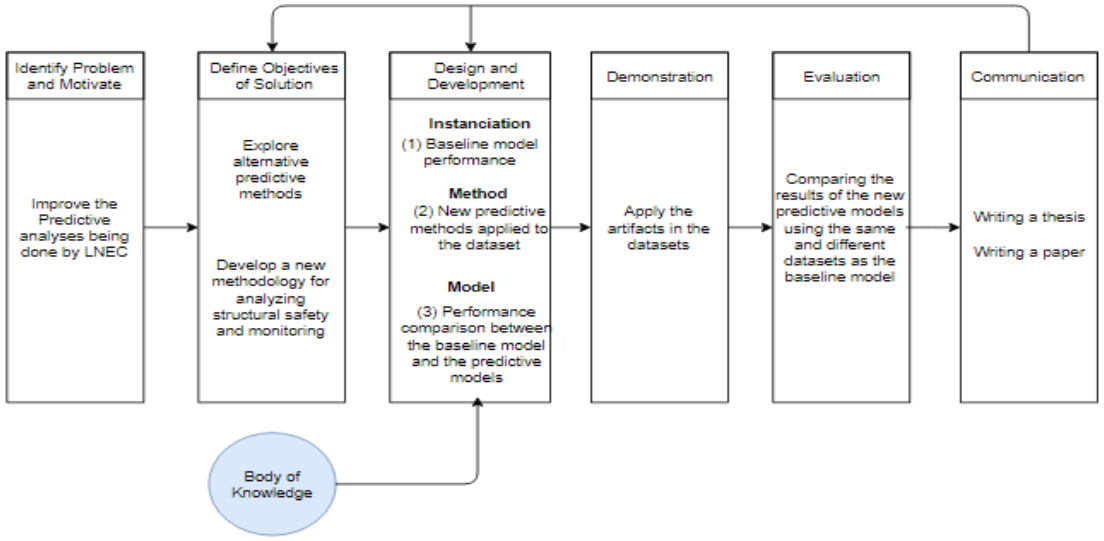


FIGURE 2.1: Adaptation Design Science Research Methodology - Related Work (extracted from (Peffers et al., 2007))

## 2.1 Big Data Analytics

Data has been growing at an exponential rate, due to constant technological advances, which in time, created the need to improve the ability to store, access, manipulate and manage data, giving the possibility for the emergence of Business Intelligence systems. Business Intelligence (BI) was first introduced in 1958 by an IBM researcher (Luhn, 1958) and has been around for decades where its definition has been reviewed and improved multiple times, focusing on changing how organizations should implement their strategies and improve their decisions. The purpose of BI is to provide insightful knowledge and find useful information to provide with the decision makers the means of detailed and summarized data through use reporting tools and dashboards (Elena et al., 2011). The reporting and analyzing requirements associated with these systems tend to maintain a similar growth rate as technology itself to allow for the most valuable and on time information as possible (Nedelcu et al., 2013). (Larson, 2012) and (Kimball & Ross, 2011) agree that BI is comprised of three fundamental stages:

- Data gathering and manipulation, retrieved from different sources and incorporated into one big repository, usually incorporated into a Data Warehouse system;

- Data analyses by use of several Data Mining techniques, Machine Learning algorithms and/or Statistical models;
- Data representation techniques, like dashboards, through access and manipulation of the data as tools for decision makers in their decision making processes.

The rising amounts of data being generated through a diverse range of industries lead to a new and interesting paradigm, which is for the moment, identified as Internet of Things (IoT), which main purpose is to enhance the potential of the data. According to (Atzori, Iera, & Morabito, 2010) data can be generated in three ways: from the Internet, from sensors or from extracted knowledge. This new data potential, quality and growing quantity allows businesses to improve their systems and by extension, their decisions, in order to gain competitive advantage over others. These amounts of data provides businesses with the realization of the importance of Big Data Analytics to support their strategies (Ularu, Puican, Apostu, Velicanu, et al., 2012); (Huisman, 2015).

Big Data Analytics adds new challenges and opportunities to BI with a similar definition, being the main difference the fact that it is used to find and retrieve value from Big Data instead of the "normal" data businesses are used to. (Gandomi & Haider, 2015) points out that Big Data is worthless unless when used to drive the process of decision making. This assertion tells us that businesses should be more focused in applying Big Data Analytics on their data to improve their decision-making process. But on the other hand it is also true that most companies have huge amounts of data but these amounts can not yet be considered as Big Data.

(Assunção, Calheiros, Bianchi, Netto, & Buyya, 2015) describes BD as being a “multi-V model” where each “V” characterizes its main aspects:

- **Variety**, refers to the different types of data being generated and can now be used. Throughout the years the generated data has been focused on structured data that its currently being used in traditional databases, but now, the biggest part of the data that is being generated is unstructured.
- **Volume**, which refers to the vast amounts of data that is being generated every second;
- **Velocity**, which refers to the speed rate that new data is being generated and moving around;

- **Veracity**, which refers to the quality and control of the volume of data being generated
- **Value**, which refers to the value that the study of this unstructured data, provides to the growth of businesses and their decision-making processes.

To retrieve knowledge from BD, (Sun, Zou, & Strang, 2015) and (Huisman, 2015) agree that BDA is comprised of three main components: Descriptive Analytics, which is often described as a summarization of historic data into knowledge and meaningful information through the discovery of existing relationships within the BD (Huisman, 2015); (Sun et al., 2015). It focuses on answering questions like "What and when did it happen?"; Predictive Analytics that according to (Buytendijk & Trepanier, 2010) is the use of several statistical, forecasting and data mining techniques to predict future events based on the descriptive data and focuses on questions like "What will happen?"; and finally, Prescriptive Analytics which tries to explain why something has happened.

## 2.2 Data Mining

Machines have become powerful instruments for providing the industries with the ability to automate processes that would be too time consuming if being done manually, even though, it is common that machines are sometimes unable to do simple tasks that humans are able to do with great ease. The Data Mining (DM) concept became relevant due to the growing availability of data, from IoT devices for instance, and the need to generate knowledge and information from these data. (Jain & Srivastava, 2013) and (Padhy, Mishra, Panigrahi, et al., 2012) define DM as a way of mining knowledge and improve decisions through processes of Machine Learning (ML) by exploring and analyzing large amounts of data or BD. The authors describe DM as a Knowledge Discovery Process (KDD) or as a way of extracting hidden information to predict trends and behaviors to gain competitive advantage. In contrast, (Fayyad, Piatetsky-Shapiro, & Smyth, 1996) refers to the KDD process as being composed of several tasks to extract knowledge where one of those tasks is DM where he considers it as the process of retrieving important and relevant information, like patterns, anomalies or any alterations made to the dataset. According to (Lei-da Chen, Frolick, et al., 2000) DM is used depending on the needs of the organization thus generating different types of information to find meaningful relationships between the data and to predict trends and patterns.

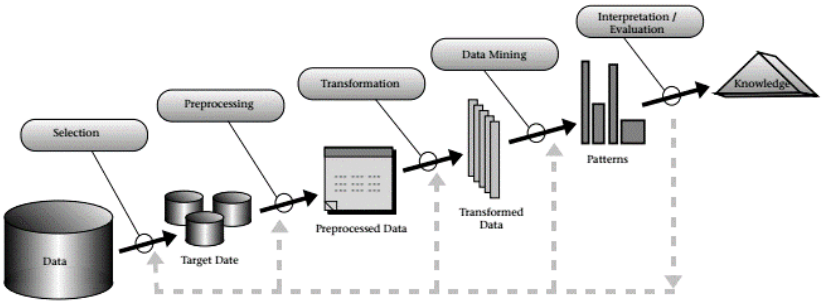


FIGURE 2.2: KDD Process (extracted from (Fayyad et al., 1996))

DM projects according to (Marbán, Mariscal, & Segovia, 2009) follow the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology. This methodology, also described as the Data Mining Life Cycle (DM-LC), demonstrated in Figure 2.3, has a comprehensibly flexible sequence of phases due to the possibility to go back to each of the previous steps to improve and modify the reasoning or the variables being used.

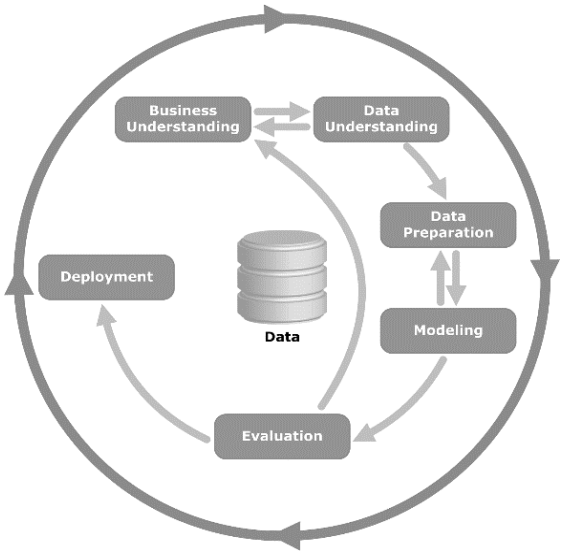


FIGURE 2.3: CRISP-DM Methodology

The Business Understanding phase or Problem Identification phase is crucial when developing DM projects, because it identifies the objectives and requirements of the business and in almost every case they are the success criteria for a good DM project. Data Understanding phase includes the initial insight to the data to describe and explore it to form ideas and to find ways of retrieving hidden information. According to (Zhang, Zhang, & Yang, 2003) the Data Preparation phase takes approximately 80% of the total project and covers the steps of creating

quality data to construct the dataset that will be used as input for the modeling phase, including data selection and data cleaning. The Modeling phase is where the different modeling techniques are selected and the model is built. The Evaluation phase is where the results produced by the models will be evaluated to see if they can achieve the business requirements and objectives that were previously identified and where the knowledge is created. The Deployment phase is the end phase of the project where the knowledge gained from the Evaluation phase is deployed.

According to (Lei-da Chen et al., 2000) DM methods are divided into two main learning groups, Statistical Learning (SL) and ML. SL is defined as a tool to build statistical models to predict outcomes by having an underlying probability model and combining different fields of computer science like Statistics, Artificial Intelligence (AI) and DM (James, Witten, Hastie, & Tibshirani, 2013). ML, according to (Mohri, Rostamizadeh, & Talwalkar, 2012), is defined as the use of efficiently designed computational methods or algorithms that are improved using experience or training, improving their performance and provide more accurate predictions.(Deshpande & Thakare, 2010) on the other hand, refers that DM should be separated into two categories, Descriptive Models and Predictive Models. This definition differs from what (Lei-da Chen et al., 2000) described in the sense that SL and ML include both Descriptive and Predictive modeling. Predictive modeling problems consist in obtaining knowledge from analysis on past experiences while Descriptive modeling problems consist in analyzing the evolution of a given dataset to increase its knowledge.

### 2.2.1 Data Preparation and Cleaning

Data Preparation is a very important part of any Data Mining project and one of the most time consuming, including several tasks. The main goal is to generate quality data from existing raw data. When generated automatically, this data is usually "dirty" or in other words, inconsistent, noisy or with missing values.

In the Data Preparation process, if applied, there is also the need for dimensionality reduction. By reducing irrelevant or redundant features or even instances of the data, the efficiency, speed and accuracy of the next DM processes can be significantly improved.

By cleaning the data and selecting the features that will be used it is also possible, by combining other, different features, to find and add hidden and undetected features (Zhang et al., 2003).

#### 2.2.1.1 Incorrect and Missing Values

Most of the predictive models make the assumption that all the input data values are present and correct, hence the need to previously identify and revise incorrect and missing values. If incorrect values reach the algorithms, as to overcome them, they are either rendered insignificant or overrated which in both cases, causes a negative impact on the model response. If incorrect values cannot be interpreted by the algorithm then they are treated as a missing values and if they do not appear that frequently on the dataset than they may be considered as irrelevant (Abbott, 2014).

Missing values cause a negative impact on the accuracy of the models and are hard to deal with and so, (Grzymala-Busse & Hu, 2001) and (Abbott, 2014), consider several approaches to mitigate them:

- Replace the missing value with a new value: By using this approach, the missing values can take on the value of either the most common attribute value, a special value (-1, for instance) or the form of a mathematical arithmetic function like the average or median of the attributes values;
- Delete missing values: The simplest approach is to delete the instances containing the missing values, either by removing an entire row or an entire column depending on what the modeler decides. The fact that an entire column or row is deleted, especially with a column, a great deal of information is also being removed and not only the missing value, which can even cause a greater impact over leaving the missing values on the data.

#### 2.2.1.2 Outliers

Outliers are defined as unusual values that do not present the same behavior as other values do. They are caused either by an anomaly caused by an equipment, like sensors for instance, or either by a real abnormal event, like an earthquake (Chen, Wang, & van Zuylen, 2010). The difficulty lies in dealing with outliers on the premise that not all outliers can be considered insignificant and can be

positively influence the outcome in terms of adding valuable information to the dataset. (Abbott, 2014) describes several approaches to deal with outliers:

- Removing the outlier from the dataset: This approach can reveal to be either good or bad, depending on the significance of the outlier. If the presence of the outlier proves significant then information is being lost, and if not then the model is improved;
- Transforming the outliers: Based on the same premise of the previous approach of removing the outlier from the dataset, changing the nature of an outlier can also compromise the model and its response;
- Keep the outliers: This approach limits the modelers choice in which models to use, being only able to use models that are not greatly affected by the presence of outliers, or penalize their existence.

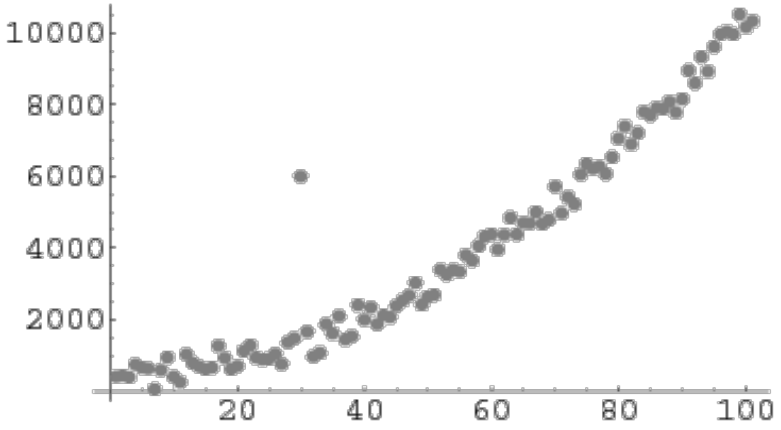


FIGURE 2.4: Exemplification of an outlier in a plot

#### 2.2.1.3 Feature Selection and Dimensionality Reduction

Reducing the dimensionality of the data allows for an model to operate faster and more effectively by removing irrelevant or redundant information. This reduction can be done through feature selection providing a better understanding and interpretation of the data or either by the application of Principle Component Analysis (PCA) (S. Kotsiantis, Kanellopoulos, & Pintelas, 2006). Another way of gaining valuable information from the data is through adding new features or as (Abbott, 2014) defines them as "derived variables". The commonality between reducing and

creating features is that when features are proven to be good they reduce the necessity for a more complex understanding of the data and they produce more valuable and trustworthy results.

#### 2.2.1.4 Principal Component Analysis

According to (Abdi & Williams, 2010) and (James et al., 2013) Principal Component Analysis (PCA) focuses on the following objectives: extracting the most important information from the dataset; and reducing the dataset with the intent of only maintaining the most important information, thus simplifying its understandability. To do this, PCA applies linear combinations to the original set of variables revealing the principal components which try to explain and reduce the existent variability of the original dataset. Before the PCA analysis can be done, all the variables must be standardized or normalized to eliminate any influences or weight that one variable might have over the rest of the variables. According to (Friedman, Hastie, & Tibshirani, 2001) PCA is computed using Singular Value Decomposition (SVD) which decomposes a matrix $X$ $(M * N)$ into three other matrices:

$$X = U * S * V^t \tag{2.1}$$

where $U$ is a $(M * M)$ matrix, $S$ is a diagonal $(M * N)$ matrix and $V^t$ is the transpose of $V$, where $V$ is a $(N * N)$ matrix. PCA is commonly used for increasing the efficiency of the analysis of the models by reducing the redundancy of the model. It does this by reducing the dimensionality with losing the minimum amount of information. And so, the original variables that presented some form of correlation between them, are transformed into uncorrelated variables or Principal Components (PC), which are linear combinations of the correlated variables.

## 2.3 Predictive Modeling

Based on (Abbott, 2014), predictive modeling algorithms are supervised learning algorithms which implies that they learn based on previous experiences, in other words, they generate a new predictive response by testing a new set of input data to a known set of inputs that are already known to produce a certain response.

(Friedman et al., 2001) refers to this as a process of "learning by example". Supervised Learning is usually divided into two main categories: Regression and Classification. In the Regression setting, response variables are usually characterized as quantitative or numerical and the main objective is to predict based on continuous measurements. In contrast, in the Classification setting, response variables are usually qualitative or categorical and the main objective is to assign a label to the response variables. The main goal of these models is to predict a given variable $Y \varepsilon \mathbb{R}$ in terms of a set of inputs $X \varepsilon \mathbb{R}^v$:

$$Y = F(X) + \varepsilon \tag{2.2}$$

where $F(X)$ is the observed value of the function in use, $\varepsilon$ is an error term and $v$ is the number of inputs. There are several predictive algorithms (regression, neural networks, decision trees, k-nearest neighbors, etc.) but throughout this dissertation the focus will be mainly on neural networks and on regression algorithms, such as: Multiple Linear Regression, Ridge Regression and Principal Component Regression, as they provide a far better comprehension of quantitative results.

### 2.3.1 Multiple Linear Regression

Linear Regression is the most basic and commonly used predictive model where its purpose is to explain the existing relationship (the weight or value of the coefficients) between a dependent variable $Y$ or response, and one or more independent variables $X$ or predictors (James et al., 2013). In other words, Linear regression provides a general description of how the inputs affect the output by weighing the coefficients (Friedman et al., 2001). Depending on the number of independent variables when applying Linear Regression, it can either be described as Simple Linear Regression when only one independent variable is used or Multiple Linear Regression (MLR) otherwise. The Multiple Linear Regression model is represented by the following equation:

$$F(X) = \beta_0 + \sum_{j=1}^{P} * X_j * \beta_j \tag{2.3}$$

Where $\beta_{0...P}$ corresponds to the coefficients, $X_{1...j}$ to the independent variables and $P$ the total number of independent variables. According to (Friedman et

al., 2001), the independent variables can be considered by taking on several forms, mainly as quantitative inputs, as transformations of those quantitative inputs, like the logarithm for example, of the polynomial representation of those inputs ($X^2$, $X^3$, $X^4$) or of arithmetic interactions between the variables, like $X_3 = X_1 * X_2$ for instance.

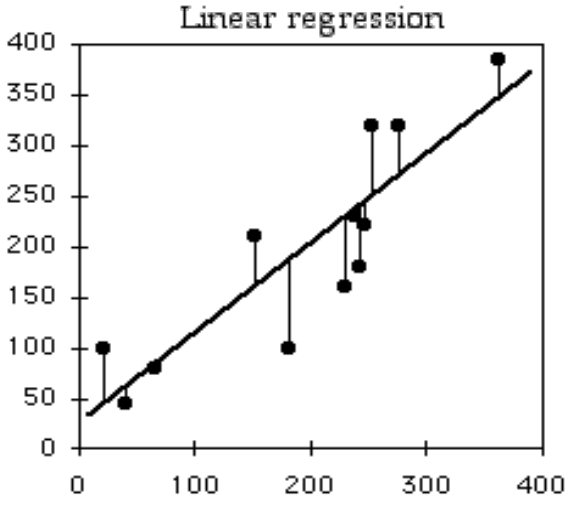


FIGURE 2.5: Example output of a Linear Regression

(Friedman et al., 2001) also refers that in most of the cases where the application of Linear Regression is prominent, the coefficients are usually calculated through the Least Squares, in which the coefficients are chosen to minimize the Residual Sum of Squares (RSS), given by Equation 4 and to find the best fit to the data.

$$RSS(\beta) = \sum_{i=1}^{N}(y_i - F(X)))^2 \tag{2.4}$$

Sometimes the Least Squares method does not provide the best accuracy or interpretation of the model due to the possibility that some predictors can present a large variance of their data which means that they provide little to almost no additional information to the model, and if large number of predictors are used, then the interpretation of the model becomes a lot more complex (Friedman et al., 2001).

The interpretability of Linear Regression models can be done in two different ways: (1) through the values of the coefficients that describe the weight that

is given to a certain predictor variable and how they will impact the estimated values; (2) by comparing the values of the response variable or estimated values, with the actual values of the model by use of error measures as will be discussed afterwards in Section 2.4 (Abbott, 2014). (Tobias et al., 1995) refers that to take full advantage of the MLR models, three different conditions should be met: (1) the predictors that are used to express a response must be few; (2) there must not be any correlation between them (which as explained in 2.2.1.4 could be achieved through the use of PCA) or in other words, there must not have a highly linear relation, and (3) they must express some sort of relationship to the responses.

### 2.3.2 Ridge Regression

According to (Friedman et al., 2001), the idea of Ridge Regression (RR), also described as the Tikhonov regularization, is to penalize the regression coefficients. This algorithm is similar to the MLR algorithm, where the only noticeable difference is in the application of a lambda ($\lambda$) parameter that controls the amount of penalty going that is going to be applied to the coefficients, thus allowing for a more controlled shrinkage of the model by shrinking the coefficients towards zero, which causes a smoothing the model. This method provides a decorrelation of the variables, without applying any sort of dimensionality reduction like PCA. Just like MLR, RR uses the Least Squares method to minimize the RSS but in this case, it considers the influence of the penalty on the coefficient to minimize the sum of the squares, as shown in Equation 2.5. If the value of the coefficients ($\beta$) is large than the value for the penalty will increase.

$$RSS(\lambda) = (y - X\beta)^T * (y - X\beta) + \lambda\beta^T\beta \tag{2.5}$$

Where $\lambda$ is the amount of shrinkage to be applied to the model and $X$ is the input matrix.

### 2.3.3 Principal Component Regression

Sometimes, there are a large number of independent variables that can sometimes be correlated between them and so Principal Component Regression (PCR) is a linear predictive model which estimates the response of the model based on the selection of Principal Components (PC) that represent the most explanatory

variables by making use of PCA that has been discussed in 2.2.1.4 Since the PCs don not present any sort of correlation between them, they are then considered as inputs for the model (Liu, Kuang, Gong, & Hou, 2003). According to (Friedman et al., 2001), the PCR starts by standardizing the inputs, and only then is the PCA algorithm applied so that there are only PC applied to the regression instead of the original predictors. By standardizing and applying the PCA to the model, reducing the dimensionality of the model, PCR takes care of any possible collinearity or correlation between the predictors. By considering Equations 2.1 and 2.5 (refer to 2.2.1.4 and 2.3.2, respectively), the representation of the PCR model becomes:

$$X = z_1 v_1^T + z_2 v_2^T + ... + z_i v_i^T + \varepsilon \tag{2.6}$$

where $z_{1...i}$ are the score values, $v_{1...i}$ are the eigenvalues of the matrix $X$ and $\varepsilon$ is the error term.

### 2.3.4 Neural Networks

Neural Networks are defined as Multi-Layer Perceptrons (MLP). MLPs are comprised of neurons, where each neuron is defined by an equation, usually referred to as a transfer function (Abbott, 2014).

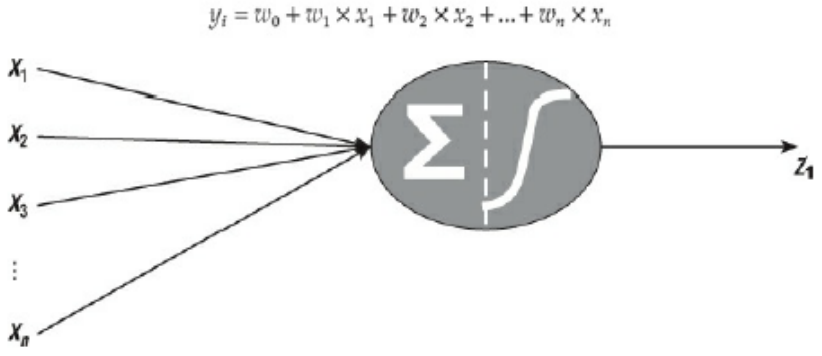


FIGURE 2.6: Example of a neuron (extracted from (Abbott, 2014))

A single neuron can produce a linear model. Using a single neuron does not provide better results than by using other predictive models because there are several linear models that have greater accuracy and are more efficient. To show improvements when compared to linear models, neural networks stack these single neurons in layers allowing for more powerful and flexible prediction algorithms.

As said, a layer is a set of stacked neurons, and can be defined either as an input layer, output layer or hidden layer. The hidden layers are usually between the input and the output layer.

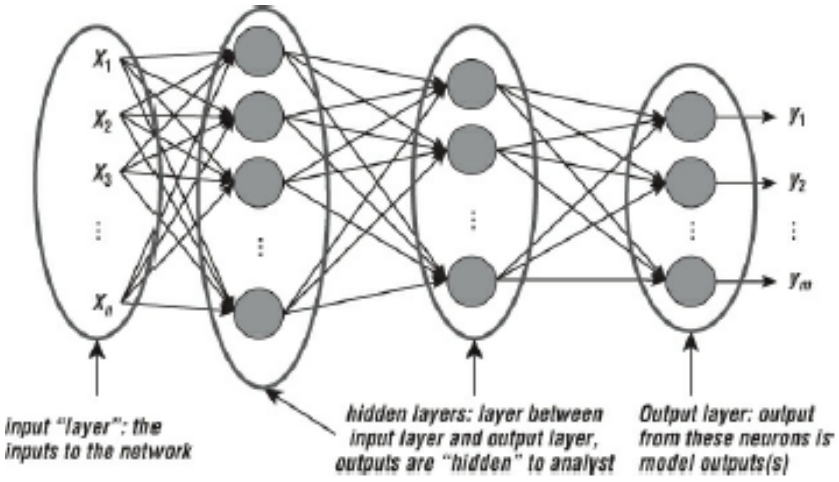


FIGURE 2.7: Example of a neural network (extracted from (Abbott, 2014))

Neural Networks are iterative learners which means that they learn step by step. In the first step the weights are randomly initialized in order to start training the algorithm, passing through the different layers of the network ending up on the hidden layers and finally in the output layer. The resulting prediction is compared by measuring the error, to the actual expected value. The weights are then adjusted with the calculated error measured and a new cycle or epoch starts. This happens until the entire dataset is trained, thus being ready to be tested with new unknown values.

## 2.4 Models Evaluation

Predictive models are developed based on past data, that will then be applied to new instances of data that have not yet been introduced in the models, or in other words, new generated data. It is necessary to evaluate and verify these in terms of accuracy and performance when confronted with new data.

The evaluation of the performance of the models is usually divided using different or a combination of several re-sampling methods into two sets, the training set and the test set (Figure 2.8). Sometimes a third set may also be considered where it is defined as the validation set. And so, this evaluation is made through the use of an evaluation measure (or error measure) in mind. To estimate the

accuracy and performance of this models is to provide a method for comparing the models between them to be able to fine tune these models.

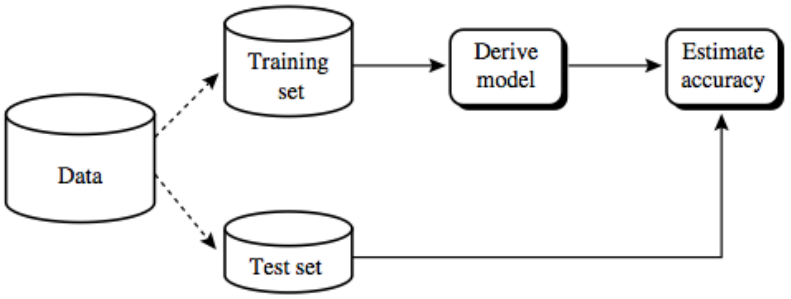


FIGURE 2.8: Models evaluation lifecycle

### 2.4.1 Criteria for Model Comparison

Many criteria, that can also be defined as accuracy measures, are used to estimate how well the models are performing, their purpose is to measure the difference between the estimated or predicted values and the actual or real values:

$$e_i = y_i - \hat{y}_i \tag{2.7}$$

Where $y_i$ is the real value and $\hat{y}_i$ is the estimated value of the model. Accuracy measures are scale dependent which means that they cannot compare values with different scales. This information is then used by the modelers, so that they can fine tune these models in order to either improve them or decide that it is not worthy using them in the long run (James et al., 2013).

#### 2.4.1.1 Correlation Coefficient

Correlation Coefficient or $R$ measures the linear relationships between the variables. It focuses on quantifying the dependence or correlation between the variables. It ranges from [-1.0,1.0] where it indicates either a negative or positive correlation between the variables, respectively. If the result equals to 1.0 then there is a perfect positive linear correlation between $x$ and $y$, presenting the same amount of variation. If, by contrast the result equals to -1.0, then there is a perfect negative linear correlation between the variables which means that they vary in an opposite way. If the result equals to 0 then there is no correlation between them.

The most common calculation for the Correlation Coefficient is through the Pearson product-moment, where its first calculated the covariance between the variables and then is divided by their standard deviations:

$$R_{xy} = \frac{Cov(v_x, v_y)}{\sigma_x * \sigma_y} \tag{2.8}$$

Where $v(x, y)$ corresponds to the real and estimated variables and $\sigma(x, y)$ corresponds to each of the variables standard deviations.

#### 2.4.1.2 Coefficient of Determination

Coefficient of Determination or $R^2$, measures how well the model can predict future outcomes, in other words, how well the dependent variables can be predicted considering the independent variables. It accounts for the variability of the model and it is calculated through the square of the Coefficient Correlation and it ranges from [0.0,1.0]:

$$R^2_{xy} = (\frac{Cov(v_x, v_y)}{\sigma_x * \sigma_y})^2 \tag{2.9}$$

An improvement can be made to the $R^2$, that is defined as the Coefficient of Determination Adjusted or $R^2_{adj}$, which gives the percentage of variation that the independent variables are really affecting the dependent variable, in other words, with provides the confidence of the model predicting the correct outcome.

#### 2.4.1.3 Mean Squared Error

Mean Squared Error or $MSE$ measures the quality of the estimated value or, in other words, how close the estimated values differ from the actual values. The $MSE$ of the predictions is the mean of the squares of difference between the estimated value and actual values and it can be defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(e_i)^2 \tag{2.10}$$

Where $n$ is the number of instances within the dataset and $e_i$ is the error exemplified in Equation 2.7. The $MSE$ ranges from [0.0,1.0] where the results

that are close to 0 are highly desirable, because if the estimate for the predictive value is 0, then the algorithms accuracy was perfect. The $MSE$ is extremely useful because it shows the variance and deviation from the estimated value to the actual value (James et al., 2013).

A variation of the $MSE$ can be defined as the Root Mean Squared Error or $RMSE$, where nothing more than the root of the $MSE$ metric. This is an useful metric due to its ability to amplify through penalization, large errors that by using only the $MSE$ would most likely pass undetected.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(e_i)^2} \tag{2.11}$$

#### 2.4.1.4 Mean Absolute Error

The Mean Absolute Error or $MAE$ is the sum of the absolute values of the errors. The advantage of applying this metric rather than $MSE$ is when dealing with outliers. Despite the outliers, $MAE$ accuracy follows the same logic as $MSE$ being that it is better when closest to zero. $MAE$ is defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i| \tag{2.12}$$

### 2.4.2 Re-sampling Methods

Re-sampling methods are used to determine and/or increase the accuracy of a model by refitting a model to retrieve new hidden information that sometimes can only be obtained by fitting the model more than one time. This process of evaluating the performance of a model is also defined as model assessment (James et al., 2013). According to (S. B. Kotsiantis, Zaharakis, & Pintelas, 2007) there are two ways for evaluating the predictive accuracy of a model: (1) by splitting the dataset into training and testing sets or Hold-Out and (2) through K-Fold Cross-Validation. (James et al., 2013) refers an additional variance of cross-validation referred to as Leave-One-Out Cross-Validation.

According to (Tashman, 2000) in order to provide a real-time assessment for forecasting, there is a need to wait a long time for data to be generated in order

to get a reliable picture of what is going to be forecast. He also refers that the Hold-Out method has become the most generally accepted re-sampling method.

#### 2.4.2.1 Hold-Out

Hold-Out is one of the simplest and easiest validation techniques by randomly splitting the training and the tests sets only once (usually dividing $\frac{2}{3}$ of the data to the training set and the other $\frac{1}{3}$ to the test set) as exemplified in Figure 2.9 (James et al., 2013). For predictive models that, to provide better results, account for most of the historical data, usually the training set takes up 80 to 95% of the records on the dataset.


FIGURE 2.9: Hold-Out Re-Sampling Method

#### 2.4.2.2 K-Fold Cross-Validation

K-Fold Cross-Validation (K-Fold CV) takes the same approach as Hold-Out but instead of splitting the dataset only once it splits it k times, usually of the same size. The k=0 subset is used as the validation set and the remaining k-1 subsets are treated as the training set. The model repeats k times where each time the validation set changes. One variation of the K-Fold Cross-Validation is the Leave-One-Out Cross-Validation (LOOCV) which follows the same logic as the K-Fold CV where the k times that the model is split corresponds to the number of elements with the dataset. In other words, the dataset is trained and tested k (number of elements minus one) times (James et al., 2013).

#### 2.4.2.3 Rolling-Origin Cross-Validation

Rolling-Origin Cross-Validation or ROCV is an out-of-sample evaluation, in which the the origin of the training set is successively being updated, much like the Sliding Window method, which results in the production of several new forecasts

(Tashman, 2000). Assuming the division of a dataset with 10 years into N=10, where N is the number of samples, each corresponding to a year of records. The ROCV maximum number of samples would be N=5, where the last sample corresponds to the test set and the other 4 to the training set. The forecast is then generated and the ROCV moves forward one until it reaches the tenth sample, thus generating 5 forecasts.

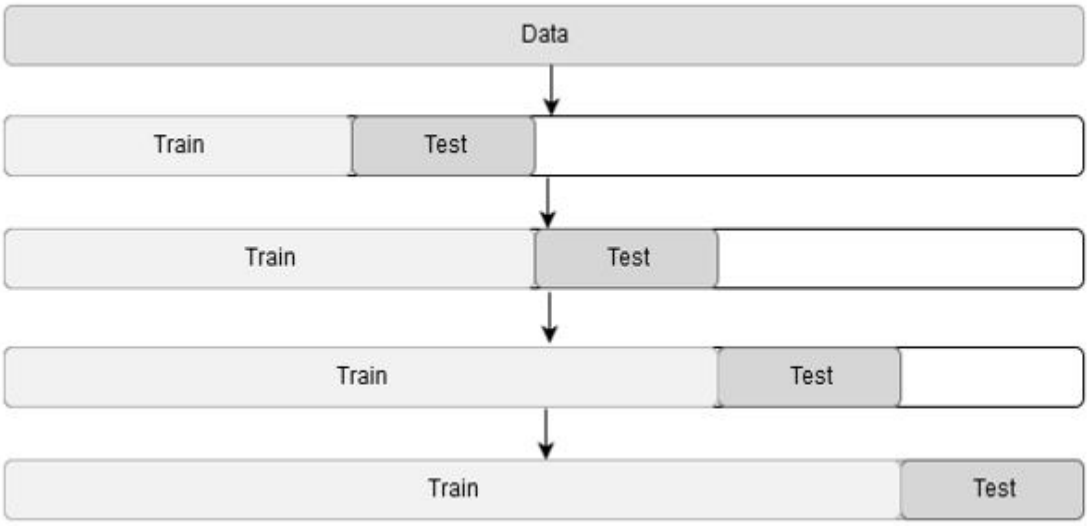


FIGURE 2.10: Example of Rolling-Origin Cross-Validation

## 2.5 Predictive Modeling for Dam Behavior

Structural Health Monitoring (SHM) has been growing and evolving over the years, mainly with the appearance and evolution of sensor technologies to identify structural damages. It offers automated methods for assessing the structural integrity and health through structural monitoring systems. These systems have been continuously growing and being improved, and are widely accepted for the detection and prediction of the behavior of the structures and are responsible for collecting the measurements from the sensors that are installed within these structures (Lynch & Loh, 2006).

GestBarragens is a software for monitoring the safety of engineering structures like dams, which supports, among others not relevant for the scope of this research, the process of manual and automatic data exploration from instruments located on the structures, the process of visual inspections as well as the ability for anomaly

detection, to ensure a good decision-making process (Silva, Galhardas, Barateiro, & Portela, 2005).

GestBarragens also supports the generation and visualization of quantitative interpretation models, numerical models as well as physical models. The quantitative interpretation models establish relations between the input values that influence the model and the structural behavior responses, as exemplified in Figure 2.11 (Portela, Pina dos Santos, Silva, Galhardas, & Barateiro, 2005).

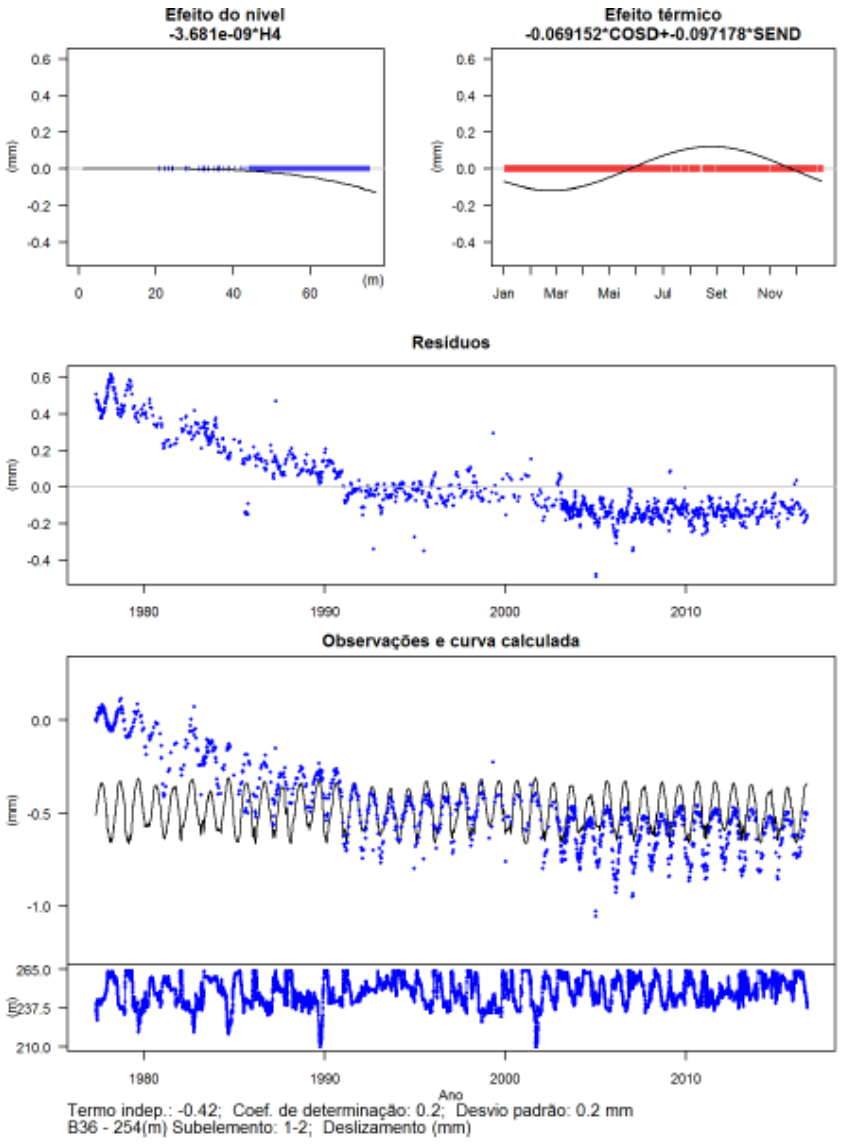


FIGURE 2.11: Example of a resulting quantitative interpretation model from GestBarragens of a structural behavior response

Other challenges like the one proposed by (Mata & Tavares de Castro, 2015) also have the intention of providing better and quality data to allow for a better further analysis. The authors propose a qualitative analyses and assessment of

paired samples of automatically and manually gathered measurements (ADAS and MDAS, respectively). Their idea is to eliminate gross measurements resulting from the difference in frequency of gathered records, pairing both the ADAS and the MDAS to determine through the use of Probability Density Functions (pdf) if they represent the same population, thus eliminating differences between the ADAS and MDAS, to successfully analyze the ADAS measurements (Figure 2.12).

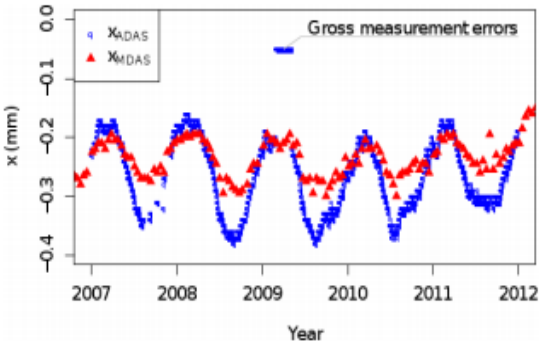


FIGURE 2.12: Plot of MDAS and ADAS Measurements over the years

There has been a significant amount of work done relative to monitoring the behavior and safety of dams. For a more general perception of what has been done in this area, several related papers have been analyzed and summarized in Table 2.1. These papers have been characterized in several dimensions: (a) the objective of the paper, (b) the models used, (c) the input attributes for the models (environmental variables), (d) the error metrics for model validation and (e) the output of the model (dam behavior).

TABLE 2.1: Survey on Related work about Predicting Dam Behavior Responses

| **Objective** | **Methods** | **Attribute** | **Error Metrics** | **Output** |
|---|---|---|---|---|
| Assessing the importance of water and thermal temperature variations on thermal displacements (Tatin, Briffaut, Dufour, Simon, & Fabre, 2015) | $HTT$ | $W_t$, $A_t$ | $\sigma$ | Radial Displacements |
| Assessing the delayed response analyses for pore pressure measurements through the effects of the water level and rainfall events (Bonelli & Royet, 2001) | $IRF$ | $W_l$, $R_f$ | - | Pore Pressure |
| Express complex relationships between the environmental variables and noise effects on the monitoring data through linear and nonlinear mapping of the variables (Cheng & Zheng, 2013) | $PCA$, $SVM$ | $W_l$, $R_f$, $A_t$ | - | Radial Displacements, Uplift Pressure |
| Determine the usefulness of a FNN (FeedForward Neural Network) model for assessing dam behavior (Ranković, Novaković, Grujović, Divac, & Milivojević, 2014) | $NN$, $MLR$ | $W_l$ | $R$, $R^2$, $MSE$, $MAE$ | Pore Pressure |
| Dam behavior analyses through the use PCA for dimensionality reduction (Yu, Wu, Bao, & Zhang, 2010) | $HST$, $PCA$ | $H$, $S$, $t$ | $R$ | Crack Opening |
| Comparison of auto regressive models for performing delayed analyses on air temperature measurements with a seasonal analysis (Bonelli & Félix, 2001) | $IRF$ | $W_l$, $A_t$ | - | Radial Displacements |
| Support Vector Regression techniques evaluation for forecasting tangential displacements (Ranković, Grujović, Divac, & Milivojević, 2014) | $SVM$ | $W_l$ | $R$, $MSE$, $MAE$ | Tangential Displacements |
| Assess how effectively an HSS can estimate the time-effect deformation on monitoring data (Li, Wang, Liu, Fu, & Wang, 2015) | $HST$ | $H$, $S$, $t$ | $\sigma$, $pdf$, $R^2$ | Radial Displacements |
| Usage of Moving PCA and Robust Regression for extracting relevant components and to detect possible anomalies on the measurements (Jung, Berges, Garrett, & Kelly, 2013) | $MPCA$, $RRA$ | $W_l$ | - | Pore Pressure |

| | | | | |
|---|---|---|---|---|
| Application of a statistical approach accompanied with a structural identification technique to provide a higher degree of accuracy in predicting and monitoring the behavior of dams (De Sortis & Paoliani, 2007) | $HST$ | $H$, $S$, $t$ | $R$, $\sigma$ | Radial Displacements |
| Application of a Feed Forward Neural Network to estimate and simulate the flow of a dam (Tayfur, Swiatek, Wita, & Singh, 2005) | $NN$ | $W_l$ | $RMSE$, $MAE$, $R^2$ | Pore Pressure |
| Performance comparison between a MLR and a NN model for assessing dam behavior (Mata, 2011) | $MLR$, $NN$ | $W_t$, $A_t$ | $MAE$, $MaxAE$, $R$ | Horizontal Displacements |
| Increase fitting accuracy and forecasting precision based on an Error Correction Model by integrating the relationships between the output and input variables (Li, Wang, & Liu, 2013) | $ECM$, $MLR$ | $H$, $S$, $t$, $error$ | $\sigma$, $pdf$, $R^2$ | Radial Displacements |
| Identification of the effect of air temperatures on the structural response of the dam based on a Fourier Transform analysis (Mata, de Castro, & da Costa, 2013) | $STFT$ | $H$, $A_t$ | $\sigma$, $R^2$ | Horizontal Displacements |
| Usage of modifications of the PLS model for mitigating the collinearity between the variables and the existence of outliers, and the selection of informative variables (Xu, Yue, & Deng, 2012) | $SIMPLS$, $GA-PLS$ | $H$, $A_t$ | $RMSE$ | Crack Opening |
| Assessing the performance of a MLR model optimized by using Genetic Algorithms (Stojanovic, Milivojevic, Ivanovic, Milivojevic, & Divac, 2013) | $MLR$ | $H$, $A_t$, $C_t$, $R_f$, $t$ | $R^2$, $RMSE$ | Radial Displacements |
| Assess the performance of hybrid models for dam deformations (Perner & Obernhuber, 2010) | $MLR$ | $H$, $C_t$, $t$ | - | Radial Displacements |

**Methods**: $HTT$=Hydrostatic Thermal Time; $IRF$=Impulse Response Function; $PCA$=Principal Component Analysis; $SVM$=Support Vector Machines; $NN$=Neural Networks; $MLR$=Multiple Linear Regression; $HST$=Hydrostatic Seasonal Time; $MPCA$=Moving PCA; $ECM$=Error Correction Method; $STFT$=Short Time Fourier Transform; $SIMPLS$=Statistically Inspired Modification of Partial Least Squares; $GA-PLS$=Hybrid Genetic Algorithm with SIMPLS.

**Attributes**: $W_t$: Water Temperature; $A_t$: Air Temperature; $W_l$: Water Level; $R_f$: Rainfall; $H$: Hydrostatic; $S$: Season; $t$: time; $C_t$: Concrete Temperature.

**Error Metrics**: $\sigma$: Standard Error of Estimate; $R$: Correlation Coefficient; $R^2$: Coefficient of Determination; $MSE$: Mean Squared Error; $RMSE$: Root Mean Squared Error; $MAE$: Mean Absolute Error; $pdf$: Probability Density Function; $MaxAE$: Maximum Absolute Error.

The main contributions provided by these authors are the identification of different models used to monitor structural damages in dams through the relationships between the environmental variables (predictors) and the behavior of the dams (response). Depending on the problem and the case study, several objectives have been defined but the commonality between them is the analyses of the monitoring data either being generated manually or by equipment within the structures (pendulums, piezometers, etc.), and the identification of responses that explain the behavior of these structures (pressures, displacements, etc.). Even though most of the authors provide different alternative models for monitoring dam behavior, most of the attributes or environmental variables that serve as inputs for these models are the same: Hydrostatic Load, Water Level, Air Temperature, Water Temperature, Rainfall, Time.

### 2.5.1 Dam Behavior variables

According to (Mata, 2011) and (Xu et al., 2012), and considering the attributes of Table 2.1, the statistical relationship between the dependent variables and the independent variables is given by:

$$Y(W,T,t) = Y_W + Y_T + Y_t + \varepsilon \tag{2.13}$$

where the $Y(W,T,t)$ corresponds to the response variable, the $W$ corresponds to the Hydrostatic Load, the $T$ to the Temperature variations, the $t$ to the time since the initial record of the structure, or in other words, the aging of the structure and $\varepsilon$ to the error component. Each of the effects of components that correspond to each of the independent variables provide different influences on the behavior of the structure.

The influence of the $Y_W$ variable can be described through the use of polynomials to scale this variable in order to provide more weight to these variable and thus giving it more influence if necessary to the models, where $\beta_{1...4}$ correspond to the coefficients to adjust and the $h = 265 - 76$, where 265 is the Crest Elevation and 76 is the Height Above Streambed, which corresponds to the water level:

$$Y_W = \beta_1 h^4 + \beta_2 h^3 + \beta_3 h^2 + \beta_4 h \tag{2.14}$$

According to (Mata, 2011) the influence of the temperature can be calculated through the use of the age of the structure, and can be considered as a sinusoidal function, extending over a period of a year or six months (In the context of this dissertation, the functions have been calculated for a period of a year). This function can be extracted in this form, especially in the case of Portugal, since the country has a sort of predictability to its temperatures, where the temperature tends to rise when approaching summer and decreasing when approaching winter. And so, the influence of the temperature can be described as follows:

$$Y_T(\sigma) = \beta_1 cos(\sigma) + \beta_2 sin(\sigma) + \beta_3 sin^2(\sigma) + \beta_4 cos(\sigma) sin(\sigma) \tag{2.15}$$

where $\beta_{1...4}$ corresponds to the coefficients to be adjusted and $\sigma = \frac{2\pi d}{365}$, where $d$ equals to the days since the beginning of a year and 365 to number of days in a year.

The influence of time or the aging of the structure is important to encompass elements which vary over time, like deterioration for instance. And so, the influence of time can be represented as:

$$Y_t = \beta_1 t + \beta_2 t^2 + \beta_3 t^3 \tag{2.16}$$

where $\beta_{1...3}$ corresponds to the coefficients to adjust and $t$ the number of days since the age of the structure since the analyses began.

# Chapter 3

# Design and Development

This chapter of the dissertation covers the Design and the Development of the artifacts phase (expressed in Figure 3.1) of the DSRM.

- **Section 3.1**: In this section it is presented the Case Study that is going to be used throughout the Demonstration and Evaluation phases (in Chapter 4) of this research, aiming to provide a generic and detailed approach of the use of each of the predictive models on engineering structures (in this case, a Dam structure);
- **Section 3.2**: In this section a development methodology for the artifacts is proposed, to allow for a more comprehensible step-by-step approach of what is going to be done throughout the Demonstration and Evaluation phase (Chapters 4);
- **Section 3.3**: In this section it is presented what the Development Language is, behind the code used to develop the applications of the different techniques to the datasets, as well as to evaluate the resulting metrics.

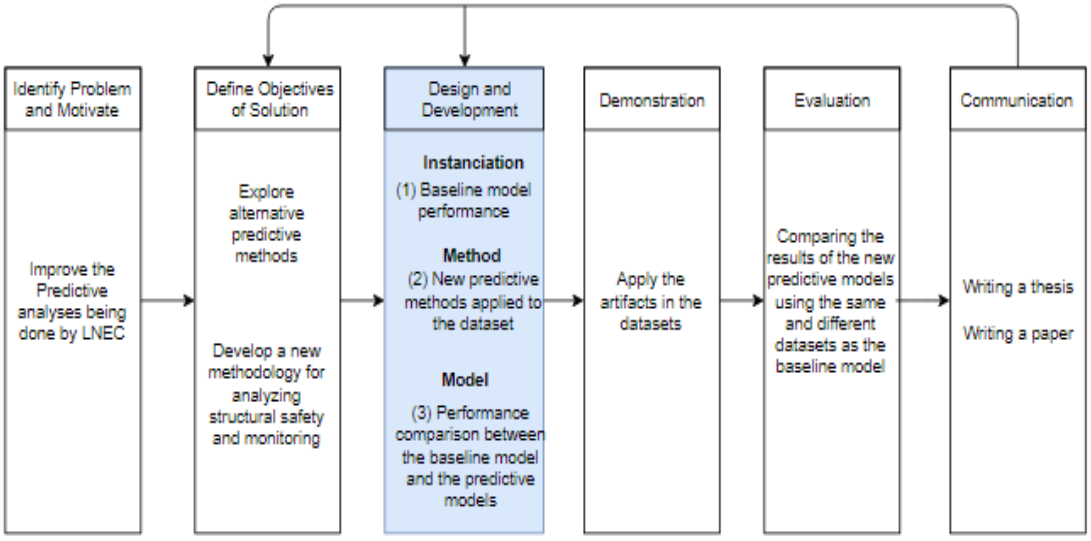


FIGURE 3.1: Adaptation Design Science Research Methodology - Design and Development (extracted from (Peffers et al., 2007))

## 3.1 Case Study - A Portuguese Concrete Dam

For this research, it has been chosen as a case study, datasets related to different types of instruments from a portuguese concrete dam. The dam used in this case study is located in the Douro river basin. It is an arch type dam with a height of more than $75m$ functioning as a hydroelectric plant and it started being explored in the decade of 1970. It can retain nearly $13^6m^3$ of water and has a concrete volume of more than $80000m^3$ (Figure 3.2).

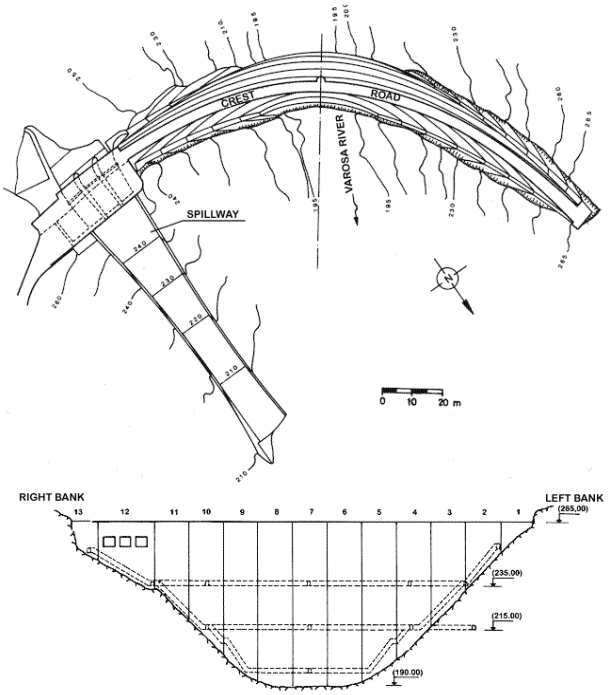


FIGURE 3.2: Structural Schema for the studied dam

The monitoring system currently implemented throughout the dam consists of a combination of several instruments that exist within the different 13 blocks, each one with the purpose of measuring different types of quantities, in most cases even measuring more than one quantity. Table 3.1 represents the different types of instruments existing within the structure, their number of records and if the instrument is gathering data either manually or automatically.

TABLE 3.1: Recording Instruments existing on the studied dam (instruments names in portuguese)

| **Instruments** | **Data Gathered Manually** | | | **Data Gathered Automatically** | | |
|---|---|---|---|---|---|---|
| | **Number of Sensors** | **Number of Records** | **Date First Record** | **Number of Sensors** | **Number of Records** | **Date First Record** |
| Base de Alongâmetro | 70 | 50383 | 11-12-1975 | 6 | 171153 | 06-10-2006 |
| Deslocamento geodésico | 24 | 630 | 01-04-1984 | - | - | - |
| Dreno | 122 | 82510 | 15-01-1977 | 2 | 22170 | 10-01-2006 |
| Escala de Nível | 1 | 85682 | 15-11-1976 | 1 | 28564 | 10-01-2006 |
| Extensómetro de Fundação | 25 | 30524 | 01-07-1976 | 3 | 83919 | 10-01-2006 |
| Extensómetro de Resistência | 68 | 47130 | 11-08-1976 | - | - | - |
| Fio de Prumo (Base) | 10 | 17871 | 05-11-1976 | 6 | 171415 | 10-01-2006 |
| Higrómetro | - | - | - | 1 | 2058 | 25-04-2008 |
| Medidor de Juntas de Resistência | 12 | 7535 | 17-01-1975 | - | - | - |
| Nivelamento geométrico de precisão | 50 | 953 | 01-04-1984 | - | - | - |
| Piezómetro | 61 | 90668 | 16-01-1977 | 4 | 114192 | 10-01-2006 |
| Termómetro de Máxima e Mínima do Ar | 1 | 13054 | 16-11-1976 | - | - | - |
| Termómetro de Resistência | 31 | 25884 | 01-07-1976 | 8 | 197541 | 10-01-2006 |
| Termómetro do Ar | - | - | - | 3 | 66811 | 10-01-2006 |
| Total | 475 | 452824 | - | 34 | 857823 | - |

The dam studied in this dissertation is not one of the oldest nor one of the biggest dams present in Portugal, but despite that, the number of instruments inside this structure reaches 509, which includes both manual and automatic data generation instrumentation. It is noticeable the difference in growth rates for the manual and automatic data gathering instruments. The first record of most of the

manual instruments is from the 1970s reaching over 452824 records from 457 instruments against the 34 instruments for the automatic instruments where the first record is from 2006 reaching almost double the value of manual instruments with 857823 records. The difference in these growing rates is related to the frequency the measurements are retrieved. For the manual instruments, measurements are only gathering data weekly while the automatic instruments are gathering data several times a day, thus providing more records over a lower period of time. The manual measurements are usually gathered and verified by one or more of the area scientists, mainly through visual inspection, where the automatic measurements are gathered through the use of sensors which also needs to be verified and analyzed, thus the need for predictive analysis to automatically monitor these structures.

The main data sources used in the context of this research have all been provided by LNEC and are organized in the traditional form of a table within a .txt file and they all correspond to instruments that are taking manual measurements and so, for each of the datasets are included the measurements taken by one sensor which corresponds to one instrument. There are several sensors taking the same measurements that are placed in different locations within the structure, not only for redundancy but for providing the most accurate information of what is happening in the entire structure. As seen in Table 3.1 there are 12 different sensors that are taking measurements manually, hence it is quite difficult and time consuming to analyze them all.

To analyze and provide a realistic approach to the identified problem there were chosen three different types of instruments (sensors): BaseDeAlongametro, ExtensometroDeFundacao and FioDePrumo(Base).

- For the BaseDeAlongametro instrument there were provided 70 sensors, each one of them reading measurements relative to the Opening and the Slippage structural responses from the dam.
- For the ExtensometroDeFundacao instrument there were provided 7 sensors, each reading measurements relative to the Displacement structural response from the dam.
- And finally, for the FioDePrumo(Base) instrument there were provided 10 sensors, each reading measurements relative to the Radial Displacements and the Tangential Displacements structural responses from the dam.

The datasets were exported from a platform developed by LNEC, called GestBarragens, and since it was already being used in a production and predictive setting, through the use of a MLR model, there was no real need to check the consistency of the data or the existence of a high number of outliers, though they could still exist. For determining and cleaning the existence of outliers, GestBarragens is already applying a method, through the use of standard deviation which removes the values that are above or under a certain threshold (parameterizable from 1 to 3 times the value of the standard deviation). Although this process of cleaning the data has been approved by engineering specialists, there could be valuable information on these values that are being removed that could prove beneficial for better understanding of the structures safety and also for their monitoring.

The datasets used consist of data gathered with a daily periodicity and range from the beginning of the exploration of the structure until the 19th of September, 2016. Table 3.2 illustrates the dependent variables that are representative of the structural responses (dam behavior) for the different datasets that have been provided and are going to be the focus of this research.

Table 3.2: Provided dependent variables

| **Name** | **Data Type** | **Data Source** | **Units** | **Measurement Frequency** |
|---|---|---|---|---|
| Opening | Numerical | GestBarragens | (mm) | Weekly |
| Slippage | Numerical | GestBarragens | (mm) | Weekly |
| Displacement | Numerical | GestBarragens | (mm) | Weekly |
| Radial Displacement | Numerical | GestBarragens | (mm) | Weekly |
| Tangential Displacement | Numerical | GestBarragens | (mm) | Weekly |

For this analyses, the independent variables retrieved from the GestBarragens software are the:

- $dateRef$, that corresponds to the day the record was taken;
- $h$, that corresponds to the Water Level present at that time.

The following independent variables have been determined considering the previously mentioned functions (2.5.1) and dam engineering practice, with the intent to add more relevant information and to increase the accuracy of the models:

- $h^2, h^3, h^4$, correspond to the scaling done to the $h$ variable to give more weight to this variable;
- $t, t^2, t^3$, where $t$ corresponds to the age of the structure;
- $cos(d), sin(d), sin^2(d), cos(d)sin(d)$, where the combination of the $cos(d)$ and the $sin(d)$ variables represent the variations in temperature and $sin^2$ and $cos(d)sin(d)$ represent transformations of those variables.

The dependent variables (see Table 3.2) from the datasets are usually characterized by temporal series where the date in which measurements are taken is uniform and successive over time. From the independent variables generated by GestBarragens and those considered through the equations on 2.5.1, the effect of the Water Level is predominant and is a significantly impacting variable. The effect of the temperature is also considered as one of the impacting variables in structural dam behavior.

## 3.2 Design and Development Methodology

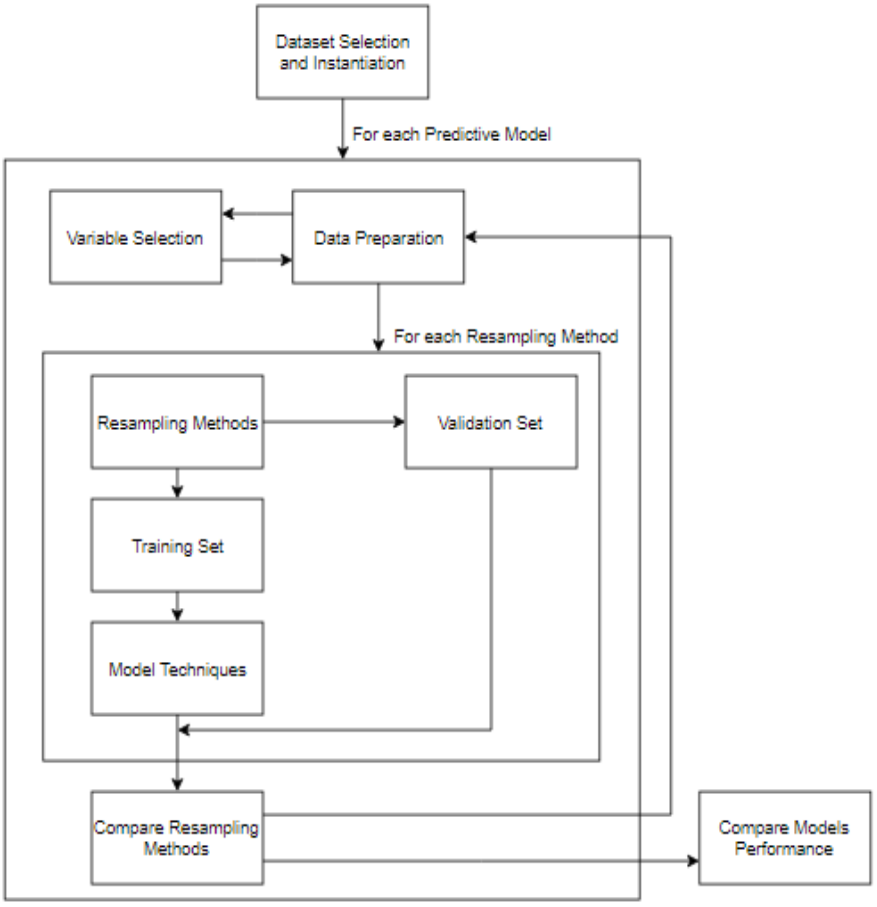


FIGURE 3.3: Design and Development Methodology Diagram

Figure 3.3 represents the overall diagram for the design and development methodology that is going to be followed throughout the demonstration and evaluation phase of the dissertation. This diagram represent the different steps followed as well as the order at which they were used. This methodology, adopted in order to develop the artifacts, is comprised of three distinct steps in order to allow for the assessment of the results provided and to allow for their comparison.

And so, the first of this methodology is the selection of which datasets to use for each of the corresponding response variables as well as the preparation of the corresponding data and selection of the combination of input variables to use (the predictors). As previously mentioned, the number of datasets for any of the response variables is varied and their corresponding sensors are often placed in different locations within the structure and can present different characteristics and so, in order to develop a viable and unambiguous comparison and evaluation

of the performance for each of the different predictive models, a first iteration on the datasets is done, where the MLR model is applied to the datasets with the $cos(d) + sin(d) + h^4$ combination of variables, which was chosen, due to being the combination that is currently being used by LNEC for their predictive analyses. The datasets that return the highest value for the $R^2_{adj}$ metric for each of the response variables are then those that are going to be selected. It was chosen the $R^2_{adj}$ metric for this "evaluation" due to its ability to provide a higher confidence of the model as explained in 2.4.1.2. This first step of the methodology is demonstrated in Figure 3.4.

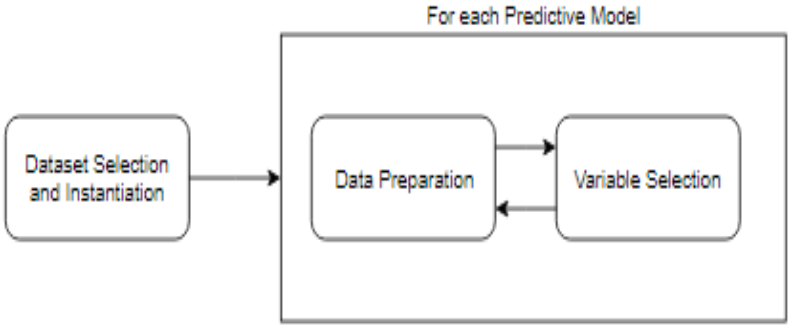


FIGURE 3.4: Development Methodology Diagram (step one)

The second step of the methodology is the separation of each of the previously selected datasets into training and testing sets using the Hold-Out re-sampling method (refer to 2.4.2.1). Each of these subsets is comprised of the combination of input variables for the model as well as the output variable to be predicted. Depending on the response variable and dataset, the number of elements contained in the training and testing sets can also be varied, either due to how long the instrument has been in use or due to the frequency of its measurements which is also different depending on the reachability of the instrument within the structure. For this comparative analysis and evaluation, the training sets include the first 80% of the records and the testing sets, the last 20%. This second step of the methodology is demonstrated in Figure 3.5.

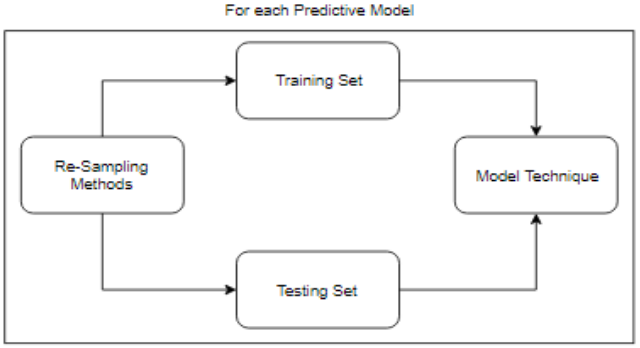


FIGURE 3.5: Development Methodology Diagram (step two)

The third step of the methodology is the analysis of predictive models. Each of the models is trained using the training sets referred in the previous step. To obtain the corresponding results, the models are then tested using the corresponding testing sets containing the data that has not yet been introduced to the model to estimate their accuracy. This step is subdivided into two phases where the first is the analyses of the Baseline model as well as the other predictive models, and the second is a comparative analysis between all of the predictive models considered (MLR, RR, PCR and NN). An analysis is done for each of the following combination of predictors variables: $h^4$, $cos(d) + sin(d)$, $h^4 + t$, $cos(d) + sin(d) + h^4$ and $cos(d) + sin(d) + h^4 + t$. The variable $t$ alone was not considered for the analysis due to its ever-increasing nature which is represented by the number of days since the structure started functioning. The $cos(d) + sin(d)$ variables are always considered in combination because of their representation of how temperature behaves, as explained in 2.5.1, in this case considering the geographical location of our case study, Portugal, which nearly follows the sinusoidal shape over the years. To evaluate the results of the models the following metrics are considered, that have been defined and presented in Chapter 2 (refer to 2.4.1): $MSE$, $RMSE$, $MAE$, $R^2$, $R^2_{adj}$ and $R$. After the comparison and evaluation of the different predictive models it is assessed if the improvement of the predictive models that have presented results. These results demonstrate how the models are behaving and which are able of being improved further, mainly using other re-sampling methods like the K-Fold CV or the Rolling-Origin CV rather than the Hold-Out method for the separation of the dataset. This third step of the methodology is demonstrated on Figure 3.6.

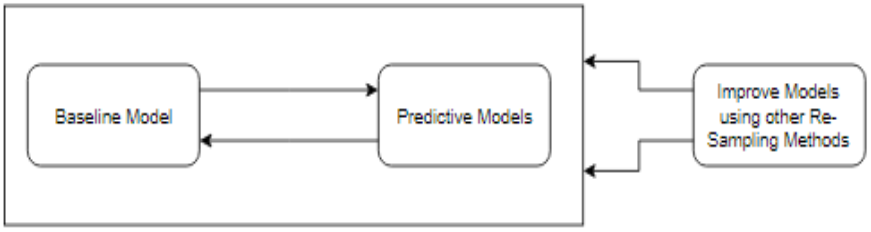


FIGURE 3.6: Development Methodology Diagram (step three)

For a better visualization and interpretation of the results, the data has been standardized to fit the plot and for the representation of the resulting metrics. This has been done, mainly due to the difference of scales in some variables, where they can either go from -15 to 15 (mm) in Radial Displacements as well as from -0,05 to 0,15 (mm) in Slippages for instance. This sort of standardization is done only for the visualization and error metrics, because if done to the prediction, the importance of the variables that are being weighted to have a higher correlation effect like the $h^4$ variable would be lost. This is not the case for NN were the data had to be standardized from the beginning before being used in the model.

## 3.3 Development Language

To analyze the different predictive models, a development language must be selected. According to the specifications that LNEC provided, the only restriction that has been imposed is that the language or platform must be open-source or in other words, not proprietary. LNEC is currently using *R* to do their data analyses and is considered to be a useful language to use in a data analysis setting. One other language that would be equally capable to deal with these types of problems would be *Python*. In comparison to *Python*, *R* functionality is developed with statistics and graphical models in mind while Python is a general-purpose type solution. *R* has several advantages over other languages, like greater features for data visualization which are a great help when dealing with predictive models, a huge user contributed documentation adopted more and more by and towards scientist, researchers and statisticians, and others.

# Chapter 4

# Demonstration and Evaluation

In this Chapter of the dissertation is presented the Demonstration and Evaluation phases of the DSRM as expressed in Figure 4.1, the different predictive methods where applied to the different artifacts that have been proposed in the Design and Development phase. To conclusively evaluate and define the success of the results, the metrics expressed in section 2.4.1 were used. This Chapter is structured as follows:

- **Section 4.1**: In this section it is demonstrated and evaluated the instantiation artifact for the Baseline model;
- **Section 4.2**: In this section it is demonstrated and evaluated the performance of the model artifact for the Predictive methods apart from the MLR, and the method artifact for comparing the different predictive methods with the baseline as well as among each other;
- **Section 4.3**: In this section it is demonstrated and evaluated the hypotheses of applying other re-sampling methods to the datasets, to evaluate if the accuracy of the models improve;
- **Section 4.4**: In this section is summarized the application of the different predictive methods to the case study of a real portuguese dam.

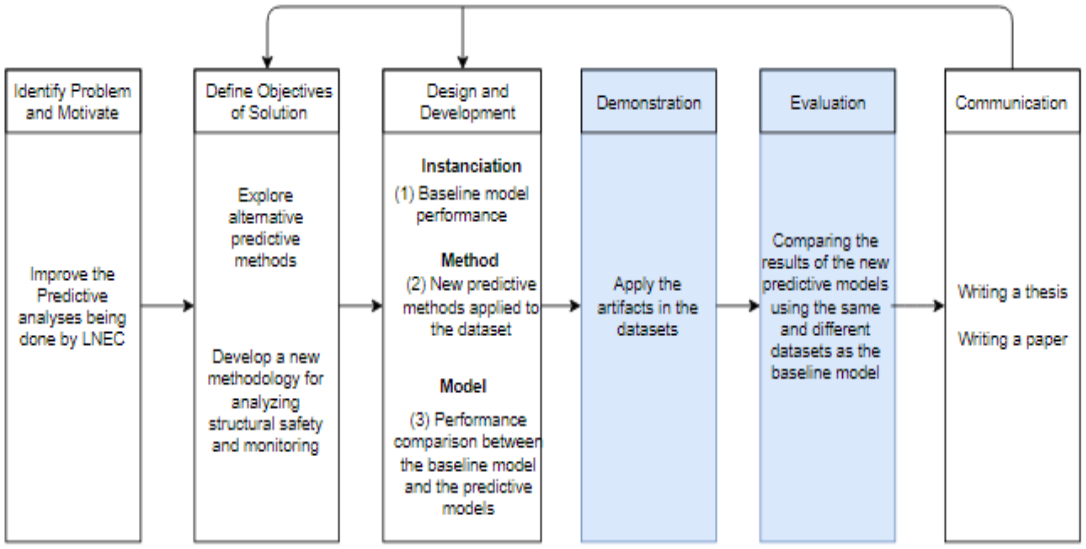


FIGURE 4.1: Adaptation Design Science Research Methodology - Demonstration and Evaluation (extracted from (Peffers et al., 2007))

## 4.1 Baseline

LNEC is currently applying MLR models for predicting the structural behavior (responses) of several structures, especially concrete dams. After using their technology for a few years LNEC has two prominent and different goals when using their data on analyzing and monitoring their structures or for investigation purposes:

1. To apply a predictive analysis, through MLR, to determine the accuracy of the models in order to estimate the structural responses of their dams;
2. To provide comparisons between the data being gathered manually and the data being gathered automatically by the embedded sensors to monitor and detect errors in their measurements in order to ensure that data is being gathered according to dam engineering practices and that there are no problems with the structure.

The application of the Baseline models to the datasets as well as the analysis of its performance, is used to prove the usefulness in understanding what is currently being done, in this case, on a predictive setting. With this analysis it will be clear what are the existing bottlenecks and possible flaws with the current implementation, as well as improvements needed to be made. The use of different input

predictor variable combinations on the same datasets provides an understanding about the correlation between each of the predictors to the response variables being analyzed as well as the effect they have on the structure. This analyses is done through the use of the evaluation metrics which will provide the resulting accuracy and confidence of each prediction.

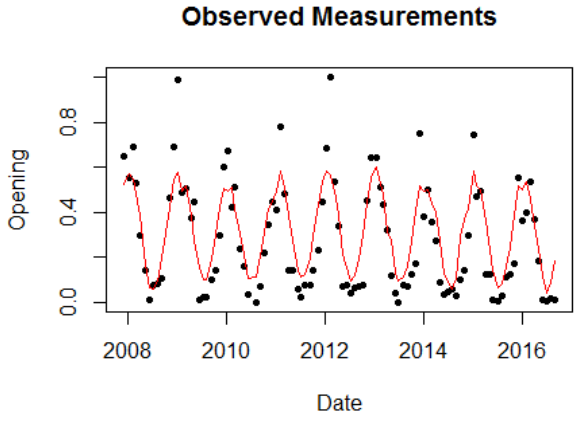


FIGURE 4.2: Opening variable for the $cos(d)+sin(d)+h^4$ predictors combination

TABLE 4.1: Metrics for the Opening Response

| **Response** | **Predictors** | $MSE$ | $RMSE$ | $MAE$ | $R^2$ | $R^2_{Adj}$ | $R$ |
|---|---|---|---|---|---|---|---|
| Opening | $h^4$ | 0,06005 | 0,24506 | 0,21123 | 0,00198 | -0,00781 | 0,04449 |
| | $cos(d) + sin(d) + h^4$ | 0,01376 | 0,11731 | 0,08496 | 0,81119 | 0,80934 | 0,90066 |
| | $cos(d) + sin(d) + t$ | 0,01765 | 0,13284 | 0,10573 | 0,76521 | 0,76291 | 0,87476 |
| | $h^4 + t$ | 0,06135 | 0,24769 | 0,21545 | 0,0000 | -0,00980 | -0,00168 |
| | $cos(d) + sin(d)$ | 0,01627 | 0,12755 | 0,09517 | 0,77095 | 0,76870 | 0,87804 |
| | $cos(d) + sin(d) + h^4 + t$ | 0,01601 | 0,12654 | 0,10131 | 0,80655 | 0,80465 | 0,89808 |

From Table 4.1 it is noticeable the differences in the correlation coefficient ($R$) metric between the $h^4$ and the $h^4 + t$ predictors, where the negative value of $h^4 + t$ shows that this combination of variables impacts the model negatively, thus not presenting any sort of relationship or dependency whatsoever to the Opening variable. The $h^4$ variable alone presents little to no relationship to the Opening variable. These two predictors combinations lack the presence of the variations in temperature ($cos(d) + sin(d)$) which for the Opening variable seem to represent most of its correlation, of nearly 88%. Just by using the $cos(d)+sin(d)$

combination, the model shows nearly 77% of goodness of fit as the $R^2_{Adj}$ metric implies, where both the $h^4$ and $h^4 + t$ impair the results. It is also noticeable that the combination of the variations in temperature and the water level ($cos(d) + sin(d) + h^4$) presents a small increase of the goodness of fit (81%), where by adding the effect of time ($cos(d) + sin(d) + h^4 + t$) it decreases, to 80%, which means that the variable $t$ also impacts the model in a negative way, as the predictors combination of $h^4 + t$ also demonstrated. Even though these variations of 1% in the $R^2_{Adj}$ metric are not that significant in terms of confidence in the model, when comparing the $MSE$ metric, the combination of predictors that gave a lower error rate estimate was the $cos(d) + sin(d) + h^4$. In summary, the best combination of predictors for the Opening variable that provide a higher goodness of fit based on the overall metrics for the MLR model is the $cos(d) + sin(d) + h^4$, presented in Figure 4.2.

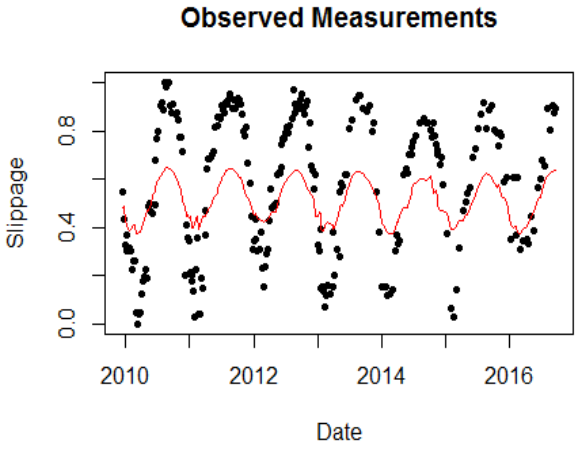


FIGURE 4.3: Slippage variable for the $cos(d) + sin(d) + h^4$ predictors combination

TABLE 4.2: Metrics for the Slippage Response

| **Response** | **Predictors** | $MSE$ | $RMSE$ | $MAE$ | $R^2$ | $R^2_{Adj}$ | $R$ |
|---|---|---|---|---|---|---|---|
| Slippage | $h^4$ | 0,07004 | 0,26466 | 0,23804 | 0,36335 | 0,36068 | 0,60270 |
| | $cos(d) + sin(d) + h^4$ | 0,04701 | 0,21681 | 0,19493 | 0,85914 | 0,85857 | 0,92690 |
| | $cos(d) + sin(d) + t$ | 0,16625 | 0,40774 | 0,35515 | 0,69319 | 0,69196 | 0,83258 |
| | $h^4 + t$ | 0,19139 | 0,43748 | 0,37229 | 0,11571 | 0,11215 | 0,34016 |
| | $cos(d) + sin(d)$ | 0,04819 | 0,21953 | 0,19702 | 0,83617 | 0,83551 | 0,91442 |
| | $cos(d) + sin(d) + h^4 + t$ | 0,16577 | 0,40715 | 0,35441 | 0,69827 | 0,69705 | 0,83562 |

The analysis of Table 4.2 is done in a similar way to that of Table 4.1 where the $h^4$ and $h^4 + t$ predictors show the least correlation to the Slippage variable, but for this response variable, the correlation of the $h^4$ variable to the model represents a 64% correlation, whereas for the Opening variable it represented only 4%. Even though it correlates 64% to the model, the $h^4$ variable alone only represents 36% of the variability of the model which means that alone is not a good variable for predicting future outcomes. The effect of the variations in temperature $(cos(d) + sin(d))$ represent an even higher correlation $(R)$ than for the Opening variable with a 91% correlation with nearly 84% of the variability of the model. This means that the Slippage response is strongly dependent of the variations of temperature. The combination of the $cos(d) + sin(d) + h^4$ seems to account for the most part of the variability of the model with 86%, where the effect of time seems to have a negative impact on the goodness of fit of the model reducing the $R^2_{Adj}$ by nearly 20%. And so, for this response variable, the combination of predictors that shows a higher $R^2_{Adj}$ is $cos(d) + sin(d) + h^4$, as presented in Figure 4.3.

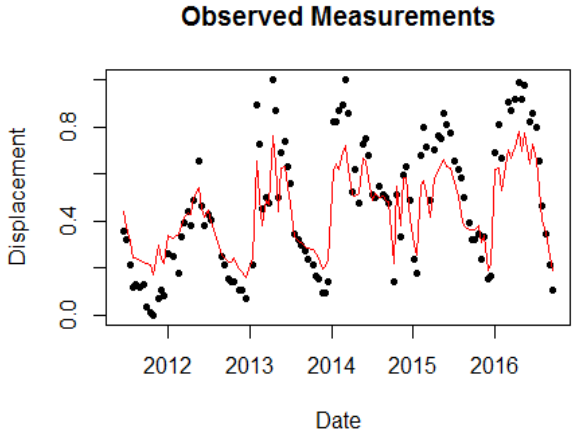


FIGURE 4.4: Displacement variable for the $cos(d) + sin(d) + h^4$ predictors combination

TABLE 4.3: Metrics for the Displacement Response

| **Response** | **Predictors** | $MSE$ | $RMSE$ | $MAE$ | $R^2$ | $R^2_{Adj}$ | $R$ |
|---|---|---|---|---|---|---|---|
| Displacement | $h^4$ | 0,02473 | 0,15726 | 0,13740 | 0,80219 | 0,80055 | 0,89565 |
| | $cos(d) + sin(d) + h^4$ | 0,01455 | 0,12061 | 0,09773 | 0,95225 | 0,95186 | 0,97583 |
| | $cos(d) + sin(d) + t$ | 0,04184 | 0,20455 | 0,17746 | 0,56504 | 0,56144 | 0,75169 |
| | $h^4 + t$ | 0,03324 | 0,18233 | 0,14693 | 0,79796 | 0,79629 | 0,89329 |
| | $cos(d) + sin(d)$ | 0,04325 | 0,20797 | 0,17857 | 0,49175 | 0,48755 | 0,70125 |
| | $cos(d) + sin(d) + h^4 + t$ | 0,02169 | 0,14727 | 0,12605 | 0,95469 | 0,95431 | 0,97708 |

The analysis of Table 4.3, to predict the Displacement response, it is noticeable that in this setting, the effect of time ($t$) has a smaller impact in the model when comparing the values from the previously seen on Table 4.1 and 4.2. The effect of the water level ($h^4$) variable has a correlation of nearly 90% for the response where the $cos(d) + sin(d)$ variable only has 70% which means that this response variable is strongly dependent of the H4 variable. And so, because the impact on the model by the effect of time is not that significant, the combination of variables that represent the highest confidence of the model is the $cos(d) + sin(d) + h^4 + t$, but by looking at the $MSE$ metric this combination gives a higher error rate than the $cos(d) + sin(d) + h^4$ which means that, in this setting, the combination of predictors that have the best goodness of fit is the $cos(d) + sin(d) + h^4$, presented in Figure 4.4, where its $R^2_{Adj}$ is 95% and its $MSE$ is 0,015.

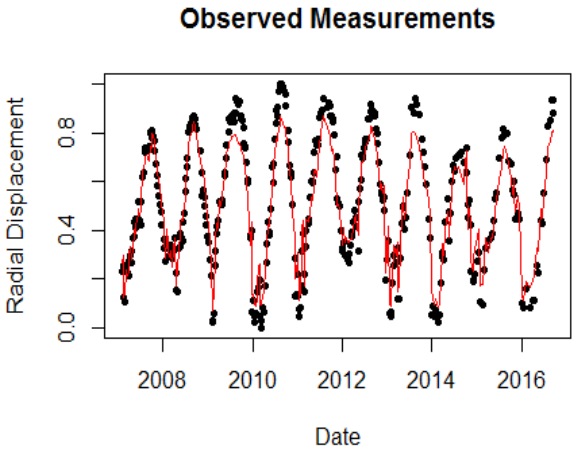


FIGURE 4.5: Radial Displacement variable for the $cos(d)+sin(d)+h^4$ predictors combination

TABLE 4.4: Metrics for the Radial Displacement Response

| **Response** | **Predictors** | $MSE$ | $RMSE$ | $MAE$ | $R^2$ | $R^2_{Adj}$ | $R$ |
|---|---|---|---|---|---|---|---|
| Radial Displacement | $h^4$ | 0,03567 | 0,18885 | 0,16755 | 0,52233 | 0,52106 | 0,72272 |
| | $cos(d) + sin(d) + h^4$ | 0,00448 | 0,06690 | 0,05553 | 0,96996 | 0,96988 | 0,98486 |
| | $cos(d) + sin(d) + t$ | 0,03255 | 0,18042 | 0,15725 | 0,80699 | 0,80645 | 0,89832 |
| | $h^4 + t$ | 0,04774 | 0,21849 | 0,18195 | 0,51029 | 0,50899 | 0,71434 |
| | $cos(d) + sin(d)$ | 0,01561 | 0,12496 | 0,10459 | 0,80279 | 0,80226 | 0,89598 |
| | $cos(d) + sin(d) + h^4 + t$ | 0,01626 | 0,12753 | 0,10816 | 0,96547 | 0,96538 | 0,98258 |

From the analysis of Table 4.4, the base combination that presents the most correlation with the response variable, the predictors with the higher correlation, is the variations in temperature $(cos(d) + sin(d))$ of nearly 90%, where the variations in the water level account for 72%. Through the comparison of the $h^4$ and $h^4 + t$ combination we can see that the effect of time impacts the model negatively. The variables combinations of $cos(d) + sin(d) + h^4$ and $cos(d) + sin(d) + h^4 + t$ present similar correlation to the model of approximately 98% and similar variability, of nearly 97% but the combination that presents the lowest $MSE$ is the $cos(d) + sin(d) + h^4$ variable, presented in Figure 4.5.

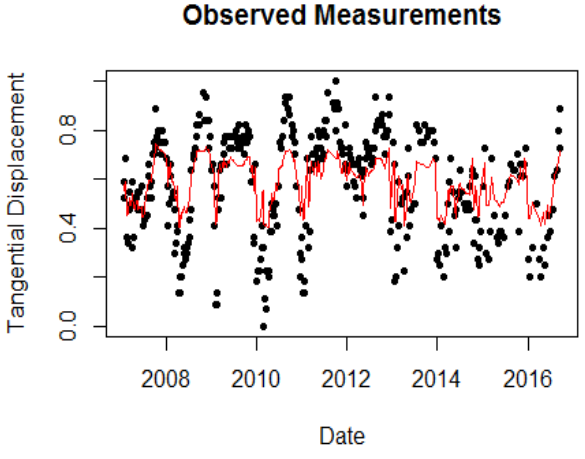


FIGURE 4.6: Tangential Displacement variable for the $cos(d) + sin(d) + h^4$ predictors combination

TABLE 4.5: Metrics for the Tangential Displacement Response

| **Response** | **Predictors** | $MSE$ | $RMSE$ | $MAE$ | $R^2$ | $R^2_{Adj}$ | $R$ |
|---|---|---|---|---|---|---|---|
| Tangential Displacement | $h^4$ | 0,01778 | 0,13335 | 0,10916 | 0,86445 | 0,86410 | 0,92976 |
| | $cos(d) + sin(d) + h^4$ | 0,01709 | 0,13072 | 0,10702 | 0,87232 | 0,87199 | 0,93398 |
| | $cos(d) + sin(d) + t$ | 0,03624 | 0,19037 | 0,16537 | 0,33169 | 0,32992 | 0,57593 |
| | $h^4 + t$ | 0,01847 | 0,13592 | 0,11453 | 0,85654 | 0,85616 | 0,92549 |
| | $cos(d) + sin(d)$ | 0,03482 | 0,18659 | 0,15596 | 0,32082 | 0,31902 | 0,56641 |
| | $cos(d) + sin(d) + h^4 + t$ | 0,01782 | 0,13347 | 0,11298 | 0,86675 | 0,86640 | 0,93100 |

From the analysis of Table 4.5 it is noticeable from the $R$ metric that the $h^4$ predictor alone accounts for nearly 93% of the correlation with the model. For the Tangential Displacement response variable, the effect of time does not appear to have significant impact as it did on others. The combination of the $h^4$ variable with the effects of the temperature, the $cos(d) + sin(d)$, accounts for slightly over than 93% of the correlation with over 87% of variability. Taking also into account that this combination presents the lowest $MSE$ value of 0,017, then the variable that provides the best goodness of fit of the model is the $cos(d) + sin(d) + h^4$ predictors combination, as presented in Figure 4.6.

## 4.2 Predictive Methods

For the new predictive methods that have been considered and explained in 2.3, the RR, PCR and the NN, the applied methodology was the same as the one applied to the Baseline model as well as the considered datasets.

To allow for a better comparison between all of the obtained results. For these new considered predictive methods the following Tables will only include the best combination of predictors and their comparison with the Baseline's best combination as well for the corresponding response variable.

### 4.2.1 Ridge Regression

The RR model is nearly the same regression model as the MLR but, as explained before, where the difference lies in applying a $\lambda$ variable that penalizes the coefficients, allowing for a controlled shrinkage of the model (refer to 2.3.2). The focus of using RR in this type of environment is to increase the accuracy of the MLR model in datasets that present a great number of outliers where, because of the penalty, the intent is to smooth the model to overcome these outliers.

Since RR is almost an identical model to MLR and since the datasets that have been chosen were those that presented the highest confidence for each of the response variables, unless there are very correlated variables or outliers, it is plausible that the results that are presented in the MLR setting are identical to the results presented in the RR setting. The $\lambda$ variable is calculated automatically for each of the coefficients, thus presenting the best results for each of the response variables.

Table 4.6: Metrics for the Response Variables for comparing MLR and RR

| **Response** | **Model** | **Predictors** | $MSE$ | $RMSE$ | $MAE$ | $R^2$ | $R^2_{Adj}$ | $R$ |
|---|---|---|---|---|---|---|---|---|
| Opening | MLR | $cos(d) + sin(d) + h^4$ | 0,01376 | 0,11731 | 0,08496 | 0,81119 | 0,80934 | 0,90066 |
| | RR | $cos(d) + sin(d) + h^4$ | 0,01382 | 0,11758 | 0,08522 | 0,81107 | 0,80922 | 0,90006 |
| Slippage | MLR | $cos(d) + sin(d) + h^4$ | 0,04701 | 0,21681 | 0,19493 | 0,85914 | 0,85857 | 0,92690 |
| | RR | $cos(d) + sin(d) + h^4$ | 0,04743 | 0,21778 | 0,19583 | 0,85869 | 0,85812 | 0,92666 |
| Displacement | MLR | $cos(d) + sin(d) + h^4$ | 0,01455 | 0,12061 | 0,09773 | 0,95225 | 0,95186 | 0,97583 |
| | RR | $cos(d) + sin(d) + h^4$ | 0,01463 | 0,12096 | 0,09804 | 0,95232 | 0,95193 | 0,97587 |
| Radial Displacement | MLR | $cos(d) + sin(d) + h^4$ | 0,00448 | 0,06690 | 0,05553 | 0,96996 | 0,96988 | 0,98486 |
| | RR | $cos(d) + sin(d) + h^4$ | 0,00448 | 0,06693 | 0,05555 | 0,96995 | 0,96988 | 0,98486 |
| Tangential Displacement | MLR | $cos(d) + sin(d) + h^4$ | 0,01709 | 0,13072 | 0,10702 | 0,87232 | 0,87199 | 0,93398 |
| | RR | $cos(d) + sin(d) + h^4$ | 0,01711 | 0,13080 | 0,10709 | 0,87242 | 0,87208 | 0,93403 |

In Table 4.6, as expected, the values appear to be identical since the datasets that have been chosen presented the highest $R^2_{Adj}$ for MLR, where the best combination of predictors remained the same, the $cos(d) + sin(d) + h^4$. It appears that RR did not clearly outperform MLR for any of the response variables, and what it can be taken from these results is that these combinations of variables are not correlated and the datasets did not contained much outliers. Despite RR worsening the models, it remains the fact that the results are practically identical which gives us the possibility to rely on a "smoothed" model (RR) rather in one that does not provide the possibility to penalize abnormalities. Figures 4.7 to 4.11 demonstrate the similarities of the each of the models using RR with their counterpart using MLR.

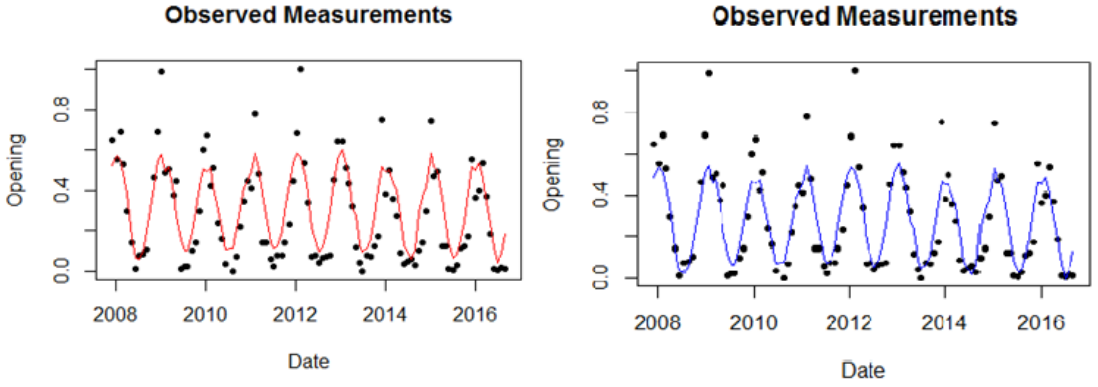


FIGURE 4.7: Opening variable comparison from MLR (on the left) and RR (on the right)

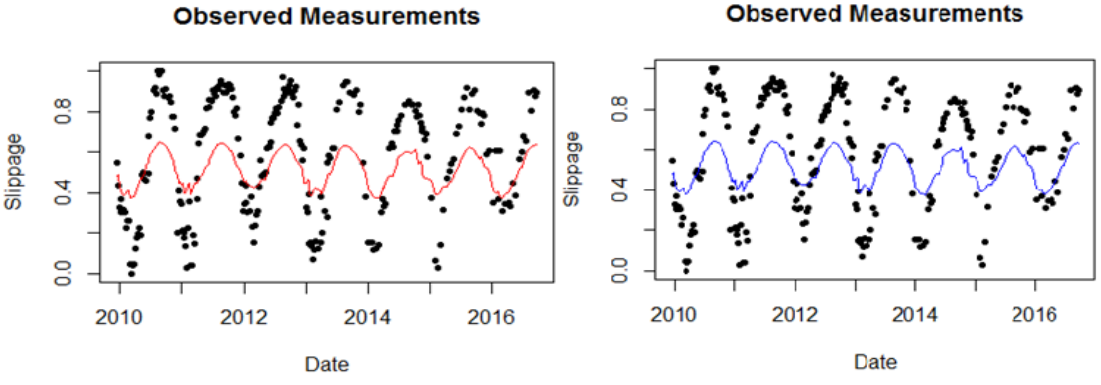


FIGURE 4.8: Slippage variable comparison from MLR (on the left) and RR (on the right)

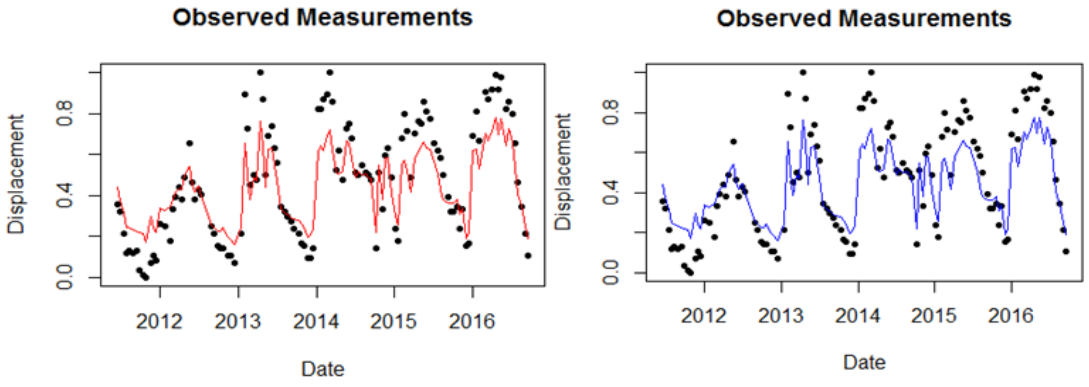


FIGURE 4.9: Displacement variable comparison from MLR (on the left) and RR (on the right)

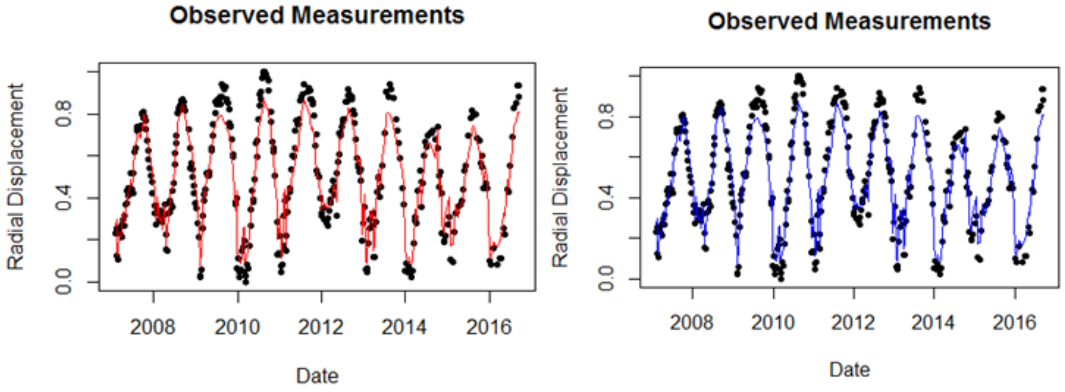


FIGURE 4.10: Radial Displacement variable comparison from MLR (on the left) and RR (on the right)

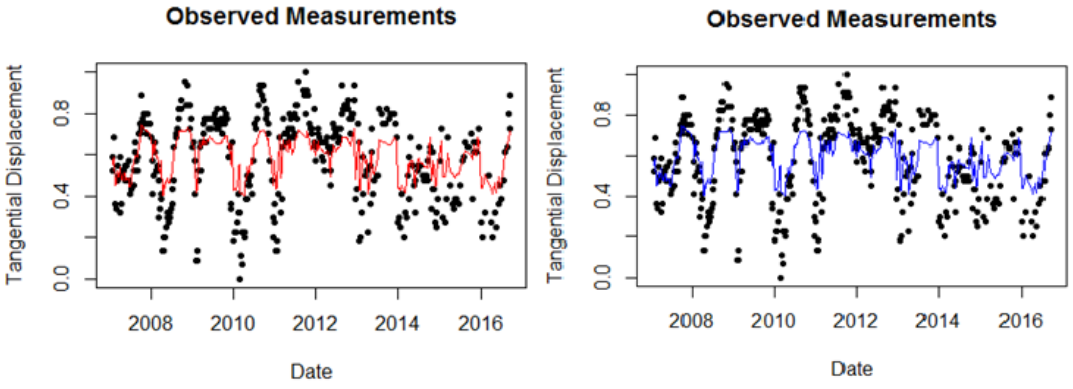


FIGURE 4.11: Tangential Displacement variable comparison from MLR (on the left) and RR (on the right)

### 4.2.2 Principal Component Regression

PCR, unlike the RR or the MLR model, applies dimensionality reduction using the PCA (see 2.2.1.4), where the RR only applies a penalty on the coefficients. PCR, as mentioned before (refer to 2.3.3), applies the resulting PC from the PCA to the model, thus preventing any redundant or non-important information to be predicted. The technique of applying the dimensionality reduction ability of the PCA before the model is a very common practice especially if the amount of input variables is high, which, in this setting, are mainly the combinations of $cos(d) + sin(d) + h^4$ and $cos(d) + sin(d) + h^4 + t$. And so, it is expected that single variable combinations would give nearly the same results as without the use of PCA. And so, it should be visible which variables are really going to have an impact on the models, in other words, which variables will give the most information about the model.

Table 4.7: Metrics for the Response Variables for comparing MLR and PCR

| **Response** | **Model** | **Predictors** | $MSE$ | $RMSE$ | $MAE$ | $R^2$ | $R^2_{Adj}$ | $R$ |
|---|---|---|---|---|---|---|---|---|
| Opening | MLR | $cos(d) + sin(d) + h^4$ | 0,01376 | 0,11731 | 0,08496 | 0,81119 | 0,80934 | 0,90066 |
| | PCR | $cos(d) + sin(d)$ | 0,05496 | 0,23444 | 0,19827 | 0,10163 | 0,09282 | 0,31880 |
| Slippage | MLR | $cos(d) + sin(d) + h^4$ | 0,04701 | 0,21681 | 0,19493 | 0,85914 | 0,85857 | 0,92690 |
| | PCR | $cos(d) + sin(d)$ | 0,05762 | 0,24003 | 0,21621 | 0,59819 | 0,59657 | 0,77343 |
| Displacement | MLR | $cos(d) + sin(d) + h^4$ | 0,01455 | 0,12061 | 0,09773 | 0,95225 | 0,95186 | 0,97583 |
| | PCR | $cos(d) + sin(d) + h^4$ | 0,02473 | 0,15726 | 0,13740 | 0,80219 | 0,80055 | 0,89565 |
| Radial Displacement | MLR | $cos(d) + sin(d) + h^4$ | 0,00448 | 0,06690 | 0,05553 | 0,96996 | 0,96988 | 0,98486 |
| | PCR | $cos(d) + sin(d) + h^4$ | 0,03567 | 0,18885 | 0,16755 | 0,52233 | 0,52106 | 0,72272 |
| Tangential Displacement | MLR | $cos(d) + sin(d) + h^4$ | 0,01709 | 0,13072 | 0,10702 | 0,87232 | 0,87199 | 0,93398 |
| | PCR | $cos(d) + sin(d)$ | 0,01778 | 0,13335 | 0,10916 | 0,86445 | 0,86410 | 0,92976 |

In Table 4.7, it seems that for the Opening, Slippage and Tangential Displacement response variables, the combination that got the best goodness of fit was the $cos(d) + sin(d)$ predictors combination, which entails that the variation of temperature is the variable that gives the most correlation and most information to

the model. For the Displacement and Radial Displacement responses the combination of predictors that gave the best goodness of fit was the $cos(d) + sin(d) + h^4$ combination. We can see that for every response variable tested, the PCR model gave a worst result than the MLR model. Figures 4.12 to 4.16 demonstrate the similarities of the each of the models using PCR with their counterpart using MLR.

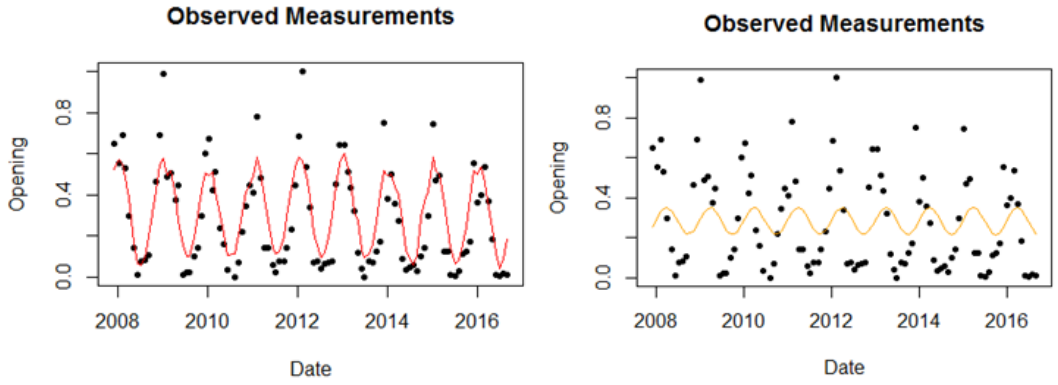


FIGURE 4.12: Opening variable comparison from MLR (on the left) and PCR (on the right)

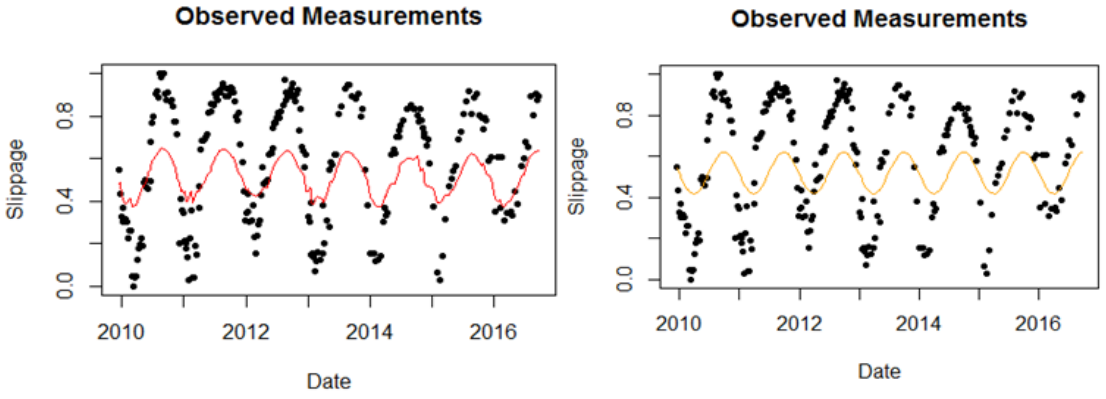


FIGURE 4.13: Slippage variable comparison from MLR (on the left) and PCR (on the right)

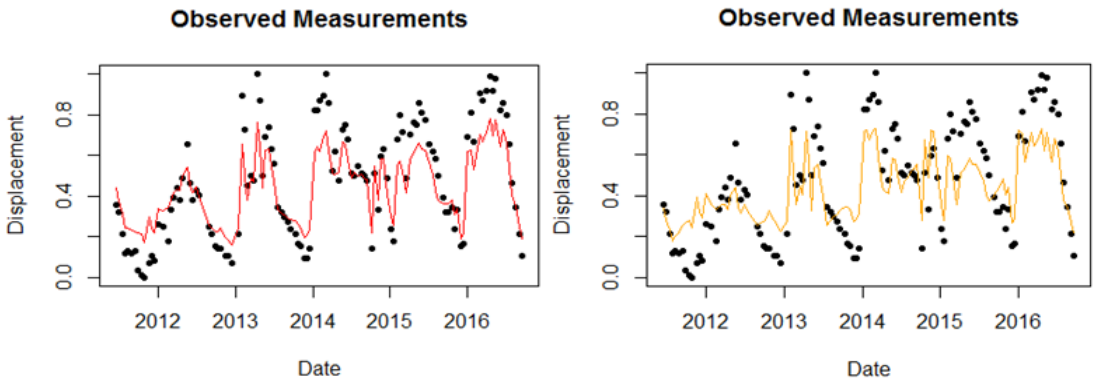


FIGURE 4.14: Displacement variable comparison from MLR (on the left) and PCR (on the right)

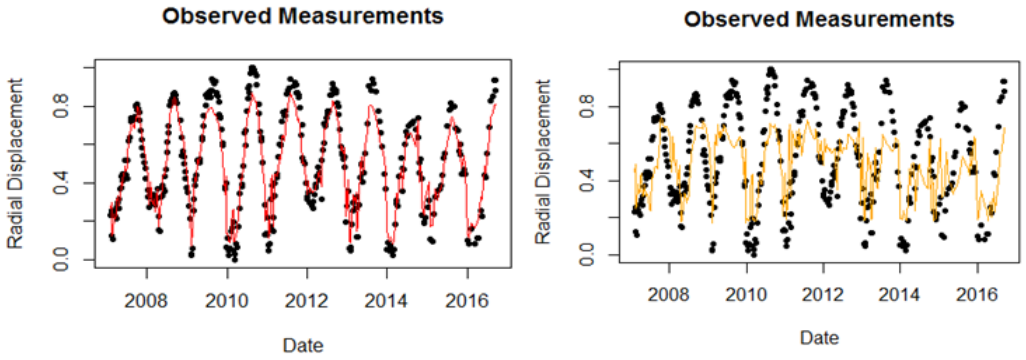


FIGURE 4.15: Radial Displacement variable comparison from MLR (on the left) and PCR (on the right)

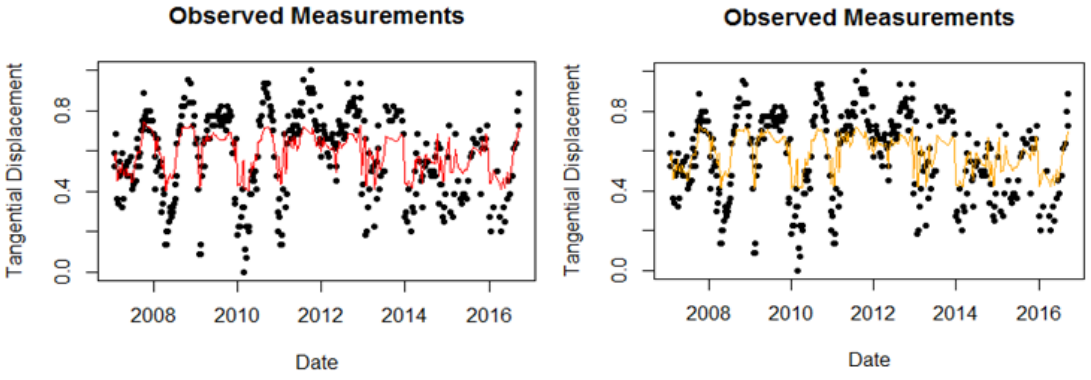


FIGURE 4.16: Tangential Displacement variable comparison from MLR (on the left) and PCR (on the right)

### 4.2.3 Neural Networks

The NN model, contrarily to the previously applied Regression models, needs the input data to be normalized in order to obtain appropriate results that are within the domain or context of the prediction. For this analysis, the NN model was parameterized with: $N$ number of nodes for the input layer and $N * 2 - 1$ number of nodes for the hidden layer and 1 node for the output layer, where $N$ is equal to the number of elements of the combination of predictors variables. The output of this model needs to be a linear output, because the input values are unbounded.

TABLE 4.8: Metrics for the Response Variables for comparing MLR and NN

| **Response** | **Model** | **Predictors** | $MSE$ | $RMSE$ | $MAE$ | $R^2$ | $R^2_{Adj}$ | $R$ |
|---|---|---|---|---|---|---|---|---|
| Opening | MLR | $cos(d) + sin(d) + h^4$ | 0,01376 | 0,11731 | 0,08496 | 0,81119 | 0,80934 | 0,90066 |
| | NN | $cos(d) + sin(d)$ | 0,01517 | 0,12316 | 0,09532 | 0,79066 | 0,78861 | 0,88919 |
| Slippage | MLR | $cos(d) + sin(d) + h^4$ | 0,04701 | 0,21681 | 0,19493 | 0,85914 | 0,85857 | 0,92690 |
| | NN | $cos(d) + sin(d) + h^4 + t$ | 0,02697 | 0,16422 | 0,13070 | 0,74544 | 0,74442 | 0,86339 |
| Displacement | MLR | $cos(d) + sin(d) + h^4$ | 0,01455 | 0,12061 | 0,09773 | 0,95225 | 0,95186 | 0,97583 |
| | NN | $h^4$ | 0,02440 | 0,15621 | 0,13436 | 0,81396 | 0,81240 | 0,90218 |
| Radial Displacement | MLR | $cos(d) + sin(d) + h^4$ | 0,00448 | 0,06690 | 0,05553 | 0,96996 | 0,96988 | 0,98486 |
| | NN | $cos(d) + sin(d) + h^4 + t$ | 0,02750 | 0,16583 | 0,13698 | 0,89137 | 0,89109 | 0,94413 |
| Tangential Displacement | MLR | $cos(d) + sin(d) + h^4$ | 0,01709 | 0,13072 | 0,10702 | 0,87232 | 0,87199 | 0,93398 |
| | NN | $h^4$ | 0,01844 | 0,13579 | 0,11055 | 0,85404 | 0,85366 | 0,92414 |

In Table 4.8 the NN showed similar results to the MLR, even improving by half, the results for the $MSE$ metric for the Slippage response variable. Considering that these datasets were chosen as the best datasets (with the highest $R^2_{Adj}$) where MLR had been applied, it is expected that with other datasets where the MLR presents bad results, NN could provide improvements to the model. It is also noticeable, from the values of the $R^2_{Adj}$ metric, that even though $MSE$ provided better results, it did not improve the confidence on the model. This could probably be due to the randomness applied by the NN model, thus giving better error measurements but worse confidence. Figures 4.17 to 4.21 demonstrate the similarities of the each of the models using NN with their counterpart using MLR.

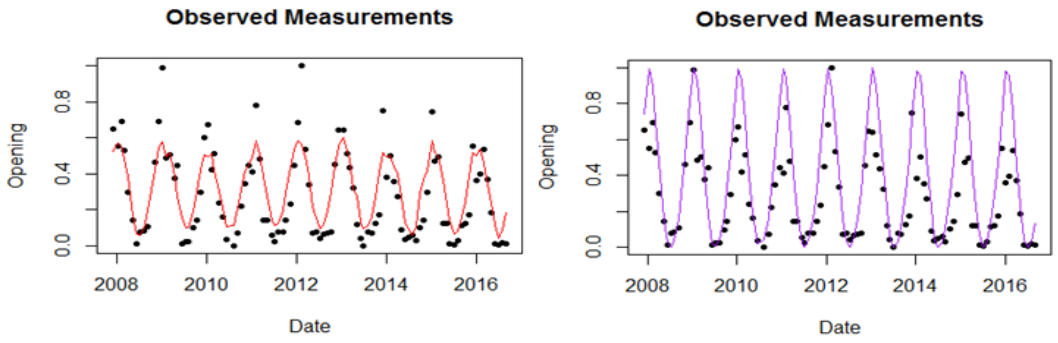


FIGURE 4.17: Opening variable comparison from MLR (on the left) and NN (on the right)

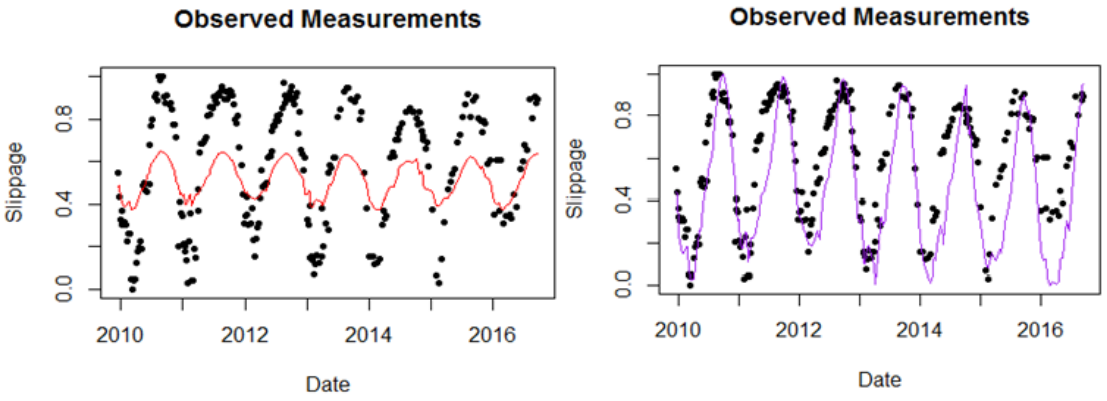


FIGURE 4.18: Slippage variable comparison from MLR (on the left) and NN (on the right)

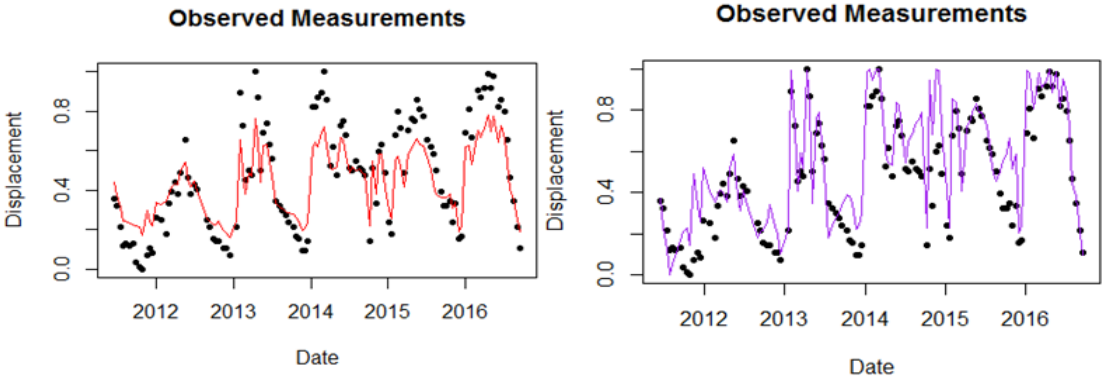


FIGURE 4.19: Displacement variable comparison from MLR (on the left) and NN (on the right)

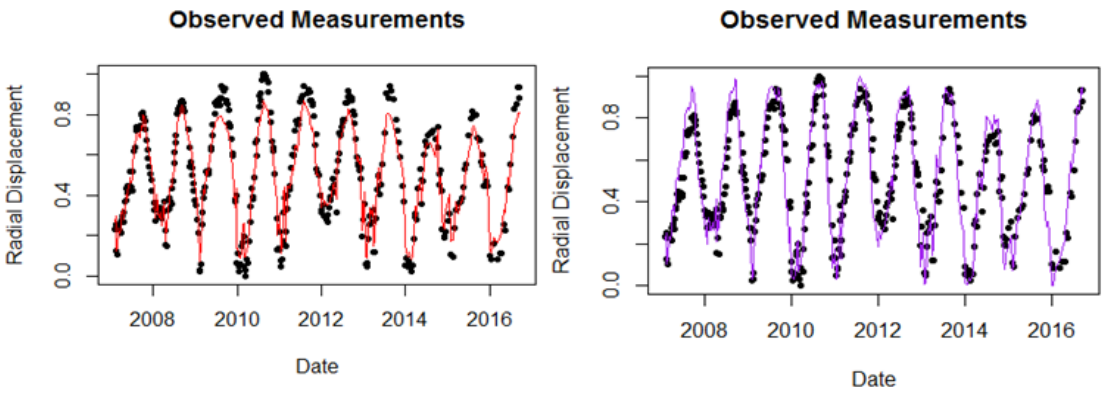


FIGURE 4.20: Radial Displacement variable comparison from MLR (on the left) and NN (on the right)

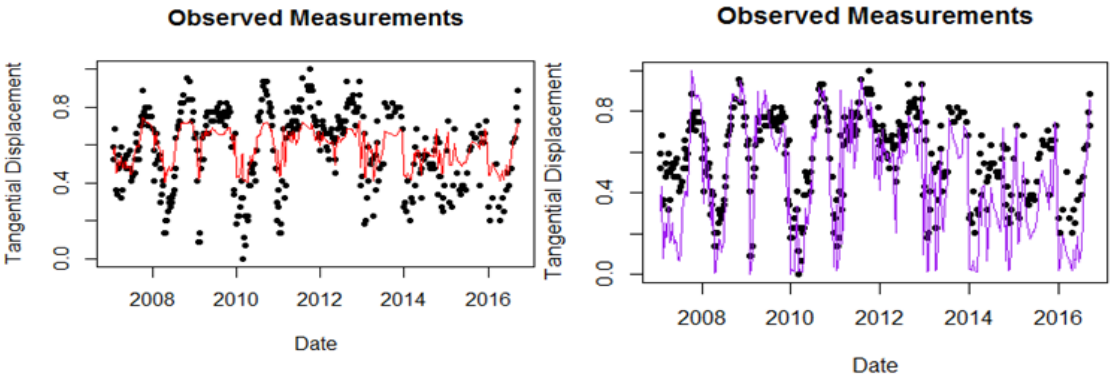


FIGURE 4.21: Tangential Displacement variable comparison from MLR (on the left) and NN (on the right)

## 4.3 Re-sampling Methods

To summarize, the predictive models that have shown promise for improvement are the MLR and the NN. The PCR model did not provide better results than MLR or NN due to the lack of input variables and lack of collinearity between them, but it did demonstrate which predictors are more correlated with each of the responses. The RR model also did not provide better results, but it could provide better results than the MLR if there were a greater number of outliers within the dataset. The usage of these models (RR and PCR) on engineering structures could prove beneficial depending on the structure they are used in, the

number of variables being gathered and the amount of discrepancies, i.e., outliers, existing on the data associated with those structures.

### 4.3.1 K-Fold Cross-Validation

K-Fold Cross-Validation, as previously mentioned in Chapter 2 (refer to 2.4.2.2), is a re-sampling method where the dataset is divided in $N$ number of samples, where $N$ - 1 of these samples are considered a part of the training set and the other single sample is considered a part of the testing set.

In the K-Fold Cross-Validation setting, for each of the datasets corresponding to each of the response variables, both the MLR and the NN models where tested using different number of samples (or "Folds"). The number of samples chosen were 5, 10 and 20. Since the samples are being generated with random elements, in other words, they are not being generated in a sequential order, these tests were repeated five times and the mean of the resulting metrics was calculated. The LOOCV variation was not considered due to the elevated number of records on each of the datasets and the number of models that would need to be generated. To demonstrate the number of models being generated, namely with the NN, which is more time consuming than MLR, for a number of samples of 20, five repetitions and a number of variable combinations of 5 ($h^4$, $cos(d)+sin(d)$, $cos(d)+sin(d)+h^4$, $h^4+t$ and $cos(d)+sin(d)+h^4+t$) we would get: 20*5*5 = 500 NN models. If the LOOCV were to be applied, considering a number of 1986 records for one response variable dataset, with the same parameters as before we would get: 1986*5*5 = 49650 NN models. Table 4.9 shows the best number of samples and the best combination of input variables for each of the MLR and NN models, for each of the response variables.

TABLE 4.9: Metrics for the Response Variables for comparing MLR and NN using the K-Fold Cross-Validation Re-Sampling method

| **Response** | **Model** | **Predictors** | **Samples** | $MSE$ | $MAE$ | $R^2_{Adj}$ |
|---|---|---|---|---|---|---|
| Opening | MLR | $cos(d) + sin(d) + h^4$ | 10 | 0,01166 | 0,07636 | 0,73480 |
| | NN | $cos(d) + sin(d) + h^4 + t$ | 10 | 0,00948 | 0,06665 | 0,78168 |
| Slippage | MLR | $cos(d) + sin(d) + h^4 + t$ | 10 | 0,00696 | 0,06731 | 0,81843 |
| | NN | $cos(d) + sin(d) + h^4 + t$ | 10 | 0,00208 | 0,03130 | 0,94590 |
| Displacement | MLR | $cos(d) + sin(d) + h^4 + t$ | 20 | 0,01283 | 0,07243 | 0,77290 |
| | NN | $cos(d) + sin(d) + h^4 + t$ | 10 | 0,00796 | 0,05829 | 0,84895 |
| Radial Displacement | MLR | $cos(d) + sin(d) + h^4 + t$ | 5 | 0,00213 | 0,03386 | 0,96422 |
| | NN | $cos(d) + sin(d) + h^4 + t$ | 10 | 0,00159 | 0,03051 | 0,97035 |
| Tangential Displacement | MLR | $cos(d) + sin(d) + h^4$ | 20 | 0,00583 | 0,04054 | 0,65093 |
| | NN | $cos(d) + sin(d) + h^4 + t$ | 10 | 0,00418 | 0,03311 | 0,73548 |

From Table 4.9, we can see that K-Fold CV behaved better than expected for either the NN and the MLR settings when comparing the $MSE$ metric with the previously results of the Baseline model (refer to Table 4.8). On the other hand, the confidence values of the $R^2_{Adj}$ metric shown a steep decrease mainly due to K-Fold CV using random sampling for splitting the data where order is an important factor, but despite that, it seems that NN shown improvements on all of the response variables even with unordered data.

### 4.3.2 Rolling-Origin Cross-Validation

K-Fold CV does not take into consideration the ordered historical values presented on the datasets, which are important to determine future outcomes or behavioral responses from the structure. And so, to overcome this issue, especially for MLR methods, a Rolling-Origin CV (ROCV) method was developed. The logic behind Rolling-Origin is presented in Chapter 2 (refer to 2.4.2.3) but to implement it, the following procedure has been developed:

1. The first step is to divide the datasets into years to find the maximum amount of years, excepting the final year, to use for the training set in order to minimize the error when using the last year as the testing set;

2. The second step is to implement the ROCV across the different datasets for each of the response variables with the number of years discovered, to obtain the mean of the resulting variables, thus providing for a more realistic approach of the use of this method from prediction.

To provide a viable evaluation the first and the last years were removed from the datasets because they could either begin or end in the middle of the year and so could provide false results, or since the case study is a Dam structure, the first year could refer to the filling of said structure and thus it could impair the results. The combination of predictors used for this analysis was the $cos(d)+sin(d)+h^4$ for the MLR method that, as seen so far, represents the best combinations of variables for this method. For the NN there was made a comparison to determine which of the combinations of variables and number of years would result in the lowest error, which resulted in the $cos(d) + sin(d) + h^4 + t$ combination of variables for each of the response variables.

TABLE 4.10: Metrics for the Response Variables using Rolling-Origin Cross-Validation

| **Response** | **Model** | **Predictors** | **Years used in Training** | $MSE$ | $MAE$ | $R^2_{Adj}$ |
|---|---|---|---|---|---|---|
| Opening | MLR | $cos(d)+sin(d)+h^4$ | 3 | 0,03701 | 0,13735 | 0,77230 |
| | NN | $cos(d) + sin(d) + h^4 + t$ | 13 | 0,05285 | 0,16047 | 0,76268 |
| Slippage | MLR | $cos(d)+sin(d)+h^4$ | 5 | 0,04058 | 0,14009 | 0,78064 |
| | NN | $cos(d) + sin(d) + h^4 + t$ | 10 | 0,08340 | 0,24568 | 0,71573 |
| Displacement | MLR | $cos(d)+sin(d)+h^4$ | 9 | 0,03224 | 0,14207 | 0,93401 |
| | NN | $cos(d) + sin(d) + h^4 + t$ | 3 | 0,06155 | 0,17257 | 0,73977 |
| Radial Displacement | MLR | $cos(d)+sin(d)+h^4$ | 11 | 0,01271 | 0,08730 | 0,97008 |
| | NN | $cos(d) + sin(d) + h^4 + t$ | 4 | 0,01239 | 0,08595 | 0,96585 |
| Tangential Displacement | MLR | $cos(d)+sin(d)+h^4$ | 12 | 0,02956 | 0,13839 | 0,90884 |
| | NN | $cos(d) + sin(d) + h^4 + t$ | 2 | 0,06356 | 0,18120 | 0,80196 |

The analysis of Table 4.10 demonstrates similarities of what has been seen so far, namely on the number of years used for training. For the MLR model it is seen that this model behaved better for the Opening and Slippage responses where the number of years was lower on these responses and quite a bit higher on the others, due to the low variability these response variables present. The metrics however demonstrated lower results over using K-Fold Cross Validation as a re-sampling technique. On the other hand, for the NN model, the number of years used in training were mostly contrary to those of the MLR. The metrics were also lower when compared to the K-Fold Cross Validation. These results were to be expected not to give higher results than other re-sampling methods due to the non existent randomness that the KFCV provided. However, it showed that, for each of the response variables, the amount of data needed for a good analysis of these types of structured can vary depending on the erratic measurements gathered by the instruments.

## 4.4 Summary

This Chapter intents to provide different perspectives of what could be done by applying different predictive models to the datasets and so, it has been purposed an Instantiation Artifact for the analysis of the Baseline model (MLR) and a Method Artifact corresponding to the New Predictive Methods applied to the Dataset.

The new predictive models that have been proposed to accomplish this where the RR, the PCR and the NN and were compared throughout. For the RR model, the results have enabled a higher understanding of the impact outliers pose in the datasets as well as for limiting possible existing collinearity between the predictor variables combinations using the penalty parameter (explained in 2.3.2). And so, taking this information into account it is clear that, by the comparison made in 4.2.1 as well as in Table 4.6, the combination of predictor variables served as input are not collinear between themselves nor are the datasets used outlier intensive. However, this is not always the case when predicting data from datasets that are mainly composed of measurements gathered manually which have a high chance of being of an outlier prone nature.

For the PCR model, the results allowed for the perception of understanding which were the PC, i.e., variables, that had the most impact, or correlation, with the response variables. The results on the other hand did not show improvements to the model which is due to the application of PCA before the application of

the predictive model which removes the clutter variables that do not present any correlation with the output but could still provide insights on the behavior of the response. The results on Table 4.7 could be expected because even though they do not present a high correlation to the output they still present a slow indication of what can be used to predict the responses.

For the NN model, the results in Table 4.8 showed some improvements where the randomness of this model was a key component. This model, by not taking into account the nature of the dataset, which is based on historical data, allowed for a different perspective on the application of random models on these types of structures and datasets.

By applying different re-sampling methods to the datasets as seen in Table 4.9, it was shown that the application of the K-Fold Cross Validation showed improvements for the NN model and not so much for the MLR, which again, the randomness of the model was key, where the MLR did not take well the changes provoked by the separation into folds and random attribution.

The ROCV method, presented in Table 4.10 showed worse results than either the Hold-Out method and the K-Fold Cross Validation but it demonstrated that, the bigger the datasets training size, the better the results, even though it showed promise for a few number of years, mainly due to the balanced gathering of the measurements and the nonexistence of outliers in the dataset.

# Chapter 5

# Conclusions and Future Work

Structural safety and monitoring has been and will continue to be a high study focused area in order to determine and achieve the most accurate and efficient ways of providing a near real-time approach to the analysis of the behavioral responses of engineering structures. Since their introduction, Machine Learning algorithms have been in the heart of the advancements made throughout these areas, but with the emergence and availability of sensors within these structures, the ability to gather more information grows exponentially higher than manual gatherings. And so, new implementations of Big Data Analytics, especially from the Predictive Analytics branch, are growing faster and more successful as knowledge keeps growing on these areas of study. However, despite this recent successful growth there is still a lot of work to be done in order to fully understand the capabilities of these methods as well as hidden value that could be retrieved from using them.

This research started by introducing the need for safety management and motorization of engineering structures as well as the theoretical background behind these concepts to investigate the proposed problem that arose. And so, in order to provide a useful response to this problems, the following Research Questions emerged:

1. Can there be a better alternate method and combination of input variables for improving the predictive accuracy of each of the different structural behavior responses of dams?

2. Can the representation of results be improved to provide new insights and help decision-makers improve their business decisions?

3. Can the application of the methodology developed for demonstrating the created artifacts be applied to other generic engineering structures, and not only for the application on dams?

In order to answer these questions, the purpose of this research has been to determine, through the use of predictive methodologies, what would be, for a certain response variable of a given engineering structure, the best combination of input variables that would provide the highest confidence for monitoring the safety of these structures. And so, to accomplish this, it was applied different predictive methods like MLR, RR, PCR and NN as well as different methods of re-sampling the data into training and testing sets to test and see if the predictive methods could be improved.

To answer RQ1, as previously seen on chapter 4, the analysis made provided us with insights on how well the models behaved when used for each of the combinations of the input variables. MLR, or the baseline model, provided the most accurate results based on the combination of the $R^2_{Adj}$ and the $RMSE$ metrics. Though, NN provided the most accurate results for the $RMSE$ results which showed that the randomness of these types of models are effective on the prediction of these patterns, i.e. behaviors. The RR model, even though it did not presented any real improvements, showed that in more correlated or outlier full environment, this technique could be more useful. The PCR model showed the capabilities of applying the PCA which in turn showed the possibilities of only using only the needed information which in this case was not useful due to some of the variables, like the temperature which are being generated automatically.

To answer RQ2, by combining some of the techniques the hypotheses of providing a more thoughtful and thorough representation of the resulting graphics. Though not fully implemented in this dissertation, the development of the methodologies applied to the development of a generic software for the monitoring of the structures behavior. Through the comparison of Figures 4.2 to 4.6 and Figures 4.7 to 4.21, the representation of the models could be further improved the unnecessary data, i.e. years, is removed from the graphics models.

To answer RQ3, and taking what has been stated in RQ2, the application of a development could improve the visualization techniques of the institution that uses it. The development methodology does not present any dependencies to any of the models used throughout the dissertation and so it could be derived that other models could also be applied to these methodologies without constraints.

The analysis and treatment done is also a generic approach which removes the outliers, as explained in SECTION Y, as well as it structures the data accordingly in order to be able to combine into the same datasets the input variables and the output behavior.

To test the developed methodology, described in Chapter 3, the case study followed in this research has been the study of a real portuguese concrete dam, due to the availability of the datasets provided by LNEC, as well as the possibility to communicate with Dam engineering specialist that would be able to provide a more comprehensive analysis on the variables being tested as well as the resulting behavioral responses given by the structure.

Furthermore, it has been concluded that the application of the Baseline model, expressed in Chapter 4, outperformed both the RR and PCR in terms of accuracy and confidence. On the other hand, both PCR and RR showed interesting responses. The PCR model allowed for the inference of which were the input variables that had the most correlation with the response variable and the RR would have given better results if the amount of outliers existent within datasets were higher or the measurements were more erratic, which was not the case due to the data treatment done to datasets prior to being given.

## 5.1 Evaluation of the Artifacts

(Von Alan et al., 2004) proposes four principles to help evaluate a DSRM artifacts, that have been considered through the entirety of the Design and Development (Chapter 3) and Demonstration and Evaluation (Chapter 4) phases of the DSRM. These principles are as follows:

- **Abstraction**: Each of the proposed artifacts must be applicable to other situations than the proposed problem.
- **Originality**: The artifacts must contribute an evolution of the body of knowledge.
- **Justification**: The artifacts must be justified comprehensibly and capable of being validated.
- **Benefit**: Each artifact must provide benefits, either immediately or in the future to the respective stakeholders.

To correlate what has been done in the context of this research with the authors proposal for evaluating the artifacts, the Design and Development phase of the DSRM methodology provided a diagram for the development of the artifacts. The main objective was to generalize the process for predicting the responses of engineering structures, rather than just to Dams, as expressed in the case study. The application of other methods than just the MLR to the datasets to Dam structures, like the RR, PCR and NN and other Re-Sampling methods, there was found that they provided a new perspective on tackling the identified problem as well as valuable contributions to the existing knowledge on the area of this research. Through the Demonstration and Evaluation phases of the methodology the models were applied in order to test the researches hypotheses, thus allowing for the gathering of conclusions about their usage on these types of structures. And so, these models have been justified, as well as their validity in being applied to these types of structures. The developed artifacts provided benefits in terms of identifying problems with the current technique and the new techniques being applied, as well as innovations on the application of different predictive models. These new models paved a way for the development of a generic way to monitor the behavior and safety of other types of engineering structures.

## 5.2 Future Work

Throughout the research a couple of bottlenecks were discovered where some of the predictive models were not as useful as they could be, thus several improvements that could be done in future work were discovered, like the need to improve the accuracy and efficiency of the current methods being applied on structural engineering monitoring and safety. Using the generic methodology for the application of the artifacts that has been developed in this research, there could be further development like a software prototype to use this methodology applying it to other engineering structures like bridges for instance. It was found that despite the previous data treatment applied to data, some further analysis on other methodologies for improving the quality of this data could be assessed like the use of other outlier detection methods, instead of that being applied in order to diminish the amount of outliers, thus allowing for more confidence on the predicting results of the models being applied. There could also be applied real temperature measurements, instead of using those being automatically generated based on the "normal" temperature fluctuations of the country in study, not only would this variable could provide a much more truthful monitoring and prediction but it could

also be applied to detect abnormalities that could be caused by this variable. An example of this abnormalities could be seen this year, 2017, where the variations in climate experienced throughout the year in Portugal have clearly been quite significant.

# Bibliography


Abbott, D. (2014). *Applied predictive analytics: Principles and techniques for the professional data analyst*. John Wiley & Sons.

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, *2*(4), 433–459.

Assunção, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A., & Buyya, R. (2015). Big data computing and clouds: Trends and future directions. *Journal of Parallel and Distributed Computing*, *79*, 3–15.

Atzori, L., Iera, A., & Morabito, G. (2010). The internet of things: A survey. *Computer networks*, *54*(15), 2787–2805.

Balageas, D., Fritzen, C.-P., & Güemes, A. (2010). *Structural health monitoring* (Vol. 90). John Wiley & Sons.

Bonelli, S., & Félix, H. (2001). Delayed response analysis of temperature effect. In *Proceedings of the sixth icold benchmark workshop on numerical analysis of dams, salzburg, austria.*

Bonelli, S., & Royet, P. (2001). Delayed response analysis of dam monitoring data. In *Icold european symposium on dams in a european context.*

Buytendijk, F., & Trepanier, L. (2010). Predictive analytics: Bringing the tools to the data. *Oracle Corporation, Redwood Shores, CA*, *94065*.

Chang, P. C., Flatau, A., & Liu, S. (2003). Review paper: health monitoring of civil infrastructure. *Structural health monitoring*, *2*(3), 257–267.

Chen, S., Wang, W., & van Zuylen, H. (2010). A comparison of outlier detection algorithms for its data. *Expert Systems with Applications*, *37*(2), 1169–1178.

Cheng, L., & Zheng, D. (2013). Two online dam safety monitoring models based on the process of extracting environmental effect. *Advances in Engineering Software*, *57*, 48–56.

Deshpande, S., & Thakare, V. (2010). Data mining system and applications: A review. *International Journal of Distributed and Parallel systems (IJDPS)*, *1*(1), 32–44.

De Sortis, A., & Paoliani, P. (2007). Statistical analysis and structural identification in concrete dam monitoring. *Engineering Structures*, *29*(1), 110–120.

Elena, C., et al. (2011). Business intelligence. *Journal of Knowledge Management, Economics and Information Technology*, *1*(2), 1–12.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer series in statistics Springer, Berlin.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144.

Grzymala-Busse, J., & Hu, M. (2001). A comparison of several approaches to missing attribute values in data mining. In *Rough sets and current trends in computing* (pp. 378–385).

Huisman, D. O. (2015). *To what extent do predictive, descriptive and prescriptive supply chain analytics affect organizational performance?* (B.S. thesis). University of Twente.

Jain, N., & Srivastava, V. (2013). Data mining techniques: a survey paper. *IJRET: International Journal of Research in Engineering and Technology*, *2*(11), 2319–1163.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 6). Springer.

Jung, I.-S., Berges, M., Garrett, J. H., Jr, & Kelly, C. J. (2013). Interpreting the dynamics of embankment dams through a time-series analysis of piezometer data using a non-parametric spectral estimation method. In *Computing in civil engineering (2013)* (pp. 25–32).

Kimball, R., & Ross, M. (2011). *The data warehouse toolkit: the complete guide to dimensional modeling.* John Wiley & Sons.

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Data preprocessing for supervised leaning. *International Journal of Computer Science*, *1*(2), 111–117.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). *Supervised machine learning: A review of classification techniques.*

Larson, B. (2012). *Delivering business intelligence with microsoft sql server 2012.* McGraw-Hill Osborne Media.

Lei-da Chen, T. S., Frolick, M. N., et al. (2000). Data mining methods, applications, and tools. *Information systems management*, *17*(1), 67–68.

Li, F., Wang, Z., & Liu, G. (2013). Towards an error correction model for dam monitoring data analysis based on cointegration theory. *Structural Safety*, *43*, 12–20.

Li, F., Wang, Z., Liu, G., Fu, C., & Wang, J. (2015). Hydrostatic seasonal state model for monitoring data analysis of concrete dams. *Structure and Infrastructure Engineering*, *11*(12), 1616–1631.

Liu, R., Kuang, J., Gong, Q., & Hou, X. (2003). Principal component regression analysis with spss. *Computer methods and programs in biomedicine*, *71*(2), 141–147.

Luhn, H. P. (1958). A business intelligence system. *IBM Journal of Research and Development*, *2*(4), 314–319.

Lynch, J. P., & Loh, K. J. (2006). A summary review of wireless sensors and sensor networks for structural health monitoring. *Shock and Vibration Digest*, *38*(2), 91–130.

Marbán, Ó., Mariscal, G., & Segovia, J. (2009). A data mining & knowledge discovery process model. *Data Mining and Knowledge Discovery in Real Life Applications*, *2009*, 8.

Mata, J. (2011). Interpretation of concrete dam behaviour with artificial neural network and multiple linear regression models. *Engineering Structures*, *33*(3), 903–910.

Mata, J., de Castro, A. T., & da Costa, J. S. (2013). Time–frequency analysis for concrete dam safety control: Correlation between the daily variation of structural response and air temperature. *Engineering Structures*, *48*, 658–665.

Mata, J., & Tavares de Castro, A. (2015). Assessment of stored automated measurements in concrete dams.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of machine learning*. MIT press.

Nedelcu, B., et al. (2013). Business intelligence systems. *Database Systems Journal*, *4*(4), 12–20.

Padhy, N., Mishra, D., Panigrahi, R., et al. (2012). The survey of data mining applications and feature scope. *arXiv preprint arXiv:1211.5723*.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, *24*(3), 45–77.

Perner, F., & Obernhuber, P. (2010). Analysis of arch dam deformations. *Frontiers of Architecture and Civil Engineering in China*, *4*(1), 102–108.

Portela, E., Pina dos Santos, C., Silva, A., Galhardas, H., & Barateiro, J. (2005). A modernizaçao dos sistemas de informaçao de barragens: o sistema gestbarragens.

Ranković, V., Grujović, N., Divac, D., & Milivojević, N. (2014). Development of support vector regression identification model for prediction of dam structural behaviour. *Structural Safety*, *48*, 33–39.

Ranković, V., Novaković, A., Grujović, N., Divac, D., & Milivojević, N. (2014). Predicting piezometric water level in dams via artificial neural networks. *Neural Computing and Applications*, *24*(5), 1115–1121.

Silva, A., Galhardas, H., Barateiro, J., & Portela, E. (2005). O sistema de informaçao gestbarragens.

Stojanovic, B., Milivojevic, M., Ivanovic, M., Milivojevic, N., & Divac, D. (2013). Adaptive system for dam behavior modeling based on linear regression and genetic algorithms. *Advances in Engineering Software*, *65*, 182–190.

Sun, Z., Zou, H., & Strang, K. (2015). Big data analytics as a service for business intelligence. In *Conference on e-business, e-services and e-society* (pp. 200–211).

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International journal of forecasting*, *16*(4), 437–450.

Tatin, M., Briffaut, M., Dufour, F., Simon, A., & Fabre, J.-P. (2015). Thermal displacements of concrete dams: accounting for water temperature in statistical models. *Engineering Structures*, *91*, 26–39.

Tayfur, G., Swiatek, D., Wita, A., & Singh, V. P. (2005). Case study: Finite element method and artificial neural network models for flow through jeziorsko earthfill dam in poland. *Journal of Hydraulic Engineering*, *131*(6), 431–440.

Tobias, R. D., et al. (1995). An introduction to partial least squares regression. In *Proceedings of the twentieth annual sas users group international conference* (pp. 1250–1257).

Ularu, E. G., Puican, F. C., Apostu, A., Velicanu, M., et al. (2012). Perspectives on big data and big data analytics. *Database Systems Journal*, *3*(4), 3–14.

Von Alan, R. H., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS quarterly*, *28*(1), 75–105.

Xu, C., Yue, D., & Deng, C. (2012). Hybrid ga/simpls as alternative regression model in dam deformation analysis. *Engineering Applications of Artificial Intelligence*, *25*(3), 468–475.

Yu, H., Wu, Z., Bao, T., & Zhang, L. (2010). Multivariate analysis in dam monitoring data with pca. *Science China Technological Sciences*, *53*(4), 1088–1097.

Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, *17*(5-6), 375–381.