Photo by Innocy BV of Kogerpolderbrug

Predicting movable bridge deck expansion – A machine learning and asset management perspective

Tejas Khedekar|4892933

MSc Thesis report

1/1/21

Delft University of Technology

MSc. Thesis (CME2001)

Predicting movable bridge deck expansion—A machine learning and asset management perspective

By

Tejas Khedekar

In partial fulfillment of the requirements for the degree of

Master of Science

Construction Management and Engineering

at the Delft University of Technology

Author

Name: Tejas Khedekar

Student number: 4892933

Graduation committee

Chair: Dr. ir. O. Morales Napoles

First Supervisor: Prof.dr. M. (Maria) Nogal Macho

Second supervisor: Dr. D.F. J. Schraven

Company supervisor: Thijs van den Eerenbeemt MSc.

This research project is done at TU Delft in collaboration with Innocy BV.

Preface and Acknowledgment

This report is created for partial fulfillment of the course, CME2001 Master Thesis from the Master program of Construction Management and Engineering at Technische Universiteit, Delft.

During the period from July 2019 to November 2019, I interned at BAAK BBV, a consortium created for Blankenverbinding project in Rotterdam. While interning with the Asset Management team of that project, I worked with employees of Innocy BV, an asset management consultancy firm based in Breukelen.

I had interned under supervision of Thijs Van Der Eerenbeemt, who later after completion of my internship discussed the possibility of working on a graduate project with Innocy BV. Every step of this thesis shaped its form under insightful guidance of Thijs and I am thankful to him for that.

Innocy BV also works on monitoring and predictive dashboard for their clients called 'JoostenTessa'. Thus, the idea of a thesis project on the use of IoTs for bridge monitoring was immediately felt useful to Thijs and Bram ten Klei, founder of Innocy. In no time, I started collaborating with the team at Innocy BV. Michiel and Oscar from Innocy gave me all the data and insights from the field of electrical engineering, data science and computer science engineering which goes behind the use of IoT and sensors and helped me at every step of research.

I hope that the results will help Innocy and its client on taking decisions on the use of sensors and IoT. The knowledge and experience sharing were crucial, and they helped me with it at every step. Overall, I believe we all indulged into a positive-sum game where we all had fun and took something or the other from each other.

I would also like to take this opportunity to thank Maria Nogal for helping me concretize my idea in productive and efficient steps. She helped me to get out of the rut of the ambiguity of topic, and I am grateful for that. Thank you, Morales Napoles and Daan Schraven, for believing in me and my process. They encouraged me that my research will bring insightful study for practitioners and that helped me to keep going during my slump days.

No acknowledgment is complete without thanking the secretary of CME, Sandra Schumann-Hagman for helping me with prompt appointments.

Last, I want to thank my friends and family who helped in these difficult times of COVID situation and helping me get through my leg surgery recovery, which happened right at the start of thesis preparation. None of these words can justify the support you have all given me through thick and thin.

Tejas Khedekar

Delft, 22nd May 2020

Abstract

Movable bridge decks experience critical expansion in summer, leading to uncertainty and unpredictability in its availability doe to improper docking and safety hazard. If the bridges are not cooled soon, the inertia of expansion stays, causing prolongation of availability problems. Structural health monitoring of such bridges with a predictive maintenance approach can help plan remedial measures on the exact day and time. For the efficient design of such a structural health monitoring system, a combination of sensor system data and weather API data is tested. A machine learning approach of Gaussian process regression which can give the results on the prediction of critical expansion of bridge deck has been evaluated in this research project. Finally, a check on the transferability of the prediction model is conducted by application on another bridge data and its performance is discussed. Scenario analysis with savings in cost per scenario is also conducted with varying levels of potential unavailability penalty costs, which could be levied on an asset manager of a bridge if the prediction models of sensor system data set or/and weather API data set gives incorrect estimation. Such an analysis is done to justify the use of the prediction models in assumed scenarios to predict expansion of movable bridge deck.

Keywords – Structural health monitoring, Weather API, machine learning, Bridge deck expansion, Feature Extraction, Gaussian Process Regression, Scenario analysis

Executive Summary

Problem in Hand

"There is a lack of research in the bridge deck expansion problem in movable bridges and efficient design of sensor system for monitoring and predicting it is lacking in current practices leading to disproportionate insightful information including predictive information coming out of huge data collected. Moreover, the unpredictability of unavailability of movable bridges results in major fines for asset owner or asset manager "

Solution space

The Literature shows that bridge deck expansion behaves to its environmental factors. Considerable useful data can be obtained from weather API which is an online open database maintained by various private and public institutions. Sensor systems and IoTs can allow data to be measured on the bridge location. Machine learned regression analysis can lead to a predictive model from this data available. Also, processes such as feature extraction and dimension reduction can help to choose the right variables coupled with literature backing. Machine learning techniques like Gaussian process regression can help to predict an expansion of bridge deck with sensor data and weather API data as input.

Central research question

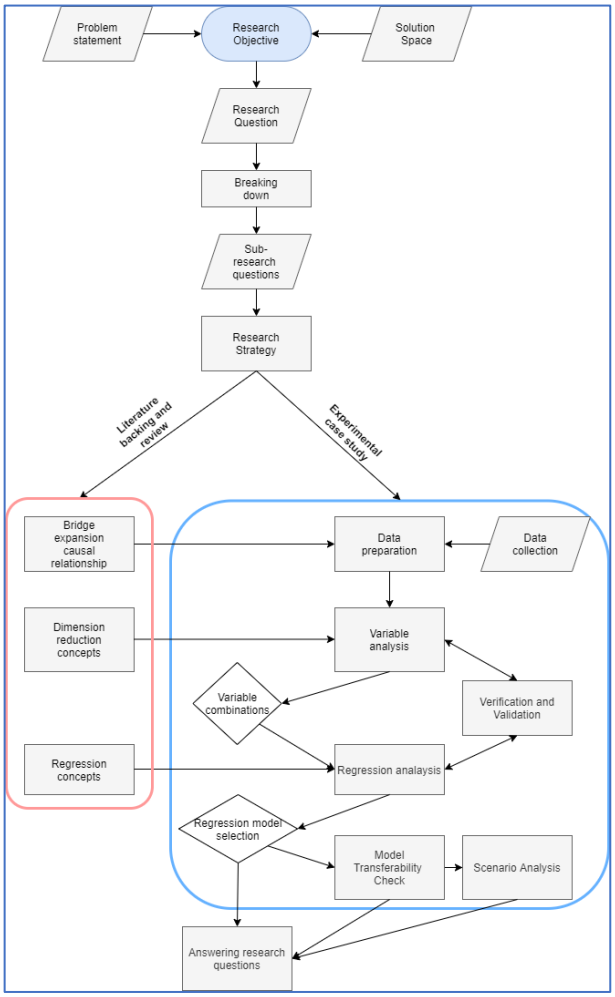
"How can data from sensor systems and weather APIs help develop prediction models of critical longitudinal expansion of bridge deck in movable bridges?"

Research Methodology adopted

The methodology to answer the research question required to follow a quantitative research approach. Thus, the strategy involves two principal parts.

The first part is a literature review and backing, in this step the physics behind the expansion of a bridge deck expansion are explored. Then, the concepts of feature extraction and regression analysis are explained and explored for the best approach to be achieved.

In the second part of the methodology, the explored in the literature review apply to the case study and data collected.
The results of the application of those procedures are presented and explained. Inferences are also made from the results, and variable combinations are made on which regression analysis will be tested to get answers to the research question.



Literature review takeaways

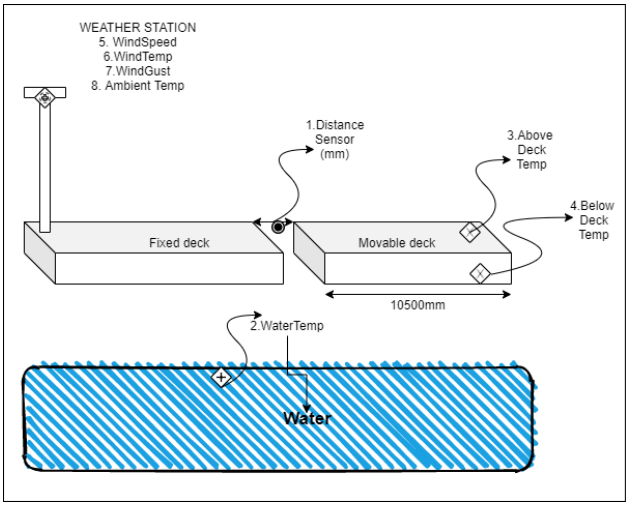
- Rain, snow, wind, ambient temperature, water temperature, wind temperature, solar radiation, humidity, etc. are the potential environmental factors that affects bridge deck expansion.
- All bridges are designed with the same temperature coefficient if they belong to the same class, still some bridges exhibit expansion problem while some do not experience critical expansion.
- Dimension reduction and feature extraction are important steps which should be done to focus on important parameters that need to be measured. It also helps in the design of the sensor system and IoT system as appropriate selection of components in the system can be done to measure the extracted parameters.
- The principal component analysis is one of the most widely used feature extraction techniques but to recognize latent factors which may cause expansion of bridge from the measured variables and to know the unique variance of variables measured then exploratory factor analysis should also be conducted.
- Linear regression is the simplest supervised regression technique in machine learning. It should be conducted first to get an understanding of the linear relationship between the measured variable and the expansion measured. Linear regression results also allow for hypothesis testing with p-value and standard error values. But it should be noted that since the linear regression model is not complex, it

can lead under-fitting problem unable to capture the non-linear relationship between the measured variable and the variable which is to be predicted.

- Gaussian process regression is a semi-supervised complex regression technique in machine learning. It does not assume one function to give the prediction model but assumes all the functions and distributes higher probability to the functions near the data set, giving a wider confidence interval and a fit which is effective. This can also lead to an over-fitting problem as it is a complex regression technique which may force-fit a relationship where no relationship between dependent variable and independent variable exists. Over-fitting and under-fitting problems should be checked with a physical relationship between the variables, and thus a literature exploration should be conducted to check the causal relationship.

Case Study sensor system and data collection

First case study is done on Alkmaarsebrug.The diagram shows the placement of the sensors and weather station sensors and thus obtaining 7 variables and one dependent variable which is the distance gap measuring between the movable bridge deck and fixed bridge deck. From weather APIs, the next 5 variables obtained are –



1. Ambient Temperature in Degree Celsius
2. Solar Radiation in Watt per meter square from a KNMI weather station 20km away from the bridge
3. Wind Temperature in degree Celsius
4. Relative Humidity index ranging from 1 to 100
5. Atmospheric Pressure in hectopascal or hPa

Kogerpolder is the second bridge case used to check the transferability of prediction model obtained from above bridge case. As for data from Kogerpolderbrug on which transferability of prediction model is done, data obtained from the sensor system include water temperature, ambient temperature, wind speed and from weather API the data include solar radiation and humidity.

Case study results

- PCA showed solar radiation as the variable with highest variance and highest importance in the first principal component. Temperature sensor above the deck and below the deck, and ambient temperature both from weather API and weather station formed a cluster of variables. Similarly, wind speed both from weather API and weather station along wind gust measured from the weather station, formed a cluster.
- EFA also showed similar result except for the fact that unlike PCA, solar radiation did not show high unique variance but showed common variance with the temperature parameters. Humidity and pressure also showed unique variance.
- All the above conclusions were verified with a correlation matrix and the pressure was removed as it showed a very negligible monotonic relationship with temperature.
- Linear regression analysis showed that the hypothesis of chosen 5 variables from the variable analysis was statistically relevant combined relationship with the dependent variable, except wind speed which did not show a linear relationship with the distance gap parameter. Linear regression also did show good fit and prediction for the values of small distance gap parameter. For critical values of expansion, the prediction from linear regression was not a good fit.
- GPR was conducted and it showed better R-squared, RMSE and MAE performance values as well as showed better fit and prediction for the critical value of distance parameter than the linear regression.

- Just a sensor data GPR and weather API showed similar prediction power but higher power than the combination of sensor data and weather API data. This concludes that for non-critical damage and fault mechanisms, the prediction model can be formed using weather API data without the use of sensors. As sensors system have different life-cycle than bridges, this decision of not applying sensors could prove to be crucial for asset management of bridges.
- Prediction model of one bridge was transferred to another bridge to evaluate its performance, while some critical values were predicted to some extent, the whole performance of the transferred model was below average. It is not advised to transfer prediction model of one bridge to another.
- Scenario analysis conducted showed that use of prediction model is justified, first with weather API data set giving highest savings on penalties due to unpredictable unavailability of bridge. Further, if prediction power of model changes, even addition of sensor is also justified clubbed with weather API data set to achieve savings on penalty costs.

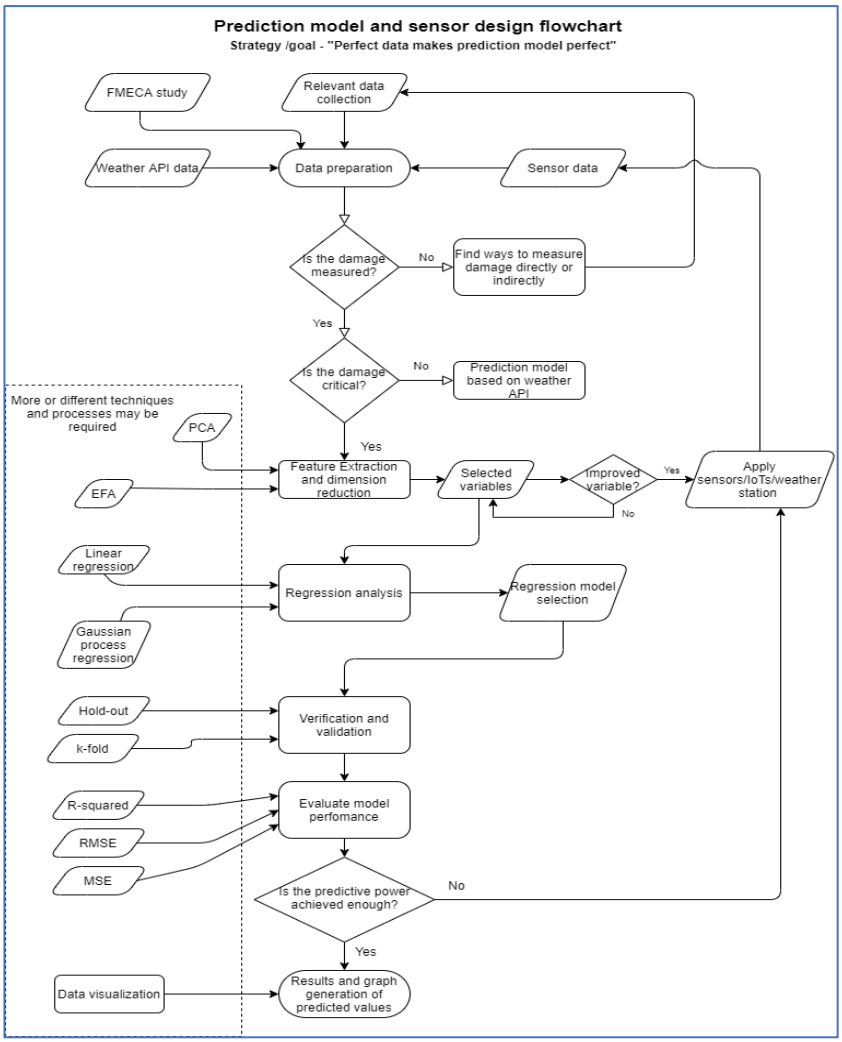
Recommended flowchart for prediction model and sensor design for general damage

The flowchart shows the general approach for sensor system design and prediction model for bridge deck expansion problem of a bridge.

It should be noted that it is a general approach and processes may change according to a specific bridge case.

Especially, there are many ways of regression analysis and dimension reduction techniques which were not explored in the research project but could come in actual practice on another bridge case.

This flowchart should also be discussed with clients, contractors and all stakeholders to make decisions for structural health monitoring of a bridge.



General steps that need to be followed according to the flowchart are

1. Through data already available and FMECA study of bridge try to get a measure of damage that is to be determined or to be predicted. In the present study case, the damage measured is the expansion of bridge deck.
2. Then, apply feature extraction and dimension reduction technique. This step helps to determine important environmental factors to focus on.
3. Further, regression analysis should be done using Gaussian process regression to form a prediction model. It is possible that other regression techniques can show better results that GPR, thus, an evaluation of various models should be conducted before finally selecting GPR.
4. Then the model should be validated and verified for a stipulated period of time. If the prediction is achieved as expected, continue using the same model by updating it in real time.
5. If the predictive power achieved is not enough and degrades in further period of time, then sensors could be deployed to captured parameters with unique variance following the results from dimension reduction technique.
6. This loop of plan-do-check-act can be continued until required level of prediction power is achieved so that complete benefits of structural health monitoring system is achieved.

Table of contents

Table of contents viii
Chapter 1 Introduction 1
1.1 Bridges and Structural health monitoring context 1
1.2 The problem context and gap in the literature 2
1.2.1 New technology, big data and double-edged sword 2
1.2.2 Bridge expansion problem 3
1.2.3 Move towards predictive maintenance over reactive maintenance 3
1.3 Problem statement 4
1.4 Solution Space 4
1.5 Starting point of research and its objective 4
1.6 Research Question 5
1.7 Research outline 5
Chapter 2 Research Design and Methodology 6
Chapter 3 Literature review 8
3.1 SHM and its challenges 8
3.2 Bridge deck expansion 10
3.3 Dimension reduction and feature extraction 13
3.1 Principal Factor Analysis 15
3.2 Exploratory factor analysis 17
3.4 Regression analysis 19
3.4.1 Gaussian process regression (GPR) 21
3.5 Verification and validation process 23
3.6 Theoretical framework building 24
Chapter 4 Case study analysis 26
4.1 Data Preparation and preprocessing 26
4.1.1 Data collection 26
4.1.2 Data tabulation 30
4.1.3 Outliers and missing values 30
4.1.4 Transform the data and enrichment 31
4.2 Variables analysis 32
4.2.1 PCA analysis 32
4.2.2 Exploratory Factor Analysis 33
4.2.3 Verification and validation for the formation of variable combinations 35

4.3 Regression analysis 37
4.3.1 Linear regression results. 37
4.3.2 GPR analysis 39
4.4 Transferability check 43
4.5 Sensor design comparison 45
4.5.1 Predictive power comparison 45
4.5.2 Scenario Analysis comparison 47
Chapter 5 Conclusions 51
Chapter 6 Discussion and Recommendations 55
Chapter 7 Drawbacks and further research alternatives 57
Chapter 8 Bibliography 59

List of Figures and List of Tables

Figure 1 - The starting point of research....4
Figure 2 - Research Methodology adopted....7
Figure 3- Structural elements of a bridge for availability consideration as per (Inkoom & Sobanjo, 2018)10
Figure 4 - Effects of the environmental parameters on bridge behavior....11
Figure 5 - Feature extraction and dimension reduction....13
Figure 6 The curse of dimensionality (Raj, 2019)....13
Figure 7 - Working of Principal component analysis....15
Figure 8 - Example of PCA biplot output....16
Figure 9 - Difference between PCA and EFA....17
Figure 10 - Simple linear regression example....19
Figure 11 - Gaussian process continuous samples depending on kernel selected a) Radial Basis Function b) Periodic Kernel c) Linear Kernel....22
Figure 12 - Change in prediction distribution upon adding of each training point....23
Figure 13 - k-fold validation process....24
Figure 14 - Theoretical framework application on case study....25
Figure 15 - Google map location of Alkmaarsebrug....26
Figure 16 - Schematic Diagram of Alkmaarsebrug sensor system....28
Figure 17 - Outliers removal....30
Figure 18 - PCA bi-plot result output with solar radiation and unstandardized data set....32
Figure 19 -PCA bi-plot result of standardized variables....33
Figure 20 -EFA bi-plot of output results....34
Figure 21 - Spearman's correlation matrix....35
Figure 22 - Predicted response vs True response of linear regression model....38
Figure 23 - Predicted response vs True response of GPR model....40
Figure 24 -Linear regression model fit of standardized observed values with 95% Confidence intervals..41
Figure 25 - GPR model fit on observed values with 95% confidence intervals....41
Figure 29 - Performance of the GPR model for transferability....44
Figure 26 - Predicted vs True response graph of Sensor data GPR....45
Figure 27 - Predicted vs True response graph of weather API data GPR....46
Figure 28 - Problems of over-fitting and underfitting....47
Figure 30 - Flowchart suggestion for general prediction model and sensor design for the expansion of bridge deck based on inferences of the research project....54

Table 1 - An example of the database collected 29
Table 2 - Example of data tabulation 30
Table 3 - Example of final enriched data tabulated 31
Table 4 - Cluster formation of the variables after PCA and EFA analysis 35
Table 5 - Final combinations of variables formed after verification and validation of variable analysis 36
Table 6 - Linear regression results 37
Table 7 - GPR results from the output 39
Table 8 - Comparison of performance of Sensor GPR model vs weather API GPR model vs GPR Combination of both 46
Table 9 - Cost savings of combination 1 (Sensor + weather API data) 49
Table 10 - Cost savings of combination 2 (Sensor data) 49
Table 11 - Cost savings of combination 3 (Weather API data) 49

List of abbreviations

SHM – Structural health monitoring

IoTs – Internet-of-things

GPR – Gaussian process regression

PCA – Principal Component Analysis

EFA – Exploratory factor Analysis

RSS – Residual sum of squares

RSE – Residual square error

RSME – Root mean square error

VIF – Variance Inflation Factor

MAE – Mean Absolute Error

SD – Standard Deviation

API - Application Programming Interfaces

FMECA – Fault mode and effect analysis

(This page has been left blank intentionally)

(This page has been left blank intentionally)

1

Chapter 1 Introduction

1.1 Bridges and Structural health monitoring context

Bridges are part of human civilization for over 6000 years. Today, bridges are a significant part of the infrastructure of the world. If the infrastructure of a city is like a human body, we consider bridges to be the veins and tissues connecting essential parts of the body. Just like the human body, bridges deteriorate because of various factors such as fatigue, overloading, cycling loading and extreme incidents such as floods and earthquakes (Balageas, Fritzen, & Güemes, 2010). Even environmental conditions like wind, rain, storm, and extreme heat can cause deterioration of steel and concrete (Wenzel, 2008). This has led to the birth of structural health monitoring of bridges right from the beginning of construction of bridges in human civilization.

Structural health monitoring or SHM is an essential activity in the maintenance of bridges. Mainly conducted to understand the relationship between the loading environment of the bridge and the consequent response by the structure (Farrar & Worden, 2007). In the SHM process, various methods are applied for damage detection. The damage in this case broadly falls in the category of material or geometric properties of a structure which affects the performance of the complete structure. Five broad working levels divide SHM into processes like confirming the presence of damage, determination of location and orientation of the damage, evaluation of the severity of the damage, the possibility of controlling or delaying the growth of damage and finally predicting the damage expected in structures. (Gopalakrishnan, Ruzzene, & Hanagud, 2011). Bridges usually fall under a high critical category of structures wherein remedial action decisions are taken after asset manager or owner examines the structure at periodic intervals. Currently, often humans examine the bridges and carry out the inspections. This process can fall prey to human error or damages can go undetected (Agdas, Rice, Martinez, & Lasa, 2016). Some old methods of SHM also include ground-penetrating radar, acoustic emission and using chain drafts. They are supposed to be time-consuming, expensive and highly dependent on human interpretation (Zalt, Meganathan, Yehia, Abudayyeh, & Abdel-Qader, 2007).

The digitalization of SHM is being done nowadays through different sensor network. A sensor network has the potential to identify problems in bridges at an early stage and consequently prolong the life of these structures and improving public safety (Zalt et al., 2007). Some of the most common types of sensors include electrical resistance strain sensor, vibrating wire strain gauges, and fiber optic strain sensor. A typical SHM system with sensors has processes such as data acquisition, data transmission, data storage, data processing, and data interpretation. Using new technology in SHM for bridges also brings us to the use of wireless sensor systems (Sonawane, Vichare, Patil, & Chavande, 2018). Wireless sensor networks and IoT (Internet-of-things) are proving to be efficient in all the above processes of the sensor monitoring system. IoTs allow interconnectedness within the sensor system's data along with the data from the internet, thus bringing in a whole new level of information and visual analytics to an asset manager.

1.2 The problem context and gap in the literature

1.2.1 New technology, big data and double-edged sword

Structural health monitoring while is beneficial to understand the real-time state of the structure's health, it comes with rigorous design and implementation problems to fully gain its benefits. It adds value as SHM allows movement of maintenance activities from scheduled intervention to on-demand inspection (Del Grosso and Inaudi, 2004). SHM costs very little compared to the total cost of the bridge from 0.5% to 3% and hence, while the return of investment is high and recovered within 10years, little thought to its design and implementation are often observed (Inaudi, 2009). SHM systems are sometimes installed without a real analysis of the owner's needs and often based on the desire to implement new technology and following a trend. Hence, the use of wireless sensors systems has increased in recent years (Vardanega, Webb, Fidler, & Middleton, 2016). As an investment, a monitoring system must prove its effectiveness, not only as qualitative improvement of the bridge performance but also as a way of reducing the overall life-cycle-cost (Radojicic et al., 1999).

Another issue that arises after design and implementation of SHM with sensors are synchronization error, data loss and dealing with large amounts of harvested data. Limited network resources such as battery power, storage capacity and bandwidth also pose huge issues (Nagayama, Sim, Miyamori, & Spencer Jr, 2007). Thus, it might render a communication problem within the whole SHM system.

New technologies such as IoT and wireless sensor systems can solve all the above problems. It can also avoid synchronization problems doe to the interconnectivity in the system. Also, low-cost IoT devices with lower power demands are assumed to reduce the cost of the SHM sensors to an extent (Kocakulak & Butun, 2017). While the claims are done by the leading IoT companies, quantitative and qualitative research in this specific field of bridge monitoring is lacking (Khan, Atamturktur, Chowdhury, & Rahman, 2016). The duality of every new technology is that the accomplishment of benefits might also cause newer risks. There are instances of risks such as lack of sufficient knowledge, IT infrastructure limitations, data management, and incorrect data capture (Hummen, Henze, Catrein, & Wehrle, 2012).

The new sensor systems currently lie in the peak of inflated expectation, especially in SHM, as it is still in the early phases of its implementation. The path to enlightenment and productivity is yet to be achieved entirely; hence, in this current phase, more research towards the plateau of productivity is needed. Thus, as the use of wireless sensor networks for bridge monitoring has become a trend, the research in it is lacking especially in its maintenance and monitoring perspective and long-term returns.

Because of the newness of the technology and subsequently lack of knowledge of new sensor systems, asset owners all around the world face the problem of investing in it and that poses as a significant question to various asset owners (Ryan & Watson, 2017). Asset owners are convinced that they need SHM and subsequent sensor networks but are unsure about its efficiency, in simple terms, they are unsure whether it is worth the buck, effort, and energy. Different stakeholders such as asset owners, designers, contractors and researchers agree on the fact that scarce resources should not be wasted in pursuit of data as opposed to information, SHM combined with wireless sensor network steps in to solve that common goal, but to what extent it is still unclear (Vardanega et al., 2016). The data sets gained from an SHM system sometimes fall into 'big data' type because of their sheer volume, complexity and diversity, thus demanding cross-disciplinary efforts in data science and bridge monitoring systems. The life cycle of bridges are longer than life-cycle of sensor systems and thus maintenance of sensors themselves causes a problem in the overall structural health monitoring process and that has to be taken into account in design of sensor systems too.

1.2.2 Bridge expansion problem

As about one-third of the Netherlands lies below sea-level there are many movable bridges in the country, as much as approximately 1500. The Netherlands has many canals and water-ways for shipping and transport. This has led to the construction of many movable bridges. These movable bridges make road transport and water transport flow smoothly with minor delays. As two ways of transport depend on movable bridges, there is a high expectation of reliability and availability from these bridges. Movable bridges are an interaction of civil, mechanical, electrical engineering, making it more difficult to maintain as components are vast and varied. The maintenance unit cost per square foot of deck of a movable bridge is estimated to be 100 times more expensive than that of a fixed bridge (Gokce, Gul, & Catbas, 2012). Research is also lacking in movable bridges compared to fixed bridges (Okasha & Frangopol, 2009). Besides, few studies provide documentation of issues related to movable bridges with proper monitoring applications specific to them (Catbas et al., 2014)

One major problem with movable bridges is the expansion of the bridge deck. Unlike a fixed bridge, critical expansion of bridge deck can cause improper docking and locking of movable parts, critically harming availability not only for road transport but also water transport as lanes might be closed for maintenance of the same. The problem is evident several times during summer in the Netherlands (DIMITROVA, 2020) ("Heatwave hits bridges, shipping in Amsterdam's waterways," 2018) (Damen). Once the bridges have crossed a certain level of critical expansion, it is not possible to bring it out of it quickly due to the inertia expansion poses. The remedy to this problem is showering the bridge deck with water or ice. Currently, this remedy is done once the bridge has reached critical expansion, which in terms of reliability engineering can be called as 'allowed to fail' or reactive maintenance. One of the principal aims of SHM is to move from reactive maintenance to preventive or predictive maintenance and currently, this problem with movable bridges is lagging to reach that goal. Extreme values because of climate change are going to affect bridge safety and availability to a great extent, and SHM of bridges should consider rising temperatures all around the world (Mondoro, Frangopol, & Liu, 2018)

1.2.3 Move towards predictive maintenance over reactive maintenance

Predictive maintenance implies the determination of the in-service condition of the structure and then further estimating when maintenance should be performed. This approach has proved to be cost saving in the long run of structures over routine or time-based preventive measures and also over reactive maintenance, i.e., allowing the structure to fail and then taking measures (Mobley, 2002). The fundamental idea behind predictive maintenance is to increase safety and avoid accidents with a negative impact on the environment through the information on the current performance of the structure and analyzing the past trends to detect early faults. Another main theme of predictive maintenance is that it allows optimization of maintenance activities and resources required for the same, in the example of bridges, if closures are required, it could be planned on off-peak hours if the exact days of closure for a specific problem are known. SHM helps achieve these themes of predictive maintenance through data collection and processing of the same. The process of capitalization of data is the key element in the evolution of maintenance towards predictive strategies (Gouriveau, Medjaher, & Zerhouni, 2016). To achieve complete realization of predictive maintenance – a culmination of people, insights and data are necessary (Daily & Peterson, 2017). This research project tries achieving the same through its results.

1.3 Problem statement

After observing the above problems, we can quote a problem statement:

"*There is a lack of research in the bridge deck expansion problem in movable bridges and efficient design of sensor systems for monitoring is lacking in current practice leading to disproportionate insightful information including prognostics coming out of large amount of data collected. Moreover, the unpredictability of availability of bridges results from bridge deck expansion, which can lead to huge penalties for asset manager or owner*"

1.4 Solution Space

The solution to getting better information from available big data lies in data science. Various analytical and machine learning regression techniques can give exceptional insights. For understanding which factors affect the expansion of the bridge deck, variable diagnostics and dimension reduction can be applied. As for data coming from the sensor, an alternative or additional data input can be obtained from weather API and weather station installed near the bridge. Use of IoTs allows this feature to be included in SHM systems. Weather API is weather application interfaces that contain an extensive database of historical and forecasted weather data obtained from various public and private weather data management sources. In the Netherlands, KNMI or Koninklijk Nederlands Meteorologisch Instituut is the Dutch international weather service, which holds an enormous amount of data on climate, air and meteorological factors. Some widely used private service providers of weather API include Accuweather and OpenWeatherAPI which are in the market providing data at competitive accuracy. A combination of these databases from APIs and sensor systems with applied data science could lead to a solution for the above problem statement.

1.5 Starting point of research and its objective

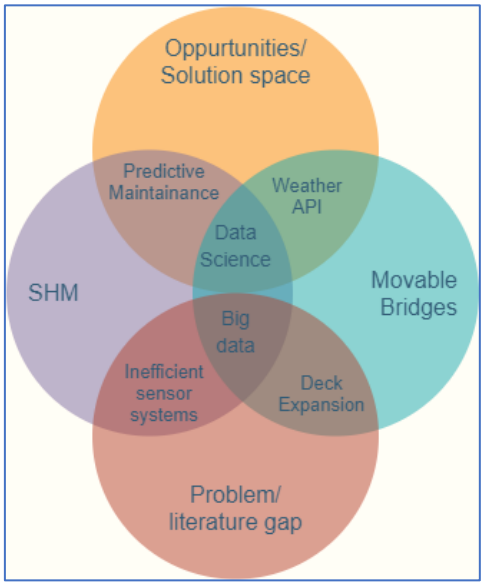


Figure 1 - The starting point of research

From the research gap mentioned above and converging all the problems to a solution, it naturally synthesized down to the following research objective for the research project–

"*To achieve an efficient design of monitoring system for bridge deck expansion in movable bridges that achieve data from the sensor systems and weather APIs into insightful information leading to predictive maintenance over reactive maintenance*"

1.6 Research Question

The research objective and following Figure 1 by combining problem and solution space, the central question of the thesis which can be formulated as follows -

"How can data from sensor systems and weather APIs help to develop predictive models of critical longitudinal expansion of bridge deck in movable bridges?"

Sub-research which follow from the central question after breaking down are -

1. Which environmental parameters affect the expansion of the bridge?
2. Which regression model shows a better fit for prediction on bridge deck expansion with sensor and weather API data as a dependent variable?
3. Are the regression models transferable from one bridge to other bridges?
4. How sensor systems can be designed economically efficient combined with data from weather API to achieve insightful information?

1.7 Research outline

Chapter 1 includes the research context and highlights the purpose of taking up this research project. Chapter 2 expands the research design and methodology adopted to achieve the research objective. Chapter 3 delves into the literature review and the theoretical framework of solving the research questions. Chapter 4 includes a case study and application of the said theoretical framework giving results on analysis conducted. Chapter 5 summarizes the results and answers the research questions. Chapter 6 elaborates the results of the case study as a part of the discussion section of the report and furnishes recommendations on said results and finally, Chapter 7 suggests further research alternatives along with drawbacks of the research project.

2

Chapter 2 Research Design and Methodology

The research objective of this thesis project requires an applied approach. An approach that can give concrete steps to solve the problem and achieve the objectives. The steps taken in research methodology chosen are inspired by Corner (2002), which gives an integrative method of applied research and quantitative approach.

Thus, first, a literature is taken up to understand the concepts and solutions which might give answers to sub-research questions. It also involves exploring the problem and solution space. This step is further divided into 2 parts. First, understanding the relationship between the bridge expansion and environmental factors causing it is explored. Then concepts of feature extraction and regression analysis are discussed and best approaches for the case study are chosen from the literature review.

The next complete part of the research is the application of concepts learned from the literature chapter on the case study available to the researcher. Following are steps taken in the case study analysis -

1. Data preparation with data collection as input.
2. Variable analysis with concepts of feature extraction from the literature review.
3. Formation of variable combinations based on inferences from the analysis.
4. Verification and validation of variable analysis.
5. Those combinations are run through regression analysis.
6. The best performing regression model is chosen and again verification and validation is kept intrinsic in the regression analysis step.
7. The selected model is transferred to another bridge case and evaluated for transferability.
8. Scenario analysis is done on consideration of sensor system design and its cost and savings in terms of penalty cost are elaborated.
9. The Final results are combined and discussed to answer the central research question and sub-research questions.
10. Further discussion and drawbacks of the research project are elaborated based on inferences from results.

The following Figure 2 is the flow diagram representation of the research strategy adopted to answer the research questions.

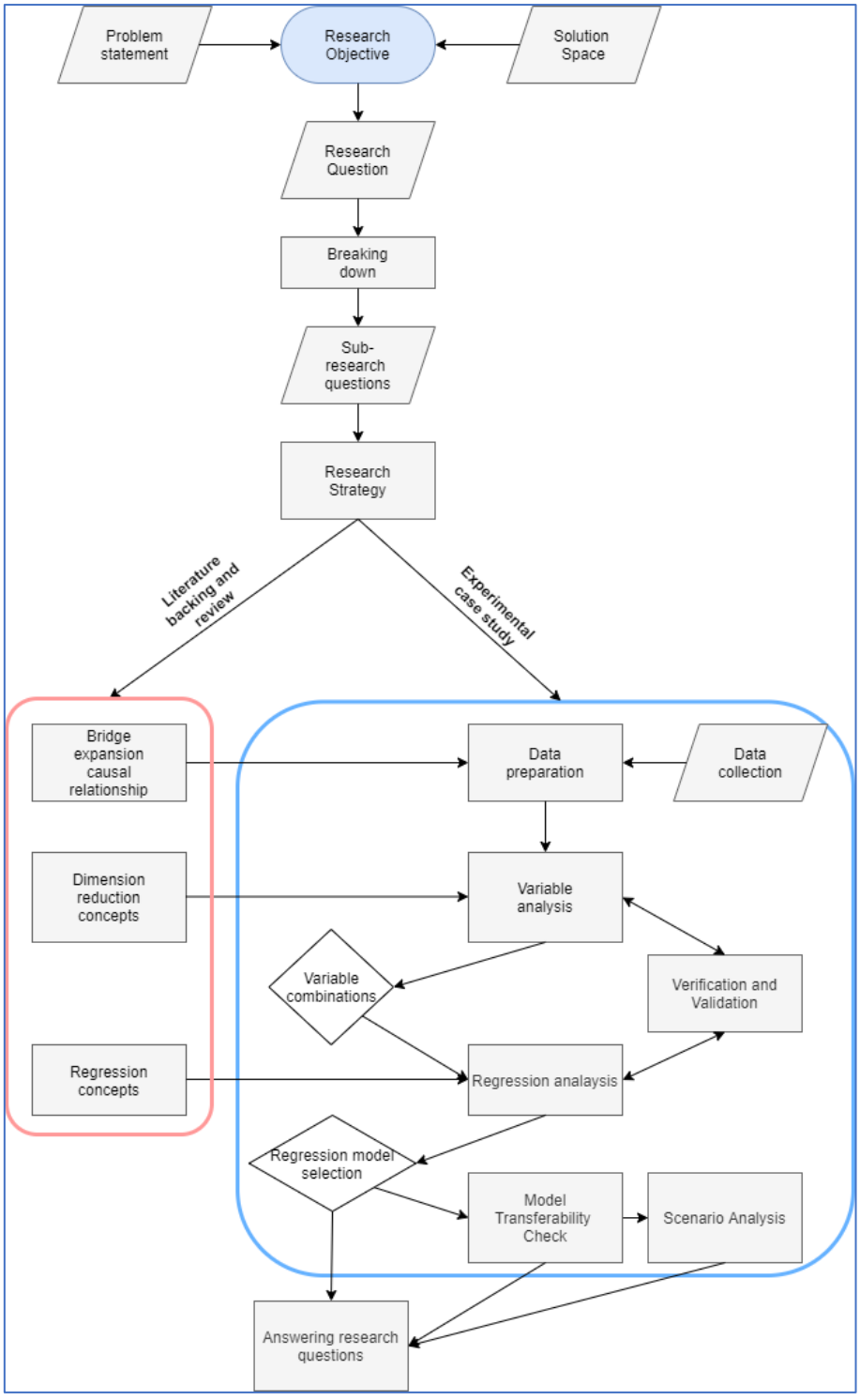


Figure 2 - Research Methodology adopted

As the very next step after formulation of literature review, the next section elaborates bridge deck expansion, feature extraction technique and regression analysis concepts.

3

Chapter 3 Literature review

3.1 SHM and its challenges

Structural health monitoring is the process of devising damage identification plan or strategy concerned with structures. While the research and implementation of SHM have increased in recent years, the idea behind it was prevalent since the earliest endeavors of humankind to worry about the deterioration of structures and then plan its repair or prolong the remaining life of the structure. Damage identification in this context refers to changes in material or geometrical properties of a structure. These damages can affect the performance of the structures adversely. Damage does not mean complete failure of a structure but representation that structure is not performing on an optimal level because of its presence. Damage can grow and reach a point where the structure cannot provide its supposed function, that point is called failure. On a larger scale of time, these damages are on the material level and growth can be linear, cyclic or non-linear, while on a shorter scale of time the damage can be referred to a discrete event like an earthquake, collision, etc. The terms analogous to SHM are condition monitoring (CM), non-destructive evaluation (NDE), statistical process control (SPC) and damage prognosis (DP). NDE is usually carried out after damage is located to find its severity, while DP is used to predict damage or failure in the life-cycle of a structure. While these tasks many years ago seemed to be difficult to achieve, now due to the progress in sensor technology, the methods of damage detection have improved. The improved damage characterization and efficient analysis techniques have made it possible to use SHM not only for operation and maintenance but to develop reliability and risk-based design techniques (Karbhari & Lee, 2009). Karbhari and Lee (2009) also pointed out that more often there is a disproportionate highlight on a collection of data instead of the management of data and using it for decision making, which is the first principal aim of data collection, i.e. better management of infrastructure. The gap between the management of data and using it for the ultimate goals of estimating or prolonging service life was also observed by Sikorsky and Karbhari (2003).

Steps involved in an SHM process are as follows. First, data acquisition takes places. This step involves the selection of tools that can detect damage detection. The human inspection was common in the earlier period, but these days sensors are most commonly used as a human inspection can become prone to human error, and sometimes it is not humanly visible to see some aspects of the damage. The newest trend in this step involves the use of a drone camera. The basic idea that sensors can detect excitation of structures forms the foundation of the data acquisition step. Data cleansing is the integrated process of this step wherein unnecessary data is cleared out. Filtering and re-sampling are the techniques of signal processing which are adopted for data cleansing. Management of missing data and outliers are also the processes taken in this step itself.

The second step involves feature extraction or dimension reduction. In this step, data features which can represent damage of the structure are identified. It might involve use of physics-related formulas or could be purely data driven. Another way it is conducted is through experimental modeling of the structure and then finding out which data features will lead to damage identification. Analytical tools like finite-element analysis also come handy in this step. For example, displacement observed by GPS sensor can show deflection in structure and eventually the amount of damage. Thus, in this case, distance measurement becomes the data feature. Various techniques of data science are used in this step to come up with a better solution for feature extraction. The later stages of SHM involves data storage, data output and data insights.

Classically, structures are designed using various design codes which specify factor of safety found through empirical and experimental methods. It is also known as parametric design method. While this method has served designers for centuries, the major disadvantages include use of a single deterministic value for statistical and probabilistic nature of structure in response to its changing environment. It also leads to structures of the same class designed with same parameters, but in reality they could have different configurations and nuances which design codes cannot accommodate. SHM can step in to resolve these disadvantages of traditional design method by in-depth analysis of real-time structural response leading to sophisticated and integrated methods of design in the future. Moreover, it enables updating risk-allocation and resource allocation in real time, thus linking design, construction and maintenance together for the betterment of structure which has its own unique configurations. All this can culminate to a new paradigm of future development of more integrated design codes, and the one based on continuous assessment and monitoring of the health of a structure. The future trends of climate change can also be accommodated in such futuristic methods as the current traditional design and maintenance activities might fail to achieve the same.

The recent SHM studies and research delve into showcasing specific sensor technology to measure factors such as seismic or wind loading on structure and show its accuracy. Less focus is seen on diagnosis and prognosis of structure damage in consideration. SHM systems themselves cannot provide better safety system or method of maintenance, but with proper design, it can achieve necessary inspection decisions. It can also ensure timely updated estimate of deterioration, performance, and remaining service life of structure. While the sensor systems are available in all shapes, sizes, methods and sensitivity, the underlying challenge of SHM remains the same, which is its appropriate selection, design and placement of sensors depending on the needs of the structure and asset owner. One of the ways to ensure the goals and actions taken are aligned is to determine performance factors of the bridge and monitor those through SHM systems. There should not be over-use of sensors, as more sensors does not necessarily mean better SHM systems. Emphasis needs to be given to interpretation and insights from the data, both in diagnosis and prognosis. Another problem associated with sensors is its reliability throughout the life cycle of structure and its integration with the harsh, changing environment. Several new sensors have proved to have high self-monitoring and self-calibration capabilities in short-term, but their long-term durability under severely changing environmental conditions is yet unknown. Civil structures have a life period of 50 years or more and the new technology in sensor systems are quite new to test their robustness, reliability and durability for such a long life-cycle period of civil structures. Improper design with a huge number of sensors not only complicates the problem of data transmission and synchronization but also increases the complexity of data analysis and its validation. Data validation is also a serious issue, as the damage could be exaggerated or underestimated if the validity of the data stream is not conducted.

3.2 Bridge deck expansion

A general arrangement of bridge elements is shown below.

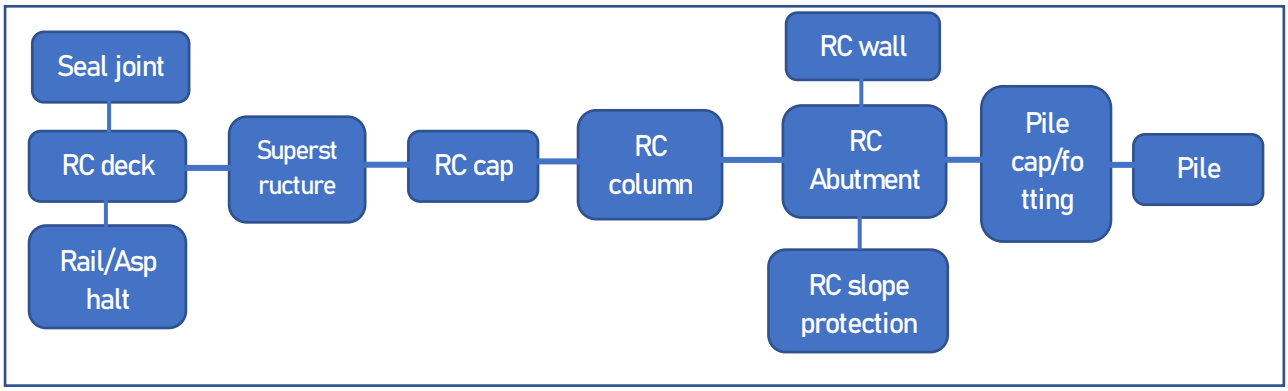


Figure 3- Structural elements of a bridge for availability consideration as per (Inkoom & Sobanjo, 2018)

For the structural health monitoring of bridges, the elements shown in Figure 3 form the critical aspects of monitoring. The other parts which do not constitute to the structural elements of a bridge but are equally important for the availability of bridges are moving parts like hydraulics, operator facilities etc. because of their downtimes and operational characteristics unlike the above structural elements that are mostly always available. This research project will focus on above structural elements and not on the moving operational parts to narrow down the topic.

It becomes important to prioritize which individual elements that are affecting the reliability of the whole bridge to a greater extent (Beeson & Andrews, 2003). Researches such as Inkoom and Sobanjo (2018), Jiang and Rens (2010), Thompson, Sobanjo, and Kerr (2003) and Nabizadehdarabi (2015) which adopted different methods to come up with criticality factors have shown that superstructure is massively important for the whole system of the bridge for availability, reliability and failure. The same problem is observed in the Netherlands, where there are a large number of movable bridges and the bridge decks tend to expand in summer and create availability problems. Just like any other material, the superstructure of bridges expands when heated and contracts when it is cool. Thermal movement in superstructure along with shrinkage matters in the overall displacement of the bridge deck, these are usually accommodated by an expansion joint. Unlike conventional bridges, wherein after the expansion joint is completely used, the lateral load because of the temperature strain is passed on to abutments; moving bridges face the problem of improper docking and thus, closing and opening of the bridge becomes almost impossible if critical expansion happens. The important factors responsible for critical expansion such as the one explained above are climatic conditions like UV, temperature, wind speed, etc. along with geometric characteristics of the bridge like the cross-section of the deck, orientation, asphalt thickness, etc. Material characteristics also matter but more or less all bridges have the same materials such as concrete, steel and composite materials unless bridges are really old and made of stone.

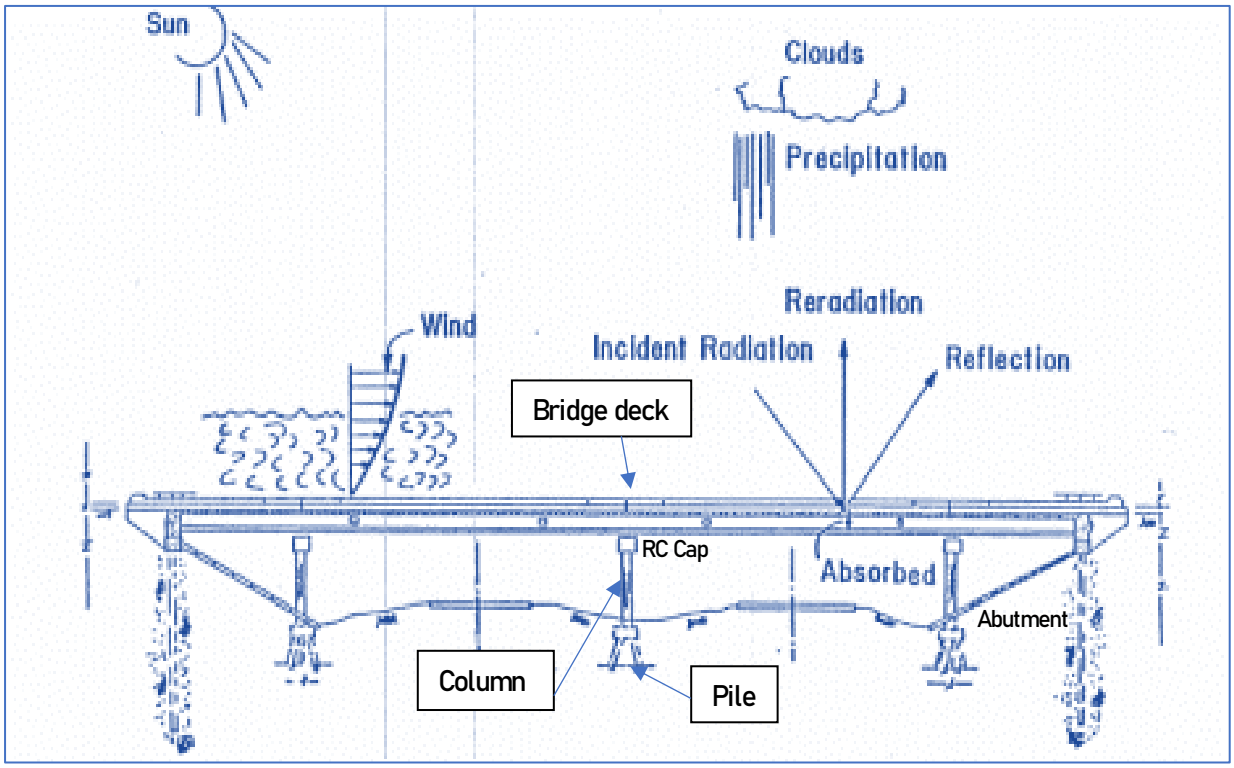


Figure 4 - Effects of the environmental parameters on bridge behavior

Figure 4 shows how environmental factors affects different elements of a bridge structure. Diurnal temperature is the change of temperature from daytime high to a night-time low and is also one of the important weather conditions which come in cycle affecting the expansion of bridge. The UV index or solar radiation is depended on the time of the day and clarity of the atmosphere. Seasonally too, UV index changes from winter to summer based on the power of sun rays. A horizontal surface like bridge deck will get hot sooner and faster than piers because of the angle of sunrays being almost perpendicular to the surface. This also causes a temperature difference throughout the bridge deck as the upper surface can get a large amount of heat more than the surface below, the same scenario alternates in the night where water radiates more heat to the surface below the bridge deck. The heat transfer from the sun to the bridge can also be distorted because of wind speed, where convection waves make more action on the bridge than conduction.

A long term fine element analysis conducted by and Min, Sun, and Dan (2009) has shown that temperature and wind speed are highly correlated to change in the behavior of bridge deck, especially for its displacement and expansion. They have also proved that the combined effect of heavy wind and temperature leads to major changes in the structural response of bridges. Another long term fine element analysis by (Cross, Koo, Brownjohn, & Worden, 2013) has shown that modal frequencies of the bridge are drastically changing because of vehicle load, temperature and wind speed. They proved that wind speed higher than 25mph hitting off the side of the bridge brings a very significant effect on bridge behaviour, while temperature changes seasonally are more effective on bridge behavior that diurnal temperature cycles. Xu and Xia (2011) pointed out that many studies are done to prove the effect of vehicle load, temperature and wind speed on the behaviour of bridges while exploratory research on the effect of

humidity and atmospheric pressure is lacking. It was also found out that the research in this field was more with a structural design engineering perspective. Fewer studies were found with asset management perspective to come up with the efficient design on sensor system to support a long term structural health monitoring.

Environmental factors are proven to affect the behaviour of bridge deck expansion if the geometrical characteristics are kept constant. Thus, AASHTO, The American Association of State Highway and Transportation Officials, the committee on bridges design and monitoring in United States of America, is therefore considering adoptions of field measurements of environmental temperatures to take action based on design and maintenance (Roeder, 2003). In such a case, sensor system which can collect data on environmental factors around the bridge can prove essential to give information on bridge behaviour based on physical causal relationship on how these parameters affect displacements in the bridge.

One of the major problems bridge deck expansion causes in SHM is its incomplete explanation through environmental factors. While literature shows these factors affect the bridge expansion, to what extent each factor brings the change and how important are their contribution to the critical damage is still ambiguous. This ambiguity arises because of many reasons. First, no two bridges are the same, they differ in shape, size and configurations. Second, the material composition of most of the modern bridges is composite, thus exact effect cannot be determined. Third, the environmental factor differs from one location to another. These factors lead to ambiguity and consequently cause a problem in designing efficient SHM system as there is no one fit solution. It should also be noted that the above relationships between environmental factors and bridge deck expansion backed by the literature were all conducted on a fixed bridge, negligible similar findings were found for movable bridge decks (Catbas et al., 2014).

3.3 Dimension reduction and feature extraction

The first step before progressing with regression is dimension reduction. Dimension reduction, feature selection and feature extraction are used interchangeably in data science. A perfect metaphor for the use of dimension reduction technique can be found in our day-to-day lives too, our television sets and laptop screen shows videos and pictures in two dimensions while in reality, those objects are in three dimensions, but the information of 3rd dimension which is height or depth in this scenario is not lost in conveying all the dimensions. As shown in the Figure 6, it is thus the process of changing data with high-dimensional variables into low-dimensional variables which retains approximately the same properties and characteristics of the original data set (Carreira-Perpinán, 1997).

Figure 5 - Feature extraction and dimension reduction

The reason to do this process is to avoid complexity in the model formation and reduce computational time. The curse of dimensionality as shown in Figure 6 explains that high-dimensional data tend to more complicated than lower dimensional ones. Those complexities are hard to discern, and the solution to it is to incorporate representative data than the complete data (Duda, Hart, & Stork, 2012). The process shows a more compact, easily interpretable representation of target goals focusing user's attention on the most important variables (Witten & Frank, 2002). It also allows visualizing data after reduction into a two-dimension or three-dimesnional figure, which helps user to get the feel of data set and their performance as predictors.

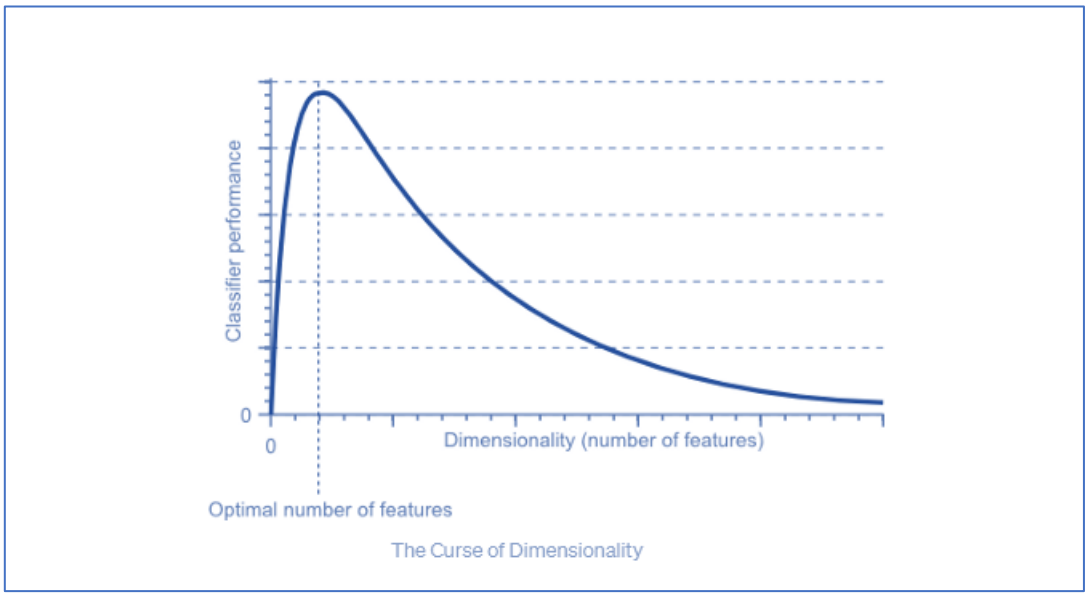


Figure 6 The curse of dimensionality (Raj, 2019)

Dimension reduction techniques highlight hidden cause and latent variables among the parameters. This becomes an exceptional tool in SHM–if the damage is recorded, various environmental factors can be measured to know which latent parameter could be the cause of damage to the structure. Also, irrelevant variables and redundant variables can be identified and removed.

It also resolves multicollinearity problem in regression analysis. The regression equation is an algebraic expression showing the prediction of the dependent variable from the values of the independent equation. It is usually shown by the following equation

$$y = \beta + \beta_1 * x_1 + \beta_2 x_2 \quad (1)$$

Where y is the predicted value and β, β_1, β_2 are parameters or coefficients of the corresponding variable that it is multiplied with which are represented by x_1, x_2.
For example, if β_1 = 2 for a regression equation, by definition it means that when all other variables are kept constant then the change in y is twice the change in x_1. The assumption that all other variables remain constant is based on the fact that the variables do not show multicollinearity or linear relationship amongst themselves. The definition of regression equation becomes futile if the variables are multi-collinear. Hence, the very first step of regression analysis is dimension reduction where variables showing multicollinearity are identified and removed to justify the definitions and assumptions of a regression equation. Multicollinearity is a phenomenon where one variable is highly linearly correlated or can highly predict another predictor variable in a regression equation with high accuracy. Thus, when multicollinearity exists, it highly affects the coefficient of an individual variable in the equation to the extent that those coefficients show absurd values. If high degrees of multicollinearity exists, the machine learning techniques of regression used cannot obtain an invertible matrix of predictor variables which is required for analysis by regression. Another consequence of multicollinearity is that standard error of estimated parameters or coefficients tends to be too large making it difficult to make concrete conclusions and inferences from the results of regression analysis. Moreover, a small change in input data could lead to huge changes in regression results even in the change of signs of parameters. One way of good regression analysis is following the thumb rule of choosing variables which are less correlated to each other while highly correlated to the dependent variable. Such models are statistically robust and referred to as 'low noise' model with lesser error values. While multicollinearity is bad for regression it is not always bad in all scenarios. If the values of parameters and coefficient are irrelevant to understand the relationship with each predictor and if the new data set on which prediction will be tested has the same multicollinearity then it is not required to pay attention to this problem and solve it. One of the ways of detecting multicollinearity is VIF or variance inflation factor.

VIF is calculated by the formula,

$$VIF = \frac{1}{1 - R^2} \quad (2)$$

Where R^2, is the coefficient of determination which measures the portion of variability explained in by statistical model, in case of linear regression it is equal to the square of the correlation coefficient. For the model to be not distorted by multicollinearity, the coefficient of determination should have small values and VIF should have values less than 5 for each variable in the model. Thus, the VIF value check should be conducted to verify and validate the regression model.

Finally, there are many ways of dimension reduction. Some of the most important ways of dimension reduction include Ratio of missing values, Low variance in column values, removal of highly correlated column values, Principal component analysis, candidates and split columns in a random forest, backward

feature elimination, forward feature elimination, linear discrimination analysis,t-distributed stochastic neighbour embedding (t-SNE) and exploratory factor analysis. Dimension reduction technique chosen for this research project is Principal component analysis and Exploratory factor analysis as they are not complex and the results can be shown in two-dimensional graphs.

3.1 Principal Factor Analysis

The principal component analysis is the most widely used technique of dimensionality reduction because of its conceptual simplicity and efficient algorithmic reduction, allowing the user to represent all the variables in a 2D or 3D plot (Jackson & Edward, 1991). Principal components are underlying structure in the dataset. A principal component axis is in the direction where the variables show high variance and are more spread out.

Steps involved in PCA are

1. The covariance matrix of all the variables are obtained.
2. The covariance matrix is processed through eigendecomposition, thus forming eigenvector and eigenvalues.
3. The components which have the highest 2 eigenvalues make up the 2 principal components of the dataset.
4. The number of principal components which captures maximum variance is selected. Usually, the 2 principal components are enough to capture 75% of the variance of the complete data set.

In more simple terms, PCA transfers the multidimensional variable data into a 2D plot based on the total variance of the complete data set. A visual representation of the simple working of PCA is shown below.

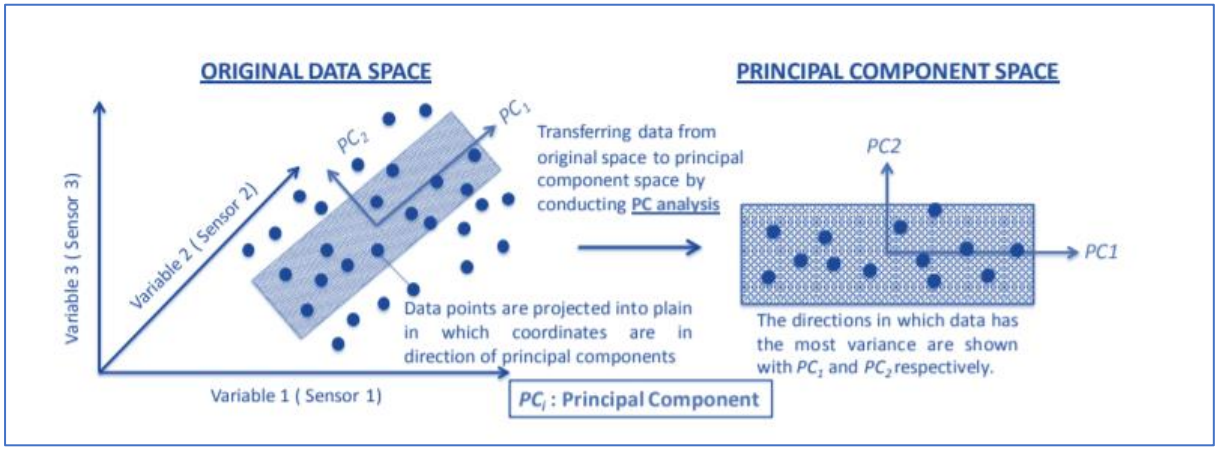


Figure 7 - Working of Principal component analysis

As shown in Figure 7, a PCA component plot can be obtained as all the variable are decomposed into zero mean and thus they are centred and orthogonally rotated. A typical PCA bi-plot can be seen below.

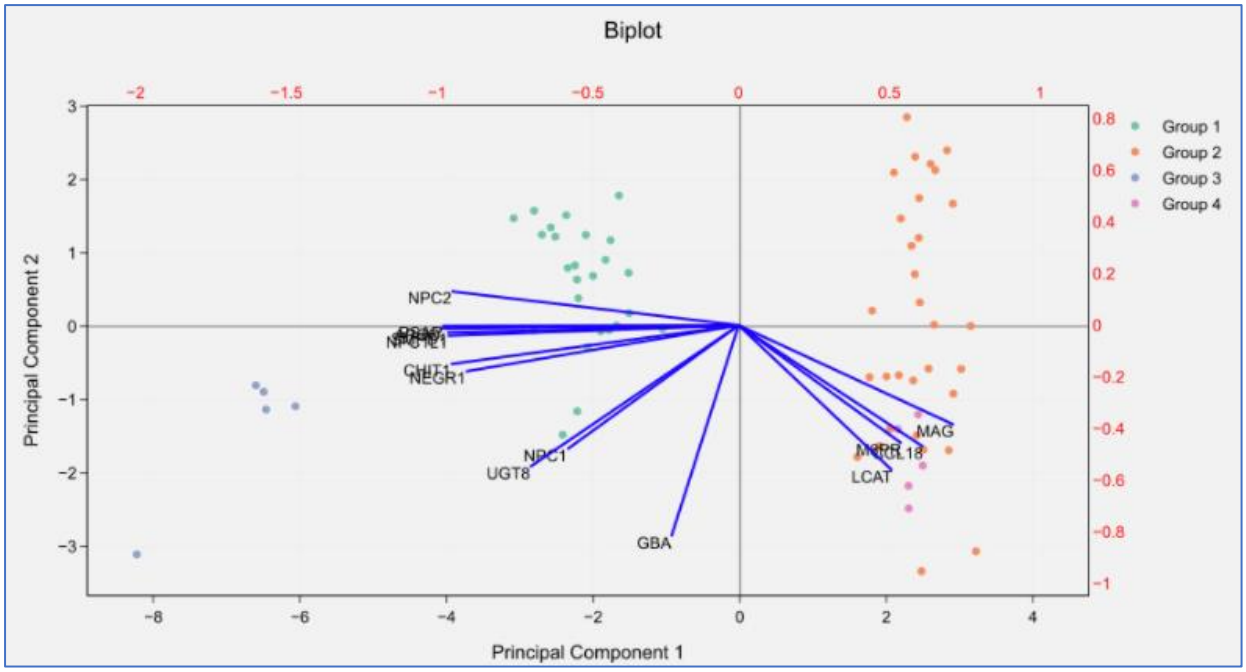


Figure 8 - Example of PCA biplot output

The X-axis shows Principal component 1 score and Y-axis shows Principal components 2 scores, which are obtained from their corresponding eigenvalues. The dots in the plot shows the PCA score of each observation of the variables involved, which is like a 2D representation of a multidimensional variable obtained by centring the mean to zero and orthogonally transforming the data.
The blue line from the centre to the variable shows the loading or score of that particular variable from the centre. Loadings plot also shows the influence of each variable on each of the principal components. The farther the variable from the centre, the more variance particular variable shows in the principal component. For example, in figure 8, MAG variable shows the highest variance in positive 1st principal component. If two variables are unrelated, they will show a 90-degree angle between them, while highly correlated variable will have the smallest angle among them or a complete 180-degree angle. When the loadings line diverges almost forming 180 degree that means they are negatively correlated. For example, the Figure 8, MAG and GBA are highly uncorrelated to each other while MAG and NCP2 show almost high negative correlation. Loadings bi-plot also show the influence of each variable on each of the principal components.

The advantages of PCA are same as that of dimension reductions–they are easy to compute and speeds up machine algorithms by counteracting curse of dimensionality.
They are some disadvantages to the usage of PCA. PCA only considers the total variance of variable and not unique variance or unexplained variance of variables. Thus, the interpretability of PCA is decreased if any variable shows a very high standard deviation. Unless the variance is justified as compared to other variables, it will give unnecessary importance to a variable to high variance. Hence, variables with different units can show abrupt results, this can be solved by normalizing the data, but it should be noted that normalization can cause information loss.
PCA is also not robust against outliers and will react highly to it if there are any. PCA also assumes a linear relationship between the variables and thus cannot capture the non-linear relationship between the variables. To counteract the disadvantages of PCA, exploratory factor analysis is also conducted in the research project.

3.2 Exploratory factor analysis

Exploratory factor analysis is a statistical method to find out underlying factors also known as latent variable from an enormous set of variables. EFA is conducted to find out the unseen relationship between the variables in each data set. This can be highly useful for SHM if the damage is known and set environmental variables are recorded. Underlying or latent factors which affect the damage can be found out by using EFA within the measured parameters. It is usually used when the user or researcher has no prior idea or hypothesis about patterns within measured variables. Steps involved in factor analysis are almost the same as that of Principal component analysis, the major difference occurs in consideration of the total variance of variables. PCA assumes that total variation is taken up by common variance for analysis while EFA splits total variation into data set into common variance and unique variance. The unobserved or latent variables that make up common variance is called a factor, similar to the PCA counterpart where it is called principal components. If the goal of analysis between variables is to simply reduce variables into a linear combination of smaller components, then PCA application is appropriate. Otherwise, if the user believes that there are some latent factors which highly influence the interrelationship between the variables, the EFA is more appropriate. A visual representation of variance consideration for EFA and PCA is shown below.

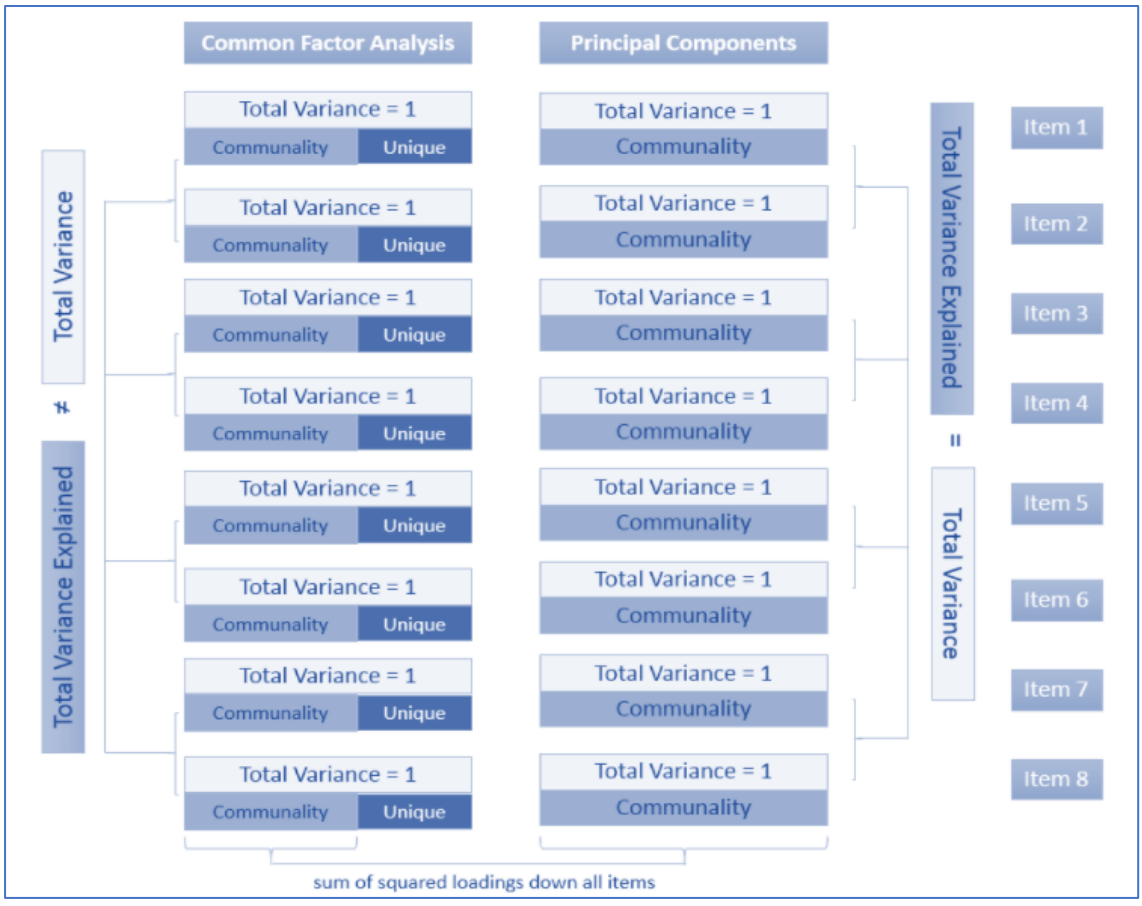


Figure 9 - Difference between PCA and EFA

As shown in Figure 9, the difference between the two approaches is just the variance consideration, the total variance explained in EFA is not equal to total common variance as unique variances is also considered unlike that of PCA analysis. The general approach for variance explanation is the extraction process which is based on Maximum Likelihood estimation. Maximum likelihood estimation is a method of

maximizing a likelihood function which gives the most probable estimate of data from observed data points and a set of parameters chosen. A bi-plot of EFA is same as that of PCA bi-plot which was explained the previous chapter.

3.4 Regression analysis

Regression analysis is a supervised machine learning tool used to form models that predict the change in the dependent variable with a given set of independent variables. The most basic form of the regression model is Linear regression. A linear regression equation includes parameters (β), independent variables shown by x, the dependent variable shown as Y and error term shown as Ɛ. It fits a line based on the ordinary least square method which is nothing but minimizing the distance of each data point from the predicted line which needs to be fitted. Essentially, regression gives the best estimate of predicted values from a given set of observed values. When there are multiple variables as input to regression analysis as independent variables, then the regression analysis is called multivariate regression analysis.

Let's consider a simple linear regression with just one dependent variable given by the following expression:

$$Y_i = \beta_0 + \beta_1 * X_i + \varepsilon_i \quad (3)$$

Where Y_i is the dependent variable. β_0 is the Y-intercept constituting the linear component of regression. β_1 is the slope co-efficient again constituting the linear component. ε_i is the random error term

First, by definition regression line is the line which fits closely to all the data points at once. This is done by minimizing the sum of all squared distance of each point to the predicted line, this method is also called the method of Least Squares. The following figure 10 shows the computation of linear regression analysis for single independent variable X and dependent variable Y wherein the line is adjusted so that squares of the distance between observed values and predicted values are minimized.

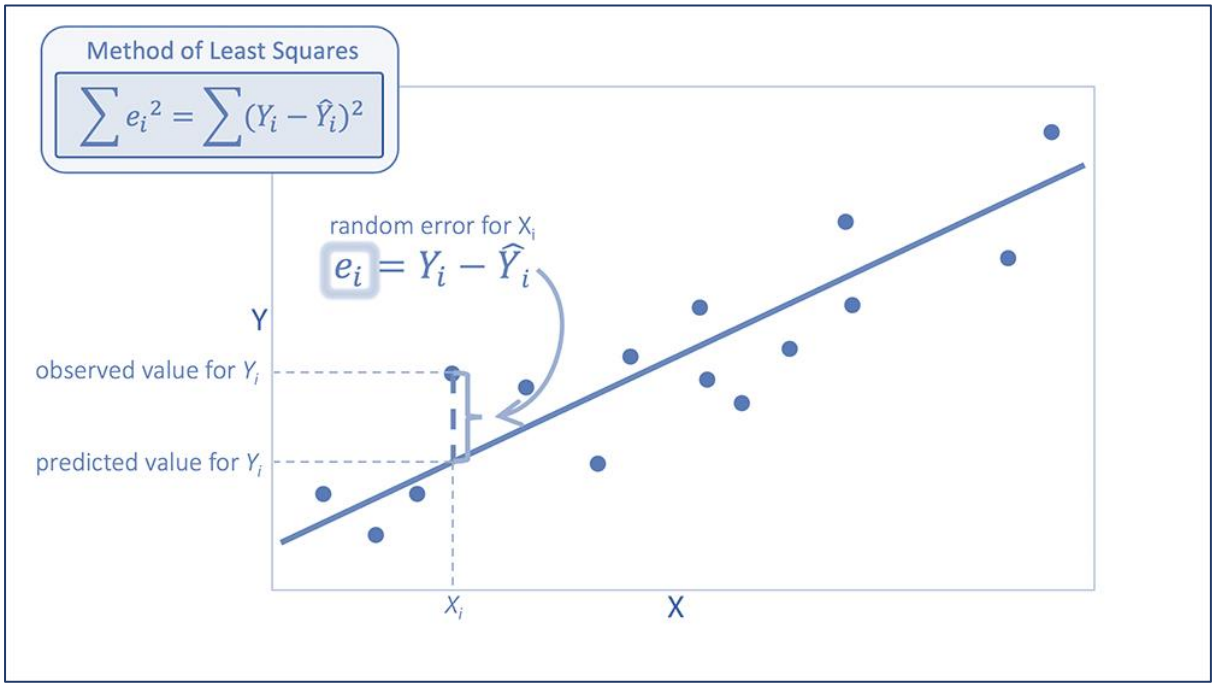


Figure 10 - Simple linear regression example

The very first step in linear regression is estimating coefficients or parameters. For the difference between the predicted value and observed values to be minimum, the following should be minimized–

$$Net\ residual\ (e_i) = \sum_{i=1}^{n} (Actual\ i-th\ value\ )\ -\ (predicted\ i-th\ value) \quad (4)$$

The i represents the individual value of the independent variable. Here residual is the subtraction value between the observed value and predicted value. But residuals can be positive or negative, leading the cancellation terms and reducing the net effect of minimization, this can be solved by using RSS or residual sum of squares

$$RSS\ = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (5)$$

where y_i is the observed value and $\hat{y}_i$ is the predicted value of i-th observation.

With a simple calculation, the value of β_0 and β_1 can be found for a minimum value of RSS as followed

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad (6)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (7)$$

where $\bar{x}$ is the mean of values of independent variable X. $\bar{y}$is the mean of values of dependent variable Y. n is the total number of observed values of dependent and independent variable.

The next step involves evaluating how the regression line has done its job of fitting the values. This can be learned through the following output results of a regression analysis.

The standard error of coefficients or SE is the error in coefficient estimates. SE of coefficient represents the average distance that observed values deviate from the regression line. If the SE is small, then the model can estimate the coefficient with great precision. The SE for simple linear regression can be found out by following formula.

$$\mathrm{SE}\left(\hat{\beta}_0\right)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right] \quad (8)$$

$$\mathrm{SE}\left(\hat{\beta}_1\right)^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad (9)$$

Where σ is the standard deviation of observed values of independent variable X.
SE can also provide Confidence interval (CI)–A 95% confidence interval gives a range where there is a 95% probability that the range of coefficients will be true from the known values of the variable. The range can be defined by a lower limit and upper limit.

SE is also used to perform hypothesis testing. If we consider H0 as a null hypothesis that there is no relation between the observed value and fitted values, then $\beta_1 = 0$. H1 can be a hypothesis where there is the relationship for which β_1 is not equal to zero. To reject the null hypothesis, β_1 should be far away from zero and there should be confidence that β_1 is non-zero for calculated SE. To perform the test, the p-value method which is obtained from SE. If the p-value is lesser than the threshold (which is 0.05 for 95% CI)

then we reject the null hypothesis and conclude that a linear relationship exists between observed values and fitted values.

The last step in regression analysis involves assessing the accuracy of the model, this done through the study of the following terms

1. R^2 or R-Squared or correlation coefficient is given by the formula

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} \tag{10}$$

where, $$RSS = \sum(y_i - \hat{y}_i) \; and \; TSS = \sum(y_i - \bar{y}) \tag{11}$$

y_i is the observed value and $\hat{y}_i$ is the predicted value while $\bar{y}_i$ is the mean value of all the fitted values.

R-Squared measures the proportion of variability in Y than can be explained by the observed values of X. The value of R-Squared ranges from 0 to 1 wherein the larger the values of R-Squared shows the quality of the relationship between fitted values and observed values. For example, if R-Squared of a model is 0.8, it means that 80% of the variability in the dependent variable Y, is explained by the independent variables, X.

2. Root Mean Square Error or RMSE is the standard deviation of the residuals, which are practically prediction errors. It represents the measure of the spread of residuals. It is found out by using following formula :

$$\text{RMSE} = \sqrt{1 - r^2} * SD_y \tag{12}$$

Where SD_y is the standard deviation of Y dependent variable values.

3.4.1 Gaussian process regression (GPR)

The Gaussian process is one of the most powerful tools in machine learning (Rasmusen & Williams, 2006). The Gaussian process has made its waves in modern machine learning as it not only useful for fitting a function to data but can also be used for classification and clustering tasks (Kim & Lee, 2007). One of the most important aspects of the Gaussian process is that it assigns a probability distribution to each function which can make a fit for predicted values which allow for flexible confidence intervals (Rasmusen & Williams, 2006). GPR is also known as a nonparametric and semi-supervised machine learning tool as there are infinite set of parameters or coefficients used in the Gaussian process and also it does not assume a relationship or function between independent variables and dependent variable unlike Linear regression and polynomial regression.

As the name suggests, GPR is based on the Gaussian distribution (also known as the normal distribution) of each random variable involved. With GPR of the multivariate dataset, each random variable is distributed normally, and the joint distribution is also Gaussian. The multivariate Gaussian distribution is defined by a mean vector (μ) and covariance matrix (Σ). Thus, the data set is converted into the following expression –

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \sim \mathcal{N}(\mu, \Sigma) \tag{10}$$

The sign $\mathcal{N}$ is the Gaussian distribution where the mean vector μ describes the expected value of the distribution, while Σ represents the covariance matrix among each dimension and determines how the variables are correlated to each other. This is because a covariance matrix consists of variance σ_i^2 of each ith variable as diagonal elements and the off-diagonal elements have correlation σ_{ij} between i-th and j-th variable.

In GPR, X becomes the training set of data points and now consider Y as a set of data points which is predicted values set. Y gives the predicted values after GPR is conducted. The Gaussian process aims to model the underlying or latent distribution of X together with Y as a multivariate normal distribution, thus, the joint probability distribution, $P_{X,Y}$, covers all the possible function values that give a predicted value from the training data X and predicted data Y. To perform the regression, Bayesian inference will be observed to update every hypothesis at every step as the new data points of X are added. Bayesian inference is a statistical method in which Baye's theorem is used for updating the probability of a hypothesis as more evidence or information becomes available. After every Bayesian inference, Y changes, thus creating a conditional probability, $P_{X|Y}$, which is also distributed normally. In order to set up a probability distribution which can predict values of test data after conditioning it with Y at every Bayesian inference, we need to define μ and Σ. In the Gaussian process, μ = 0 is done which is also known as centring of the data. The interesting setting up of a covariance matrix , Σ , is done by evaluating Kernel function (k). Kernels and kernel functions are widely used in machine learning which returns the similarity measure between points. Thus, a chosen kernel function ultimately determines the character of the function which will do the prediction. Kernels can be of two types -stationary kernel such a Radial basis function or RBF and periodic kernel which returns a measure of similarity depending on the relative position of two points while non-stationary kernel like linear kernel does not have this constraint and returns a measure depending on the absolute location of the data point. An example of shape observed in prediction function depending on the type of kernel chosen is shown below in Figure 11.

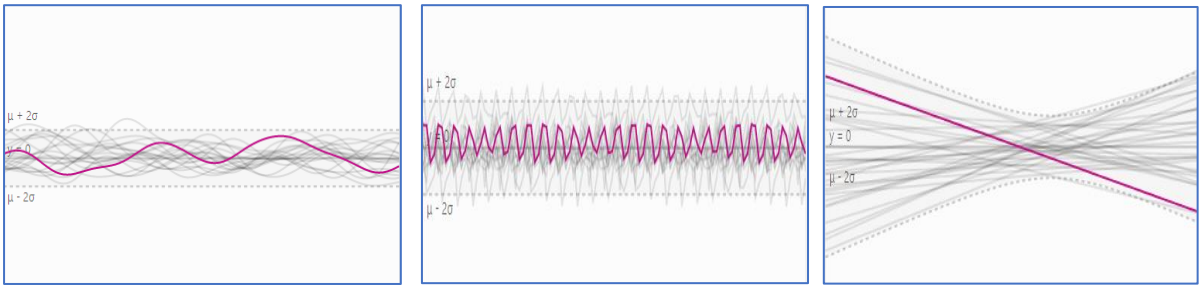


Figure 11 - Gaussian process continuous samples depending on kernel selected a) Radial Basis Function b) Periodic Kernel c) Linear Kernel

The last step of GPR is forming a Posterior distribution. Joint distribution, $P_{X,Y}$, is formed first between the training point X and predicted point Y, further using conditioning $P_{X|Y}$ can be found out. Here, conditioning leads to a new derived version of the mean and covariance matrix: $X|Y \sim \mathcal{N}(\mu', \Sigma')$. The idea behind conditioning is that the training points constrain the set of functions to those that pass through the training points. This could lead to unnecessary complex fitted functions and also in real-world scenarios it is an unrealistic assumption as data could be filled with measurement errors uncertainty. The Gaussian process offers a solution to it by modeling error of measurements by adding an error term in the joint distribution of X and Y. Then by another property of Gaussian process called marginalization, from each random variable, the respective mean function value and standard deviation can be extracted for every i-th training

point. This not only allows to form meaningful prediction, but also allows to form a confidence interval of the prediction. The above process can be shown in the following Figure 12 for 2-data points.

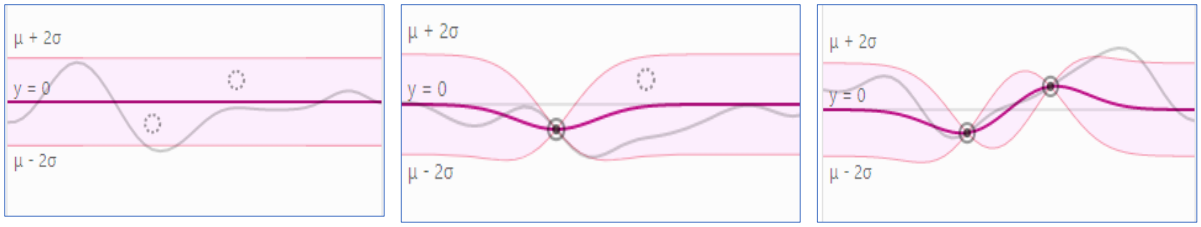


Figure 12 - Change in prediction distribution upon adding of each training point

In figure 12, at first, there are no predicted points observed, thus the mean stays at 0 and the standard deviation is the same for each test points (due to centring). A new predicted points are observed from the conditional probability of X|Y, it leads to a constrained distribution, thus shifting the mean and standard deviation and covariance matrix into a newly derived form. This process is continued following a Bayesian inference and as expected, the uncertainty of prediction is small in regions close to predicted data points and expands further away as we move far from the predicted data points.

Another property of GPR is combining kernels which allows much wider and flexibility in choosing the shape of our prediction depending on user's knowledge of data points or an optimization process can be followed to achieve the best combination of kernels to use. This makes GPR extremely versatile machine learning tool.

3.5 Verification and validation process

Validation of the trained regression model can be done in various ways. Some of the most important ways of validation are hold-out, k-fold and leave-one-out validation. Hold out validation involves the splitting of data set into 70-30 or 80-20 percentage sets for training and testing the model respectively. K-fold and leave-one-out are types of cross-validation. Cross-validation is a method of model validation which splits the data in various ways to get better estimates of model performance by minimizing the validation error. In this research topic, k-fold validation and verification process is used. This is because the data set is small and there are some missing values. The way k-fold cross-validation is carried out is by shuffling the dataset into training and test sets into k-number of folds. The following Figure 13 shows the representation of how the shuffling is carried out.

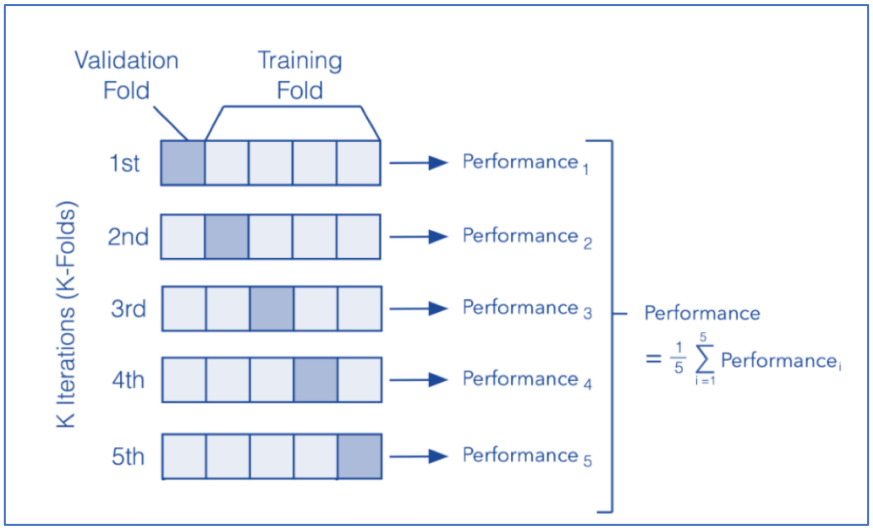


Figure 13 - k-fold validation process

In Figure 13, it can be seen that the process starts with taking one k-th group as test data set and other (k-1) groups as a training set and noting its performance. The process is repeated by moving the test set to the next k-th group and thus exhausting each group as test data. The average performance of the iterations is shown as the comprehensive performance of the model. The selection of the number of folds can influence the result of performance as an increase in k, the bias decreases but computational needs and variance of validation error grows. Cross-validation process of k-fold validation allows no information to be missed in the final performance of the model as each value is passed through training and test set. This counteracts the disadvantage of hold-out validation where information loss is possible due to the splitting of data and not exchanging it as training and test data sets.

3.6 Theoretical framework building

Thus, for answering the first research question and second research question, first combination of variable (Combination1) will be formed with a combination of extracted variables from both the sensor data and weather API data for variable analysis. This will be followed by regression analysis for choosing the prediction model by evaluation of their predictive power.

Further, for answering the third research question, Combination 2 and Combination 3 will be formed from the variable analysis which will represent extracted sensor data set and weather API data set, respectively. A comparison will be done on their predictive power achieved to conclude on efficient design of sensor systems for movable bridges. Scenario analysis is also conducted to completely conclude on cost and savings of the sensor system for bridge expansion.

Finally, a transferability check is done with the combination 1 of variables with the same dependent variable and the independent variable on the data set that of another bridge to test the performance of the transferred prediction model.

Thus, after reviewing the literature the framework shown in Figure 14 is applied to the case study to achieve answers to research questions.

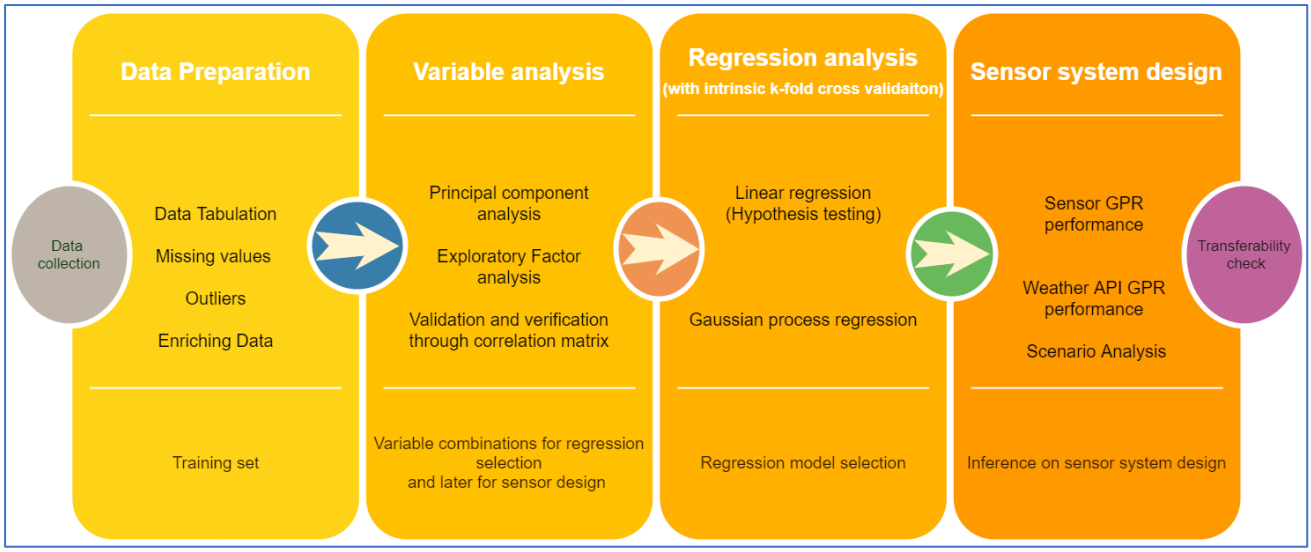


Figure 14 - Theoretical framework application on case study

This framework is in line with the methodology adopted and elaborates the whole procedure applied for the case study. The bottom text shows the results obtained from the procedure applied in the middle text on the data collected.

The literature review section is completed with theoretical framework building as the concepts and technique that will be applied are established. The next section contains the case study analysis where the theoretical framework will be applied in its completeness and results will be showcased.

4

Chapter 4 Case study analysis

4.1 Data Preparation and preprocessing

The basis of data preparation is to form a physics-driven data, thus capturing the essence of all environmental parameters simultaneously.

4.1.1 Data collection

Alkmaarsebrug is a 2-lane movable bridge which lies on N242 provincial road in Middenmeer town of Noord-Holland province in the Netherlands. The N242 road runs from A9 highway at Alkmaar at ends at A7 highway in Middenmeer right after crossing the Alkmaarsebrug. The bridge was constructed in 1990 and was put to use in 1992 after replacing an old bridge. It spans approximately 18m in length and goes over a canal called Westfriese ("Nieuwe brug in Middenmeer," 1992). The length of the movable bridge deck is 10.5 meters. Figure 15 shows the Google map location of the bridge while Figure 16 shows the typology of the bridge also obtained from Google street view.

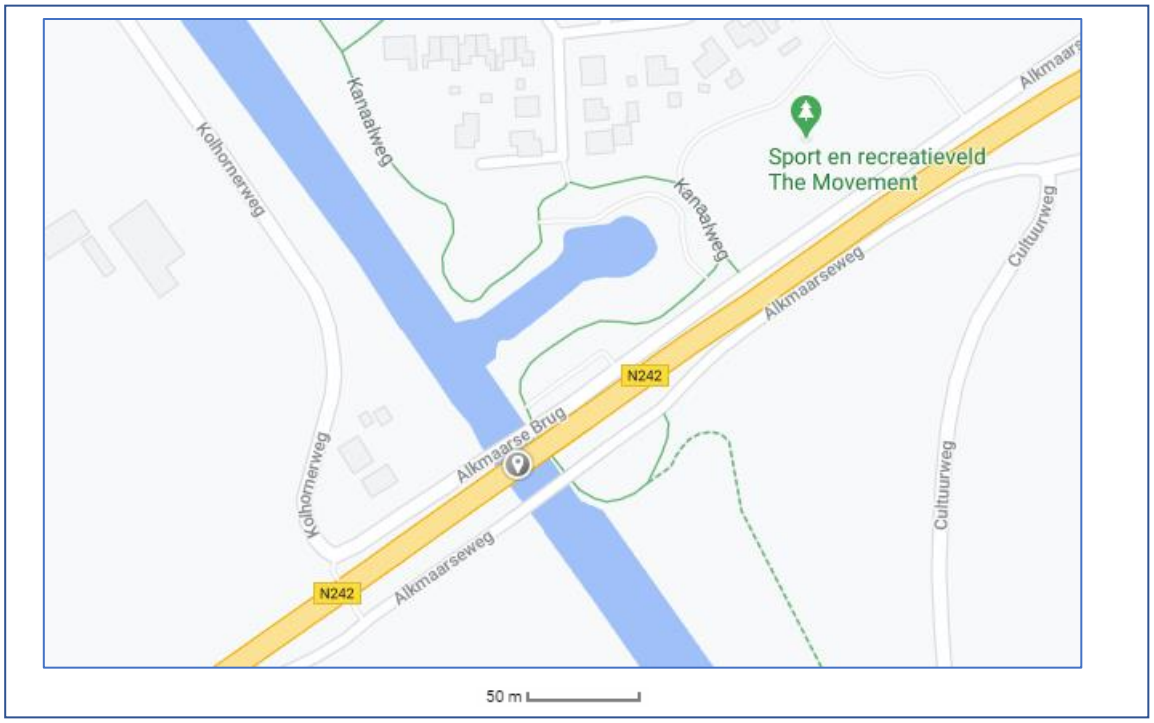


Figure 15 - Google map location of Alkmaarsebrug

Figure 16 - Typology of Alkmaarsebrug obtained from Google street view

Alkmaarsebrug face the problem of critical bridge expansion and hence, the following sensors has been installed on the bridge for monitoring purpose.

1. Distance gap sensor–The expansion joint gap between the movable deck and fixed deck is measured in millimeters by a piezoelectric sensor.
2. Water temperature sensor measuring the temperature of the water below the bridge. It is installed one meter below the water level.
3. Temperature sensor measuring the temperature of the upper side of the bridge deck installed just a few centimeters above the surface of the deck.
4. Temperature sensor measuring the temperature of the underside of bridge deck installed just a few centimeters below the movable bridge deck.

A weather station has been installed a couple of meters above the bridge, which acts like a multi-sensor measuring

5. Wind speed in meter per second
6. Wind temperature in degree Celsius
7. Wind gust in meter per second, which is the highest wind speed in a brief interval of time
8. Ambient temperature in degree Celsius

A schematic diagram of the sensor system at Alkmaarsebrug is shown below in figure 16.

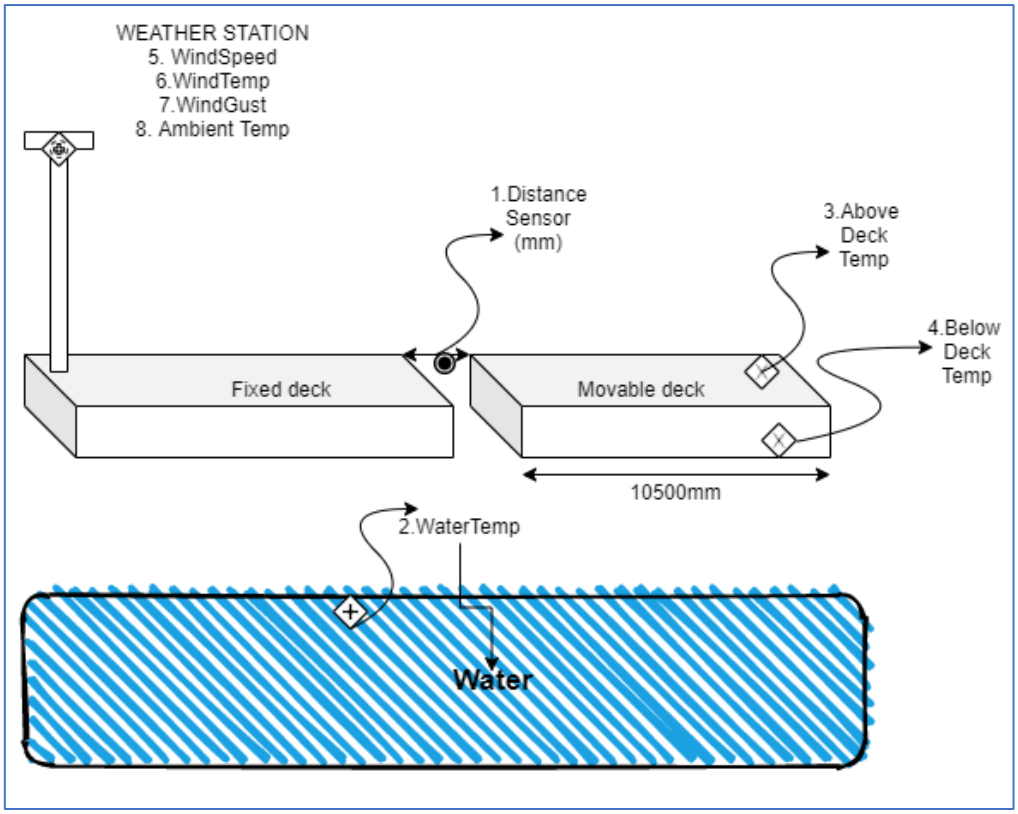


Figure 17 - Schematic Diagram of Alkmaarsebrug sensor system

The weather data considered for this project acquired from API sources and KNMI database include:

9. Ambient temperature in degree Celsius.
10. Solar tadiation in Watt per meter square from a KNMI weather station 20km away from the bridge.
11. Wind temperature in degree Celsius.
12. Relative humidity index ranging from 1 to 100.
13. Atmospheric pressure in hectopascals or hPa.

The data available of all the above parameters are available from November 2018 to September 2020 with few missing days erratically and two months of June 2020 and July 2020 completely missing.

Data is sent from the sensor to cloud database with a timestamp and with the name of the parameter which has been recorded. Thus an example of data obtained from the database is shown below in Table 1.

Table 1 - An example of the database collected

1	added date	added time	data	measured value
2	3/18/2020	12:01 AM	[weather]uv_index	2.55
3	3/18/2020	12:01 AM	[weather]temperature	8.5
4	3/18/2020	12:01 AM	[weather]pressure	1026
5	3/18/2020	12:01 AM	[weather]humidity	71
6	3/18/2020	12:01 AM	[weather]wind_speed	4.6
7	3/18/2020	12:03 AM	WaterTemp	8.3
8	3/18/2020	12:03 AM	AmbientTemp	8.4
9	3/18/2020	12:12 AM	[PNH]0059AC00001526DC_VBat	3624
10	3/18/2020	11:56 PM	Distance	55.5375
11	3/18/2020	12:12 AM	Distance1	2.23
12	3/18/2020	12:16 AM	[weather]temperature	8.49
13	3/18/2020	12:16 AM	[weather]pressure	1026
14	3/18/2020	12:16 AM	[weather]humidity	52
15	3/18/2020	12:16 AM	[weather]wind_speed	5.7

This data is gathered by the data management system like SQL. The measured values are automatically transforming from an analogue signal into digital and thus giving the right value of the parameter in the units mentioned earlier.

4.1.2 Data tabulation

The data is then tabulated with parameters as columns. The timestamp at which paratmers are recorded was sorted in the rows as shown in Table 2. There is a total of 13 variables considered, 7 data points from sensors and 5 from weather AP as independent variables.

Table 2 - Example of data tabulation

Average of dbl_v	Column Labels						
Row Labels	avgCorrectedDistance	OWTemp1	Temp2	wind_chillfactor	wind_gust	wind_speed	wind_temperature
⊞1-Dec	28.54196539	4.953333333	6.230232558	2.958635997	6.908037623	6.301928251	6.698627104
⊞2-Dec	28.04420885	5.789361702	10.64491682	8.326546497	6.969615883	6.285800866	11.04291667
⊟3-Dec	27.96348783	7.151764706	9.710056926	8.223464373	6.202374429	5.0910947	10.32784708
⊞12 AM	27.81942935	6.575	11.01764706	9.00989011	7.992708333	7.040425532	11.61926606
⊟1 AM	27.868125	6.675	10.79545455	8.60733945	8.173958333	7.101075269	11.39263158
:00	26.82625		10.9				
:01				8.64	7.86	7.04	11.36
:03	28.91	6.65	10.9				
:04	26.82625	6.6	10.8	8.5	8.066666667	7.066666667	11.4

4.1.3 Outliers and missing values

All the parameters taken from the database are taken at face value, which means it is considered that there are no errors in the measured value. An exception was considered for distance measured between the deck. There were many instances in which distance parameter showed high values and peaks as shown below in Figure 17.

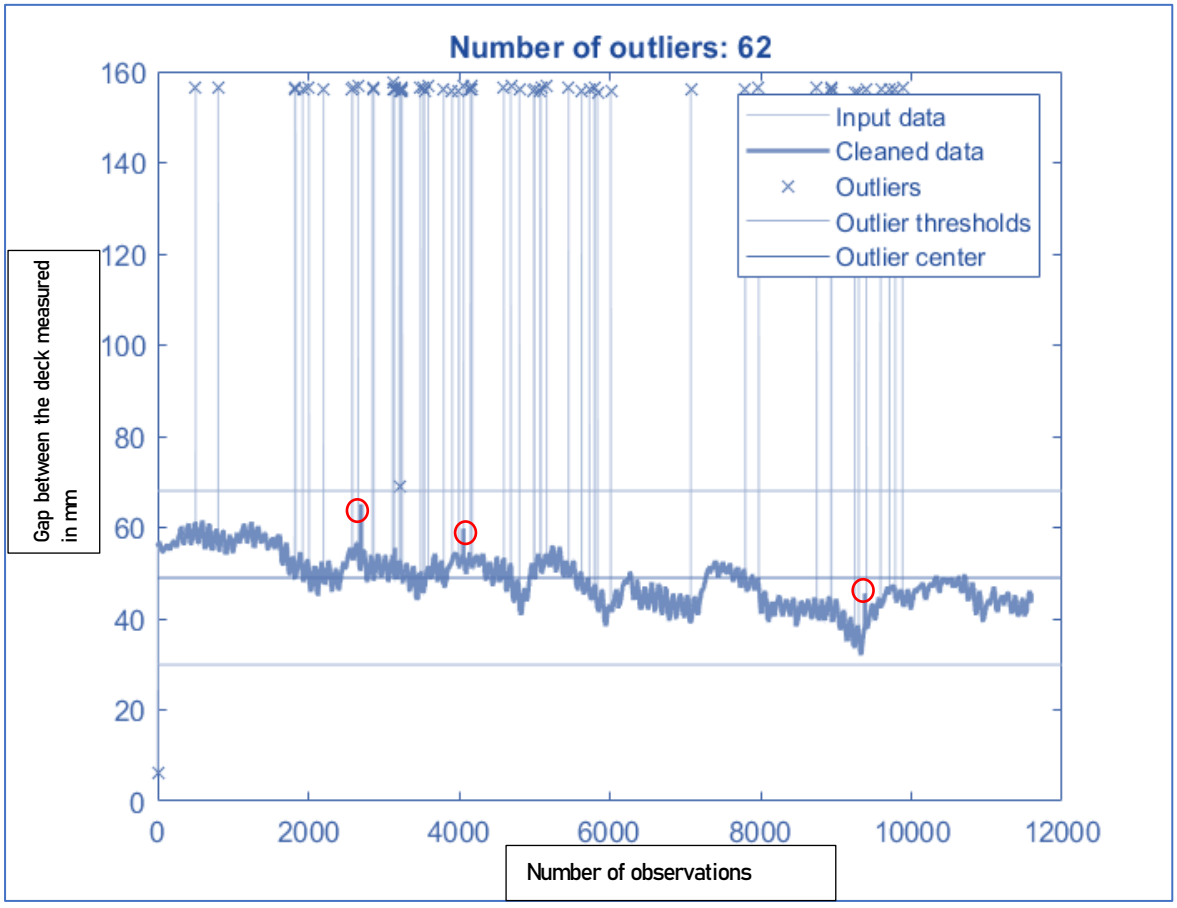


Figure 18 - Outliers removal

After observing these outliers in the distance measurement, it was found that the timestamp matched with the opening of the bridge. As these values are not a representation of the expansion of the bridge deck, these values were removed. Any value which was showing 5mm or more expansion than previous recorded value was removed considering the incorrect measurement of gap and expansion of the bridge deck.

Missing values at any timestamp of any parameter were removed from the entire database. The idea was to form a physics-driven analysis rather than data-driven analysis and hence; the way missing values were managed was just by omission.

4.1.4 Transform the data and enrichment

After meticulously observing the data, it was found that the parameters show little change in brief intervals of time. It was necessary to concise the data as the data log number was easily ranging somewhere in millions. This is not only difficult to fathom for the researcher, but any machine learning tools or software to be used will have long hours of processing time. Thus, the data was enriched and merged in 6-hour intervals wherein all the parameters were averaged to form each interval of data except for gap between the decks parameter for which minimum values are considered to not lose the information of critical expansion. Thus, there were 4 intervals each day, namely 12 am to 6 am (morning), 6 am to 12 pm (afternoon), 12 pm to 6 pm (evening) and 6 pm to 12 am (night). If at any interval there were missing values of any parameter, the whole interval was omitted because simultaneous values of all the parameters were the essence of the dataset which was needed to be preserved. Table 3 shows an example of the final enriched dataset below.

Table 3 - Example of final enriched data tabulated

timestmp	Distance	WaterTemp	AboveDEck	BelowDeck	OutsideTemp	WindChillFctor	WindGust	WindSpeed	WindTemp	SolarRadn
11/26/2018 12:00	28.73949384	3.591666667	4.371544715	4.156923077	4.951470588	1.882228916	1.761587302	1.355182927	3.031578947	87
11/26/2018 18:00	28.79495124	3.429166667	1.878723404	1.465467626	1.923076923	0.313924051	1.990641248	1.558813839	1.941080196	1
11/27/2018	28.799375	2.890909091	2.208955224	1.80703125	2.121212121	1.880525502	1.130292599	0.826553672	2.324574961	1
11/27/2018 6:00	28.79502216	2.8	2.717037037	2.387692308	3.116176471	2.464239482	0.836141907	0.409854604	2.492109501	112
11/27/2018 12:00	28.76964349	2.877777778	4.323021583	4.186764706	5.079365079	2.727133106	1.445892857	1.024879615	3.544772727	155

Also, for a transferability check of regression models, the distance gap considered was per meter length of movable bridge deck multiplied by 100. Given below is the formula used to measure gap between the decks parameter. This parameter is a representation of expansion, smaller the values means the expansion of larger.

$$Distance\ paramter = Distance\ gap\ \%$$

$$= \frac{Actual\ gap\ measured\ in\ mm}{Lenght\ of\ the\ movable\ bridge\ deck\ in\ mm} \times 100 \tag{11}$$

For Alkmaarsebrug, the length of the movable bridge deck was 10500mm or 10.5m. This scalability of distance parameter can help the model to be transferred to another bridge and check the performance of prediction models.

4.2 Variables analysis

For variable analysis, the distance gap parameter was not considered, it was done to understand the relationship between the explanatory or predictor variables. This was necessary to understand how the variables would have affected the expansion of the bridge if there was no distance measurement done. In that way, the literature hypothesis of the relationship between environmental parameters and bridge deck expansion can also be tested to some extent.

4.2.1 PCA analysis

When the very first PCA analysis was done, solar radiation showed extreme variation explanation in the first principal component. The loading bi-plot with the first principal on X-axis and 2nd principal on Y-axis is as shown below in Figure 18.

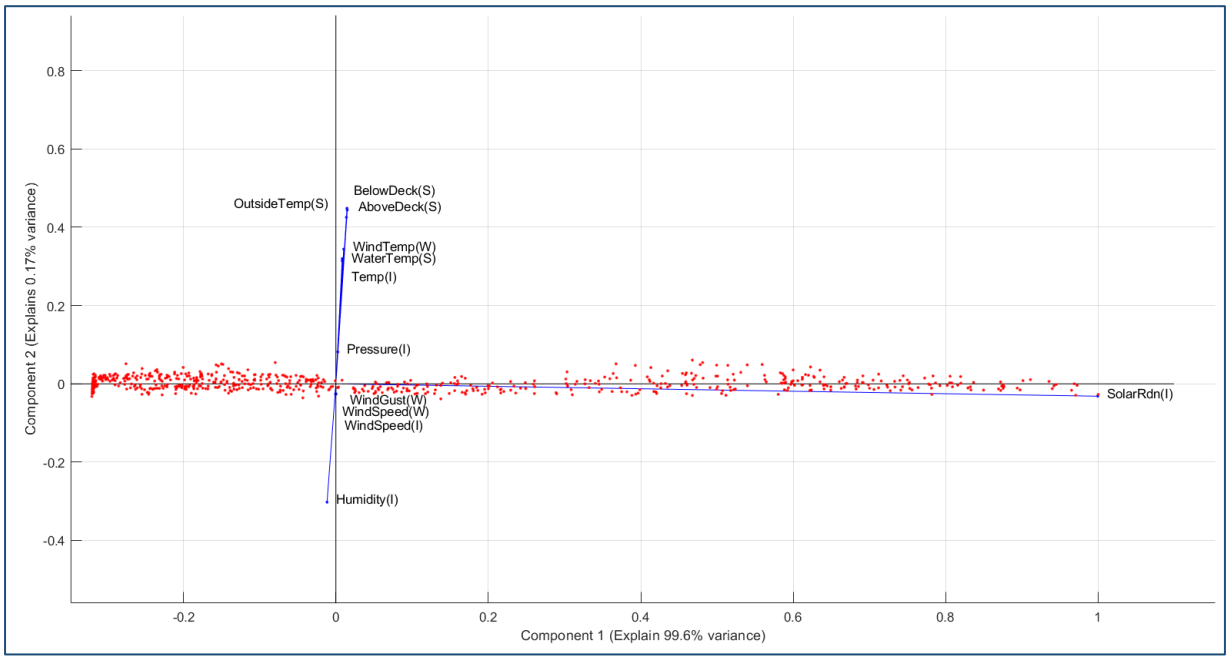


Figure 19 - PCA bi-plot result output with solar radiation and unstandardized data set

The red dots show the centered and orthogonal representation of each observation interval. The blue line shows the loading of each variable. In the first result, solar radiation is shown having maximum importance in explaining the variation in the dataset. The first principal component captures 99.6% of the variance of the entire data set. All other variables fall on the second principal component axis, but this graph does not show the importance of variables among the remaining other variables as the variance of solar radiation overpowered the individual total variance of other parameters. Thus, this result demanded standardization of variables for better interpretation of PCA analysis. Hence, a second PCA analysis is done with standardization of variables to understand the variance explanation of other variables. Another reason to opt for standardization over normalization is to not lose information of spread of the variables.

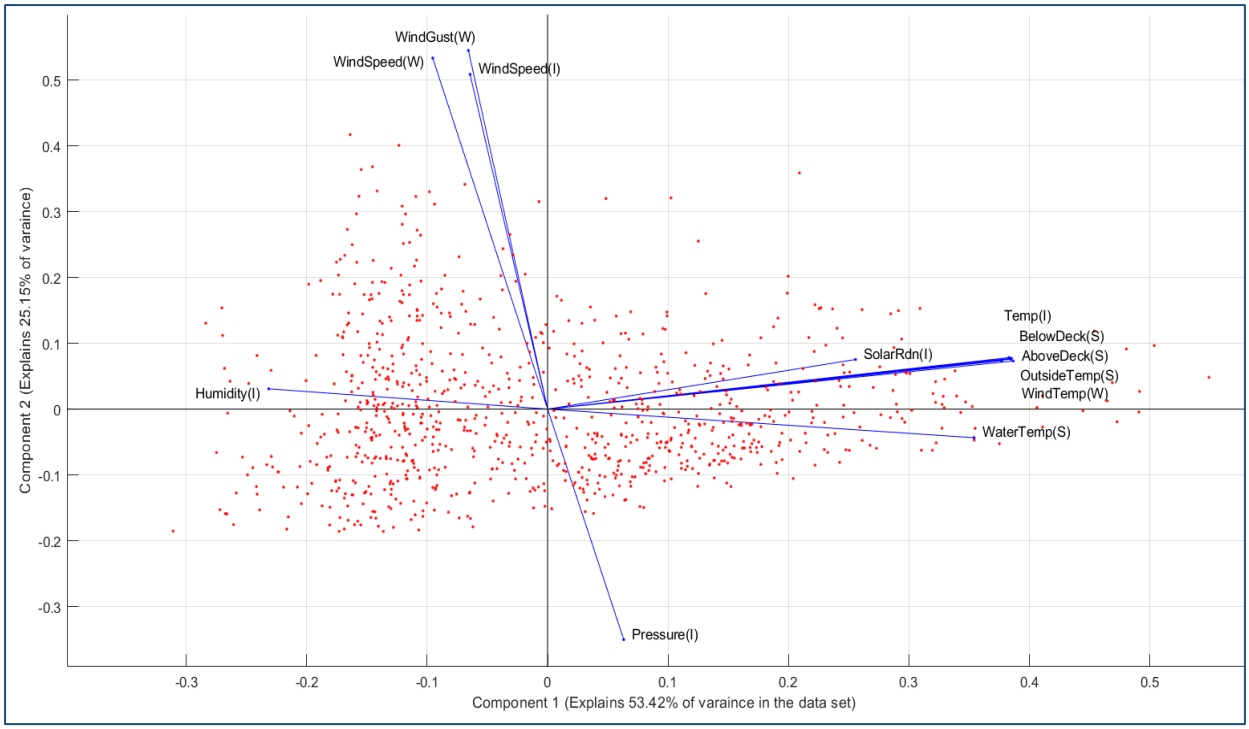


Figure 20 -PCA bi-plot result of standardized variables

Figure 19 shows PCA bi-plot result of standardized variables. This graph could explain how each variable relates to each other. The first principal component explained 53.42% variance in the whole data set, while the second principal component explained 25.15% variance. This graph shows that pressure and humidity index stand out from all the parameters considered and they are unrelated to each other.

After inspecting, we understand wind speed and wind gust measured both by sensors and weather API have a small angle between them on loadings line, this means they are correlated. They show little importance in first principal component as they lie closer to Y-axis. On X-axis and first principal component, temperatures measures above the deck, below the deck and outside temperature showed very little angle between themselves. Thus, they are forming a cluster showing a high correlation between themselves. One major inference to take from PCA analysis is the formation of these clusters of variables. Using just one variable from the cluster for further consideration is enough to represent the whole cluster. Identifying such clusters helps to reduce unnecessary variables. PCA is further verified and validated by an exploratory factor analysis.

4.2.2 Exploratory Factor Analysis

Just like the PCA graph, even the EFA graph shows loadings of variables by a blue line. After EFA analysis of all the variables, the graph looks like as shown in Figure 20.

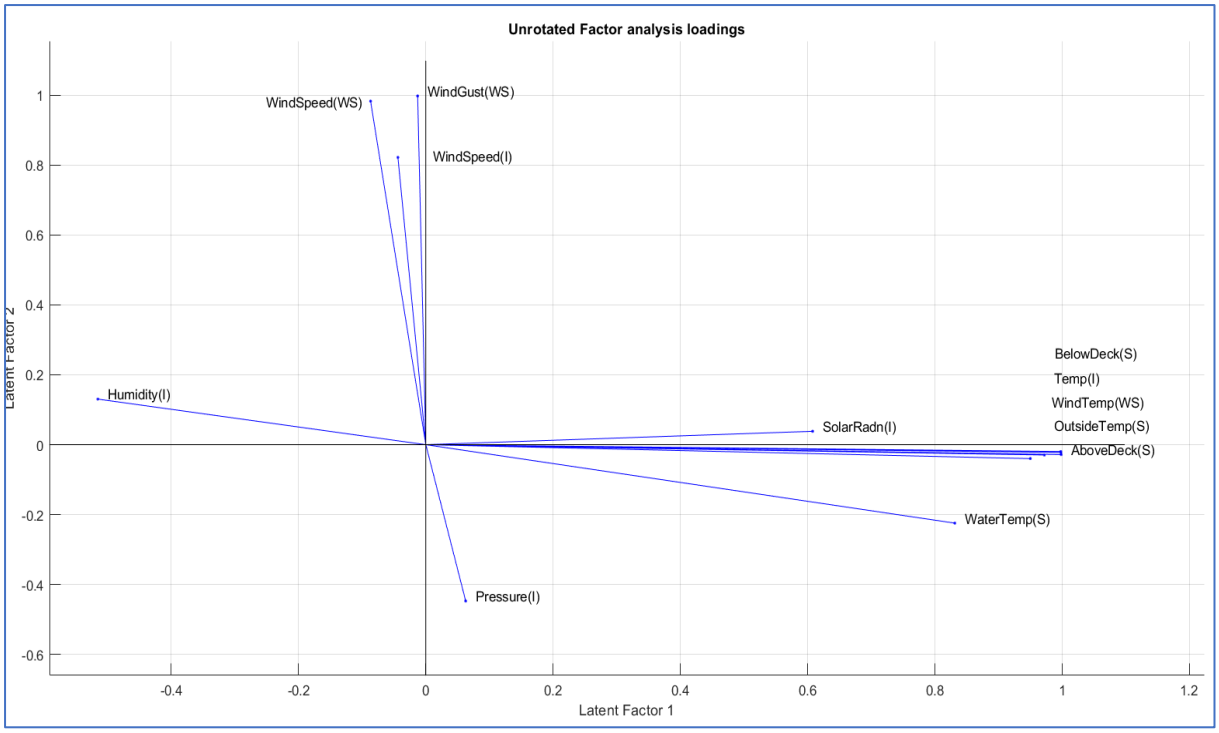


Figure 21 -EFA bi-plot of output results

Unlike PCA, there is no need for a 2-step analysis as solar radiation does not show outrageously high unique variance. This is because EFA splits total variance of a variable into a unique variable and common variable. Thus, in this case, solar radiation has a very high total variance, but the amount of unique variance may not be that high and hence, it fits in one graph unlike bi-plot of Figure 18. Thus, EFA analysis is better at showing unique variance and just does not consider only total variance as in finding out latent factors. It also does not need standardization of input variables as it can be seen that Figure 19 and Figure 20 are almost same but PCA analysis needed standardization of variables approach for better understanding of analysis. In EFA also, cluster of wind speed and wind gust both from internet and weather API formed a cluster. Solar radiation, water temperature, humidity, pressure showed a unique variance not forming a cluster. Above deck temperature, below deck temperature, ambient temperature and wind temperature observed from both weather API and sensor formed a clear cluster. One major inference formed from EFA and PCA is that there are two obvious latent factors, one is related to temperature while the other is related to wind. The results of standardized PCA and EFA are almost similar and give same inferences.

Thus, Table 4 as shown below concises the formation of clusters of observed variables.

Table 4 - Cluster formation of the variables after PCA and EFA analysis

	Cluster 1	Cluster 2
PCA	Outside Temp, Below Temp, Above Temp, Temp from weather API, Wind Temp	Wind Speed (From sensor and weather API) And Wind Gust
EFA	Outside Temp, Below Temp, Above Temp, Temp from weather API, Wind Temp	Wind Speed (From sensor and weather API) And Wind Gust

4.2.3 Verification and validation for the formation of variable combinations

The variable combination is formed from cluster information and further validated and verified by studying Spearman's correlation matrix of the whole dataset. In this validation, even the correlation with distance has been considered in forming final variable combinations.

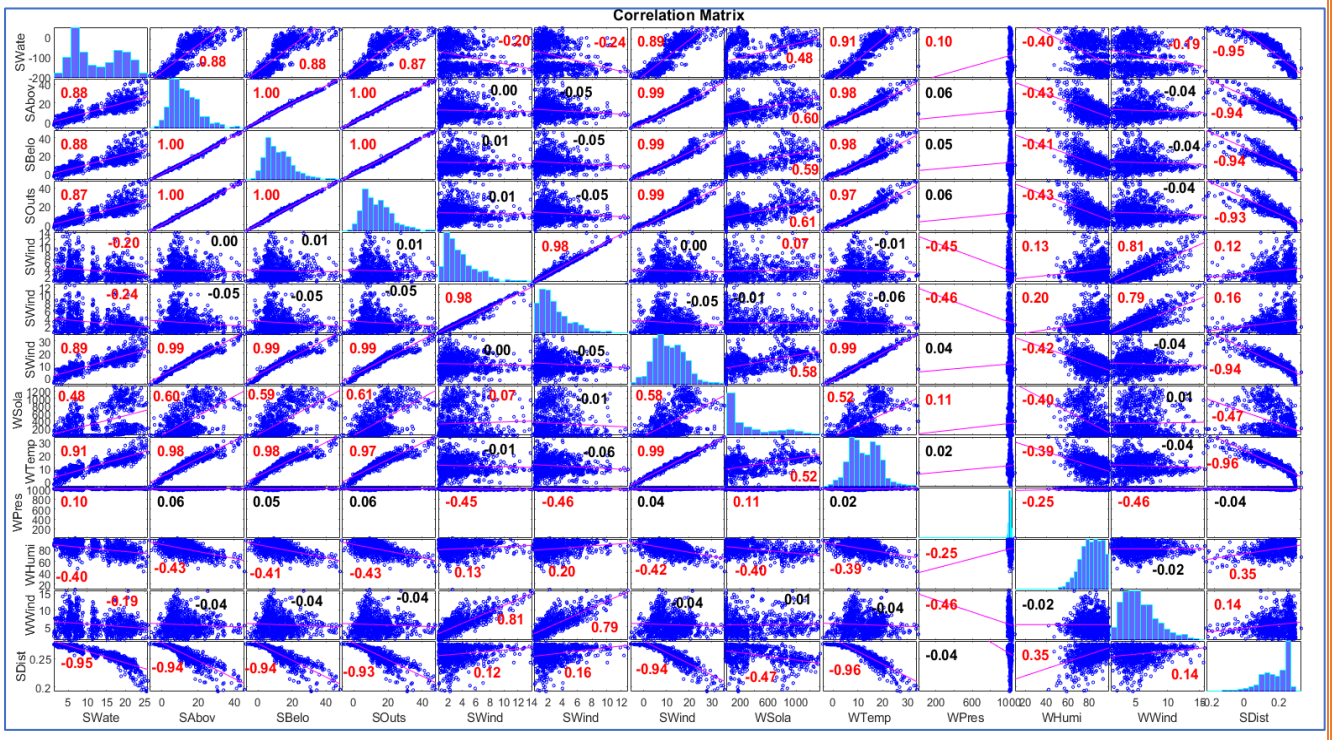


Figure 22 - Spearman's correlation matrix

Figure 22 shows the Spearman's correlation matrix of independent variables within themselves as well as with the dependent variable. The parameters recorded via sensors are marked with "S" at the beginning of their labeles while the ones recorded via weather API are marked with "W".

As the variables do not show exact normal distribution, Spearman's correlation is used over Pearson's correlation to understand the monotonic relationship between two variables and make inferences for the final formation of variable combinations.

Following inferences are made from the correlation matrix from figure 22 are as follows :

1. Outside temperature from the sensor and deck temperatures are highly related to each other. The temperature obtained from weather API also shows a high correlation to all above 3 temperatures. Thus, the cluster formation 1 from PCA and EFA are validated.
2. Wind speed from sensor and weather API and wind gust shows a high correlation. Even wind speed from the internet shows 80% correlation with wind speed from the sensor. And both the weather API wind speed and sensor wind speed shows 0.19 and 0.22 correlation with the distance parameter. This shows that both wind speed values can be used interchangeably for regression analysis, making the other redundant.
3. Pressure parameter obtained from weather API shows 0.07 correlation with distance parameter which is extremely weak and hence, can be allowed to be omitted.

Thus, after analyzing the inferences on the variables from above 3 processes, following variable combinations are made for the purpose of regression analysis study and sensor design consideration.

Table 5 - Final combinations of variables formed after verification and validation of variable analysis

Combination 1 (For hypothesis testing and regression model selection)	Combination 2 (For sensor system design inference representing only sensor data)	Combination 3 (For sensor system design inference representing only weather API data)
Above the deck temperature(sensor)	Above the deck temperature(sensor)	Temperature (Weather API)
Water Temperature (sensor)	Water Temperature (sensor)	
Solar Radiation (weather API)		Solar Radiation (weather API)
Wind Speed (Weather API)	Wind Speed (sensor)	Wind speed (Weather API)
Humidity (weather API)		Humidity (weather API)

Before feature extraction and dimension reduction, the variance inflation factor (VIF) was around 100 for each of the variables, thus proving the problem of multicollinearity existing within the data set. The new VIF for 5 variables selected is less than 5, which verifies that the problem of multicollinearity has been removed from the variable set after removal of redundant variables from the total set of 13 variables.

Thus, this section completes the complete feature extraction or dimension reduction step of the research methodology. In the next section the combinations formed are run through regression analysis and the results are discussed.

4.3 Regression analysis

As it was seen in the first step of PCA analysis, the data shows high variance for some variables because of its measure of units. To avoid further problems in interpreting the results of regression analysis and to exercise feature scaling, which is an important first step of regression analysis, standardization of variable is done.

The results of regression analysis will also be done based on its performance of critical values. For research purpose and with the practical assumption, the critical value of gap between the bridge deck can be considered to 20mm. This critical value represents that once this threshold has reached, the expansion momentum has set-in and thus after this point the bridge has to be closed to avoid any hazards. Also, once critical values are reached, the bridge deck needs to be cooled down immediately for it to be functional again. Thus, when the distance between the bridge deck is 20mm or less, we can say that remedial action needs to be taken to avoid further thermal inertia setting in the bridge, causing further problems in availability. The distance parameter is measured in percentage per length of the bridge deck, hence, in those terms, 20mm shows 0.21% gap distance per length of the bridge deck. The mean of the distance gap variable is 0.25% while the standard deviation is 0.01352%. Thus, finally, the critical value in terms of the number of standard deviation from the mean is -3 standard deviations. This critical space of values will be highlighted in the results of the regression analysis to understand its performance in predicting critical values. Also, all regression steps taken are done with 10-fold cross validation, thus making validation of the prediction model an integral step of the analysis.

4.3.1 Linear regression results.

First, linear regression was applied with 5 variables of combination 1 as the independent variables and distance gap percentage as the dependent variable. The output of regression analysis is shown below in table 6, it should also be noted that results are 10-fold cross-validated. The estimate shows the value of the coefficient for the corresponding variable, SE shows its standard error of co-efficient while p-value shows value for hypothesis testing of each parameter.

Table 6 - Linear regression results

	Co-efficient Estimates	*SE*	*pValue*
(Intercept)	*-2.8589e-11*	*0.013197*	*1*
WaterTemp in °C (sensor)	*-0.59864*	*0.022308*	*<0.05*
AboveTemp in °C (sensor)	*-0.33071*	*0.024219*	*<0.05*
SolarRdn in W/m² (Weather API)	*0.17806*	*0.01661*	*<0.05*
Humidity (0-100) (Weather API)	*-0.058671*	*0.015321*	*0.00013234*
Windspeed in m/s (Weather API)	*0.094164*	*0.0141*	*<0.05*
Root Mean Squared Error: 0.6			
R-Squared: 0.642			
p-value of the model = 0			

From this result, we can see that p-values of each variable are less than 0.05, which means all the variables show a statistically linear relationship with gap parameter. It can also be seen that the regression line can explain 64.2% variation in the dependent variable through the 5-variable input as R-Squared is 0.642. The p-value of model show that the model can be considered passing the hypothesis test. Thus, rejecting the null hypothesis which says that the model shows no relationship between the dependent variable and independent variable.

The accuracy of the linear regression model is evaluated with the help of predicted vs true response graph which is shown below in Figure 22.

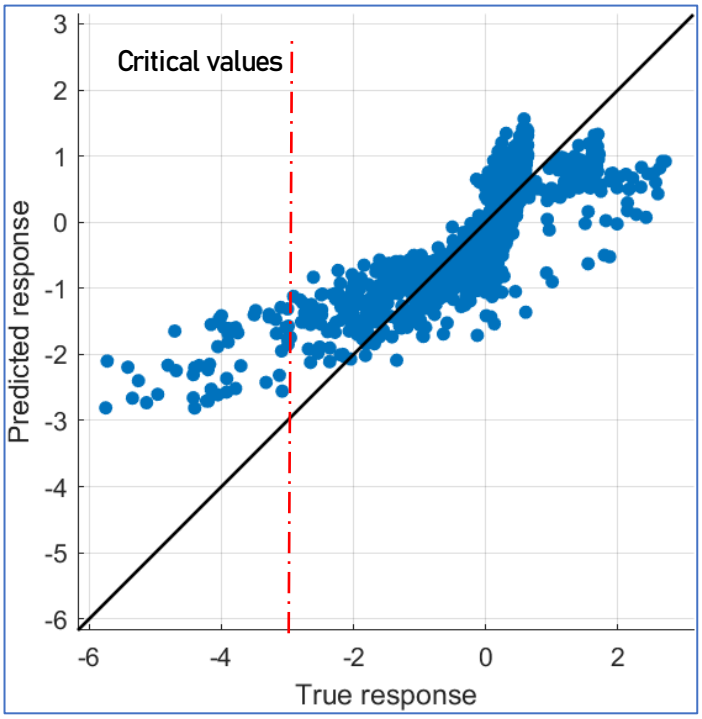


Figure 23 - Predicted response vs True response of linear regression model

As seen in Figure 22, for lower values of the distance gap parameter, the predicted values are underestimating the severity of the distance gap. For example, when the true value is -4 SD, the linear model shows a prediction of -2 SD. All the values below -3 are critical values and linear regression is proving to not do a good prediction of those values as the blue dots are not falling on the black line where the predicted response is equal to the true response. Thus, the distance gap predicted is underestimating the true response for critical values. The predicted value of the distance gap is more than the actual gap present. Underestimation of values could lead to a wrong decision by the asset manager, as critical expansion could happen, but prediction won't be able to capture it. This could lead to not closing the bridge when it requires to be closed. This could also lead to not pursuing remedial measures quickly, which would be required to stop expansion inertia. Unexpected penalty can also be levied on asset manager, as unplanned closure of lanes has to be done once the actual critical expansion is observed.

4.3.2 GPR analysis

Table 7 shows the results of GPR analysis done with 5-variables as independent variable and distance gap parameter as the dependent variable with 10-fold cross-validation. Table 7 also shows the type of Kernel function used and other information about the GPR model type.

Table 7 - GPR results from the output

Results	
RMSE	0.42032
R-Squared	0.82
MSE	0.17667
MAE	0.26228
Prediction speed	~12000 obs/sec
Training time	87.816 sec

Model Type

Preset: Matern 5/2 GPR
Basis function: Constant
Kernel function: Matern 5/2
Use isotropic kernel: true
Kernel scale: Automatic
Signal standard deviation: Automatic
Sigma: Automatic
Standardize: true
Optimize numeric parameters: true

Thus, RMSE of GPR is 0.42 which is better than RMSE of the linear regression which was 0.642. In terms of millimeter gap distance between the bridge deck, the results show an error of approximately 0.55mm and 0.83mm from GPR and linear regression, respectively. The R-squared also performs better in GPR with 82% explanation of variation in expansion from the 5-variable input as an independent variable.

But R-Squared and RMSE should not be the only reason to select the GPR model over linear regression. Hence, how the model performs on actual true response vs predicted response graph which is shown below in Figure 23.

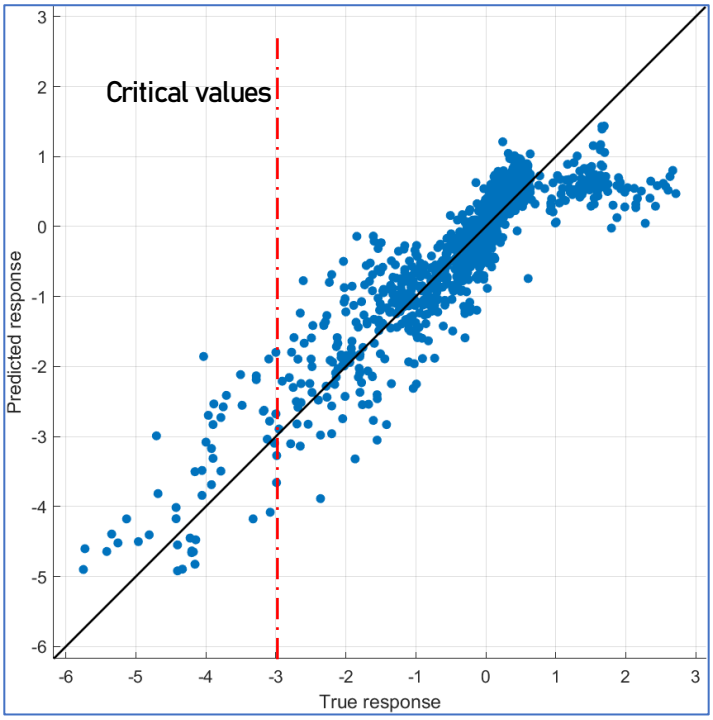


Figure 24 - Predicted response vs True response of GPR model

Here, for critical values lower than -3 SD, the model performs better than the linear model. There are still some values where underestimation is seen, but it is much better than the performance of the linear regression model.

A final visual presentation of regression model performance on observed value by plotting a confidence interval of 95% as shown below in Figure 24 and 25 of the Linear regression and GPR respectively. On X-axis is the number of observation. Green light depicts the predicted value while the grey-shaded area shows the confidence interval around the predicted value.

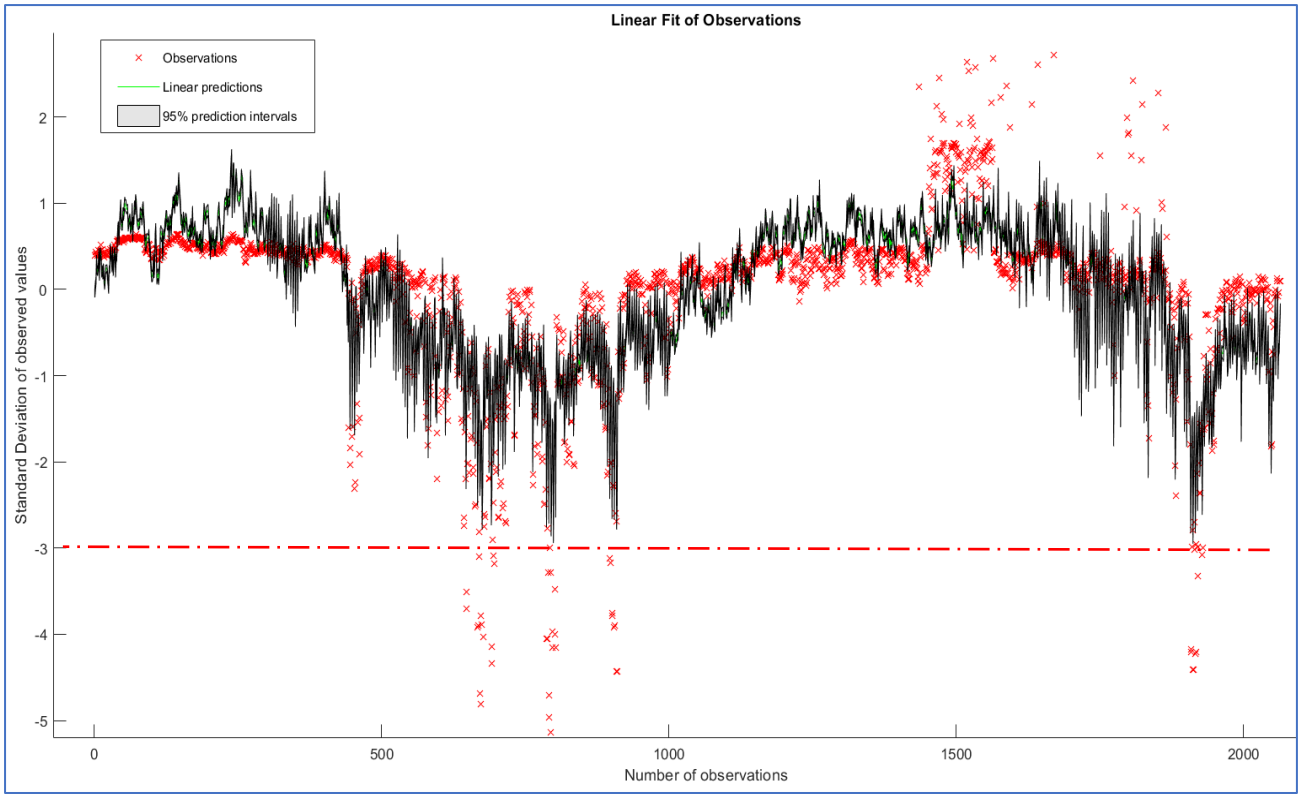


Figure 25 -Linear regression model fit of standardized observed values with 95% Confidence intervals

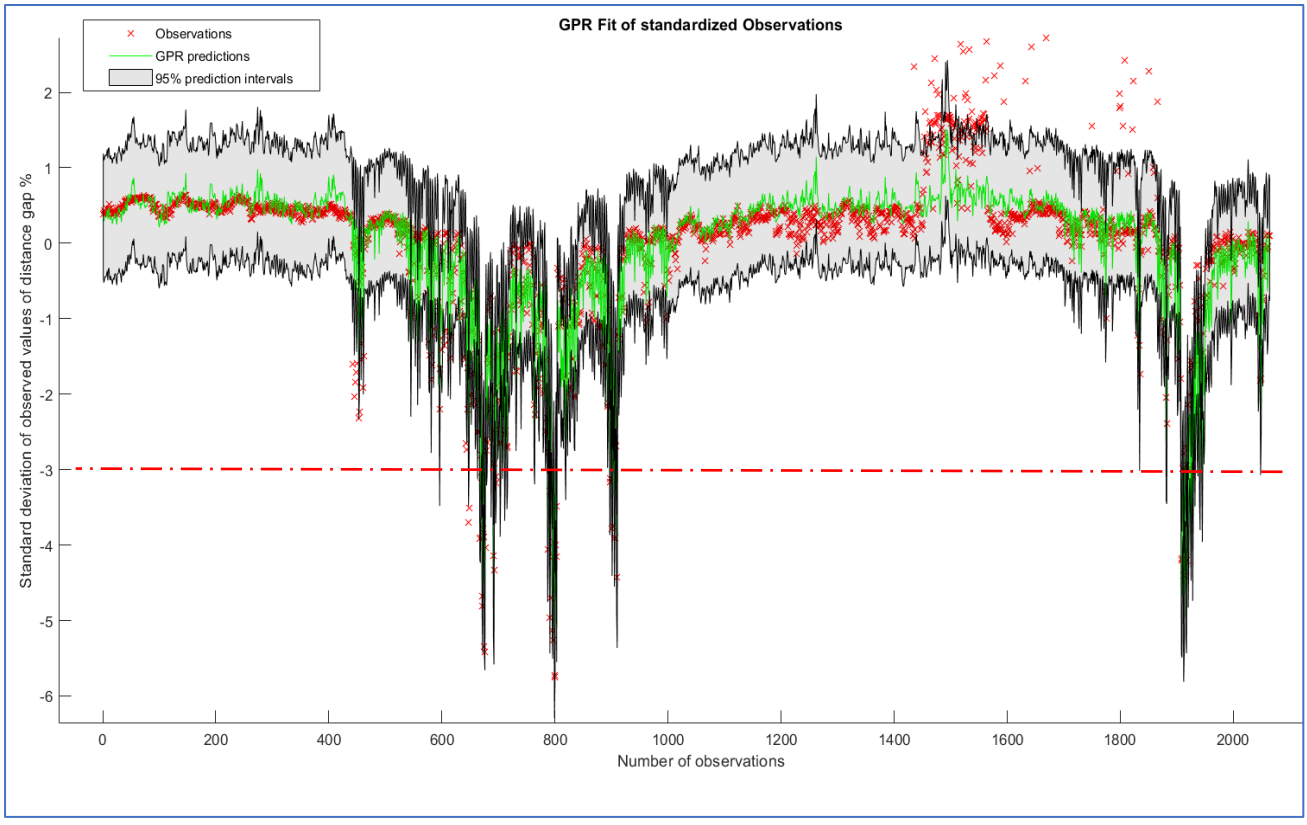


Figure 26 - GPR model fit on observed values with 95% confidence intervals

As observed in Figure 24 and Figure 25, the confidence interval of GPR is wider than that of seen in linear regression. This is due to advantage of GPR wherein the space where data points are less, the GPR gives a wider confidence interval. This is ideal in this scenario, as even though the predicted value underestimates the critical value, the confidence interval of GPR compensates for underestimation by giving a more conservative value. This is a good prediction graph for an asset manager, as the conservative values allow for quick remedial measures or closing of the bridge before critical expansion has happened.

Thus, from the above conclusions, the GPR model is selected for prediction of the distance gap in the movable bridge from the environmental parameters chosen in combination 1. The next section checks for transferability of above chose GPR model on data of another bridge.

4.4 Transferability check

The GPR model with the first combination was checked with another bridge case, called Kogerpolderbrug, which is in Kogerpolder town of Noord-holland province in the Netherlands. The movable bridge deck of Kogerpolderbrug is approximately 17.5 meters

Figure 27 and Figure 26 shows the location and the typology of the bridge respectively obtained via Google maps and street view.

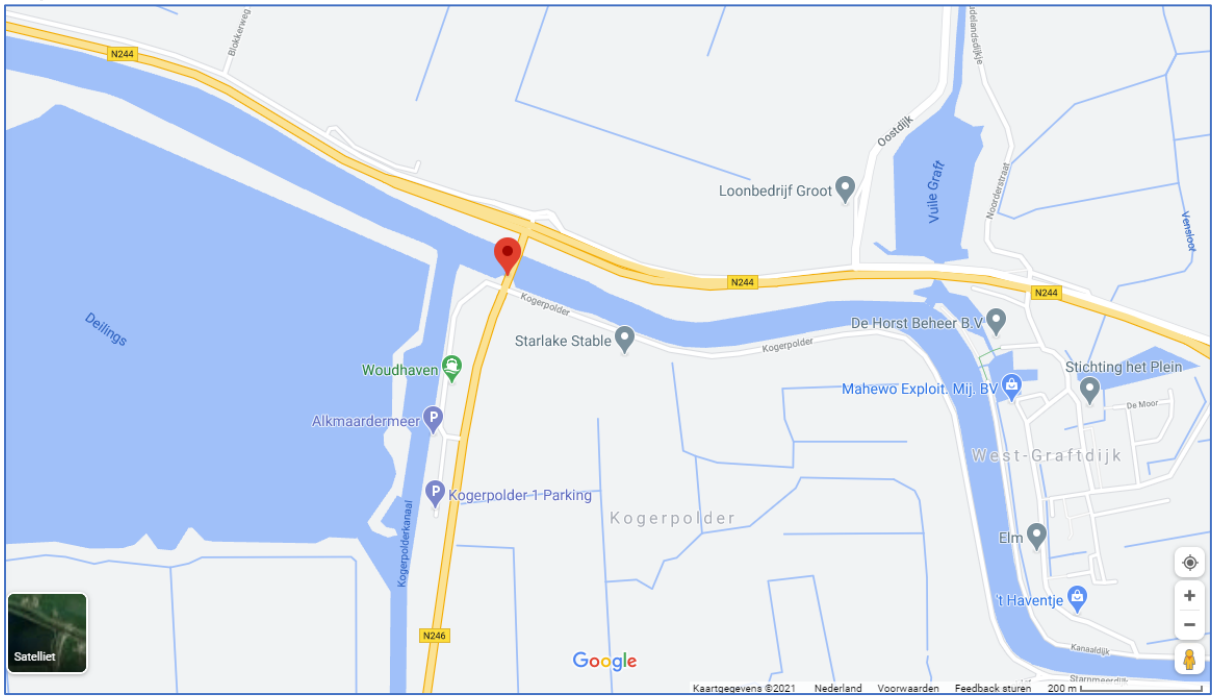


Figure 27 - Google map location of Kogerpolderbrug

Figure 28 - Typology of Kogerpolderbrug obtained from Google street view

The data collected from this bridge were the same as that of 5 variables selected in the first bridge, and the GPR model from Alkmaarsebrug was applied on 5 months of data available to us from April 2020 to September 2020. Again a standardized variable GPR analysis approach is followed, critical values are observed 3.5 deviations below mean as mean and length of movable bridge deck changed for Kogerpolder bridge case. The performance of the transferred GPR model is shown as below

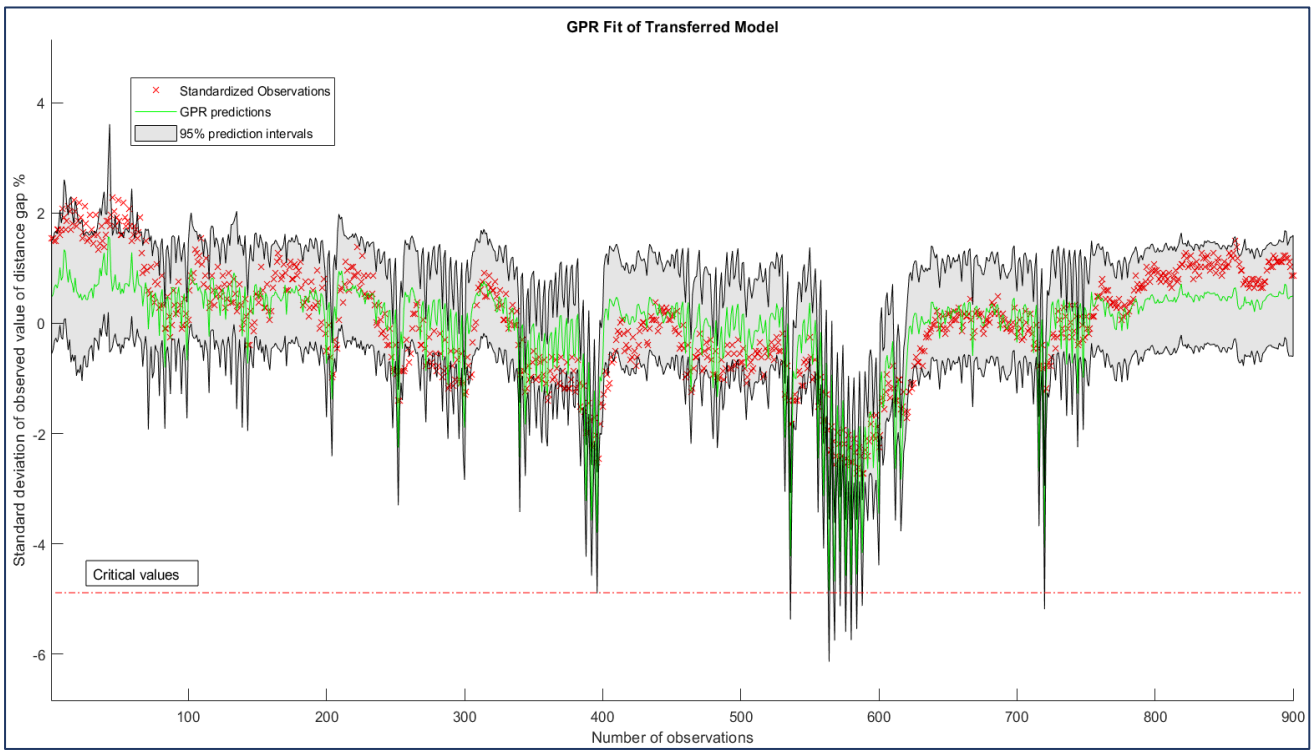


Figure 29 - Performance of the GPR model for transferability

It can be seen that the prediction model performed well for some values, while some values were not predicted well, even showing predictions out of the confidence intervals. Overall, a decent prediction model is obtained from the transferred model, but validation over a year or two is not confirmed in this result. Also, the critical values were not observed in Kogerpolderbrug even though the temperature of summer of 2020 has been considered. Thus, the correctness of prediction of the transferred model for critical values can't be ensured.

4.5 Sensor design comparison

4.5.1 Predictive power comparison

The predicted vs true response graph of GPR on combination 2 of just sensor data is shown below in Figure 26.

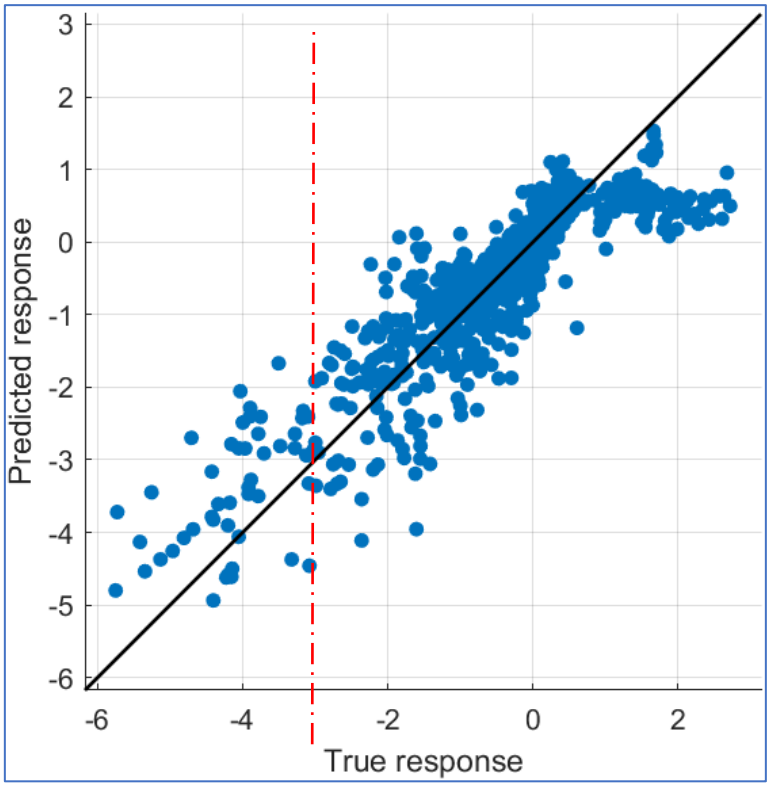


Figure 30 - Predicted vs True response graph of Sensor data GPR

The predicted vs true response graph of GPR on combination 3 of just weather API data is shown below in Figure 27.

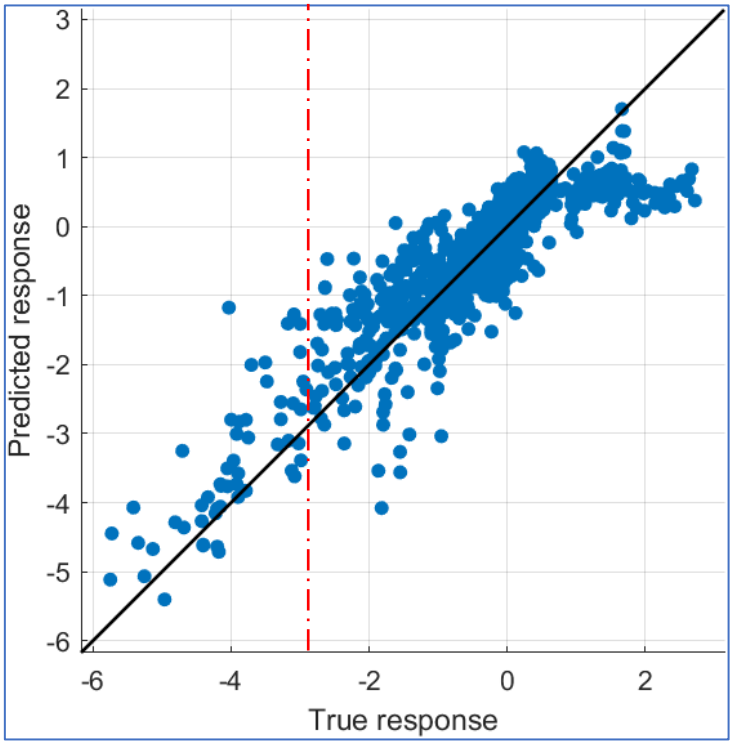


Figure 31 - Predicted vs True response graph of weather API data GPR

From Figure 26 and Figure 27 , it can be noted that performance of both the GPR model is approximately the same for prediction of critical values of gap parameter.

The comparison of RMSE and R-Squared of sensor GPR model and weather API GPR model with combination 1 GPR performance is shown in Table 8 below.

Table 8 - Comparison of performance of Sensor GPR model vs weather API GPR model vs GPR Combination of both

	Combination 1 GPR	Combination 2 (Sensor GPR)	Combination 3 (Weather API GPR)
RMSE	0.42	0.45044	0.45162
R-squared	0.82	0.8	0.8
Error percentage of critical values ((percentage underestimated)	25%	31.25%	20.8%

From Table 8, the results show that sensor data can be interchangeable with weather API data, if a combination of both is not used. This means that the combination of sensor data and internet data gives better result in terms of RMSE and R-Squared than Combination 2, but Combination 3 has better results on critical values than Combination 1. If in case there are no sensors installed yet on a bridge case, a collection of weather API data of bridge location can also help to produce a predictive model on bridge deck expansion. This is one of the huge decision in asset management, but these results should be taken with a grain of salt because of the over-fitting problem of GPR.

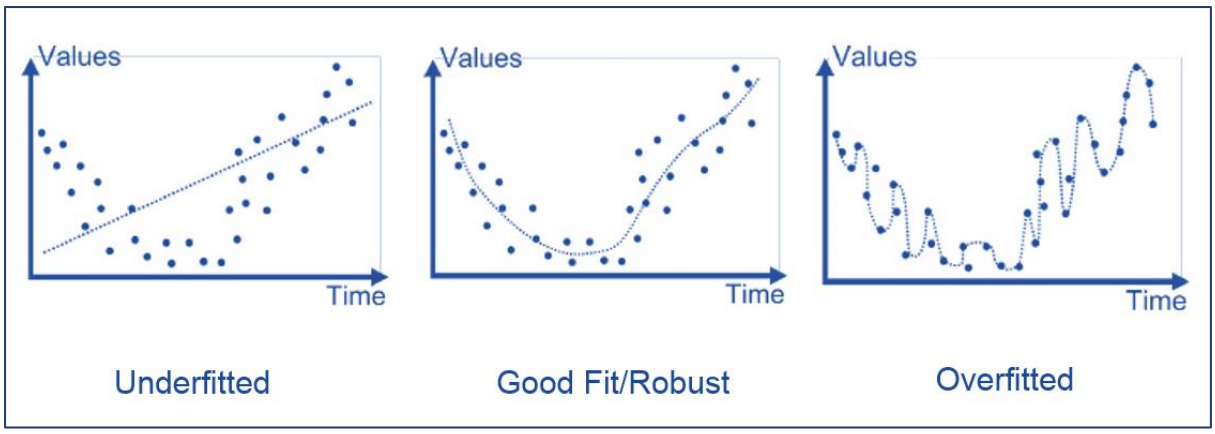


Figure 32 - Problems of over-fitting and underfitting

Figure 28 shows over-fitting and under-fitting problem of a regression model. Overfitting is a problem in predictions model which try to make sense or force the relationship between independent variables and dependent variable when there is no relationship existing. Hence, although GPR gives a better relationship, the overfitting problem could be an unseen effect in those smaller RMSE values and larger R-Squared values. Thus, the context in which GPR is used should be noted, in the above case study, the over-fitting problem is nullified due to backing by the literature of environmental parameters used and causal relationship already being established in the literature review. But in any another case, full context should be given of GPR used to check if the over-fitting problem exists.

This section concludes that weather API GPR model can also give good predictions if sensors are not installed based on predictive power.

4.5.2 Scenario Analysis comparison

Scenario Assumptions

The above 3 combinations are compared based on 4 scenarios. The assumptions of these 4 scenarios are based on type 1 and type 2 errors of prediction model achieved from each of the combinations. Type 1 error occurs due to rejection of null hypothesis also known as false positive, while type 2 error occurs due to non-rejection of a false null hypothesis also known as false negative. From Table 8, the error percentage shown is a type 1 error, where the prediction model is accepted but it underestimates the critical expansion. Based on these types of errors and probability of its occurrence, the following scenario assumptions are made to compare the cost saving of the prediction model of each combination of data sets.

From the regression analysis done in the previous section, it was observed that in a year there was approximately 20 days of critical expansion experience in an entire year. The assumptions per scenario for medium-sized bridge with a moving deck of 5 to 10 meters and experiencing 20 days of critical expansion in a year is as follows :

Scenario 1—Actual critical bridge expansion happens, and the prediction model also predicts it accurately considering confidence intervals too. This allows for planning of remedial measure on time and rearranging traffic accordingly well within permission time. This scenario avoids penalty for closing lanes like an emergency scenario without prior notice. The cost of closing lanes in such a scenario is minimal as asset manager is allowed a few days in a year to use it for maintenance purpose. This scenario is basically all the times the prediction model did correct prediction and did not showcase underestimation, and from Table 8 the probability of this scenario is 100% minus the error percentage shown.

Scenario 2–Actual critical bridge expansion happens, but the prediction model from the combination used does not predict it. Thus, the asset manager has no idea of such a critical expansion. This calls for an

immediate closure of lanes as soon as the expansion is realized and proper planning of rearranging the traffic is not achieved. This calls for huge penalty to an asset manager as such an expansion is not observed during the closure of lanes done for maintenance period. The probability of this error is taken from Table 8 where underestimation is occurring from the prediction models.

Scenario 3–No actual critical expansion of bridge happens, but the prediction model from the combination of sensors used predicts a critical expansion. Thus, rearranging of traffic is done, and remedial measures are taken place but for no apparent reason and prove to be futile as critical expansion does not happen. The closure of lanes for which permission is obtained happens unnecessarily and same goes for traffic diversion. The probability of this happening is the time when critical values where overestimated. From the observation of Figure 26 and Figure 27, when the model does not exhibit underestimation, all the values were within the confidence interval of predicted values. Thus for all the critical days an overestimation was observed and hence, for all 20 days of critical expansion, planned diversion of traffic is assumed.

Scenario 4–No critical expansion of bridge happens and the prediction model from the combination of sensor used also predicts no critical expansion. The consequences of this scenario are general case and observed for most days of the year. This is normal functioning of the bridge without calling for any action required. Although monitoring needs to be done in real-time to look for change in the scenario.

Cost assumptions

The cost comparison is done for each year of operation of the combinations. The assumption made are as follows.

1. Average Cost of installation, maintenance, replacement and operation of sensors each year (Agdas et al., 2016) - €3000
2. Average cost of cloud service for storage of data per year (aws.amazon.com, 2020)- €1000
3. Average cost of weather API data source per year (OpenWeather, 2020) - €500
4. Average penalty per day of closing lanes without prior intimation - €15000
   For sensitivity analysis, 50% range of the above amount is considered for a non-critical bridge and a highly critical bridge. Thus, penalty per day for highly critical bridge - €22500
   and penalty per day for a non-critical bridge - €7500
5. The cost of planned traffic diversion can be considered to be €1000 per day.
   Such planned maintenance activities can be conveyed by asset manager to traffic coordinator and get it approved in advance. Although it should be noted that approval of such closures may or may not be sanctioned.

The cost of combination is calculated by adding cost (1), (2), (3) , (5) and adding cost of all scenarios. The cost of each scenario is calculated by multiplying the number of days of occurrence of the scenario with the cost of penalty per day corresponding to the scenario.

Cost savings of combinations

25% of prediction is expected to be incorrect from Table 8. Thus, the penalty will be observed for 5 days out of 20 days of critical expansion. As the cost (1) already involves the cost of operation of combination used, the associated cost of scenario 4 can be assumed to be 0 for further combinations.

Table 9 - Cost savings of combination 1 (Sensor + weather API data)

	Non- critical bridge	Critical bridge	Highly Critical bridge
Cost of installation (1)	€3000	€3000	€3000
Cost of cloud service (2)	€1000	€1000	€1000
Cost of weather API data (3)	€500	€500	€500
Cost of scenario 1 (profit)	€15*7500	€15*15000	€15*22500
Cost of scenario 2 (loss)	€5*7500	€5*15000	€5*22500
Cost of scenario 3 (loss)	€20*1000	€20*1000	€20*1000
Cost of scenario 4	0	0	0
Total savings of combination 1 per year	€50.500	€125.500	€200.500

Similarly, cost benefits of other two combinations are also calculated.
For combination 2, 31.25% of critical values were incorrectly estimated. Thus from 20 days of critical expansion, 6.25 days of penalty will be observed. The traffic diversion will be planned for all 20 days, out of which 13.75 days will be correctly estimated.

Table 10 - Cost savings of combination 2 (Sensor data)

	Non- critical bridge	Critical bridge	Highly Critical bridge
Cost of installation (1)	€3000	€3000	€3000
Cost of cloud service (2)	€1000	€1000	€1000
Cost of weather API data (3)	€0	€0	€0
Cost of scenario 1 (profit)	€13.75*7500	€13.75*15000	€13.75*22500
Cost of scenario 2 (loss)	€6.25*7500	€6.25*15000	€6.25*22500
Cost of scenario 3 (loss)	€20*1000	€20*1000	€20*1000
Cost of scenario 4	0	0	0
Total savings of combination 2 per year	€32.250	€88.500	€144.750

For combination 3, approximately 20% of critical values were incorrectly estimated. Thus from 20 days of critical expansion, 4 days of penalty will be observed. The traffic diversion will be planned for all 20 days out of which 16 days will be correctly estimated.

Table 11 - Cost benefits of combination 3 (Weather API data)

	Non- critical bridge	Critical bridge	Highly Critical bridge
Cost of installation (1)	€0	€0	€0
Cost of cloud service (2)	€1000	€1000	€1000
Cost of weather API data (3)	€500	€500	€500
Cost of scenario 1 (profit)	€16*7500	€16*15000	€16*22500
Cost of scenario 2 (loss)	€4*7500	€4*15000	€4*22500
Cost of scenario 3 (loss)	€20*1000	€20*1000	€20*1000
Cost of scenario 4	0	0	0
Total savings of combination 3 per year	€68,500	€158.500	€248,500

Thus, on comparing Table 9,10 and 11, it can be observed that just the weather API data can prove to have better cost benefits that the combination of sensor data and weather API. The weather API data set also has better cost-benefits over just the sensor data set.
This conclusion is based on the assumption that error percentage of the prediction models does not change in coming years and same performance of the prediction is observed throughout the life-cycle.
If at any point, the performance of the prediction model changes, it will be advised to add sensors over the base data set of weather API as the cost savings of combination of data as observed in Combination 1 is more than sensor data observed in Combination 2. Such decision can be made by evaluating the prediction models in real-time throughout the life cycle of the bridge.
It should be noted that in above comparison the essence of use of such prediction model is mainly in savings of penalty, which can arise due to uncertain unavailability of the bridge. The cost of operation and installation of all the combinations is little compared to the savings any of the prediction model with its accurate estimation can achieve. Thus, the main inference from scenario analysis is the use of prediction models or a move towards predictive maintenance is highly advised to an asset manager. The prior estimate of imminent critical expansion for a bridge with little underestimating error is still an incentive for asset manager and thus either via weather API data or installation of sensors, prediction models prove to be useful for the bridge deck expansion problem.

5

Chapter 5 Conclusions

This section deals with conclusions by answering the research questions stated in the introduction chapter. The results obtained from the case study are discussed and concluded to answer each research question. From the inferences of each sub-research question, the central research question is answered with a generalized suggested flowchart of decisions for sensor design and predictive model of bridge deck expansion.

1. Which environmental parameters affect the expansion of the bridge?

The literature review shows that temperature, wind speed, humidity, solar radiation, rains, snow, etc. almost every possible environmental factor could affect bridge decks differently depending on the material property of the bridge deck. From Table 4, it can be concluded that temperature and wind factors showed high importance in PCA and EFA analysis, while humidity showed unique variance for consideration of predicting bridge expansion which was observed in Figure 19 and Figure 20.
The direct individual monotonic correlation of environmental factors with bridge deck expansion was seen from temperature parameters having Spearman's correlation approximately higher than 0.85. The individual wind parameters showed rather a weak direct correlation with bridge expansion as the Spearman's correlation was approximately 0.25. Although in further regression analysis, the correlation is strong with a combination of temperature and wind parameters as R-Squared or correlation coefficient values ranged from 0.642 to 0.8. It can also be further established by analyzing linear regression and GPR results that the combination of temperature and wind parameters are more non-linearly correlated to bridge deck expansion than a linear correlation.
Finally, it can also be concluded that to completely answer this question for any other bridge case, various environmental factors should be measured especially temperature and wind.Then, applying feature extraction process to test which environmental factor affects the most and is important amongst all the measured environmental parameters should be established. Then, continue with regression analysis to finally evaluate the extent of relationship.

2. Which regression model shows a better fit for prediction on bridge deck expansion with sensor and weather API data as a dependent variable?

From Table 6 and Table 7, RMSE value of GPR model is 0.42 and that of the linear regression model is 0.6. Further, the R-Squared value of GPR model is 0.82 and that of the linear regression model is 0.6. Also, through visual performance of models from Figure 22,23,24 and 25, it was observed that GPR model performed better than linear regression for critical value prediction. Also, GPR had wider confidence giving conservation and overestimated values of prediction for an asset manager to be prepared for remedial action. Thus, GPR model shows a better fit than simple linear regression, although the problem of over-fitting should be investigated closely while selecting the GPR model. Any unnecessary and redundant variable which may show no relationship with the dependent variable should be excluded. Thorough

physics-based research should be applied before selecting variables for GPR and results should be shown in the context it is used. In some cases, it may happen that other regression model can show better results than GPR, but the versatility of the GPR model could prove to be useful to get the right fit for this particular problem. It was seen that in summer the relationship was linear but for other values, no particular linear relationship was observed and this can be understood from the linear regression results. Thus, to get a one-fit solution where the shape of fit can not be determined, GPR with its adaptability option by optimizing hyper-parameters could prove to solve the regression fit of bridge deck expansion from extracted environmental factors.

3. Are the regression models transferable from one bridge to other bridges?

From visual performance presentation shown in Figure 29, the transferred model did a decent job of predicting expansion values of another bridge case. It is not advisable to transfer prediction model of one bridge and use it for another, as the literature study has also shown that different bridges behaving differently to their environment. This proved to be true, as no critical values were observed in the new bridge case study referred to in the research project and hence no conclusion can be made on its performance on critical values. The results obtained also do not have good validation of at least a complete year of data. Thus, models if transferable, should also be done with an adequate amount of validation process before directly transferring from one bridge to another. Due to the literature backing that each bridge case behaves differently to its surrounding and not all bridges show critical values of bridge deck expansion, a general conclusion can be made that data collection should be done from the bridge location itself and transferring prediction models should be avoided.
As the scenario analysis showed that cost savings of using sensors and weather data collection from the bridge location can prove to have positive returns on investment, transferring model can also be dismissed on monetary grounds as the accuracy of transfer model is highly questionable.

4. How sensor systems can be designed economically efficient combined with data from weather API to achieve insightful information?

From table 8, it can be seen that value of RMSE and R-Squared is approximately equal for sensor data GPR model and weather API data GPR model. The predictive power comparison from Table 8 also shows that the combination 3 of weather data API GPR gives the lowest error percentage of underestimation of approximately 20% out of all the critical expansion values. It can be established that prediction of critical values is possible with just weather API data set. Thus, it can be concluded that sensors can become redundant from this research case study. Although it should be noted that such decision of moving away from sensors should be done after enough validation is achieved. If validation and predictive power are not up to the mark established by the asset management team of the project, then the combination of sensor data and weather API data is suggested until stipulated predictive power and validation is achieved.

Further, Table 9,10 and 11 solidifies the predictive power conclusion that weather API data set can give higher cost savings on penalties that are expected from unplanned closure of bridge due to critical bridge deck expansion. The cost savings can be as high as approximately €250.000 for a highly critical bridge. If the predictive performance of models changes, the second best prediction model is that of a combination of data sets as the cost savings of Combination 1 is higher than Combination 2, which is just the sensor data prediction model. The scenario analysis has also shown how accuracy of predictive models can save costs in the complete life-cycle of a bridge. Thus, an aim towards collecting the appropriate data should be done to get higher accuracy from the prediction models.

From the conclusion of above sub-research question, central research question can be answered as follows -

"How can data from sensor systems and weather APIs help to develop predictive models of critical longitudinal expansion of bridge deck in movable bridges?"

The approach learned from the conclusion of results is to first measure the gap between the decks. Then gather the data from weather API for one to two years, a predictive model can already be achieved with a good amount of prediction power. If more prediction power is required, then the addition of a sensor system is advisable after knowing important environmental parameters from PCA and EFA analysis of already available weather API data. For example, if temperature shows the highest importance in feature extraction analysis, then the temperature sensor of the deck and water will be advised as both sensors will capture unique variances related to temperature parameter. Thus, the decision should be made on the latent and principal factors obtained from EFA and PCA. After the best combination of a variable input is selected, the GPR machine learning tool should get the best fit for critical expansion. In some other cases, maybe different techniques of feature extraction and regression could give better results but above general steps applying to any medium-sized movable bridge.

A scenario analysis conducted showed that sensors cost little as compared to the benefits the accuracy of the prediction model can achieve. Although, if the critical expansion is not observed, the use of sensors are not justified based on the central result of this research project which is that weather data itself from weather API can be used to form prediction models with similar predictive power and higher cost savings on penalties.

From the learnings and inferences on the results obtained from the research project conducted, a generalized flowchart of decision-making on design of sensor systems and to obtain prediction model for expansion of bridges can be done as shown in the figure below –

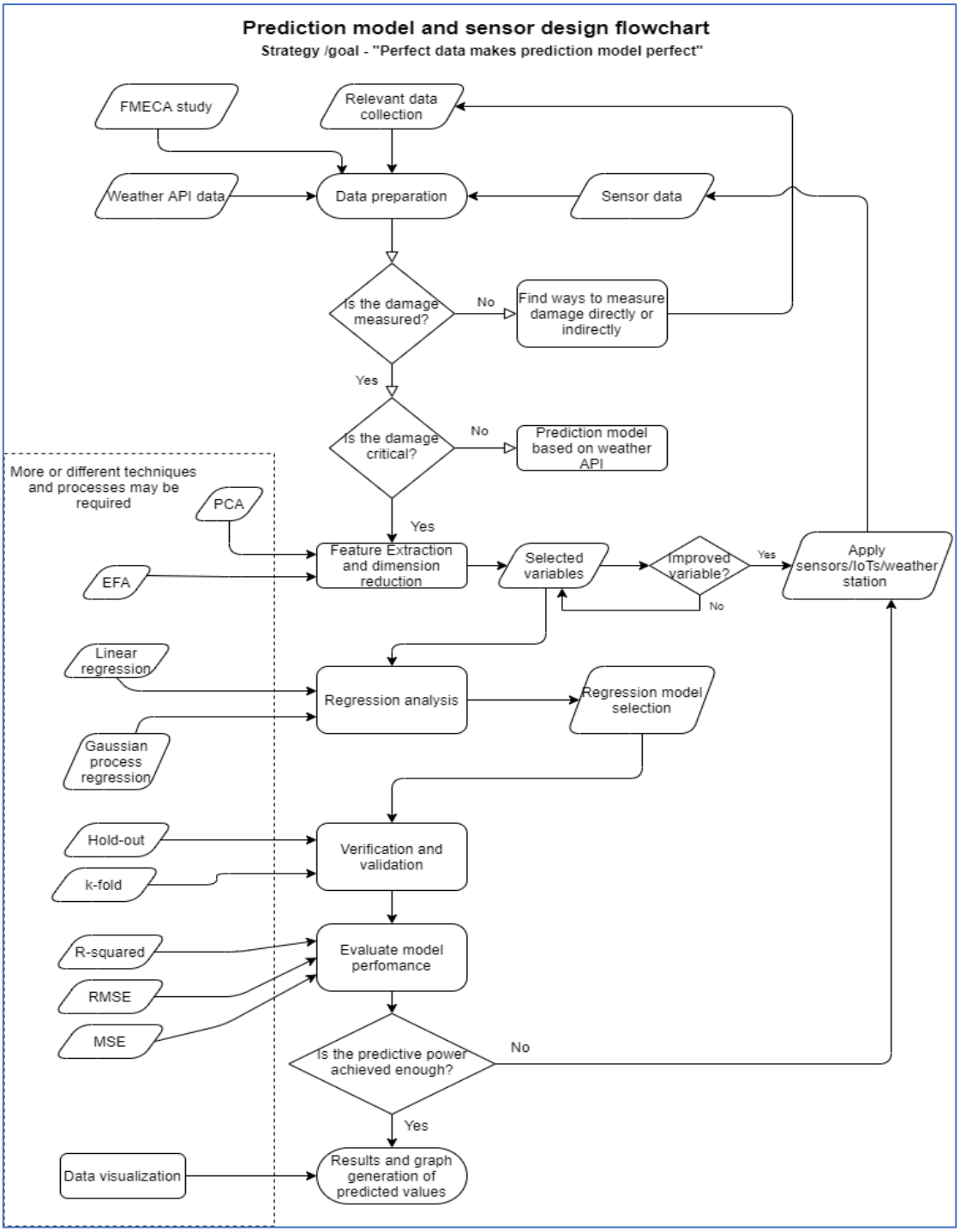


Figure 33 - Flowchart suggestion for general prediction model and sensor design for the expansion of bridge deck based on inferences of the research project

6

Chapter 6 Discussion and Recommendations

This chapter delves into discussing some interpretations and inferences of the results observed above before conclusions on the research questions.

1. From the Figure 19 and 20, the two latent factors and principal components evident from the PCA and EFA analysis were temperature and wind speed. This was also known from the literature review, hence, sensors or weather API data which can capture unique variances of these factors will be beneficial to predict the expansion of bridge deck in possibly any other bridge case too. This can be based on the results of this research project as well as literature backing. It should be noted that the linear relationship directly between wind speed and expansion is weak, so a non-linear relationship may exist between expansion and combination of wind and temperature effect as the whole prediction model had p-value lesser than 0.05 making the model statistically relevant. Some ways to capture unique variances of these factors would be installing temperature in the water below the bridge, installing accurate wind speed station, capturing amount of direct sunlight and solar radiation on bridge surface along with its intensity and power values. It should be noted that wind speed varies as per height from the ground it is captured, the ideal height should be taken into consideration for placement of sensors. Wind direction, wind maps, wind forecasts are some data already available in weather APIs for any location which could prove useful to find data with unique variances.
2. From Figure 22 it was observed that simple linear regression did not exhibit a good prediction model, but it should be noted that there are various other methods of supervised regression which could show better result. In the same figure it can be observed that below 0 standard deviation the predictions were skewed, thus splitting of equation could have also resulted in better prediction, but it was not done as the aim of research was to find one fit prediction model. Some other types of regression are logistic regression, polynomial regression, stepwise regression, ridge regression, lasso regression, poison regression. It was not practically possible to try each method of these regressions in the time frame of this research project, but depending on different variables, the above regression techniques could give better results. Fitting the right regression technique is a matter of big debate in the data science community, the only way to do it to get the feel of the data through feature extraction and dimension reduction and then apply the trial-and-error method of fitting various regression techniques. Machine learning softwares in R, MATLAB and python allow simultaneous evaluation of various regression models which should be exploited to the fullest. It should also be noted that instead of focusing on larger R-Squared number and smaller error value, visual representation of how the fit performs should also be evaluated, for which various data visualization techniques should be applied. Visual presentation can help to detect over-fitting and under-fitting problems of regression models.
3. One of the main outcomes of the research project is to move from the sensor to weather API data. This is an important decision in asset management as monitoring can potentially become logistic and cost-efficient. This outcome should be tested and validated on various bridge cases especially where critical

damage may not be observed. There may be no need of a sensor measuring a data point if the damage linked to that data point is harmless or shows less criticality in risk analysis. The resources used for such sensor system could be transferred toward measuring potentially critical damage. Although the cost of installation and operation of sensor system is low, scarce resources such as man-hours and energy spent on it could be transferred to focus on other critical damage expected in a bridge. To be wary of such allocation of resources, a risk analysis of damage expected through FMECA study can always help.

4. It was seen in transferability check in Figure 29 that the second bridge case did not show critical expansion, even though values of summer 2020 were included in the data set. This proves the hypothesis in the literature review which states that each bridge behaves differently to its environment due to various factors bringing in the ambiguity. The new bridge case does not show as much critical expansion as the earlier case-study. The current design codes also cannot acknowledge this issue in designing of bridges, as the same temperature coefficient (parametric design) is used for a design for the same class of bridges. A suggestion can be made to here to bridge designers of making a move towards the efficient design of the bridges by factoring in weather parameters of the location during the design process. Also, to make future robust design codes, the ambiguity factors should be studied by installing efficient SHM systems with or without a combination of data from sensor system and weather API parameters on the current bridges and create learnings from each bridge case. An emerging pattern to resolve the current SHM problems may emerge or may not emerge, but efforts towards it will make a vast difference as speculations will lead to specifics.
5. The results also show that sensors are not always the solution for efficient SHM systems. If sensors are required, more thought could be given to their design with prior knowledge of the behaviour of critical damage within the environmental factors which are acting on it. Thus, while recent research studies shows the usage of sensors and its accuracy with various types, very little research was found in its efficient design for a complete bridge case. This research project shows that data might be already in abundant with asset owner through various means like weather API and combining some machine learning and data science techniques with it, we can already establish a knowledge base for the efficient design of sensor systems for any bridge. The early intuition of installing sensor because it is cheap compared to the cost of maintenance of the bridge leads to abundant data with little to no insights and information obtained from it. The logistic problem of maintaining sensors for the whole life-cycle of the bridge is also realized later. This research would like to add to the conscience of asset owner that more data does not necessarily mean more insights and information, but rather perfect data leads to insightful information, including prediction of the damage. The goal and strategy of data collection should be focused towards capturing the appropriate parameter with required quality rather than quantity. The very first step towards such an approach is feature extraction, and it is advisable to follow that step for every critical fault or damage expected in a bridge.
6. Another benefit that can not be quantified in monetary units is the readiness of the maintenance team. The prior information of critical expansion throughout the year is useful to plan remedial action of cooling the bridge well in advance. There are no last-time or reactive actions towards an uncertain event. The preparedness of all the stakeholders involved in maintenance of bridge is also an incentive for asset manager to make the most of the data available and crunch the data for prediction models. The entire supply chain of maintenance activities including sub contractors feel the ease of regular operations due to insights of prediction models which includes prior intimation of imminent expansion of bridge deck.
7. Another benefit of prediction models is potential avoidance of a breakdown scenario. If critical expansion extends beyond breakage point where a complete failure of bridge deck is observed, the potential cost of breakdown will be a vast amount. Those scenarios of complete failures can also be avoided with the use of prediction models.

7

Chapter 7 Drawbacks and further research alternatives

Some drawbacks in the research project conducted are as below -

1. The value measured by the sensor and taken from weather API was taken at face value and considered that there are no errors in their measurement except for distance parameter where outliers were removed. To counteract this drawback, sensors were checked for correctness of measure, even so, minute errors cannot be guaranteed. The temperature parameters were also checked amongst themselves for any unexplained values and no such outliers were found.
2. There is missing data from June and July 2020 for Alkmaarsebrug. These are the months were critical expansion was seen in the year 2020. Unavailability of these values could have led to the disproportionate prediction of critical values on which central research results are based. A cross-validated approach of prediction model counteracts this missing values drawback to some extent.
3. This study is highly case-based. Some inferences from this research may not apply for other bridge cases, and the conclusion section clarifies application of the inferences. Although a solid validation process can negate this drawback. If the prediction models of this research project is validated for 2-3 years on a new case study of another bridge, then inferences of this research project can become generalized. The result of using weather API data for prediction models can be used for another bridge too, with adequate validation process.
4. There exist many other environmental factors which could lead to change in expansion of bridge as explained in the literature review but not all of them were explored to due to limited time-frame of the research project.
5. The data set of 2-year could not be said to be enough to make concrete conclusions. The current model is also not future-proof. If the models are not kept stagnant and updated with current monitored data, it can become robust in the coming years.
6. These results are strictly in asset management and machine learning perspective. It cannot be applied for structural engineering design perspective directly. For exact damage detection structurally in pure sense, temperature above the deck and below deck both are required to study differential temperature in vertical cross-section of the bridge. For asset management purposes, it was advisable to not use one of the deck sensor data as it was redundant, but for structural design research, it could be prove to be crucial data. The only structural design inference taken from this study is that design codes

cannot consider the behaviour of bridges in actual environmental factors it may face at that particular location, thus causing few bridges to experience expansion problem while others do not face the same problem.

7. Hold-out validation is still a widely used practice in machine learning, but because of lack of data, this research project lacks validation power due to the omission of hold-out validation. Disadvantages of hold-out validation are counteract via cross-validation in the research project.

Taking inferences and conclusion from this research project, further research recommendations are as followed -

1. A rather exploratory research can be done to completely advise on the move towards weather data API data over sensor data. For which faults and damages this general conclusion can be extended has to be explored.
2. As the life-cycle of sensors and bridge do not match, prediction models which achieved adequate prediction power can make the sensor systems redundant once they have reached sufficient confident levels of predictive power. Thus, instead of managing sensors for the complete life-cycle of the bridge, which is a tremendous challenge, maybe after 2 to 3 life-cycle of sensor systems, they can be removed and data can solely be obtained from weather API. At what point in the complete life-cycle of the bridge can this step be taken to optimize both sensor system life-cycle and prediction models can be researched further. Formation of predictive power thresholds can also be a step in the research strategy in this approach.
3. For the model to become climate-change and future proof, a time-series GPR can be used over a normal GPR as used in this research as the temperature which will be seen in coming years will be potentially higher and could be unprecedented than the training set used until now.
4. GPR is a complex machine learning tool with many optimization options. An exploratory study of how to optimize hyper-parameters and Kernel function of GPR for powerful predictions is also suggested.
5. As seen in the transferability check, the second bridge did not exhibit critical bridge deck expansion. Another exploratory research can be done on ambiguous factors that lead to such different bridge deck expansion behavior of bridges, even though they are of the same class.

8

Chapter 8 Bibliography

Agdas, D., Rice, J. A., Martinez, J. R., & Lasa, I. R. (2016). Comparison of visual inspection and structural-health monitoring as bridge condition assessment methods. *Journal of Performance of Constructed Facilities, 30*(3), 04015049.

aws.amazon.com. (2020). Amazon web service pricing. Retrieved from https://aws.amazon.com/pricing/services/.

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

Beeson, S., & Andrews, J. D. (2003). Importance measures for noncoherent-system analysis. *IEEE Transactions on Reliability, 52*(3), 301-310.

Carreira-Perpinán, M. A. (1997). A review of dimension reduction techniques. *Department of Computer Science. University of Sheffield. Tech. Rep. CS-96-09, 9*, 1-69.

Catbas, F. N., Gul, M., Gokce, H. B., Zaurin, R., Frangopol, D. M., & Grimmelsman, K. A. (2014). Critical issues, condition assessment and monitoring of heavy movable structures: emphasis on movable bridges. *Structure and Infrastructure Engineering, 10*(2), 261-276. Retrieved from https://doi.org/10.1080/15732479.2012.744060. doi:10.1080/15732479.2012.744060

Corner, P. D. (2002). An integrative model for teaching quantitative research design. *Journal of Management Education, 26*(6), 671-692.

Cross, E., Koo, K., Brownjohn, J., & Worden, K. (2013). Long-term monitoring and data analysis of the Tamar Bridge. *Mechanical systems and signal processing, 35*(1-2), 16-34.

Daily, J., & Peterson, J. (2017). Predictive maintenance: How big data analysis can improve maintenance. In *Supply Chain Integration Challenges in Commercial Aerospace* (pp. 267-278): Springer.

Damen, T. Amsterdam begonnen met koelen bruggen. *Het Parool*. Retrieved from https://www.parool.nl/nieuws/amsterdam-begonnen-met-koelen-bruggen~b49f3e03/.

DIMITROVA, A. (2020). Amsterdam cools down its bridges on hot summer days. Retrieved from https://www.themayor.eu/en/amsterdam-cools-down-its-bridges-on-hot-summer-days

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification*: John Wiley & Sons.

Farrar, C. R., & Worden, K. (2007). An introduction to structural health monitoring. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 365*(1851), 303-315.

Gokce, H. B., Gul, M., & Catbas, F. N. (2012). Implementation of structural health monitoring for movable bridges. *Transportation research record, 2313*(1), 124-133.

Gopalakrishnan, S., Ruzzene, M., & Hanagud, S. (2011). *Computational techniques for structural health monitoring*: Springer Science & Business Media.

Gouriveau, R., Medjaher, K., & Zerhouni, N. (2016). *From prognostics and health systems management to predictive maintenance 1: Monitoring and prognostics*: John Wiley & Sons.

Heatwave hits bridges, shipping in Amsterdam's waterways. (2018). *DutchNew.nl*. Retrieved from https://www.dutchnews.nl/news/2018/07/heatwave-hits-bridges-shipping-in-amsterdams-waterways/.

Hummen, R., Henze, M., Catrein, D., & Wehrle, K. (2012). *A cloud design for user-controlled storage and processing of sensor data.* Paper presented at the 4th IEEE International Conference on Cloud Computing Technology and Science Proceedings.

Inaudi, D. (2009). Structural health monitoring of bridges: general issues and applications. In *Structural health monitoring of civil infrastructure systems* (pp. 339-370): Elsevier.

Inkoom, S., & Sobanjo, J. (2018). Availability function as bridge element's importance weight in computing overall bridge health index. *Structure and Infrastructure Engineering, 14*(12), 1598-1610.

Jackson, J., & Edward, A. (1991). User's guide to principal components. John Willey Sons. *Inc., New York*, 40.

Jiang, X., & Rens, K. L. (2010). Bridge health index for the city and county of Denver, Colorado. II: Denver bridge health index. *Journal of Performance of Constructed Facilities, 24*(6), 588-596.

Karbhari, V. M., & Lee, L. S.-W. (2009). Vibration-based damage detection techniques for structural health monitoring of civil infrastructure systems. In *Structural health monitoring of civil infrastructure systems* (pp. 177-212): Elsevier.

Khan, S. M., Atamturktur, S., Chowdhury, M., & Rahman, M. (2016). Integration of structural health monitoring and intelligent transportation systems for bridge condition assessment: current status and future direction. *IEEE Transactions on Intelligent Transportation Systems, 17*(8), 2107-2122.

Kim, H.-C., & Lee, J. (2007). Clustering based on gaussian processes. *Neural computation, 19*(11), 3088-3107.

Kocakulak, M., & Butun, I. (2017). *An overview of Wireless Sensor Networks towards internet of things.* Paper presented at the 2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC).

Min, Z.-h., Sun, L.-m., & Dan, D.-h. (2009). Effect analysis of environmental factors on structural modal parameters of a cable-stayed bridge. *Journal of Vibration and Shock, 28*(10), 99-105.

Mobley, R. K. (2002). *An introduction to predictive maintenance*: Elsevier.

Mondoro, A., Frangopol, D. M., & Liu, L. (2018). Bridge adaptation and management under climate change uncertainties: A review. *Natural Hazards Review, 19*(1), 04017023.

Nabizadehdarabi, A. (2015). Reliability of bridge superstructures in Wisconsin.

Nagayama, T., Sim, S.-H., Miyamori, Y., & Spencer Jr, B. (2007). Issues in structural health monitoring employing smart sensors. *Smart Structures and Systems, 3*(3), 299-320.

Nieuwe brug in Middenmeer: wegvervoer blij. (1992). *Niuewsblad Tranport.* Retrieved from https://www.nieuwsbladtransport.nl/archief/1992/02/20/nieuwe-brug-in-middenmeer-wegvervoer-blij/?gdpr=accept

Okasha, N. M., & Frangopol, D. M. (2009). Lifetime-oriented multi-objective optimization of structural maintenance considering system reliability, redundancy and life-cycle cost using GA. *Structural Safety, 31*(6), 460-474.

OpenWeather. (2020). OpenWeather Pricing. Retrieved from https://openweathermap.org/price.

Raj, J. T. (2019). A beginner's guide to dimensionality reduction in Machine Learning. Retrieved from https://towardsdatascience.com/dimensionality-reduction-for-machine-learning-80a46c2ebb7e

Rasmusen, C. E., & Williams, C. (2006). Gaussian processes for machine learning. In: MIT press Cambridge, MA:.

Roeder, C. W. (2003). Proposed design method for thermal bridge movements. *Journal of Bridge Engineering, 8*(1), 12-19.

Ryan, P. J., & Watson, R. B. (2017). Research challenges for the Internet of Things: what role can OR play? *Systems, 5*(1), 24.

Sikorsky, C., & Karbhari, V. (2003). *An Assessment of structural health monitoring capabilities to load rate bridges.* Paper presented at the Proceedings of the First International Conference on Structural Health Monitoring and Intelligent Infrastructure.

Sonawane, M. A. D., Vichare, M. P. P., Patil, M. S. S., & Chavande, P. (2018). Bridge Monitoring System Using IOT. *Journal of Advances in Electrical Devices, 3*(2).

Thompson, P. D., Sobanjo, J. O., & Kerr, R. (2003). Florida DOT project-level bridge management models. *Journal of Bridge Engineering, 8*(6), 345-352.

Vardanega, P. J., Webb, G. T., Fidler, P. R., & Middleton, C. R. (2016). *Assessing the potential value of bridge monitoring systems.* Paper presented at the Proceedings of the Institution of Civil Engineers-Bridge Engineering.

Wenzel, H. (2008). *Health monitoring of bridges*: John Wiley & Sons.

Witten, I. H., & Frank, E. (2002). Data mining: practical machine learning tools and techniques with Java implementations. *Acm Sigmod Record, 31*(1), 76-77.

Xu, Y. L., & Xia, Y. (2011). *Structural health monitoring of long-span suspension bridges*: CRC Press.

Zalt, A., Meganathan, V., Yehia, S., Abudayyeh, O., & Abdel-Qader, I. (2007). *Evaluating sensors for bridge health monitoring.* Paper presented at the 2007 IEEE International Conference on Electro/Information Technology.