POLITECNICO
MILANO 1863

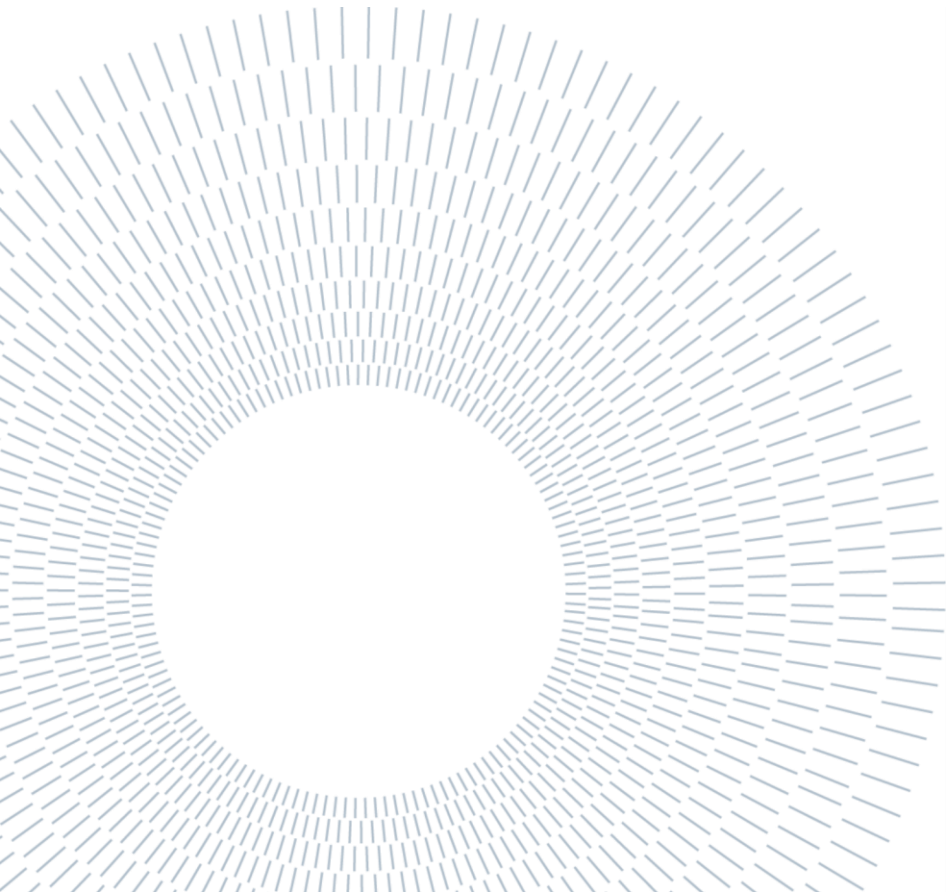
SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Mitigating the influence of environmental variability by using machine learning for Structural Health Monitoring applications

TESI DI LAUREA MAGISTRALE IN
MECHANICAL ENGINEERING-INGEGNERIA MECCANICA

Author: **Aniket Pande**

Student ID: 10792153
Advisor: Prof. Simone Cinquemani
Co-advisor: Luca Radicioni, Gabriele Cazzulani, Lorenzo Bernardini
Academic Year: 2022-2023



Abstract

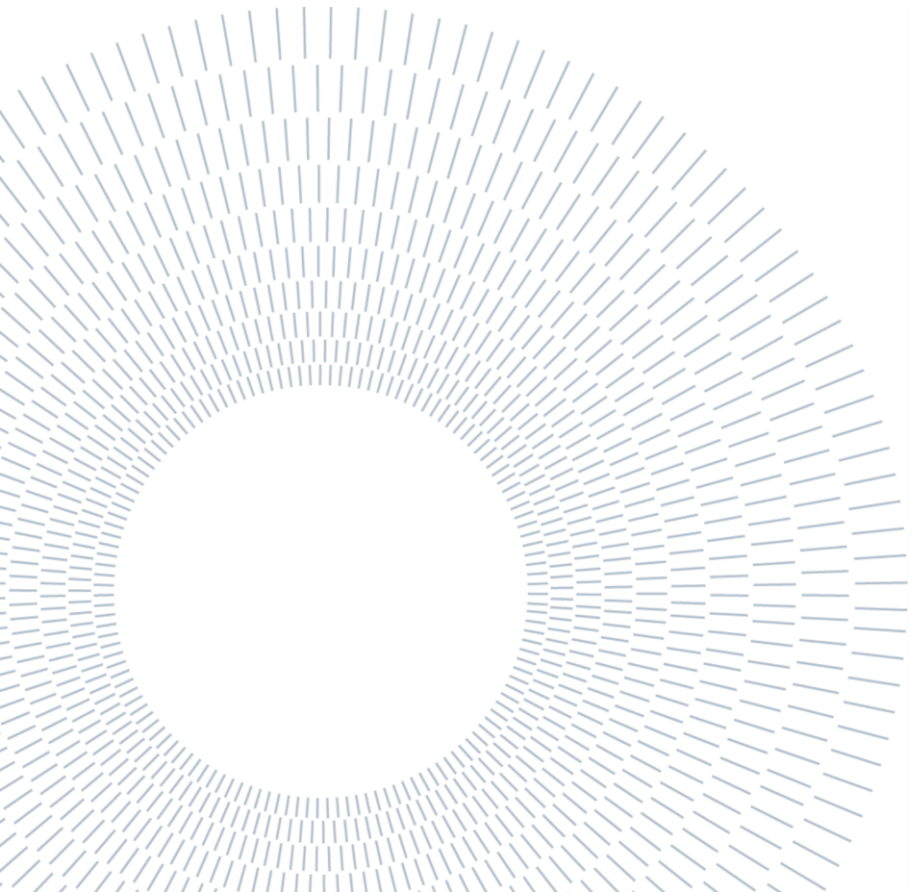
Structural Health Monitoring (SHM) is a critical process that involves various sensing and data analytics techniques to evaluate the current state of a structure's health and detect any damage at the earliest possible stage. However, structures are constantly subjected to varying environmental and operational conditions that induce changes in their dynamic response, posing significant challenges for accurate SHM damage detection. One possible solution by developing hybrid machine learning techniques to isolate indicators of structural damage from environmental impacts in SHM data is proposed, and demonstrated on a case study of a Livenza railway bridge monitoring system. A local outlier removing function is used to mitigate the operational variability and Dynamic regression with time lag of 24 hours prove optimal in predicting bridge response from temporal environmental patterns thereby removing environmental variability. Further filtering residuals with principal component analysis removes remaining unmeasured variability. Simulated damage scenarios validate the integrated methodology's ability to extract damage-reflecting residuals by separating environmental and mechanical variability.

Keywords: Structural health monitoring (SHM), machine learning, environmental variability, damage detection, Long Short-Term Memory (LSTM) networks, Principal Component Analysis (PCA), Dynamic Regression.

Abstract in lingua italiana

Il monitoraggio dello stato di salute delle strutture (Structural Health Monitoring, SHM) è un processo critico che coinvolge varie tecniche di rilevamento e analisi dei dati per valutare lo stato di salute attuale di una struttura e rilevare eventuali danni nella fase più precoce possibile. Tuttavia, le strutture sono costantemente soggette a condizioni ambientali e operative variabili che inducono cambiamenti nella loro risposta dinamica, ponendo sfide significative per un accurato rilevamento dei danni SHM. Viene proposta una possibile soluzione sviluppando tecniche ibride di apprendimento automatico per isolare gli indicatori di danno strutturale dagli impatti ambientali nei dati SHM, dimostrata su un caso di studio del sistema di monitoraggio di un ponte ferroviario Livenza. Per attenuare la variabilità operativa viene utilizzata una funzione di rimozione degli outlier locali e la regressione dinamica con un ritardo temporale di 24 ore si dimostra ottimale nel prevedere la risposta del ponte dai modelli ambientali temporali, eliminando così la variabilità ambientale. Un ulteriore filtraggio dei residui con l'analisi delle componenti principali rimuove la variabilità non misurata rimanente. Gli scenari di danno simulati convalidano la capacità della metodologia integrata di estrarre i residui che riflettono il danno, separando la variabilità ambientale da quella meccanica.

Parole chiave: Structural health monitoring (SHM), machine learning, environmental variability, damage detection, Long Short-Term Memory (LSTM) networks, Principal Component Analysis (PCA), Dynamic Regression.



Contents

Introduction and Research Objectives

Structural Health Monitoring (SHM) utilizes various sensing and data analytics techniques to evaluate the current state of a structure's health and detect any damage at the earliest possible stage. However, structures are constantly subjected to varying environmental and operational conditions that induce changes in their dynamic response. For example, temperature fluctuations cause expansions and contractions that alter vibration characteristics. Similarly, traffic loads on a bridge lead to different strain patterns compared to periods of low use.

This variability poses significant challenges for accurate SHM damage detection. Environmental and operational changes can mask underlying structural damage or produce false indications of damage where none exists. Without properly accounting for these effects, SHM systems are likely to suffer from frequent false alarms or missed damage events.[1]

Several types of environmental factors impact structural response:

- Temperature - Thermal expansion/contraction alters stiffness and induces strains. Daily and seasonal variations are common.
- Wind - Wind loads directly add dynamic forces and can amplify vibrations. Speed and direction are often variable.
- Humidity - Moisture absorption in some materials affects stiffness and mass. Rainfall causes sharp humidity spikes.
- Solar radiation - Thermal gradients and deck deformation can develop under solar loading. Cloud cover leads to rapid fluctuations.

Operational variability arises from changes in loading and usage patterns:

- Traffic - Vehicular traffic adds moving dynamic loads. Congestion causes stress peaks during rush hours.
- Special events – Construction work may alter bridge response.

Research Gaps and Objectives

While numerous SHM algorithms have been developed, the critical problem of environmental and operational variability has received less focused attention. Additionally, most numerical simulations lack modelling of real-world variability, which limits their utility in validating SHM techniques.

This thesis aims to address these research gaps by mitigating the influence of environmental and operational variability on SHM damage detection. The objectives are:

1. Review existing methods, such as data normalization, that help reduce variability effects.
2. Implement and evaluate suitable techniques on a case study of the Livenza Railway Bridge using real monitoring data.
3. Draw general conclusions and recommendations for enhancing SHM robustness to variability.

Although performed for a specific bridge, the research will focus on developing techniques broadly applicable to SHM systems in general.

This thesis has been divided into four chapters. Chapter1 will begin with an introduction of Structural Health Monitoring and Machine Learning methods. Chapter2 will review the existing literature on machine learning models used to detect the structural damage. The main contribution is presented in Chapter3, which focuses on the case study of the Livenza bridge. This chapter describes the bridge and its structural health monitoring system. The performance of proposed machine learning models on data from the Livenza bridge is then assessed. This model aims to reduce the influence of changing environmental and operational conditions on damage detection. The thesis concludes by summarizing key findings and suggesting directions for further research.

1. Structural Health Monitoring and Machine Learning Methodologies

It is crucial to continuously monitor the structure in order to enable early damage detection and give the ability to stop potential future structural failures because proper functioning of structures is fundamentally important from the perspectives of both user safety and economics.

Condition monitoring (CM), non-destructive testing (NDT), damage prognosis (DP), statistical process control (SPC), and finally structural health monitoring (SHM) can be used to carry out this damage detection procedure.

Condition monitoring (CM) involves monitoring the condition of machinery through various sensors and analysis techniques in order to detect potential faults or failures early. The key advantage is that it allows for predictive maintenance and avoiding catastrophic failures. Some disadvantages are the cost of sensors and data analytics systems, as well as the need for expertise in analyzing the data. Overall, condition monitoring provides critical insights into machine health to minimize downtime and maintenance costs when properly implemented.[2]

Non-Destructive Testing (NDT) utilizes specialized techniques like x-ray, ultrasound, and eddy current to examine materials and structures without causing damage. NDT can reliably detect flaws, cracks, corrosion, and other damage through inspection. The drawback is that it usually requires direct physical access to the structure. There is also a need for skilled technicians to correctly conduct testing and interpret results.[3]

Damage prognosis is the process of estimating the remaining useful life of an engineered system by assessing its current damage state through structural health monitoring, estimating future loading conditions, and predicting through simulation models how damage will accumulate over time. The main advantage of damage prognosis is that it enables more proactive maintenance and safety assessments. However, it requires developing and integrating several complex technologies like sensing systems, data analytics, and physics-based models, which is challenging.[4]

Statistical process control (SPC) is a statistical method for monitoring and controlling a process to ensure it operates at its full potential. It uses control charts to analyze variation in a process and signal when a process is not in control. Advantages of SPC include detecting early signs of process variation and reducing waste by minimizing over-adjustment. Potential disadvantages are it requires historical data to determine control limits, control charts may fail to detect small process shifts, and it can be labor intensive to monitor charts.[5]

Structural Health Monitoring (SHM) utilizes integrated systems of sensors, data acquisition, and analytics to provide real-time damage detection. While extremely capable, SHM can be prohibitively expensive to scale across very large structures. Expert knowledge of sensors, data science, and structural analysis is also recommended to implement SHM effectively.[1]

SHM utilizes an integrated network of sensors permanently installed across the structure to provide real-time monitoring data. The spatial distribution and continuous collection under diverse conditions provides extensive operational and environmental data. This enables robust baseline modelling and separation of environmental factors from damage effects. The customized sensor networks and analytical tools in SHM systems are specifically tailored to monitor relevant environmental factors and detect damage for the structure.[1]

The long-term, rich data from a broad sensor network makes SHM the ideal platform for parsing environmental variability. Combined with physics-based and data-driven analytical capabilities, SHM provides the scale, customization, and analytical power needed for reliable damage detection in large, monitored structures with extensive sensor data. The integrated nature of SHM makes it a more holistic approach compared to standalone methods.

1.1 Structural Health Monitoring

In most cases, SHM refers to the process of monitoring a structure or mechanical system over time using periodically spaced dynamic and static response measurements, followed by the extraction of damage-sensitive features from these measurements, and the statistical analysis of the obtained features to ascertain the current state of system health.

The SHM system is implemented in four stages as shown in Figure 1.1 below.

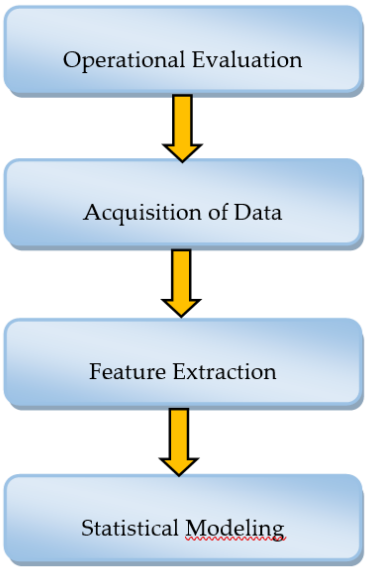


Figure 1.1: Stages of SHM system

The operational evaluation phase, this first phase is to figure out how to define damage to the system currently in examination, for example, measuring cracks in concrete, corrosion levels in steel, etc. One also need to look at any limits on how to collect data on the system. For example, one may not be able to easily access part of the system to take measurements. The main goal at this stage is to specify what kind of damage one need to spot. So, there's a need to set up ways to quantify and measure the damage. Later one can pick good sensors, collect useful data, and build models to detect that damage.[6]

The process's data collecting phase can begin once the operational evaluation phase is over. In this stage, the type and frequency of data collection, as well as the sensors and technology for storing, processing, and transmitting it, are chosen. This must be accomplished while making sure the data gathering system is adequately robust with regard to variations in environmental conditions and while also minimizing expenditures.[6]

Feature extraction phase takes care of extracting the features that are dependent on damage. These features are measurements that were taken from structural response data and are related to the presence of structural deterioration. The damage sensitive feature should, ideally, alter consistently with the degree of damage to the structure. However, the more sensitive a feature is to damage, the more susceptible it is to shifting operational and environmental conditions, which might obscure damage-related changes.[6]

Once the three aforementioned procedures have been completed, the process for developing the statistical model can begin using a variety of machine learning methods. [6]

But, as mentioned above, due to the presence of environmental and operational variability, the damage related changes will get hidden. To overcome this problem, data normalization methods can be applied. This can be done using some machine learning or Deep learning algorithms which has been described in the following sections. [7]

1.2 Data Normalization methods

The normalization of data becomes crucial to the SHM process since data can be measured under various circumstances. The new technological advancements in this area aim to address the operational and environmental SHM problems listed above.

There are three different scenarios for data normalization where operating or environmental variability is a problem.

First, several types of regression and interpolation analyses can be carried out to relate measurements relevant to structural damage and those associated with environmental and operational variation of the system when direct measurements of the varying environmental or operational parameters are available. When a large number of extracted features and measured environmental factors are available, regression analysis, as illustrated in Figure 1.2a[8], can be used, for instance, to approximate the dependency of two-dimensional features on some environmental variable T. It should be noted that there may be some damage scenarios that, unless the environmental variable is observed, cannot be separated from the undamaged conditions. [8]

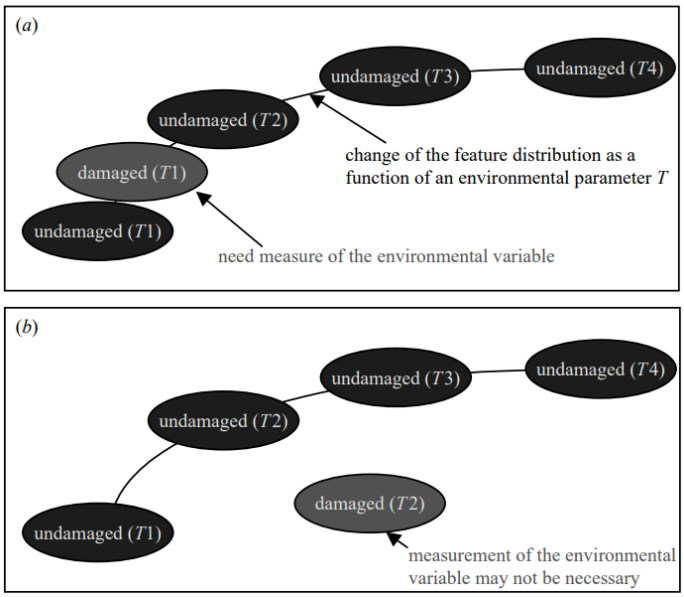


Figure 1.2: Two conceptual situation for data normalization a) Environmental variables available, b) Environmental variables not available.[9]

On the other hand, there are circumstances in which it is impractical or challenging to obtain direct measurements of these operational and environmental parameters, and damage results in changes in the extracted features that are "orthogonal" to the changes brought on by the operational and environmental variation of the system (Figure 1.2b[8]), i.e. the changes in the extracted features due to damage are independent of or unrelated to the changes caused by operational and environmental parameters. Without measuring the operational and environmental parameters, it could be possible to separate the changes brought on by damage from those brought on by the system's operational and environmental variations in this case[9].

Some of the data normalization methods are described in the further sections.

1.3 Regression

Regression analysis is a statistical method used to uncover insights about how a response variable (dependent) relates to one or several predictor variables (independent). The purpose of regression analysis is to express the response variable as a function of the predictor variables.[10] The data used determines the fit's duality and the accuracy of the regression. As a result, non-representative or incorrectly prepared data result in poor fits and results. Thus, in order to perform regression analysis effectively, one must first evaluate the data gathering process, identify any limits in the data acquired, and limit results accordingly.

Once a regression analysis relationship has been established, it can be used to forecast response variable values, identify factors that have the greatest influence on the response, or confirm proposed causal theories for the response. Through statistical analyses of the estimated coefficients (multipliers) of the predictor variables, the value of each predictor variable can be determined.

Regression analysis has three applications:

1. Prediction - To forecast future results, regression analysis is frequently utilized. Following the creation of a regression model that illustrates the link between independent variables (X) and a dependent variable (Y), modifying the independent variables' values and using the model to forecast the predicted value of Y.[10]

2. Defining the model - Another key use of regression analysis is quantifying the mathematical relationship between independent and dependent variables. The regression equation defines the model - it specifies the exact functional relationship between X and Y. Analysis of the regression coefficients provides insight into the strength and direction (positive or negative) of the impact of each independent variable on the dependent variable. Defining this mathematical relationship is useful for understanding causal connections and key drivers.[10]

3. Estimating the parameters - Important model parameters can also be estimated via regression analysis using sample data. The intercept term, coefficients, and error term are only a few of the parameters that regression algorithms estimate. These estimates were chosen to reduce the sum of squared residuals between the values of the dependent variable that was observed and those that the model predicted. Based on the available data, these parameter estimations aid in describing the actual connection between X and Y. When using the model on fresh data, accurate parameter estimates are crucial for producing predictions that can be trusted.[10]

Dynamic Regression

By adding time series elements including trend, seasonality, and autoregression, dynamic regression expands on the normal regression model for time series forecasting. Dynamic regression methods allow the relationship between the input variables and the output variable to change over time, in contrast to classic regression models that assume the input variables are independent of time.[10]

The fundamental principle of dynamic regression is to predict the future value of the output variable using the historical values of the input variables and the output variable. Lagged values of the input variables and the output variable are used as predictors in the regression model to achieve this. While the lagged values of the output variable account for the autoregressive effect, the lagged values of the input variables account for the impact of past values on the current value.[10]

The adaptability and flexibility of dynamic regression to various forms of time series data is one of its key benefits. It can deal with data that exhibit varying degrees of trend and seasonality as well as data that exhibit evolving relationships between the input and output variables across time. Exogenous factors may also be included, which could add to the information available, and boost forecast accuracy.

Dynamic regression does, however, have significant drawbacks that must be taken into account. One of the major drawbacks of the model is its susceptibility to outliers and extreme data values, which might have an impact on the estimation of the model's parameters and the forecast's accuracy. Its reliance on the stationarity assumption, which might not hold for some types of time series data, is another drawback. Furthermore, dynamic regression models may require a significant amount of computer power, particularly when working with large datasets or intricate models.[11]

Regression analysis approaches come in a wide variety, and the use of each method depends upon the number of factors. The kind of target variable, the pattern of the regression line, and the quantity of independent variables are some examples of these factors.

The different regression approaches are discussed in sections below.

1.3.1 Linear Regression

One of the most fundamental kinds of regression in machine learning is linear regression. A predictor variable and a response variable that are linearly related to one another make up the linear regression model. Multiple linear regression models are linear regression models with multiple independent variables included in the data. [10]

$$y = ß_0 + ß_1 x_1 + ß_2 x_2 + \ldots\ldots + ß_n x_n + \varepsilon \qquad \text{(Eq 1.1)}$$

Where $ß_0$, $ß_1$, $ß_2$…. $ß_n$, are regression coefficients (model parameters), and ε is the error due to variability.

Determining the model parameters:

Considering a linear relationship between two variables x(independent) and y(dependent), one can suppose the relation, $y = f(x)$ as:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad \text{(Eq 1.2)}$$

Where, β_0 and β_1 are model parameters (to be determined), and ε is random error component. The errors are assumed to have zero mean and uncorrelated.[10]

Here, linear regression can be regarded as an optimization method.

Least square estimation of the parameters:

One approach to figuring out the values of the two parameters β_0 and β_1 based on the dataset is the least-squares approach. The goal is to figure out the values of β_0 and β_1 that match the lowest possible sum of squared errors.[10]

This is how an error is defined:

$$\epsilon_i = y_i - y(x_i) = y_i - (\beta_0 + \beta_1 x_i) \qquad \text{(Eq 1.3)}$$

Where: y_i is the exact value of y, and $y(x_i)$ is the predicted value of y.

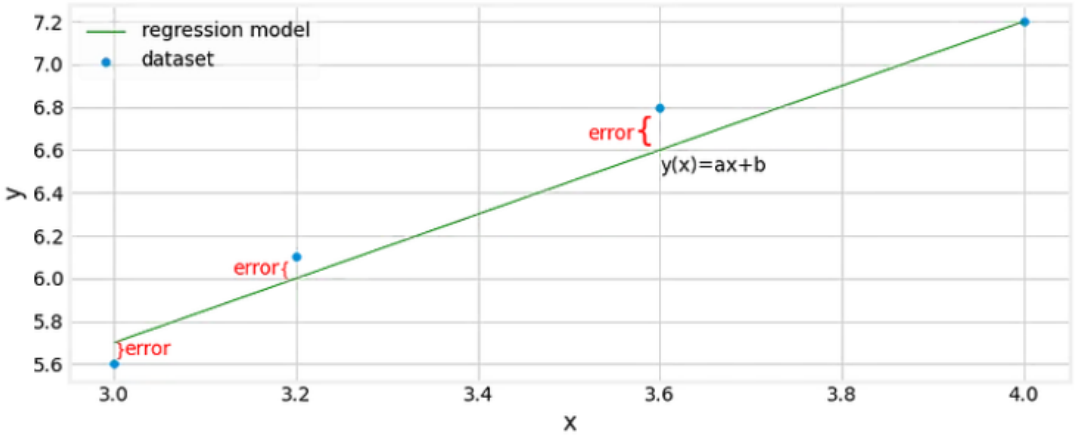


Figure 1.3: A line that fits to minimize the error.

The sum of squared errors is defined as:

$$E(\beta_0, \beta_1) = \sum_{i-1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \qquad \text{(Eq 1.4)}$$

Where n is number of observations in the dataset.

Applying minimization problem to find the extremum $E(\beta_0, \beta_1)$ to find β_0 and β_1. Hence β_0 and β_1 can be given as:

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \qquad \text{(Eq 1.5)}$$

Where $\bar{x}$ and $\bar{y}$ are mean values of x and y respectively:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$
$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \quad \text{(Eq 1.6)}$$

1.3.2 Normality test

Checking for normality is an important preliminary step in statistical analysis, especially before applying parametric regression models. This involves evaluating whether the data follows a normal distribution, also known as a Gaussian distribution. The assumption of normality is essential for many statistical methods, including parametric regression, because violating this assumption can lead to biased estimates, inaccurate p-values, and unreliable predictions.

Importance in Parametric Regression:

Parametric regression models, such as linear regression, assume that the errors (residuals) are normally distributed[10]. When the data are approximately normally distributed, the model's assumptions are more likely to be satisfied, leading to valid inferences and reliable predictions. Violation of normality assumptions can lead to biased coefficient estimates and incorrect hypothesis testing, affecting the integrity of the regression analysis.

Methods for Normality Testing:

Several methods are available to test for normality. Some of the commonly used ones include:

1. **Graphical Methods:** Histograms, Q-Q plots (quantile-quantile plots), and P-P plots (probability-probability plots)[12] provide visual insights into the distribution's departure from normality.

Histogram: A histogram (Figure 3.11(a)) is a bar chart that displays the frequency distribution of data values in intervals or "bins." In a normal distribution, the histogram should exhibit a bell-shaped curve, with data points concentrated around the mean and tapering off towards the tails. Skewed or non-normal distributions may show uneven or asymmetric histogram shapes.

Q-Q plots: A Q-Q plot is a scatterplot that compares quantization of observed data with those of the theoretical normal distribution. If the data points are approximately along a straight line, it indicates that the data is normally distributed. Deviation from a straight line represents deviation from the normality.[12]

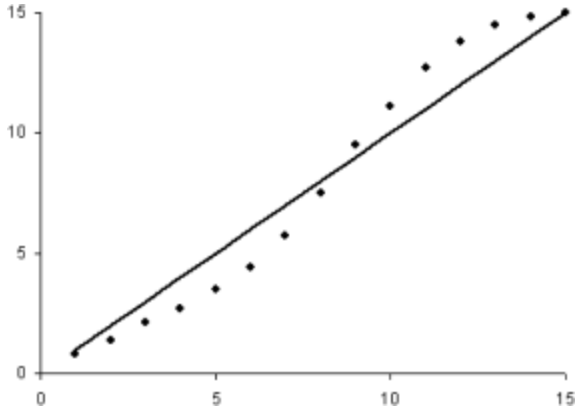


Figure 1.4: Q-Q plot showing deviation from the normal distribution.[12]

It is important to note that these methods provide visual cues but do not draw firm conclusions. Some deviations from the rule can be difficult to detect, and graphical methods can help identify potential problems that require further investigation. It is also important to use a combination of graphical methods and statistical tests for a more complete assessment of normality.

2. **Statistical Tests:** Shapiro-Wilk test, Anderson-Darling test, and Kolmogorov-Smirnov test are formal statistical tests that assess the deviation of data from a normal distribution. These tests provide p-values, which indicate the level of confidence in accepting or rejecting the null hypothesis of normality.

Shapiro-Wilk Test for Normality:
The Shapiro-Wilk test is a widely used statistical test to evaluate the normality of a data set. It is based on the idea of comparing observed sample data with what would be expected in a normal distribution. The experiment provides test statistics and p-values that help determine if the data can be considered to come from a normal distribution.[13]
Test Assumptions:

1. The test assumes that the data are independent and identically distributed.
2. The null hypothesis (H0) assumes that the data follow a normal distribution.
3. The alternative hypothesis (Ha) assumes that the data do not follow a normal distribution.

Test statistic:
Shapiro-Wilk test statistics are calculated using sorted sample data and some constants are obtained from the order statistics covariance matrix. The test statistic formula involves the sum of squares of the deviations between the observed values and the expected value under the assumption of normality. The statistical value of the test is compared with the critical values to determine the test result.[13]

$$W = \frac{(\sum_{i=1}^{n} a_i x_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where $x(i)$ is the i^{th} ordered observation, $\bar{x}$ is the sample mean, and a_i are constants derived from the covariance matrix of the order statistics.

Interpreting the Results:
If the p-value is greater than the chosen significance level (usually set to 0.05), then one cannot reject the null hypothesis. This implies that there is not enough evidence to conclude that the data deviate significantly from the normal distribution.
If the p-value is less than or equal to the significance level, reject the null hypothesis. This indicates that the data deviates significantly from the normal distribution.[13]

Non-parametric regression is a type of regression analysis that does not assume any specific form of the relationship between the dependent and independent variables. Non-parametric regression can be useful when the data is not normally distributed, as it does not rely on the assumptions of parametric methods, such as normality, homoscedasticity, and linearity. Further section describes some of the non-parametric regression methods.

1.3.3 KNN regression

KNN regression, commonly referred to as k-nearest neighbor regression, is a kind of machine learning technique used for regression problems. The fundamental principle of this technique is to forecast the output value for a new data point using the distances between the input data points and their k-nearest neighbors. KNN regression is a non-parametric technique that can discover intricate nonlinear correlations between input and output variables without relying on any presumptions regarding the distribution of the underlying data.

The number of closest neighbors that the KNN regression method will take into account when creating a forecast for a new data point, k, must first be chosen. The algorithm calculates the distances between the new data point and each of the training data points after k has been determined. Then, the distance metrics are used to determine the k-nearest neighbors.[14]

The anticipated output value for the new data point is then determined by averaging the output values for these k-nearest neighbors. Using the same value of k, this procedure is repeated for each additional data point that needs to be forecasted.[14]

Figure 1.5 shows a scatterplot of housing prices (y-axis) versus total square feet (x-axis). The blue data point is the point whose price is to be forecasted. The three nearest neighbours to the blue data point are the three red points. The distance between each data point is calculated using a distance metric, such as the Euclidean distance.

In KNN regression, the predicted price for the blue data point is calculated as the average of the prices of the three nearest neighbours.

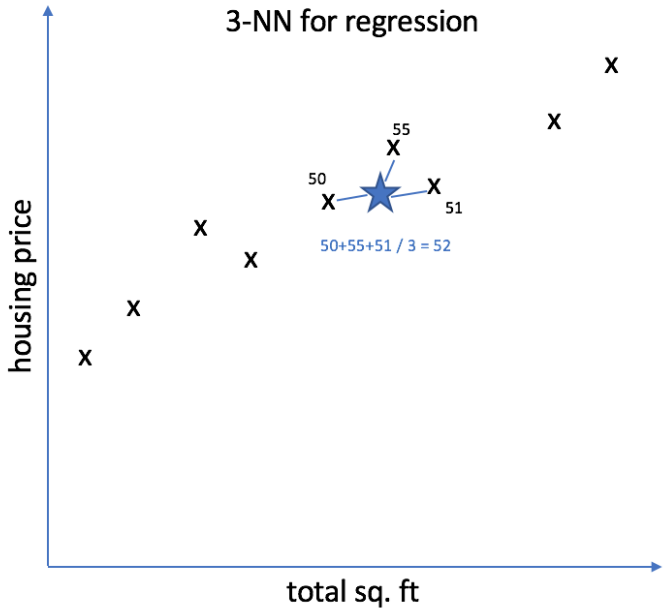


Figure 1.5: A KNN Regression Model for Predicting Housing Prices.[15]

KNN regression has the benefit of being a straightforward, easy-to-implement technique that can be applied to both small and big datasets. Additionally, it can tolerate missing values and is robust to noisy data. KNN regression also has the benefit of being able to capture nonlinear relationships between the input and output variables, which makes it applicable to a variety of regression issues.

KNN regression is a versatile machine learning approach that may be applied to a variety of regression problems. Its handling of nonlinear relationships and its tolerance to noisy data are just two of its many benefits. It does have some drawbacks, though, such as its sensitivity to the choice of distance metric and its potential computational cost for large datasets.[14]

1.3.4 Random Forest

Random forest regressor is a non-parametric machine learning algorithm used for regression tasks that involve predicting continuous output values. A more precise and reliable prediction is made with this ensemble learning technique by combining the results of various decision trees. The fundamental concept of the random forest regressor is to build a collection of decision trees, each of which is trained using a random subset of the input features and a portion of the training data. To arrive at the final prediction, the output of each decision tree is then averaged.

First, a random subset of the training data is chosen to be used for each tree in the random forest method. This method, known as bagging (bootstrap aggregation, contributes to the development of different, independent trees. Each node's split criteria for each tree are likewise chosen at random from a subset of the available input features. Through this procedure, the trees' individual accuracy is improved while the correlation between them is decreased.[16]

As shown in Figure 1.6, the random forest regressor uses the newly constructed trees to forecast the output value for a fresh data point. The algorithm accomplishes this by passing each fresh piece of input through a tree and averaging the results to arrive at the final prediction. This procedure can help to lessen the effects of overfitting and data noise and can also add a certain amount of uncertainty to the forecast.[16]

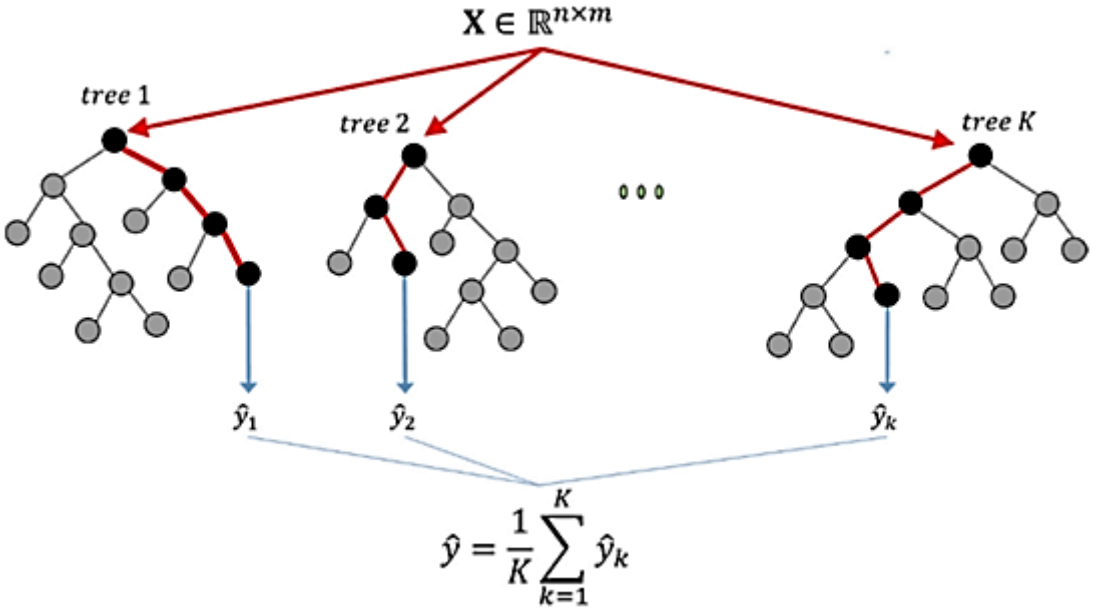


Figure 1.6: Representation of a random forest.[16]

Therefore, the Random Forest Regressor is a strong machine learning technique that may be used in a variety of regression applications. It can handle high-dimensional data with intricate nonlinear relationships and is resilient to noise and missing values, among other benefits. However, it also has certain drawbacks, such as the possibility of overfitting and the cost of calculation. Also, random forests are complex black-box models. Lack interpretability compared to simpler linear models.[16]

1.4 LSTM

The Long Short-Term Memory (LSTM) network is a type of Recurrent Neural Network (RNN) designed to address the vanishing gradients problem that can arise when training RNNs to model sequences with long-range dependencies. Unlike standard regression models, which assume the input variables are time-independent, LSTM networks utilize gated cell architectures that allow gradient information to flow unattenuated over many time steps. This enables LSTM regression models to capture time-dependent correlations between input and output variables, making them well-suited for time series forecasting problems.[17]

Long Short-Term Memory (LSTM) is an abbreviation for the type of memory units utilized in the model. These memory units can recall or forget information from previous time steps selectively, allowing the model to represent both short-term and long-term dependencies in time series data.[17]

An LSTM regression model's basic design consists of numerous layers of LSTM cells as shown in Figure 1.7, each of which has a set of memory units and gates that govern the flow of input. The model's input is a series of time steps, with each time step consisting of a collection of input features and the matching output value.[18]

During training, the model learns to modify the weights and biases of the LSTM cells in order to minimize the difference between the expected and true output values. Typically, a gradient-based optimization technique such as stochastic gradient descent (SGD) or Adam is used for optimization.[18]

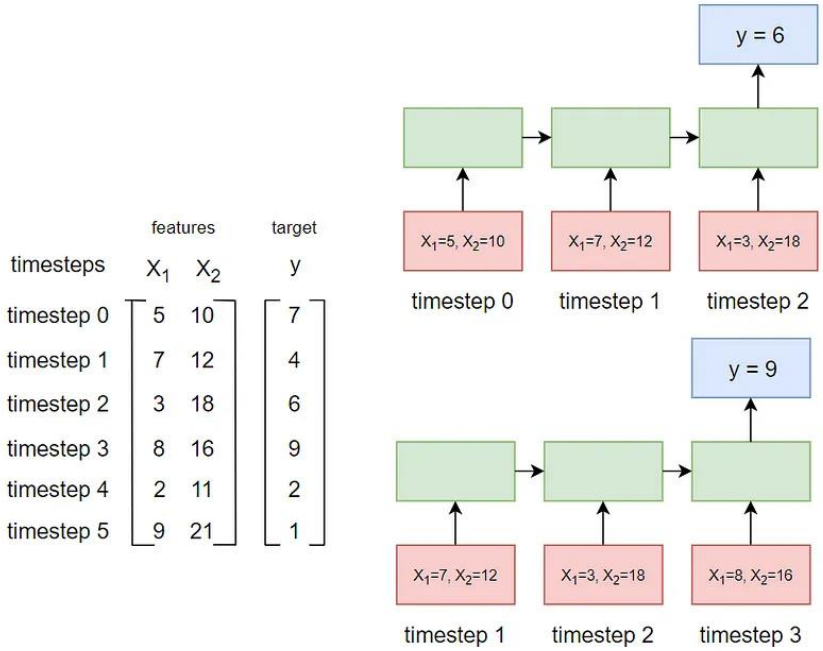


Figure 1.7: Flowchart of the genetic algorithm used to train the LSTM model for time series regression.[18]

LSTMs can capture long-range dependencies in time series data due to their gated architecture that allows gradient information to flow unattenuated over many time steps. This makes them well-suited for modeling sequences with long-term correlations. Traditional regression models frequently fail to capture these dependencies because they only evaluate the input variables' most recent values.[17]

However, there are several limits to LSTM regression models that must be recognized. One of the most significant disadvantages is their computational complexity, which can make it difficult to train and deploy on big datasets or on resource-constrained devices. In order to avoid overfitting and obtain acceptable generalization performance, they also require a substantial amount of training data. Also, determining the optimal hyperparameters (e.g., number of gates, memory units, layers) for an LSTM model can be difficult and problem specific.[17]

1.5 Kalman Filter

The Kalman filter is a powerful and widely used mathematical technique for estimating the state of a dynamic system from a series of noisy and incomplete measurements. It offers a methodical approach to combining ambiguous data with forecasts of how a system will behave over time, producing a more precise and trustworthy estimate of the real state.

Initially, control systems and navigation were envisioned to be applications for the Kalman filter. Since then, a wide range of disciplines, including aerospace, robotics, economics, signal processing, and more, have found use for it. When working with systems that are impacted by noise, uncertainty, and shifting situations, the filter is extremely helpful.

Basic concept and methodology

This dynamic estimating approach works by combining the prediction step with the update step. These actions, supported by a number of crucial components, allow the filter to produce accurate state estimations despite noisy observations and uncertainty.

If one considers a system:

$$\dot{x} = A \cdot x + B \cdot u$$
$$z = C \cdot x + n \qquad \text{(Eq 1.7)}$$

Components of the Kalman Filter:

i. **State Vector (x):** This vector encapsulates the variables that describe the state of the system. It might encompass quantities such as position, velocity, orientation, and more. In present work, this is bridge response sensor value.

ii. **State Transition Matrix (A):** The A matrix captures the system's dynamics by representing how the state evolves over time. It combines the current state with external influences or control inputs ($B \cdot u$) if they exist.

iii. **Control Input (B):** When present, this matrix accounts for external influences, like forces or commands, that affect the state transition.

iv. **Observation Matrix (C):** The C matrix maps the state vector to the expected measurements. It outlines the relationship between the state and the measurements obtained from the system.

v. **Measurement Noise Covariance (R):** The R matrix signifies the uncertainty associated with measurements. It captures the variance or covariance of measurement errors, reflecting their inherent imprecision.

$$R = E[(z - y)(z - y)^T]$$

vi. **Process Noise Covariance (Q):** This matrix captures the uncertainty stemming from the system's dynamics not accounted for in the state transition. It represents external influences, disturbances (w), or other factors.

$$Q = E[ww^T]$$

vii. **Error Covariance (*P*):** The error covariance matrix P is a crucial component in the Kalman filter that quantifies the uncertainty associated with the state estimate. It reflects how confident the filter is in the accuracy of its predictions and updates.

$$P = E[(x - \hat{x})(x - \hat{x})^T]$$

1. Prediction Step:

In the prediction step, the Kalman filter leverages the current state estimate and the system's dynamics model to predict the state at the upcoming time step.

- **State Prediction:** Utilizing the A matrix and, when applicable, the control input $B \cdot u$, the predicted state at k+1 is formulated:

$$\hat{x}_{k+1|k} = A \cdot \hat{x}_{k|k} + B \cdot u_k \quad \text{(Eq 1.8)}$$

- **Error Covariance Prediction:** The error covariance P prediction portrays the uncertainty linked to the state prediction. It evolves using the state transition matrix A and the process noise covariance Q:

$$P_{k+1|k} = A \cdot P_{k|k} \cdot A^T + Q \quad \text{(Eq 1.9)}$$

2. Update Step:

In the update step, the Kalman filter integrates a fresh measurement to correct and refine the forecasted state estimate.

- **Kalman Gain Calculation:** The Kalman gain K blends the prediction's uncertainty and the measurement's uncertainty, regulating the measurement's influence on the update. It's determined through the error covariance prediction, the observation matrix C, and the measurement noise covariance R:

$$K_{k+1} = P_{k+1|k} \cdot C^T \cdot (C \cdot P_{k+1|k} \cdot C^T + R)^{-1} \quad \text{(Eq 1.10)}$$

- **State Update:** Incorporating the Kalman gain-scaled difference between the actual measurement z_{k+1} and the predicted measurement $C \cdot \hat{x}_{k+1|k}$, the updated state estimate emerges:

$$\hat{x}_{k+1|k+1} = \hat{x}_{k+1|k} + K_{k+1} \cdot (z_{k+1} - C \cdot \hat{x}_{k+1|k}) \quad \text{(Eq 1.11)}$$

- **Error Covariance Update:** The updated error covariance takes into account the Kalman gain's impact on diminishing uncertainty due to measurement integration:

$$P_{k+1|k+1} = (I - K_{k+1} \cdot C) \cdot P_{k+1|k} \quad \text{(Eq 1.12)}$$

The equations Eq 1.8 – Eq 1.12 combined makes Kalman forecaster algorithm.

Terminology and Notation:

k: Time index

u_k: Control input at time *k*

x: State vector

A: State transition matrix

B: Control input matrix

Q: Process noise covariance matrix

C: Observation matrix

R: Measurement noise covariance matrix

P: Error covariance matrix

z_{k+1}: Actual measurement at time *k*+1

I: Identity matrix

Matrix A and C are determined using N4SID algorithm[19], which uses the observation to build the matrices.

Predict Future Values: Once the filter has been initialized and the model parameters are set, one can also use it to predict future values of the time series by performing only the prediction step. The filter will incorporate new observations and adjust its estimates as new data arrives.

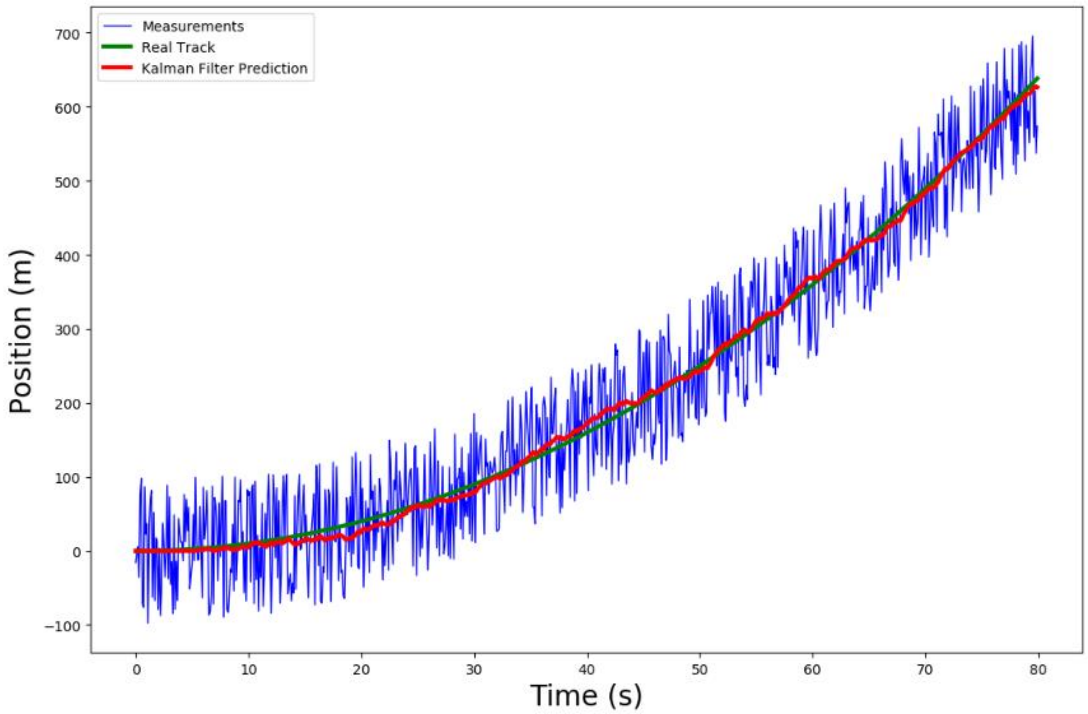


Figure 1.8: Kalman filter for tracking moving object.[20]

1.6 Principal Component Analysis

Definition

Principal Component Analysis (PCA) is a multivariate statistical procedure that is used to analyze data by minimizing its dimensionality. It is frequently used in disciplines like machine learning, data analytics, and finance to reveal patterns and relationships in data that can be challenging to see in other ways. PCA enables to convert a high-dimensional dataset into a lower-dimensional space while maintaining most of the important data.

In a dataset there are as many numbers of principal components as there are number of features. Each principal component explains a certain amount of variations in the dataset. The principal components are ranked according to how much variance they account for and are orthogonal to one another, making them uncorrelated. The first principal component, followed by the second principal component, and so on, explains the most variance.

In order to prepare the data for machine learning algorithms, PCA is frequently utilized. It can assist in locating redundant or pointless data characteristics, which can result in overfitting and subpar generalization. The accuracy and effectiveness of

the machine learning model can be increased by using PCA to assist reduce the number of variables.

Methodology

PCA works by determining the eigenvectors and eigenvalues of the data's covariance matrix. The covariance matrix explains how the variables in the dataset are related to one another. The covariance matrix's eigenvectors reflect the highest variance directions in the data, while the accompanying eigenvalues represent the amount of variance explained by each eigenvector.[21]

The method begins by normalizing the data so that each variable has a zero mean and a unit variance before using PCA as shown in Figure 1.9. This is crucial since the analysis will be dominated by variables with higher variances. The covariance matrix of the standardized data is subsequently computed. Each pair of covariances between the variables in the dataset are contained in the covariance matrix, which is a square matrix.[21]

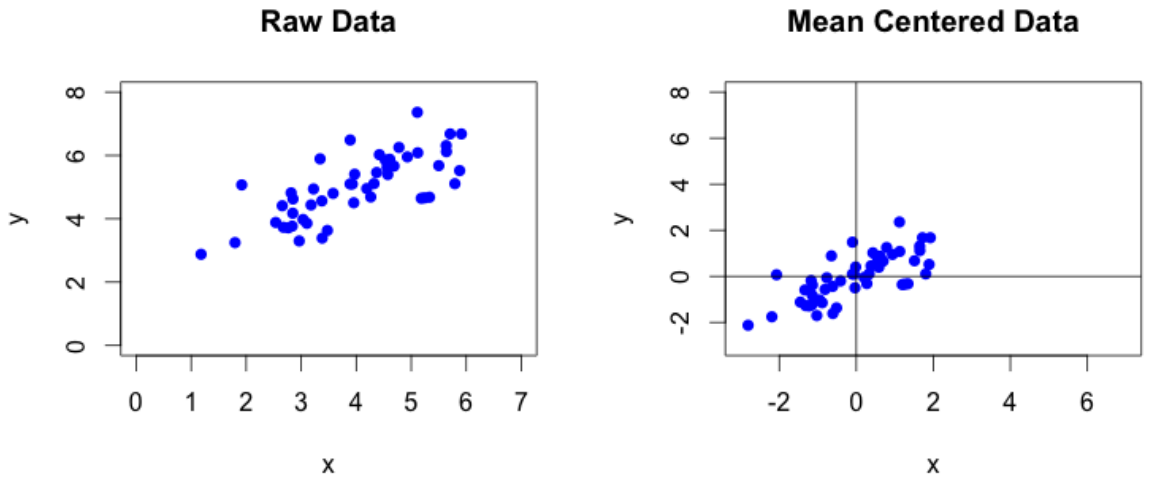


Figure 1.9: Normalizing data before using PCA.[22]

The covariance matrix is known for being square symmetric. It is also known that the covariance matrix should ideally be a diagonal matrix with variances on each diagonal and zero values for all values off the diagonal. Here, the aim is to diagonalize this covariance matrix. [22]

From linear algebra it is possible to create a matrix E such that the following equation is true for a symmetric matrix S:

$$S = EDE^T \quad \text{(Eq 1.13)}$$

where E is a matrix with the eigenvectors of matrix S stored in its columns and D is a diagonal matrix. Eigenvalues and eigenvectors are always presented in pairs. The direction of the maximum variance is indicated by the covariance matrix's eigenvectors, while the proportion is indicated by the eigenvalues. In other words, the more information can be found in the direction of an eigenvector the higher its eigenvalue.[22]

So, finding B such that X is transformed into $\acute{X}$:

$$XB = \acute{X} \quad \text{(Eq 1.14)}$$

where X is the data matrix that is to be transformed by means of the vector basis that is kept in the columns of matrix B. A new matrix $\acute{X}$ emerges from the transformation.

Covariances of both X and $\acute{X}$ are related by:

$$C_{\acute{X}} = B^T C_X B \quad \text{(Eq 1.15)}$$

Since, C_X is symmetric, it follows:

$$C_{\acute{X}} = B^T C_X B = B^T (EDE^T) B \quad \text{(Eq 1.16)}$$

One may demonstrate that $C_{\acute{X}}$ is diagonal if the matrix B is chosen such that its columns contain the eigenvectors of the covariance matrix C_X i.e., $B = E$, One can also use the property of eigenvectors $E^T = E^{-1}$.[22]

Thus, the eigenvectors of the covariance matrix of the measured data can be used to compute the principal components.

Hence the next step after normalizing the data is to find the covariance matrix's eigenvalues and eigenvectors. The principal components of the data are represented by the eigenvectors, and the variance explained by each principal component is represented by the eigenvalues. The first eigenvector, which accounts for the most variation in the data, is followed by the second eigenvector, and so on, in the order of the eigenvectors and their corresponding eigenvalues.[22]

By projecting the data onto the eigenvectors, one can then convert it into the principal component space as shown in Figure 1.10. The final dataset will contain fewer variables, each of which will correlate to a principal component, but it will also contain the same number of observations as the initial dataset.[22]

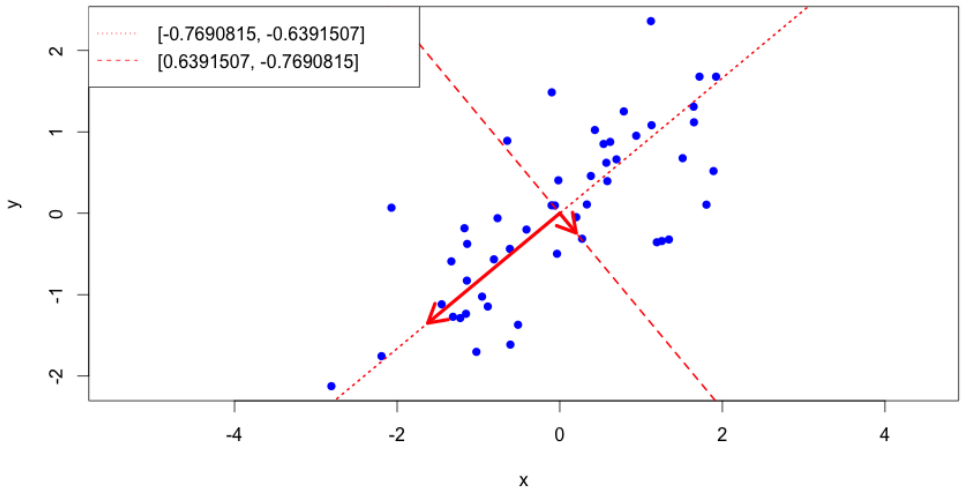


Figure 1.10: Dataset transformed into Principal component space.[22]

1.7 Darts Python Library

Darts is a Python machine learning library for time series forecasting that offers a variety of models, from classics such as ARIMA to state-of-the-art deep neural networks. The library focuses on providing modern machine learning functionalities, such as supporting multidimensional series, fitting models on multiple series, training on large datasets, incorporating external data, ensembling models, and providing a rich support for probabilistic forecasting.[23]

One of the key features of Darts is its user-friendly and easy-to-use API design. All models in Darts support the same basic fit()/ predict() interface, similar to scikit-learn, making it easy to train and forecast with different models without having to know their inner workings. Darts also supports training one model on a potentially large number of separate time series, which is beneficial for ML models that work best when trained on datasets containing multiple time series.[23]

Here are some key features and functionalities of Darts:

1. Time Series Representation: Darts has its own TimeSeries data container type, which represents one time series. TimeSeries are immutable and provide guarantees that the data represents a well-formed time series with correct shape, type, and sorted time index. TimeSeries can be indexed either with Pandas DatetimeIndex or RangeIndex.
2. Unified High-Level Forecasting API: All models in Darts support the same basic fit(series: TimeSeries) and predict(n: int) --> TimeSeries interface to be trained on a single series and forecast n time steps after the end of the series.

This unified API makes it possible to seamlessly compare, backtest, and ensemble diverse models without having to know their inner workings.

3. Training Models on Collections of Time Series: Darts supports training one model on a potentially large number of separate time series. The library provides various classes implementing different ways of slicing series into training samples. All neural networks in Darts are implemented using PyTorch and support training and inference on GPUs.
4. Support for Past and Future Covariates: Several models in Darts support covariate series as a way to specify external data potentially helpful for forecasting the target series. Darts differentiates between past covariates, which are known only into the past, and future covariates, which are known into the future. The models accept past covariates and/or future covariates arguments, which make it clear whether future values are required at inference time.
5. Other Features: Darts also offers additional features such as transformers and pipelines for data preprocessing, backtesting, hyperparameter search, extensive metrics, dynamic time warping module, ensemble models, and filtering models such as Kalman filters and Gaussian Processes.

Overall, Darts provides a comprehensive and user-friendly environment for time series forecasting, with support for a wide range of models and functionalities.

Due to its comprehensive features and user-friendly API, Darts library has been chosen for all the machine learning models in the present thesis work.

1.8 Residual method

In the context of machine learning, residuals are defined as the difference between the observed values of the dependent variable and the values predicted through regression or predictive modeling techniques. Formally, the residuals are defined as shown in Figure 1.11:

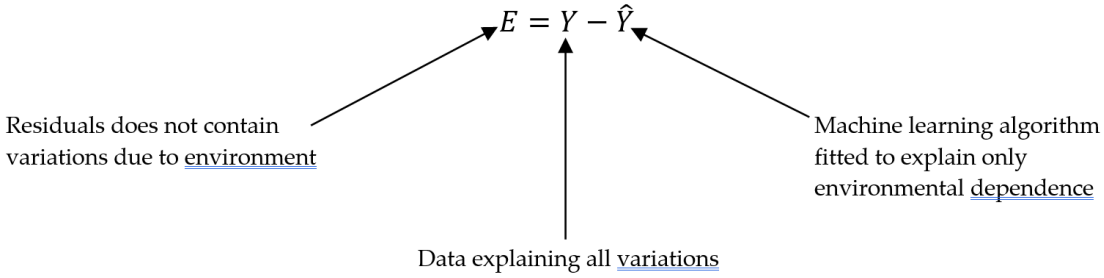


Figure 1.11: Residual calculation

Residual Method's Importance:

Accurate Structural Assessment: The residual approach makes it possible to examine the structural behavior of the bridge more precisely by removing environmental variables. This enables more accurate assessments and well-informed decision-making by enabling engineers and researchers to isolate and examine the response brought on by actual structural loads, degradation, or damage.

Early Damage Detection: Because environmental factors can hide or obfuscate the presence of structural deterioration or damage, it can be difficult to spot warning signals of impending collapse. The residual approach aids in separating the impacts of damage from the overall reaction, making it easier to identify structural problems early on. Engineers can spot potential issues quickly by keeping an eye on how residuals vary over time.

Cost-Effective Monitoring: The residual technique makes use of the already-existing sensor network and makes use of predictive models, so there is no longer a need for more sensors that are only used to monitor the environment. This method makes monitoring the health of bridges more affordable by lowering the expenses associated with sensor installation, upkeep, and data collection.

2. Literature review on damage detection

The four fundamental components of data-based Structural health monitoring are [6]:

(i) defining and quantifying the damage that needs to be identified.
(ii) an ongoing network of sensors,
(iii) an automated process for instantaneous or periodic feature extraction, and
(iv) an effective novelty detection system.

The second component has drawn a lot of attention over the past ten years, and thanks to tremendous strides in sensor and instrument technology, it is now possible to install very large sensor networks on structures and collect the measured data in central recording units at high sampling rates. The third component is still a challenge today and is a current area of research for the most often utilized features (eigenfrequencies and mode shapes). Several methods from statistics and machine learning have been used for the fourth component; the most popular ones are control charts, hypothesis testing, and outlier analysis using the Mahalanobis squared-distance.

This literature work focuses on the fourth component of a data-based Structural health monitoring system in the presence of environmental and operational variability. Several studies [24]–[27] have found significant variation in the dynamic characteristics of structures subjected to ambient vibrations. These researches highlight the fact that variations in dynamic features caused by confounding factors (temperature, humidity, traffic, solar radiation) can be of the same order of magnitude as, or higher than, variations caused by damage, making detection of the onset of damage difficult.

In this chapter, a literature review pointing out different techniques used to filter out these environmental effects is carried out. The most basic methods rely on determining the linear subspace to which the environmental and operational factors belong in order to remove their influence on the monitored properties. Such methods

are appropriate when the dimension of the feature vector is large enough to allow the identification of a linear subspace to which the confounding effects belong.[28]

2.1 The Mahalanobis squared-distance outlier analysis.

This section describes the mathematical basis of the approach suggested by Deraemaeker et el. 2018 to eliminate confounding effects using the Mahalanobis squared distance.

Considering N healthy state observations of a structure with n features $\{y_i\}_{n\times 1}$ with $(i = 1.\ldots.N)$, the covariance $[C]_{n\times n}$ is calculated as:

$$\{\bar{y}\} = \frac{1}{N}\sum_{i=1}^{N} y_i \qquad \text{(Eq 2.1)}$$

$$[C] = \frac{1}{N-1}\sum_{i=1}^{N}(\{y_i\} - \{\bar{y}\})(\{y_i\} - \{\bar{y}\})^T \qquad \text{(Eq 2.2)}$$

The features taken from the vibration data, such as a set of eigenfrequencies, mode shapes, FRF or transmissibility functions at specific frequencies, etc., can be represented by the multivariate feature vectors. The basic idea behind outlier analysis is to compute the Mahalanobis squared distance provided by, for each sample of the multivariate feature vector y_j.

$$D_j^2 = \left(\{y_j\} - \{\bar{y}\}\right)^T [C]^{-1}\left(\{y_j\} - \{\bar{y}\}\right) \qquad \text{(Eq 2.3)}$$

One can set a threshold and if D_j of a new sample y_j is above this threshold, it is considered as an outlier.[28]

2.1.1 Spectral decomposition

The covariance matrix is typically not diagonal, thus feature transformation is carried out to diagonalize the covariance matrix.

$$\{\eta_i\} = [U]^T y_i \qquad \text{(Eq 2.4)}$$

The new mahalanobis squared-distance is given by:

$$D_j^2 = \sum_{i=1}^{n} \frac{1}{\sigma_i^2}\left(\eta_{ji} - \bar{\eta}_i\right)^2 \qquad \text{(Eq 2.5)}$$

This demonstrates how the changed variables can each contribute independently to the Mahalanobis squared-distance. The weights of the contributions are determined by the inverse of the corresponding eigenvalues, σ_i^2, which are the variances of the

newly converted variables. The contribution to the distance is little if the variation is high.[28]

2.1.2 Filtering environmental effects

The full variability in the feature vector collected from the healthy condition may frequently be explained by a smaller number of transformed features, which are typically referred to as the principal component, when the number of features is sufficient. In mathematical terms, this happens when part of the eigenvalues of $[C]$ equal zero. The training data's null-space is made up of the related eigenvectors. In reality, the eigenvalues are not absolutely equal to zero because of noise and problems with numerical precision, but a noticeable decline in the eigenvalues can be seen and used to determine how many principal components are responsible for the majority of the variability. The following indicator can be used to discover how many principal components can be used:

$$I = \frac{\sum_{i=1}^{p} \sigma_i^2}{\sum_{i=1}^{n} \sigma_i^2} \quad \text{(Eq 2.6)}$$

A threshold can be set such that $I > e\%$, which means p principal components are needed to explain $e\%$ of the variance. Thus, the mahalanobis squared distance can be decomposed into two parts,

$$D_j^2 = \sum_{i=1}^{p} \frac{1}{\sigma_i^2} \left(\eta_{ji} - \bar{\eta}_i\right)^2 + \sum_{i=p+1}^{n} \frac{1}{\sigma_i^2} \left(\eta_{ji} - \bar{\eta}_i\right)^2 = D_{1j}^2 + D_{2j}^2 \quad \text{(Eq 2.7)}$$

the Mahalanobis squared-distance of y_j projected on the principal components is D_{1j}^2, while the Mahalanobis squared-distance of y_j projected on the null-space of the principal components is D_{2j}^2.

Assuming environmental factors now provide a relatively high amount of variability in the feature vector retrieved from the healthy condition, if this variability is more significant than other causes like noise, it will fall within the category of the first p principal components. The distance will be particularly insensitive to environmental changes since the Mahalanobis squared-distance scales each independent component with respect to the inverse of its variance. The Mahalanobis squared-distance is rendered insensitive to the environmental conditions by adding the feature vector measured in all conceivable environmental conditions in the construction of the covariance matrix.[28]

Application of the mahalanobis distance for outlier analysis can be found in [28].

2.2 Principal component analysis for damage identification

In this section, a different strategy based on principal component analysis (PCA) is suggested by Yan et el. 2005. This method does not rely on the measurement of environmental factors or the knowledge of underlying physical quantities. Environmental influences are instead considered embedded variables. The method's fundamental premise is that environmental fluctuations in measured features can be accounted for using PCA and that these variations are distinct from those caused by structural damage. Consequently, it is possible to identify them. The PCA model's prediction errors could be used as a damage indicator. To determine whether the features point to a divergence from previously predicted normal conditions or not, novelty analysis is used. In this study, the principal components linked to environmental influences were removed, and an outlier analysis was performed on the minor components to find damage. [21]

2.2.1 Methodology

It is well-known that changes in environmental factors (such as temperature, temperature gradients, humidity, wind, etc.) have a significant impact on vibrational characteristics. Environmental factors are typically not quantified; instead, their effects are just seen from variations in the measured properties.[21]

PCA transforms data from original dimension n to a lower dimension p:

$$X = TY \tag{Eq 2.8}$$

Y is original data containing N number of observations and n number of features.

X is transformed data with lower p number of features.

Typically, the principal p eigenvectors of the covariance matrix of Y can be used to calculate matrix T. But a more practical alternative is to use singular value decomposition of feature covariance matrix.

$$\begin{aligned} YY^T &= U\Sigma^2 U^T \\ UU^T &= I \\ \Sigma &= \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \end{aligned} \tag{Eq 2.9}$$

where U is an orthonormal matrix, whose columns are principal components, whose active energy is given by diagonal terms of matrix Σ.

$\Sigma_1 = diag(\sigma_1, \sigma_2, \dots, \sigma_p)$ and $\Sigma_2 = diag(\sigma_{p+1}, \sigma_{p+2}, \dots, \sigma_n)$

Also, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq \sigma_{p+1} \geq \cdots \geq \sigma_n \to 0$

According to PCA, the structure's vibrational characteristics fluctuate mostly along the principal component directions that are linked to the highest energies. To put it another way, the vibrational features roughly stay on the hyperplane established by the p principal components chosen. The choice of an acceptable dimension p is not as important as it might first appear. It is possible to acquire steady monitoring results with various values of the order p as long as the relative change of this hyperplane from the reference to the current states is taken into consideration.

Thus, the matrix T in Eq. (2.8) may be constructed using the first p columns of U to project the observed features into the space described by environmental factors. By re-mapping the projected data to the original space, the amount of information lost in this projection can be evaluated.

$$\hat{Y} = T^T X = T^T T Y$$

And residual error matrix is given as:

$$E = Y - \hat{Y} \qquad \text{(Eq 2.10)}$$

The Novelty Index (NI) is calculated using the prediction error vector E_k collected at time t_k and is defined either using the Euclidean norm:

$NI_k^E = ||E_k||$

Or mahalanobis norm:

$NI_k^M = \sqrt{E_k^T R^{-1} E_k}$

Where R is feature covariance matrix.

It is possible to do statistical analysis if it is additionally assumed that the Euclidean or Mahalanobis indices are normally distributed.

A centerline (CL) at NI and two additional horizontal lines (UCL and LCL) versus the identification numbers are drawn to create an X-bar control chart, with $\overline{NI}$ and σ defined as the mean value and standard deviation of NI for the prediction in the reference state respectively. A confidence interval of 99.7% can be chosen.[21]

The hyperplane spanned by the vibration features of the reference state should contain the vibration features corresponding to the present data if there is no damage. Therefore, the current data's outlier statistics value should stay at the same level as for the reference data. In contrast, structural damage should result in a departure from the original hyperplane and a sharp increase in the outlier statistics

of the damaged state. The ratio $\frac{NI_d}{NI_r}$ (d and r stand for, respectively, the damaged and reference states) may also be employed as a quantitative indicator of damage level in addition to the outlier statistics.[21]

2.2.2 Geometric Interpretation

A geometric interpretation of two-dimensional data with two characteristics (y_1 and y_2) is used to illustrate the method under discussion. The characteristics are shown in Figure 2.1 as circles spaced out from their geometric center (point O'). Environmental differences are thought to be the main cause of the features' dispersal. This data set is subjected to PCA analysis, which produces PC-I and PC-II as the two principal components. The dominant environmental component or a mix of several factors, PC-I has the largest unique value and accounts for the majority of the feature variance. PC-II, on the other hand, symbolizes the impact of secondary factors. This method offers valuable insights into the underlying relationships and patterns in the data set.[21]

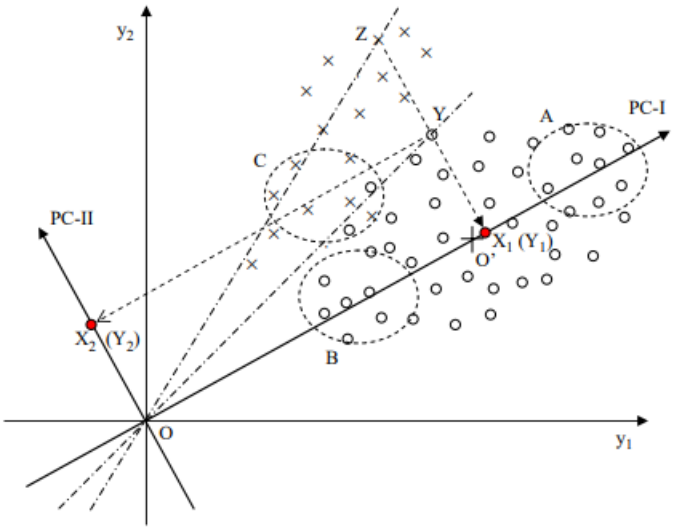


Figure 2.1: Geometric Interpretation.[21]

First, this 2D data is projected into the 1D space spanned by PC-I using Eq. (1), using point Y as an example. It yields a scalar with length OX_1 as its value. This data point is remapped into the original 2D space to produce point Y_1, and the length of segment Y_1Y is used to calculate the residual error.

Using point Z as an example for a damaged state, the residual error (Y_1Z) greatly rises in comparison to Y_1Y using the same projection method as for point Y. The

effect of environmental factors in such a comparison between healthy and damaged states has been roughly eliminated.

It should be noted that the traditional PCA approach typically requires a data normalization process to produce variables with a zero-mean and a unit standard deviation.

$$y_k^* = (y_k - \bar{y})/\sigma_y \quad \text{(Eq 2.11)}$$

Where $\bar{y}$ is mean, σ_y is standard deviation of dataset.

Damaged-state data are normalized by subtracting always the reference data set's mean value rather than the damaged data's own mean value. Taking a closer look at the two distinct data sets in Figure 2.2 that represent the healthy (reference) marked with ∘ and damaged states marked with ×, respectively, to better understand this. Figure 2.2(a) demonstrates that the features belonging to the damaged state are combined with the features corresponding to the starting structure when the mean value of each data set is removed, rendering damage undetectable.[21]

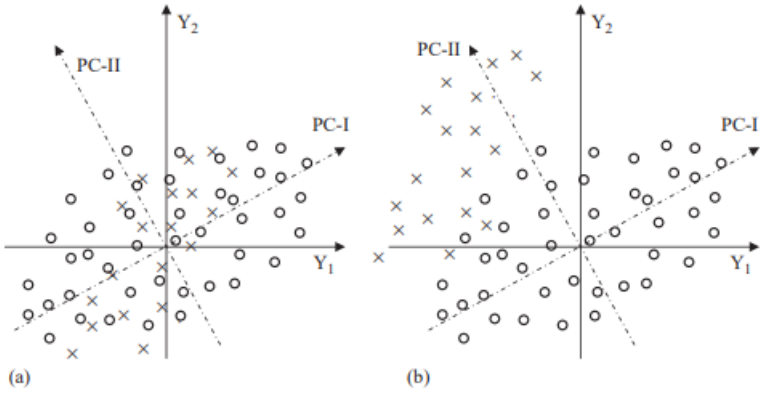


Figure 2.2: PCA geometric interpretation with data normalization: (a) Elimination of mean from both set separately; (b) Elimination of mean of damaged set from reference set.[21]

Application of the mahalanobis distance for outlier analysis can be found in [21].

2.3 Linear regression for damage identification

This section introduces a linear filter proposed by Sohn et el. 1999 for a large-scale bridge's damage detection system that adapts to environmental variations. This system predicts the underlying response variables of the structure based on a time-environmental profile. In doing so, the system is able to distinguish between

response variable changes brought on by environmental changes and those brought on by structural degradation. For instance, the system can reliably indicate that structural changes are probably caused by variables other than environmental influence when the measured response departs from the expected confidence intervals.

2.3.1 Model Formulation

The first aim for this study is the prediction of the response variable. The changing of the response variable is thought to be mostly caused by changes in the bridge's temperature, and is assumed to be linearly correlated.

A linear predictor is selected as the system architecture in light of these presumptions. Simply put, a linear filter makes an input-output mapping that is linear and one-to-one. The alternative coefficients can be explicitly calculated using a straightforward matrix calculation, and they can be modified in the future using adaptive least-mean-squares-error minimization. Training and prediction are the two modes in which the filter functions.[29]

2.3.2 Training linear filter model

The architecture of the linear filter produces a single output that corresponds to the estimated or expected response variable from a subset of temperature profiles as inputs. The filter, which is likewise a multiple linear regression model in this sense, is more frequently referred to as a predictor or estimator. The method of Least-Mean-Squares (LMS) error reduction is used to calculate the coefficients of the predictor, and the variable selection problem, which is defined as selecting the suitable subset of the available temperature profiles, is explained in Section 2.3.3.[29]

The linear filter creates a linear function to represent the relationship between the observed response variable, y, at the selected bridge temperature inputs, x, a column vector of r inputs.

$$y = x^T w + \varepsilon \qquad \text{(Eq 2.12)}$$

Where,

$$x = [1 \quad x_1 \quad x_2 \quad \cdots \quad x_r]^T$$
$$w = [w_0 \quad w_1 \quad w_2 \quad \cdots \quad w_r]^T$$

w is a coefficient vector to weight temperature inputs, and ε is residual error.

The filter used to construct this model is shown in Figure 2.3. The temperature readings at the present time T_i and the preceding time T_i' are utilized as input

variables in order to take into account both the temporal and spatial change of temperature. That means $x = [1 \quad T_1 \quad \cdots \quad T_9 \quad T_1' \quad \cdots \quad T_9']$. In Figure 2.3 strict linear mapping is mandated by the filter.

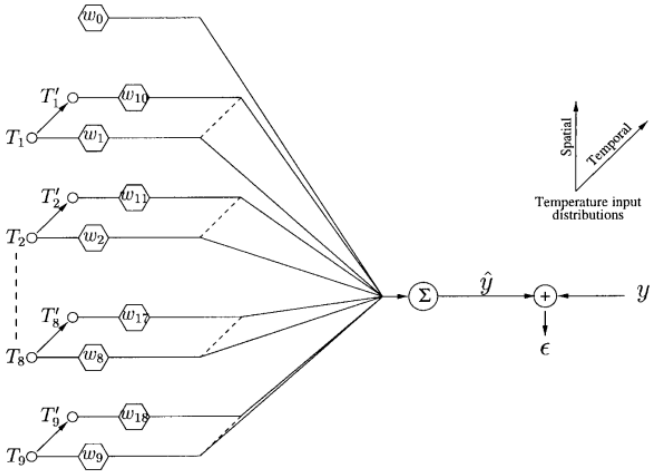


Figure 2.3: A linear adaptive filter.[29]

Let's say there are n observations available, and the i^{th} input-output pair is represented by $x(i)$ and $y(i)$.

Matrix notation can be used to write Eq. (2.12):

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \varepsilon$$

Where,

$$y = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(n) \end{bmatrix}; x = \begin{bmatrix} 1 & x_1(1) & x_2(1) & . & x_r(1) \\ 1 & x_1(2) & x_2(2) & \cdots & x_r(2) \\ . & \vdots & . & \ddots & \vdots \\ 1 & x_1(n) & x_2(n) & \cdots & x_r(n) \end{bmatrix}; \varepsilon = \begin{bmatrix} \varepsilon(1) \\ \varepsilon(2) \\ \vdots \\ \varepsilon(n) \end{bmatrix}$$

The filter coefficients are estimated using the LMS error minimization method. The aim is to find filter coefficients vector that minimizes the predicted value of the square of the filter error.

$$\min_{w} E[\varepsilon(\mathrm{i})^2]$$

where $E[\varepsilon(\mathrm{i})^2]$ represents the average of the filter errors brought on by the n observations. It is possible to rewrite $E[\varepsilon(\mathrm{i})^2]$ as follows. For ease of notation, the index i is dropped after the first line.

$$\begin{aligned} E[\varepsilon(i)^2] &= E\left[\left(y(i) - \boldsymbol{w}^T\boldsymbol{x}(i)\right)^2\right] \\ &= E[(y - \boldsymbol{w}^T\boldsymbol{x})^2] \\ &= E[y^2 + \boldsymbol{w}^T\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{w} - 2y\boldsymbol{x}^T\boldsymbol{w}] \end{aligned}$$

$$= E[y^2] + \boldsymbol{w}^T E[\boldsymbol{x}\boldsymbol{x}^T]\boldsymbol{w} - 2E[y\boldsymbol{x}^T]\boldsymbol{w}$$
$$= E[y^2] + \boldsymbol{w}^T \boldsymbol{R}\boldsymbol{w} - 2\boldsymbol{p}^T\boldsymbol{w} \quad \text{(Eq 2.13)}$$

Where $p = E[y\boldsymbol{x}^T]$ is the cross-correlation between the intended output and the input vector and $R = E[\boldsymbol{x}\boldsymbol{x}^T]$ is the autocorrelation of the random input vector x. Here, it is clear that $E[\varepsilon^2]$ is quadratic with w and can be calculated for a single extremum (minima) with regard to w. Eq. (2.13) is differentiated with respect to w to find the estimated coefficients, $\boldsymbol{w}$, and the resulting value is set to zero:

$$\nabla(E[\varepsilon^2])\frac{\partial E[\varepsilon^2]}{\partial \boldsymbol{w}} = 2(\boldsymbol{R}\hat{\boldsymbol{w}} - \boldsymbol{p}) = 0$$
$$\hat{\boldsymbol{w}} = \boldsymbol{R}^{-1}\boldsymbol{p} \quad \text{(Eq 2.14)}$$

The Wiener-Hopf equation, or Eq. (2.14), is used to calculate the estimated coefficients, $\hat{\boldsymbol{w}}$, for a set of input-output pairs.

All input variables are taken into account in the derivation of Eq. (2.12) in order to predict the output response. However, in the majority of real-world applications, the analyst must assess the significance of each input and select an ideal subset from a set of potential inputs. This approach is covered in the following subsection and is equal to removing unnecessary or duplicated inputs from the filter of Figure 2.3.[29]

2.3.3 Input variable selection

Before estimating the filter coefficients, the choice of input variables should be made to minimize the size of the filter. A model with fewer input variables is often preferred because the variance of the prediction $\hat{y}$ rises as the number of inputs does. Additionally, the expense of gathering data and maintaining the model rises when additional inputs are added.

First, the relationship between the measured response variable and the nine temperature sensor readings is looked into. The outcome of the correlation matrix is shown in Table 1. The correlation matrix demonstrates a strong relationship between (T_3) and (T_4) temperatures. The temperature (T_6) and (T_8) have a high correlation. Because T_3 has a stronger correlation with the observed output y than T_4, T_4 is removed from the filter model.[29]

	y	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_7	T_8
y	1.000									
T_1	-0-097	1.000								
T_2	0.435	0.835	1.000					Sym.		
T_3	0.608	0.684	0.941	1.000						
T_4	0.580	0.707	0.943	0.997	1.000					
T_5	0.485	0.787	0.969	0.966	0.966	1.000				
T_6	0.130	0.949	0.901	0.839	0.853	0.916	1.000			
T_7	0.741	0.396	0.750	0.910	0.909	0.807	0.605	1.000		
T_8	0.065	0.968	0.883	0.804	0.820	0.886	0.996	0.556	1.000	
T_9	-0-232	0.886	0.641	0.518	0.540	0.668	0.870	0-283	0.889	1-000

Table 1: Correlation of the measured fundamental frequency and the thermometer readings

2.3.4 Prediction

The response variable of the bridge is estimated using the adaptive filter set up in the preceding section. The measured response value is then employed to separate the variations in the response variable brought on by temperature impacts from variations brought on by other potential structural problems. Let x_0, for instance, stand for a vector of fresh temperature values. The response at the temperature profile is predicted at y_0 as follows:

$$\hat{y}_0 = x_0^T \hat{w} \qquad \text{(Eq 2.15)}$$

Where $\hat{w}$ is weight vector.

However, a perfect match between the predicted and measured modal parameters cannot be anticipated due to the model's shortcomings, a lack of training data sets, errors in the actual testing and measurements, and other factors.

One may believe with reasonable confidence that some variations in the underlying structural characteristic are caused by damage or other factors if the response variable is outside the confidence interval.[29]

Figure 2.4 shows the predicted ∘ and measured × response variable as a function of temporal temperatures.[29]

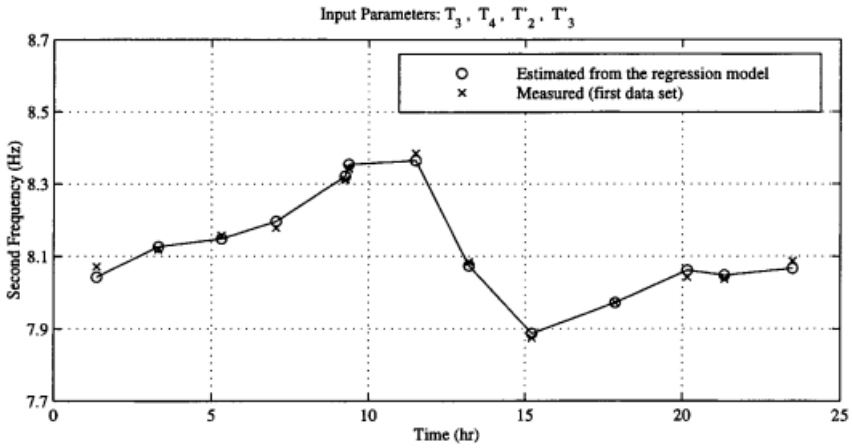


Figure 2.4: Reproduction of frequency using linear filter.[29]

2.4 Combination of MLR and PCA

MLR and linear PCA are coupled to possibly improve damage detection results. In the MLR-PCA combined technique, an MLR model is used first, and then PCA is applied to the residuals from the MLR model's Eq. (2.18). The MLR-PCA method is based on the premise that MLR removes the impacts of quantifiable predictors, and PCA removes the residual effects of unmeasured environmental and operational factors. It should be noted that a combination PCA-MLR model, with the order of PCA and MLR reversed, might also be proposed. However, if the operational and environmental factors that have the greatest impact on frequency data are accurately measured, then PCA-MLR technique would not perform any better than the MLR method alone because the PCA would remove the same amount of variance as regression analysis does.[30]

Figure 2.5 shows the flowchart to make control chart from the residuals obtained from the PCA analysis. The control chart is a statistical tool that can be used to monitor the values of features that are insensitive to operational and environmental factors, in order to detect abnormal occurrences. The control chart consists of data plotted in the time order and horizontal lines, designated control limits, which indicate the amount of variation due to common causes. An observation outside the control region is considered to be an out-of-control observation, or in other words, an

observation suggesting a special cause of variation. This cause of variation may be linked to the occurrence of damage in the context of structural health monitoring.[31]

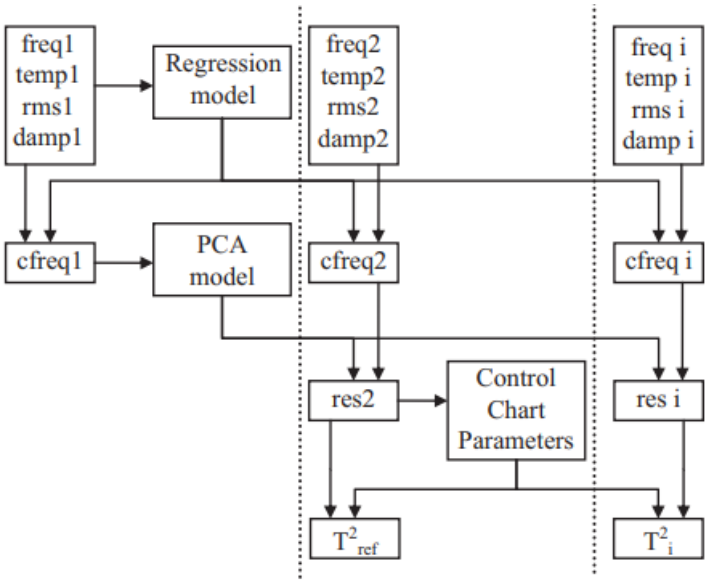


Figure 2.5: A flowchart of MLR-PCA method that can be used to create control chart.[31]

To make a control chart using a combination of regression analysis and PCA, the following steps can be taken:

1. Use regression analysis to establish a model relating observed environmental or operational factors with estimated natural frequencies. This will help to eliminate the influence of these factors on the natural frequencies, so that small changes due to damages can be detected.
2. Use PCA to reduce the dimension of the problem, by substituting a group of correlated variables by a new smaller group of independent variables, which are designated principal components. The original variables can be transformed into another set of variables by the application of an orthonormal matrix that applies a rotation to the original coordinate system.
3. Apply the transformation expressed by the orthonormal matrix to new observations and calculate the residues following a specific equation. The residues can be used to detect abnormal occurrences that might be justified by the existence of damaged zones.
4. Use control charts to monitor the values of the features obtained from the PCA analysis in order to detect abnormal occurrences. Control charts can be used to set a control region for future observations, taking into account the properties

of previously collected data. The verification of future observations can be performed by checking if each new observation lies within a previously defined 'safety' region.

The residuals from MLR can be given as:

$$E_R = Y - \beta^T Z^T \quad \text{(Eq 2.16)}$$

Where Y is observation matrix, Y has dimensions $n \times N$, where n is the total number of observations and N is the total number of features. $\beta \in \mathbb{R}^{N\times(P+1)}$ is weighing matrix estimated using least square minimization method. Z is a matrix containing independent variables.

The final residuals after MLR-PCA can be given as:

$$E = \hat{T}_E^T \hat{T}_E E_R$$

Where $\hat{T}_E$ is the reduced loading matrix of the residuals of the MLR model. This is computed by only retaining part variance.[30]

It should be noted that the MLR filter's parameters and the appropriate selection of the number, l, of retained PCs are the key parameters influencing the combined method's results and necessitating some initial adjustment.

3. Livenza Bridge, Results and Discussion

3.1 Bridge Description and monitoring system characterization

The bridge under observation is a railway steel truss bridge with two spans that spans the Livenza River in Northern Italy. The primary span, with a length of 60 m is shown in Figure 3.1 and a model is represented in Figure 3.2, is used for this research without losing generality. The bridge is monitored by a permanent system that includes multiple sensors, as shown in Figure 3.3: the latter does not report all of the installed sensing devices, but only those that are located on the span of interest. Tiltmeters and displacement transducers (LVDTs) have sensor IDs reported.[32]

Figure 3.1: Livenza Bridge.[32]

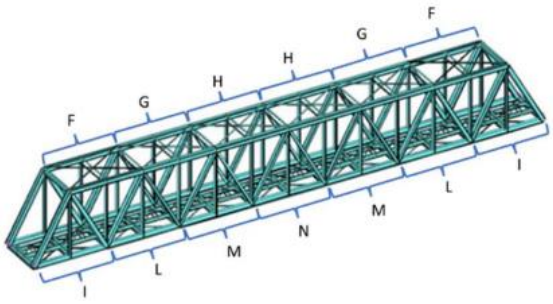


Figure 3.2: Structural model of the main span of Livenza bridge (length of 60.5 m). The letters identify different lower/upper chords sections.[32]

The installed monitoring system additionally offers capabilities for the assessment of external variables (temperature, air humidity, etc.): therefore, the influence of seasonal fluctuations on sensor measurements (for inclinometers and LVDTs) may possibly be eliminated. The data from all sensors is continually logged at a low frequency (1 Hz).[32]

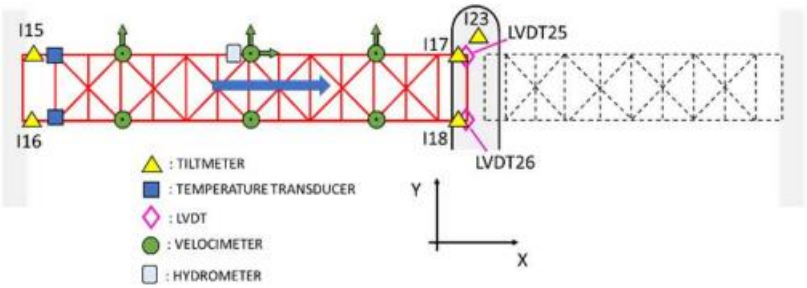


Figure 3.3: Scheme of the sensors composing the monitoring system on the span of interest for the present work.[32]

3.2 Susceptibility to Environmental and Operational factors

From the Figure 3.4(a) and (b) which shows time series of Bridge response sensor – E35 and Environmental sensor – T39 respectively over a period of 1 week, it can be shown that the variations in bridge response sensor E35 are due to the fluctuations in temperature. This can also be proved statistically by analyzing the correlation matrix shown in Figure 3.12, which shows high correlation coefficient between the sensors E35 and T39. Similarly, it can also be shown for other bridge response and environmental sensors as well.

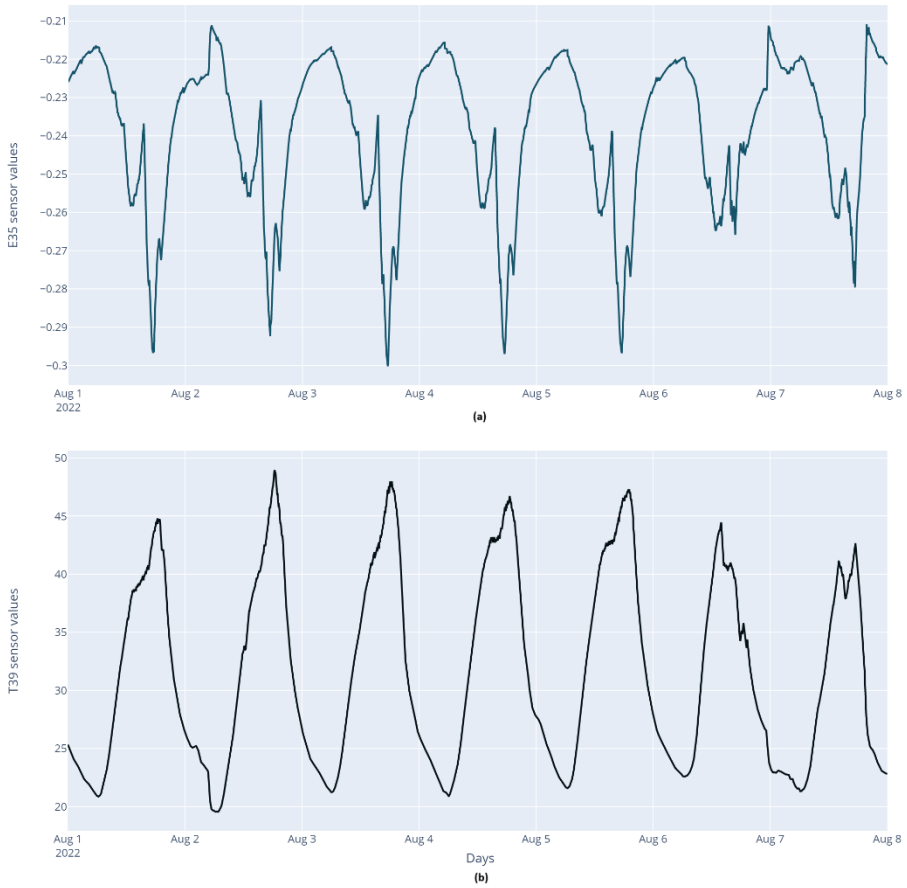


Figure 3.4: Time series of (a) Bridge response sensor E35; (b) Environmental sensor T39

Moreover, some spikes can also be observed which might be due to the operational variabilities as discussed earlier. These can also be referred as local outliers and are addressed in Section3.4.2.

3.3 Methodology

Model building and residual calculation are the two steps used in the residual approach. In order to create correlations between the environmental variables (temperature, humidity, solar radiation) and the structural response variables (strain, displacement, and inclinometer readings), historical data is examined throughout the model building phase. To create predictive models that forecast the predicted values of the sensor data based on the known conditions at hand, a variety of statistical and machine learning techniques can be used. These models accurately reflect the innate relationship between environmental variables and structural response.

The sensor measurements that would be anticipated primarily owing to environmental factors are predicted using the models after they have been built. The residuals are then produced by subtracting these anticipated values from the corresponding real sensor data as shown in Figure 1.11 of Section1.8. The residuals represent the genuine structural response, free of environmental influences, and they offer valuable information about the condition and behavior of the bridge.

A predictive model can be created using a variety of machine learning techniques in order to calculate the residuals. These techniques can be roughly divided into four categories: static regression, dynamic regression, principal component analysis (PCA), and PCA and regression analysis combined.

Static Regression: Static regression techniques use a fixed or static model to represent the link between external parameters and bridge response sensor readings. Static regression methods including linear regression, KNN regression, and random forest regression are frequently employed.

- Linear Regression: Linear Regression presupposes that environmental conditions and bridge sensor values have a straight-line relation. In order to forecast sensor measurements based on environmental parameters, it predicts the coefficients that best suit the data.
- KNN Regression: KNN regression is a non-parametric technique that makes predictions about sensor readings by taking into account the k nearest neighbors in the training dataset. The average or weighted average of the nearest neighbor observations is used to get the expected values.
- Random Forest Regression: Using a combination of different decision trees, random forest regression is a collective learning technique. It is appropriate for capturing intricate patterns in the data because it can manage nonlinear correlations and interactions between variables.

Dynamic Regression: When the relationship between environment parameters and bridge response sensor readings demonstrates temporal dynamics, dynamic regression approaches are very helpful. Long Short-Term Memory (LSTM) regression is a dynamic regression method that is frequently employed.

- LSTM Regression: Recurrent neural networks (RNNs) with the ability to detect long-term dependencies and temporal patterns in sequential data. It is useful for forecasting sensor measurements based on time-varying environmental conditions since it is well-suited for modeling time series data.

Principal Component Analysis: The principal component analysis is primarily used for dimensionality reduction. In this case, it can be used to identify the most significant principal component that causes environmental variation. This

component can further be deleted and the remaining can be retained using residual method.

Combination of PCA and Regression Analysis: For better environmental compensation, the residuals from the regression analysis can be fed into principal component analysis to remove the effects of some unknown variations and thus the residual method by this combination can be made more efficient.

But before a model can be created, the data to be fed to the model should be good, i.e., it should be consistent, free from outliers and missing values, to ensure accurate and meaningful results.

3.4 Data Preprocessing

Any project involving data analysis or machine learning must start with data preprocessing. It entails converting unstructured raw data into an organized format that algorithms and models can use to their fullest potential. This procedure is essential for raising the data's quality and dependability, which raises the precision and effectiveness of any future analysis or prediction jobs.

3.4.1 Data visualization and cleaning

A crucial part of data preprocessing is visualization and cleaning, also known as data cleansing or data scrubbing. To make sure that the data is accurate, full, and dependable for further analysis, it entails locating and fixing flaws, inconsistencies, and inaccuracies within a dataset. The accuracy and effectiveness of subsequent analysis or prediction jobs strongly depend on the cleanliness of the data, hence data cleaning is crucial to raising the quality of the data.

Data is prone to a wide range of problems and inaccuracies, which might come from several sources. For instance, human mistakes during data entry could lead to typos, missing numbers, or inaccurate data inputs. Additionally, defective sensors or other data collection equipment might compromise data, resulting in outliers or unusual values. Data analysis might also be complicated by incomplete records, inconsistent formats, or differences in measuring units.

Finding and fixing these problems is the main goal of data cleaning in order to provide a trustworthy and consistent dataset. This entails using a variety of strategies and procedures that are suited to the particular dataset and the types of mistakes that are there.

But to do data cleansing, visualizing the data is crucial. With the use of data visualization, one can immediately understand the meaning of the data. Charts,

graphs, and maps are examples of visual representations that offer a clear and succinct overview of the data rather than digging through rows and columns of figures. When data is presented visually, it is easier to spot patterns, trends, and anomalies, which aids in understanding the underlying data.

Data visualization might assist in locating any potential data issues, such as inconsistencies, outliers, or missing values. It can also assist in better understanding the relationship between the dependent and independent variables.

Sections below show different data visualization methods.

3.4.1.1 Time series plots

To see how data evolves over time, time series line graphs might be helpful. They can be used to spot patterns, outliers, missing data, and trends. Additionally, they can be used to compare various time series.

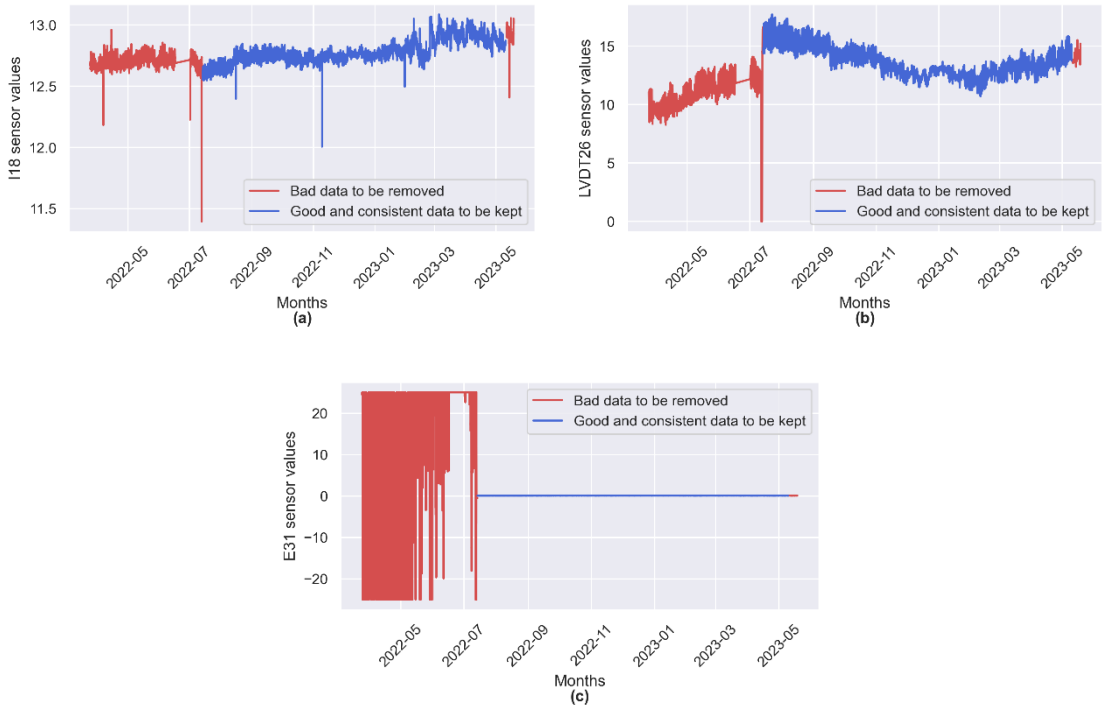


Figure 3.5: Time series visualization for (a) I18 inclinometer; (b) LVDT25 displacement; (c) E31 strain sensors

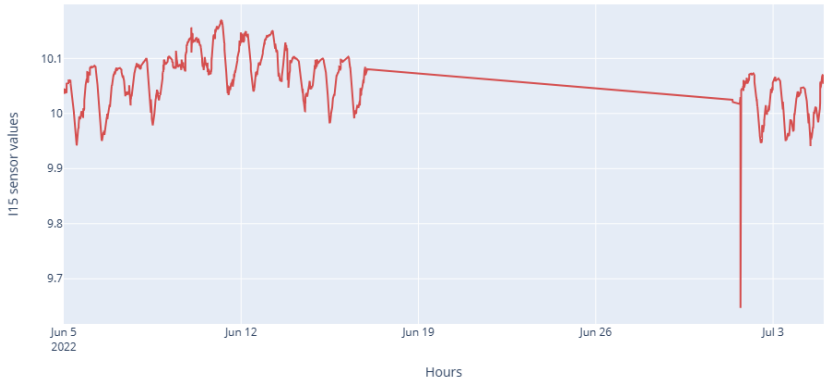


Figure 3.6: I15 sensor time series enlarged view.

As it can be seen from Figure 3.6 that the data points for almost 15 days (from June 15 to June 30) are missing in I15 inclinometer sensor, hence it's not continuous, and in sensor E31 as shown in Figure 3.5(c) there are many anomalies before July and a sudden inconsistency in the data of LVDT25 sensor at around mid-July as can be seen in Figure 3.5(b).

Imputing too many values can result in a number of potential issues.

Bias in the data: When many missing values are imputed, the dataset may be biased since the imputed values may not be a genuine representation of the missing true values. As a result, the machine learning model may produce incorrect or misleading findings by distorting the relationships and patterns in the data.

Overfitting: When imputing a lot of missing values, there is essentially addition of new data to the dataset. This might result in overfitting, when the model exhibits excessive sensitivity to the imputed values and exhibits poor performance on fresh, unforeseen data.

So, it is better to remove this large inconsistent data before applying any machine learning method. This method of removing the data is referred as Exclusion.

Thus, the data selected for the study was 14th July 2022 to 10th May 2023.

Out of ['I15', 'I16', 'I17', 'I18', 'I19', 'I20', 'I21', 'I22', 'I23_X',

'I23_Y', 'LVDT25', 'LVDT26', 'LVDT27', 'LVDT28', 'IDR41', 'E29',

'E30', 'E31', 'E32', 'E33', 'E34', 'E35', 'E36', 'T37', 'T38',

'T39', 'T40', 'Patm', 'Int temp', 'Ext Temp', 'Int hum', 'Ext hum',

'W speed', 'W dir', 'Rain rate', 'Solar rad\n']

Only good performing sensors were selected, and faulty sensors were dropped.

Faulty sensors are I16', 'I17', 'I19', 'I20', 'I21', 'I22', 'I23_X', 'I23_Y', 'LVDT27', 'LVDT28', 'IDR41'.

Also, the sensors which are not useful in calculation of residuals can also be dropped, i.e., the sensors which have constant values over the period.

'Patm', 'W speed', 'Rain rate' are sensors with almost constant value over the period of time considered as shown in Figure 3.7.

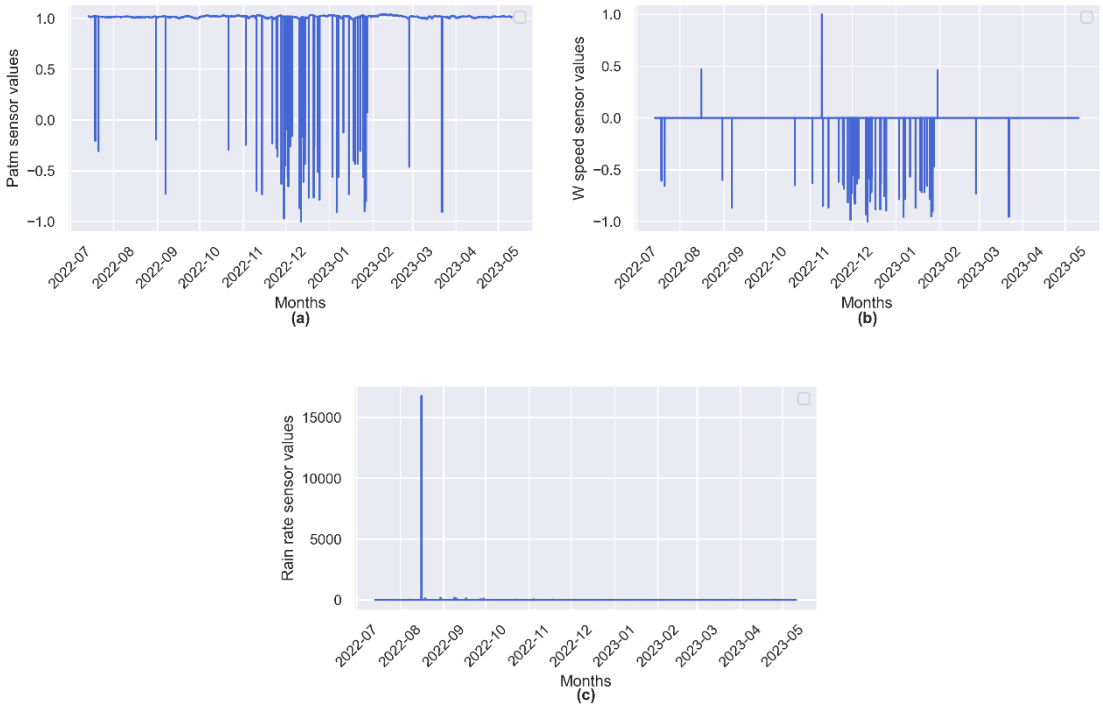


Figure 3.7: Sensors with almost constant value over time period

3.4.1.2 Scatter Plots

Graphs that depict the relationship between two variables are called scatterplots. In a scatterplot, each data point is represented by a dot. Scatter plots allows to uncover relationships between two variables, helping discern correlations, clusters, or outliers.

As can be seen from Figure 3.8, the pearson correlation factor is quite high between all the dependent and independent sensors, which indicates a strong linear relationship between these variables.

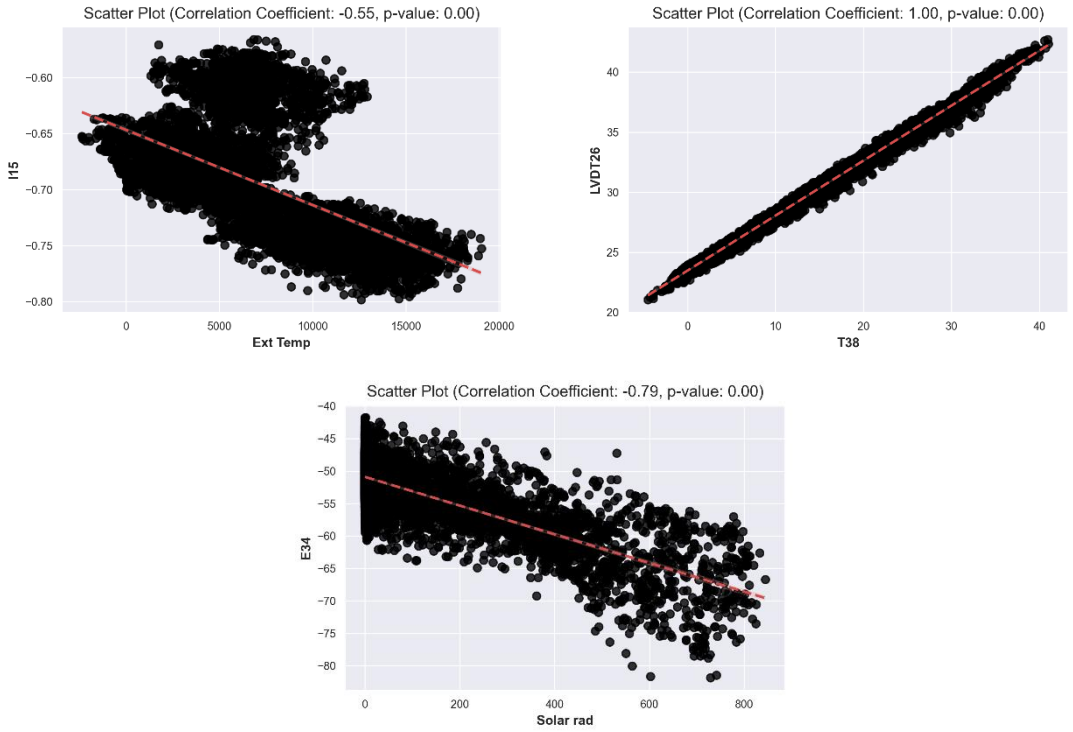


Figure 3.8: Scatter Plots

The elimination of duplicate records is also a frequent task in data cleansing. Due to the fact that they essentially give the same information more than once, duplicates can skew analysis results and cause biases. Each data point is only used once in the analysis thanks to the detection and removal of these duplicates.

3.4.2 Handling Outliers and Missing Values

Outliers are data points that dramatically vary from predicted patterns or trends. Measurement errors, data entry problems, or truly extreme numbers in the dataset can all be causes of outliers. In the present case, the outliers are also said to be due to the operation of the bridge itself. This can be due to passing of a train which in turn changes the dynamic response of the bridge and can be seen in Figure 3.9(a) with red curve having spikes. These outliers can skew statistical measures, affect how variables relate to one another, and skew the findings of analyses. Therefore, it is crucial to identify outliers and deal with them effectively.

The z-score method uses the mean and standard deviation to calculate a z-score for each data point. The z-score is calculated as:

$$z = \frac{x - \bar{x}}{\sigma}$$

Data points with absolute z-scores greater than 3 are typically considered potential outliers. This rule is based on the empirical rule that 99.7% of observations from a normal distribution should lie within 3 standard deviations of the mean.[33]

The interquartile range (IQR) method[33] is another approach that requires figuring out the range between the data's first and third quartiles (Q_1 and Q_3). Outliers are data points that are outside the normal distribution of:

$$Q_1 - 1.5 \times IQR \; or \; Q_3 + 1.5 \times IQR.$$

Several strategies can be employed to deal with outliers once they have been identified. Outliers are replaced with the nearest non-outlier values using the widely used Winsorization technique[34]. By reducing the impact of extreme numbers while maintaining the data's normal distribution, this strategy aids. Alternatively, if outliers are determined to be data anomalies or if they significantly impair the analysis results, they can be eliminated from the dataset. However, care must be used when eliminating outliers because doing so may result in the loss of important data or skew the study.

Because this is a time series data, it is preferable to analyze local outliers rather than global outliers, so an approach that involves replacing outliers with more typical values based on adjacent data was used. A function was defined to remove the outliers. The steps to remove and impute outliers are mentioned below:

1. Iterating through the Time Series: Ensuring that sufficient neighboring data points are available for comparison and replacement.
2. Retrieving Neighboring Values: By considering a window of neighboring values, the function aims to capture the local context and assess whether the current value is an outlier or not.
3. Comparing with the Mean: The function calculates the mean of the selected neighboring values. This mean value represents the local average and is used as a benchmark for comparison.
4. Detecting and Replacing Outliers: The current value is then compared to the calculated mean. If the absolute difference between the current value and the mean exceeds the defined threshold, it is considered an outlier. In such cases, the outlier is replaced with the calculated mean value.
5. Returning the Modified Time Series: Once all the iterations are complete, the function returns the modified time series with outliers replaced by the corresponding means.

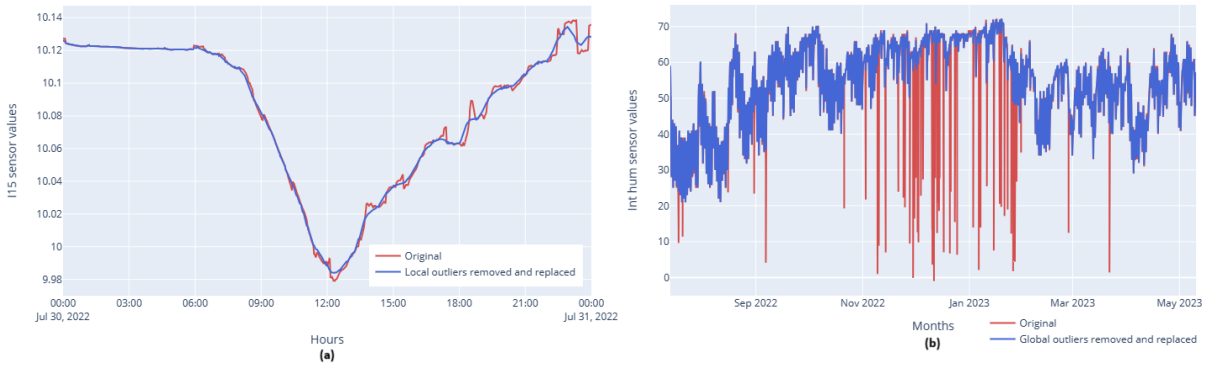


Figure 3.9: (a) Local outlier removal; (b) Global outlier removal

This function provides a simple method for removing outliers from a time series collection. It finds and substitutes probable outliers by taking into account local context and comparing values to the local mean, resulting in a more representative dataset. The pseudo algorithm for the function can be reported as in Algorithm 1 below.

Algorithm 1: Imputing local outliers/ operational data function
Input: Unfiltered dataframe
Output: Filtered dataframe
Defining Function: Replace Outliers (Input arguments – time series, number of neighbors, threshold)
for (*i* from number of neighbors to (size of time series − number of neighbors − 1) **do**
Slicing timeseries
if (absolute value of (current value−average of sliced timeseries)>threshold) **then**
Replace the value with the mean value of the window
end if
end for
return modified timeseries

However, a few assumptions while applying this function have been considered. To begin, the function is assumed to have a symmetric distribution of values around the mean. This strategy may not be appropriate if the distribution is severely skewed or has complex patterns. Furthermore, the function assumes that the given threshold accurately captures the extent of outlier departure. Setting an improper threshold value may result in the wrong removal or retention of outliers.

Handling missing values is an additional essential component of data preprocessing. Data might be missing for a number of reasons, including the fact that it was not gathered or recorded for certain variables or circumstances. Missing values can significantly alter the outcome of analysis, producing skewed findings or insufficient insights. There are various strategies that can be used, depending on the percentage of missing values and the type of data.

Exclusion, is a strategy for dealing with missing data which was already discussed in the Section3.4.1.1 and applied to the data, and entails deleting instances or variables with missing values from the analysis. However, if the missing data exhibits a pattern, this strategy should be utilized cautiously since it could result in the loss of important data or biased findings.

Whereas Imputation is a popular method for dealing with missing values. It entails substituting estimated or expected values for the missing variables using the data that is already available. The process of imputation can be carried out using a variety of techniques, including mean imputation (replacing missing values with the variable's mean), regression imputation (forecasting missing values based on other variables), and sophisticated methods like multiple imputation that produce multiple imputations that are plausible and take uncertainty into account.

Interpolation, which includes guessing the missing values based on the observed data patterns, is a useful method for handling missing data. The second-degree polynomial interpolation method is used to fill the missing data points as shown in Figure 3.10.

A mathematical method called polynomial interpolation enables to approximate a function using a polynomial equation. When the data points display a curved pattern, second-degree polynomial interpolation, commonly referred to as quadratic interpolation, is especially helpful. It entails guessing the missing values based on a curve that is fitted to a quadratic equation using the data that are already available.

For imputing missing data, the second-degree polynomial interpolation approach has a number of benefits. It can identify intricately curved patterns in the data and offer reasonably precise estimates for values that are absent.

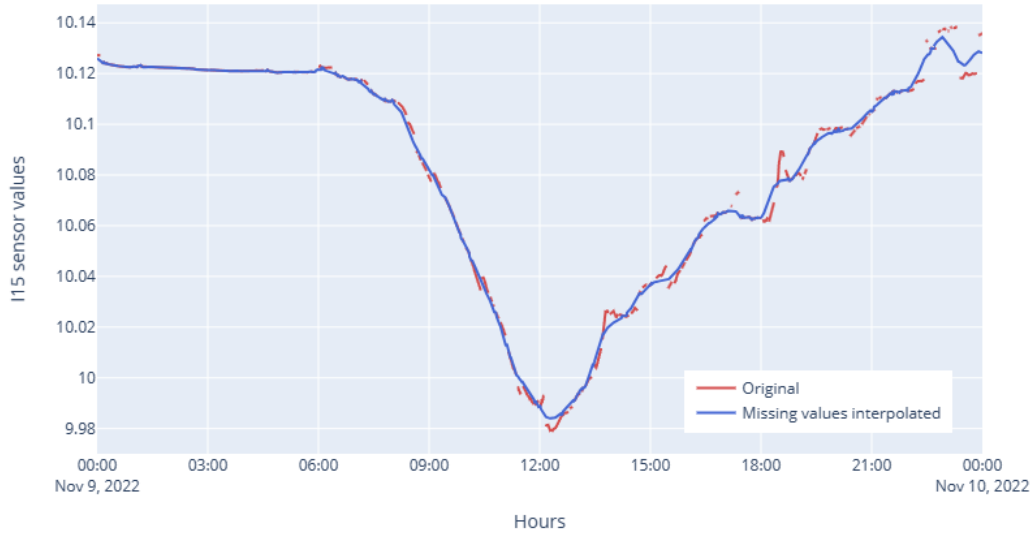


Figure 3.10: Missing timestamp interpolation using second-degree polynomial.

3.4.3 Feature transformation

The raw values that come directly from the sensors are electrical voltages, so there is a need to transform these values to their respective physical values and units. So, an electrical to physical transformation was applied, which converts all data to physical units. Also, since large dataset close to 1 year is considered, a 1 min frequency of observations can be changed to 1 hour frequency by averaging the data over 1 hour. Further analysis is done considering dataset frequency of 1 hour.

3.4.4 Normality test

As discussed in Section1.3.2, the normality test was performed on the dataset to know the type of distribution.

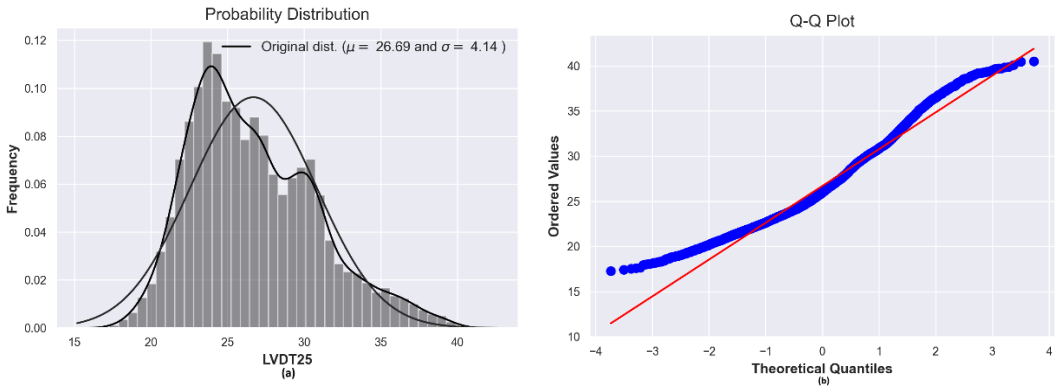


Figure 3.11: (a) Histogram probability plot; (b) Q-Q plot both showing deviation from normal distribution.

But it is not necessary to have a normal distribution to apply linear regression if the dataset is large because the Central Limit Theorem states that the regression coefficients will be normally distributed for large enough samples, regardless of the distribution of the response variable. This is because the regression coefficients are weighted averages of the response variable values.[35]

3.5 Regression combined with PCA.

3.5.1 Dimensionality Reduction and Feature Selection

Predictive models might encounter issues including overfitting, increased computing cost, and diminished interpretability when working with datasets that have a lot of features or dimensions. Prior to executing these models, dimensionality reduction and feature selection approaches are frequently used to address these problems.

The term "curse of dimensionality" describes the situation in high-dimensional spaces where the amount of data grows exponentially as the number of dimensions rises. Data becomes sparse as a result, which makes it difficult to identify significant patterns and linkages. This problem can be addressed in turn boosting the model's efficiency by lowering the number of dimensions.

Finding the most pertinent subset of features from the initial feature space is the goal of feature selection. Along with reducing dimensionality, this procedure also makes data more interpretable and less susceptible to overfitting. Typical methods for feature selection include:

a. Filter Methods: Filter methods place features in a certain order depending on statistical characteristics like mutual information or correlation with the target variable. Independent of the selected learning algorithm, features are chosen according to predefined criteria. Chi-square, information gain, and correlation-based feature selection are a few examples of filtering techniques.

The reduced feature set is fed into the regression model when the dimensionality reduction or feature selection phase is finished. The model can concentrate on the most useful variables by removing irrelevant or redundant features, which will enhance prediction accuracy, decrease overfitting, and improve interpretability.

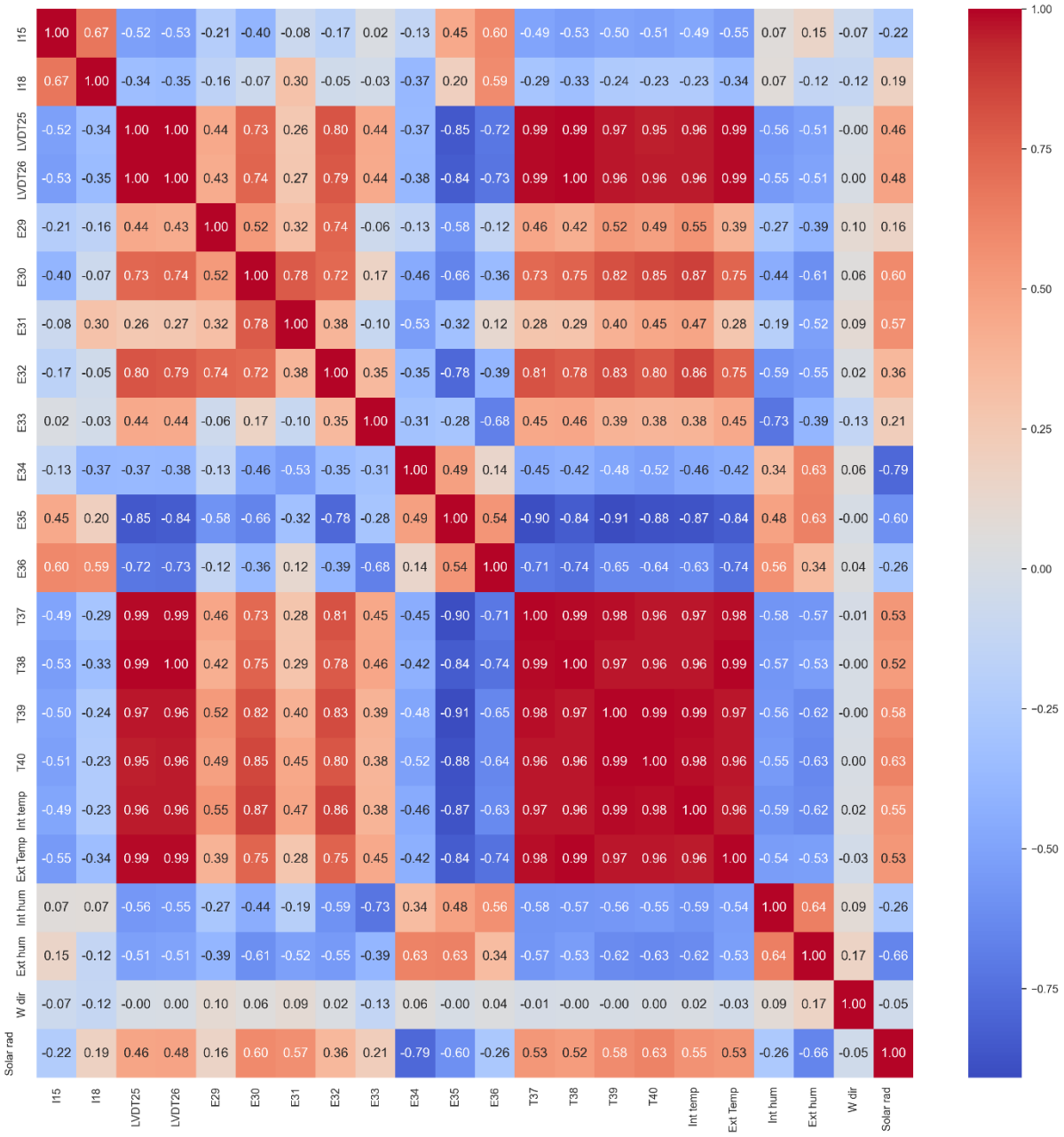


Figure 3.12: Correlation matrix

Observed from the correlation matrix, 'W dir' sensor data can be straightaway neglected since it shows very less correlation with all the sensors. Also, a threshold of |0.3| can be set and any correlation below this threshold can be neglected.

So, these are the dependent and independent features considered:

Dependent Variables	*Independent Variables*
I15, I18	Ext Temp, T38
LVDT25, LVDT26	T38, Int hum, solar rad
E30, E31	Int temp, Ext hum, solar rad
E34, E35	T40, Ext hum, sola rad
E33, E36	Ext Temp, T38, Int hum
E29, E32	Int temp, T39, Ext hum

Table 2: Features corresponding to the predictor variables.

3.5.2 Selection of best regression model

The selection of the best regression model plays an important part in accurately estimating the true structural response by effectively removing environmental effects. This section discusses the methodology employed for choosing the most suitable regression model and presents an evaluation framework based on the root mean square error (RMSE) metric.

RMSE is preferred over other performance metrics like mean absolute error (MAE), mean square error (MSE), mean absolute percentage error (MAPE) for several reasons:

Sensitivity to Deviations: The average deviation between the anticipated and actual values is measured by the RMSE. It considers both the magnitude and the direction of the errors. The RMSE emphasizes greater errors by taking into account the squared differences, which makes it sensitive to significant variations between the expected and actual values. The accuracy of the regression models in capturing the relation between environmental effects and bridge measurement data must be evaluated in light of their sensitivity.

Familiarity and Interpretability.

Application to Outliers: In real-world situations, it is usual to run into outliers or extreme data values that can have a big impact on how well regression models work. By taking into account the squared errors, RMSE gives these outliers more weight,

thereby capturing their influence on performance as a whole. As a result, RMSE is robust and trustworthy when analyzing data that contains unusual or extreme observations.

Comparison Among Models: RMSE makes it easy to compare and rank several regression models. It is simple to identify the model that, by successfully removing temperature influences, produces the most accurate estimations of the underlying structural response since lower RMSE values imply superior performance. Due to its comparability, RMSE can be used to choose the optimal regression model when attempting to eliminate external influences from bridge monitoring data.

First, a wide range of models are taken into account. These models include popular and well-known machine learning methods such as linear regression, KNN regression, random forest regression, dynamic regressions, and also deep learning methods like LSTM. A Kalman forecaster is also taken into consideration. A dataset that includes environmental measurements and bridge response measurements gathered from the bridge monitoring system is used to train each model.

The selection of independent and dependent features is already discussed in section 3.5.1.

Then the dependent variables are used to predict the target variables. And RMSE is calculated for each bridge response sensor using every regression model. A median RMSE is then calculated for each regression model.

This median RMSE is used to assess the models' performance. The average difference between the true and predicted values is quantified by RMSE, which provides a measurement of prediction accuracy. A model that performs better at capturing the link between environmental data and bridge response data has a lower median RMSE value. One can determine which regression model performs best at identifying the true structural response after accounting for temperature influences by comparing the median RMSE values generated from each model.

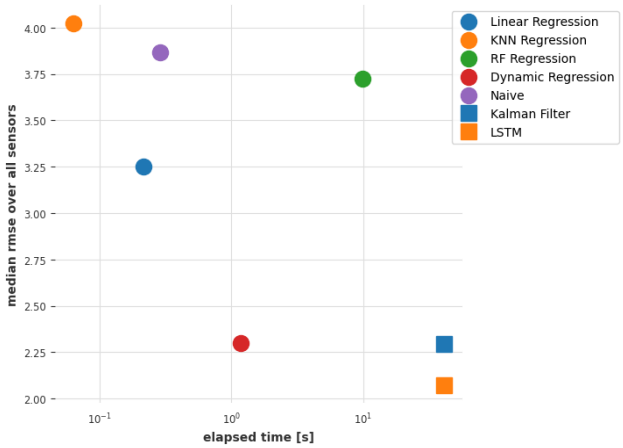


Figure 3.13: Comparison of Model Performance and Efficiency.

Figure 3.13 shows median RMSE versus time taken to fit and predict each model. This plot shows the medians of model performance and their respective running times for various regression models, aiding in model selection.

This shows that the Dynamic regression (●) considering time lag of 24 hours has the optimum performance.

The evaluation of forecast accuracy might not, however, give a complete picture of how well the model eliminates environmental changes. Evaluation of the residual standard deviation (RSD) is crucial to ensuring the removal of environmental effects.

The low RSD indicates that the dispersion of the residuals, which represent the remaining variation in the bridge response data after removing the environmental effects, is minimal. A reduced spread in the residuals suggests that the environmental-related variations have been successfully eliminated, as the residuals primarily capture the remaining sources of variation unrelated to environment. This observation reinforces the notion that the selected regression model is capable of effectively isolating and removing the environmental effects, resulting in a mechanical dataset that is more representative of the true structural response.

Model→	Linear	KNN	RandomForest	Dynamic	NaiveSeasonal	KalmanForecaster	LSTM
I15	0.018	0.025	0.026	0.022	0.015	0.017	0.017
I18	0.022	0.028	0.029	0.013	0.021	0.020	0.020
LVDT25	0.449	1.595	0.497	0.385	2.535	0.441	0.449
LVDT26	0.323	1.574	0.365	0.292	2.382	0.347	0.344
E30	1.655	2.902	2.013	1.801	3.457	1.708	1.718
E31	2.368	2.827	2.718	2.230	3.510	2.179	2.190
E34	2.999	3.354	3.978	2.162	5.782	3.489	3.502
E35	4.100	5.639	4.529	3.276	7.733	4.283	4.330
E33	9.205	12.331	11.164	3.405	5.908	3.262	3.080
E36	14.205	19.552	19.163	2.059	3.635	1.610	1.510
E29	5.885	7.753	7.117	3.401	8.366	4.740	4.739
E32	3.716	5.871	4.405	2.632	7.284	3.858	3.885

Table 3: Residual standard deviation (RSD) of each bridge response sensor for different models.

Table 3 shows that the RSD of mostly all sensors for Dynamic regression considering time lag of 24 hours has the lowest values which suggests this method is best in effectively removing the environmental effects.

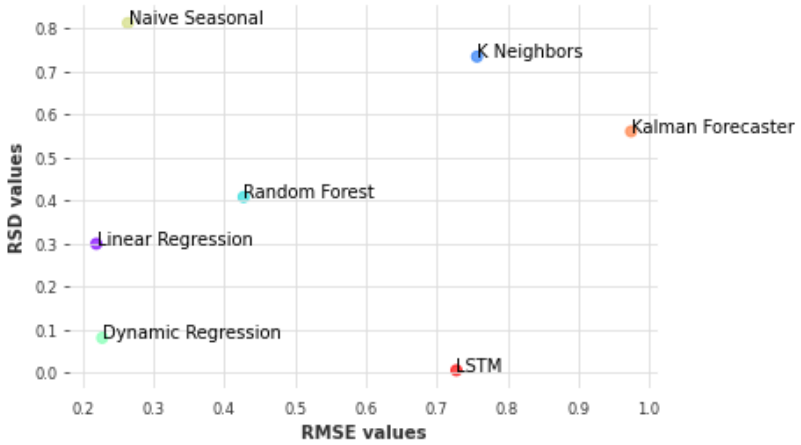


Figure 3.14: Scatter plot median RMSE vs median RSD

From Figure 3.14 it can be inferred that the selected regression model not only accurately estimates the true structural response but also successfully removes the temperature effects from the bridge monitoring data. This finding reinforces the validity and reliability of the chosen model in mitigating the influence of temperature variations, ultimately leading to improved accuracy in assessing the structural health of bridges.

3.5.3 Selection of Principal component parameters

Another machine learning method that can be utilized to remove the environmental effects from the bridge monitoring data is by removing the principal component which shows high variability in the data due to environmental effects.

The processed dataset considering only bridge response sensors are scaled and then a PCA transform is applied to it, which transforms original data from original subspace to Principal subspace.

Theoretically, the first principal component should explain the maximum variation due to environmental effects in the dataset, which can be proved by plotting a correlation matrix of all principal components and environmental data as shown in Figure 3.15.

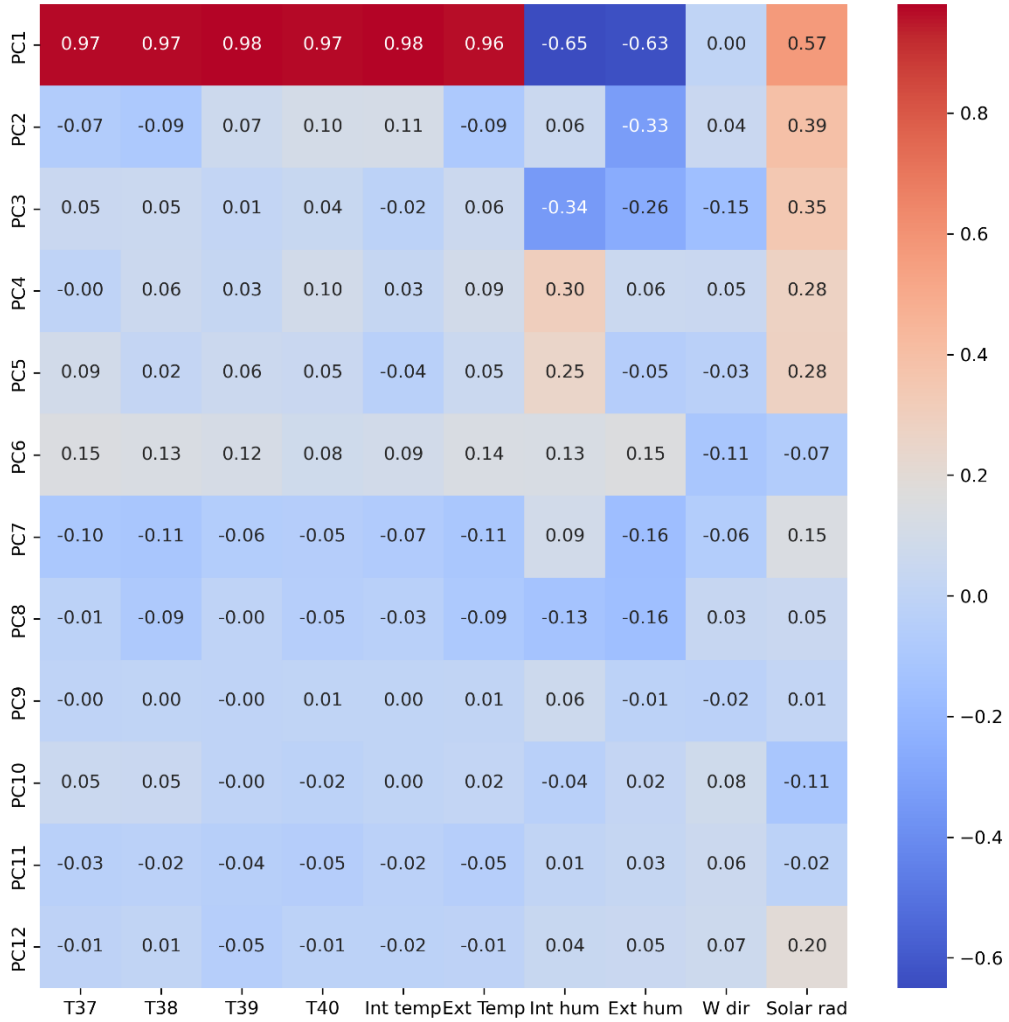


Figure 3.15: Correlation matrix for different principal components vs Environmental sensors

Figure 3.15 shows that the 1st principal component has highest correlation with the environmental data.

By retaining this principal component and inverse transforming the 1st principal component from Principal subspace to original subspace, gives projected data which only has environmental variations. The residuals are calculated by subtracting the Projected data with actual data. Thus, these residuals will only have variations other than environmental effects.

The RSD of each sensor using PCA is given in Table 4 below, and it is clear after comparing this with Table 3 that the dynamic linear regression performs better.

	PCA
I15	0.017
I18	0.024
LVDT25	1.061
LVDT26	1.105
E29	5.904
E30	1.705
E31	2.881
E32	3.701
E33	11.543
E34	4.838
E35	4.071
E36	19.959

Table 4: Residual standard deviation of sensors for only PCA

3.5.4 Applying PCA to regression residuals

It is critical to distinguish between unmeasured variations arising from factors like sensor noise or environmental conditions and variations linked to structural damage when examining residuals obtained after subtracting temperature effects from bridge monitoring data. The Principal Component Analysis (PCA) transform is used to further clean up the residuals and pinpoint the changes linked to structural damage. So, the intention is to remove unmeasured variations while keeping the changes associated with structural damage by identifying the dominating component by PCA and subtracting it from the residuals.

There are various benefits of using PCA-based residual analysis. First off, it makes it possible to find and remove unmeasured variations that could otherwise make residual analysis interpretation difficult. The improved residuals concentrate entirely on the differences attributable to structural damage by eliminating these unmeasured variations, improving the accuracy and clarity of the study.

In this approach, PCA is applied to the residuals obtained after removing temperature effects using the selected regression model. By performing PCA on the residuals, the principal components that capture the majority of the variations are extracted.

One can evaluate the explained variance ratio of each major component in order to determine which one is the most significant. The percentage of the overall variance in the residuals that is quantified by each primary component is known as the explained variance ratio. Priority is given to the principal component that explains the most significant variations in the residuals by selecting the principal component with the highest explained variance ratio, in this case the 1st principal component as shown in Figure 3.16.

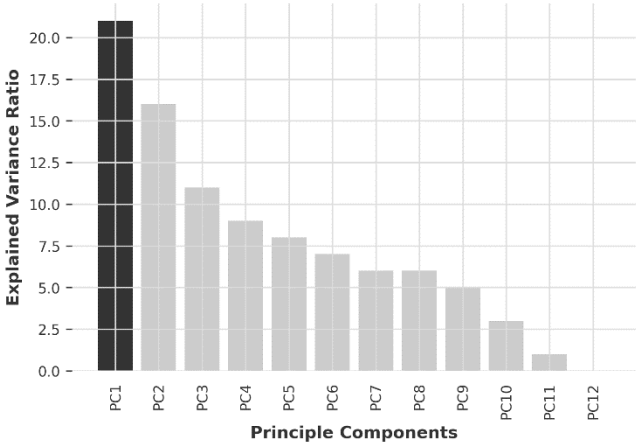


Figure 3.16: Scree plot showing explained variance ratio for different principal components.

Inverse transform is used to generate the projected data corresponding to the dominating principal component after identifying it. The projected data is rebuilt using this inverse transform into the original feature space. The changes related to the dominant component are then effectively eliminated by subtracting this projected data from the initial residuals.

To evaluate the effectiveness of the proposed PCA-based method, a comparison is made between the refined residuals and the original bridge response sensor variations. This can be done visually by plotting the refined residuals against the original bridge response sensor variations as shown in Figure 3.17.

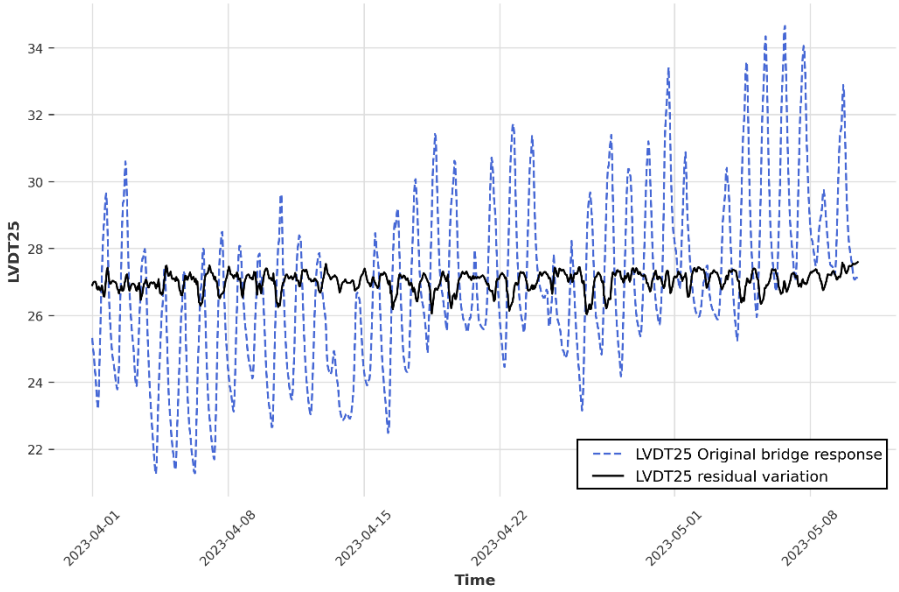


Figure 3.17: Variation of residuals and original bridge response for sensor LVDT25

3.6 PCA combined with regression.

Applying PCA to the data from the bridge response sensor is the first stage in this methodology. PCA enables to capture the most important variables while removing noise and redundant data from the dataset by portraying the data in a lower-dimensional space. By using PCA, a set of orthogonal principal components are generated that are ordered according to how much variance in the bridge response data they contribute overall.

The correlation coefficients between each PC and the environmental data are determined same as done before in section 3.5.3 in order to choose the PC that is most consistent with the environmental sensors. The PC with the highest correlation as shown in Figure 3.15 is picked as the main part for the analysis to come. The bridge response fluctuations that are significantly impacted by environmental variations are anticipated to be captured by this PC.

Next, the best regression model is identified using Figure 3.18(b) which suggests that dynamic regression with 24-time step lag gives the optimum performance. The methodology to select the optimum regression model is already discussed above. The only difference is now the dependent variable is the selected principal component. Different regression models are trained with the selected principal component as the target variable and the environmental sensors as the feature variables. The RMSE is used as the performance metrics to capture the

environmental effects and RSD as shown in Figure 3.18(a) is used to evaluate the effectiveness of the model to remove the environmental effects.

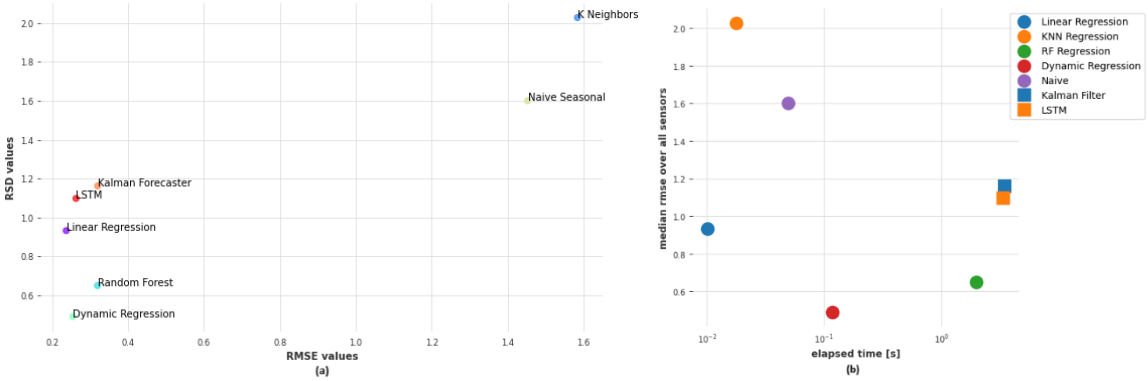


Figure 3.18: (a) RMSE vs RSD scatter plot; (b) Performance vs efficiency plot

After selecting the dynamic regression model, the residuals are calculated by subtracting the predicted values from the actual values of the selected PC. The bridge response data that cannot be fully explained by environmental changes is represented by these residuals. The influences of environment have been eliminated using the regression modeling, therefore the residuals obtained through this procedure mostly represent the fluctuations related to structural degradation.

To convert the residuals back to the original subspace, an inverse PCA transform is applied. This inverse transform reconstructs the residuals into the original feature space, allowing to obtain the refined residuals that reflect the variations attributed to structural damage. These refined residuals provide valuable information for damage detection and analysis.

Refined residuals can be analyzed using various statistical analysis techniques, visualization methods as shown in Figure 3.19. The refined residuals enable the detection and interpretation of damage indicators within the bridge monitoring data. By focusing on the variations associated with structural damage and removing the confounding effects of temperature, the proposed PCA-based method enhances the accuracy and reliability of damage detection in bridge monitoring systems.

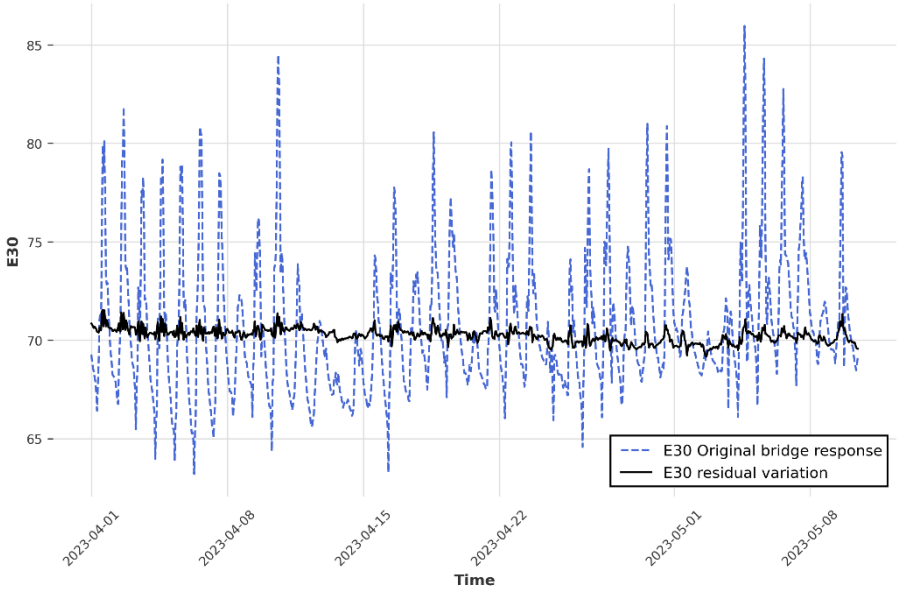


Figure 3.19: Variation of residuals and original bridge response for sensor E30

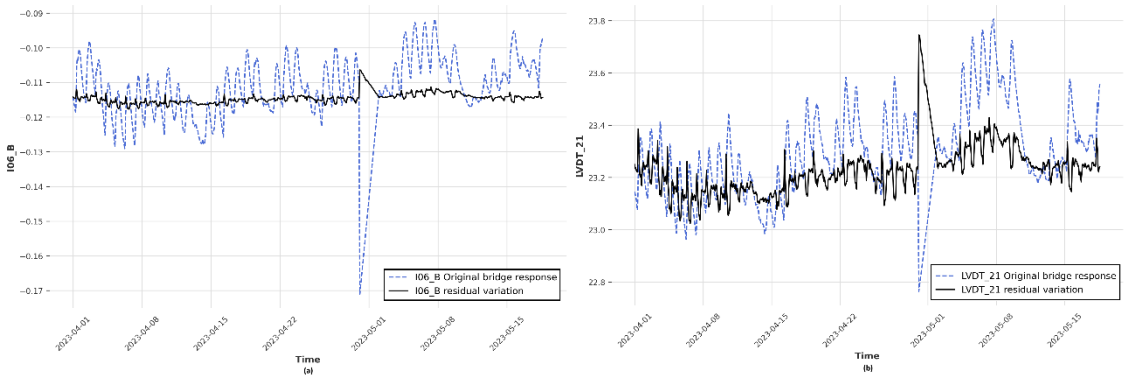
	REG→PCA	PCA→REG
I15	0.020	0.003
I18	0.016	0.001
LVDT25	0.278	0.429
LVDT26	0.261	0.411
E29	3.179	0.431
E30	1.272	0.356
E31	1.515	0.129
E32	2.561	0.648
E33	4.832	1.068
E34	2.138	0.244
E35	2.246	0.687
E36	1.790	2.539

Table 5: A comparison of two methodologies

As from the Table 5 the first column contains RSD of each bridge response sensor for the methodology described in section 3.5, whereas the second column contains RSD of the methodologies described in section 3.6. It is clear that the method where regression model is applied to the principal components performs better.

Therefore, this methodology will be used to validate simulated damage scenario.

The trained hybrid model when used on the sensor data from another railway bridge "Piave bridge" was also able to remove the environmental effects, this can be shown in figure below.



3.7 Introduction of simulated damage

The design and appropriate tuning of the structural health monitoring system for damage detection and identification are based on a simulation of the bridge's structural response under operating circumstances, taking into account the occurrence of various forms of structural damage. Notional damage scenarios with varying damage magnitude and extent over the structure, in particular, have been investigated in order to capture the role of the various bridge components and characterize the order of magnitude of the expected variations of structural response indicators, such as displacements and rotations as measured by sensors, as well as natural frequencies and suitable norms of modal eigenvectors.[32]

A virtual damage is induced to all of the mechanical sensors in the bridge monitoring system in order to assess the effectiveness of the Principal Component Analysis (PCA) and regression modeling combination that has been suggested for environmental effects removal. One may evaluate if the refined residuals obtained by selected method predominantly reflect the fluctuations caused by structural damage while successfully reducing the impacts of environmental variations thanks to this simulated damage.

3.7.1 Damage induction

The virtual introduction of damage involves altering the mechanical sensor measurements in a way that mimics the existence of structural damage. Different methods, such as changing strain values, sensor readings, or the relationships between sensor data, can be used to introduce this simulated damage.

The investigated damage scenarios involve corrosion of single or multiple structural elements based on various corrosion propagation assumptions. In reality, corrosion may cause a considerable loss in load-bearing capacity as well as excessive displacements and rotations during the life of a bridge. Corrosion has been modelled

in this study by lowering the thickness of structural parts while considering varying levels of corrosion penetration. The notional damage scenarios have also taken into account possible member failures owing to local instability, as well as the potential collapse of bridge components and connections. Furthermore, the impacts of locking the support devices have been investigated in order to account for probable gradual degradation of the support devices owing to aging and deterioration. Finally, probable pile abutment settlements were investigated by employing a set of forced displacements at supports.[32]

One possible scenario of the simulated damage that is used in the present work is shown in the Table 6 below.

Livenza Bridge Response Sensors			Intact Structure	SCENARIO #14: Bottom, Top, Diagonal damage (δ=10%)	
Measurements	**Sensor ID**	**Model ID**	**Value (δ=0)**	**Value**	**Δ**
Rotation[deg]	I15	1	-4.67×10^{-2}	-4.81×10^{-2}	-1.41×10^{-3}
	I18	56	4.67×10^{-2}	4.81×10^{-2}	-1.41×10^{-3}
Longitudinal Displacement [mm]	LVDT25	28	4.64	4.80	0.16
	LVDT26	56	4.64	4.80	0.16
Strain (Delta in microepsilon)	E29	1	-609.38	-581.94	9.79
	E31	1	-609.38	-581.94	9.79
	E30	41	-609.28	-581.84	9.79
	E32	41	-609.28	-581.84	9.79
	E33	33	974.16	918.67	-10.12
	E35	33	974.16	918.67	-10.12
	E34	73	974.15	918.65	-10.12
	E36	73	974.15	918.65	-10.12

Table 6: Simulated Damage scenario

A new dataset is generated by addition of these recognized virtual damage delta to the original bridge response values, which may be utilized for validation and comparison.

3.7.2 Application of Model

The previously learned PCA and regression model from Section3.6 is reapplied to this new dataset after the virtual damage dataset has been generated. On the data from the damaged bridge response sensors, the PCA step is used to identify the predominant modes of variations. The chosen principal component, which shows the best correlation with the environmental sensors, is then subjected to the regression model, which was trained on the original, undamaged data. The transformed damaged dataset is then projected back into original subspace, in a manner similar to the method employed for the undamaged data. These steps enables to anticipate and eliminate the environmental impacts from the damaged data.

The final phase compares the original undamaged residuals with the damaged residuals derived from the damaged dataset. While the effects of the environment have been successfully eliminated, the changes caused by the simulated damage should primarily be reflected in the damaged residuals. One may evaluate the effectiveness of this method for identifying the differences related to structural damage by contrasting the damaged residuals with the undamaged residuals.

The plot below in Figure 3.20 shows the time series of original bridge response vs Damaged Bridge response without any processing of the data. As can be seen from the plot, the damage in the data cannot be seen.

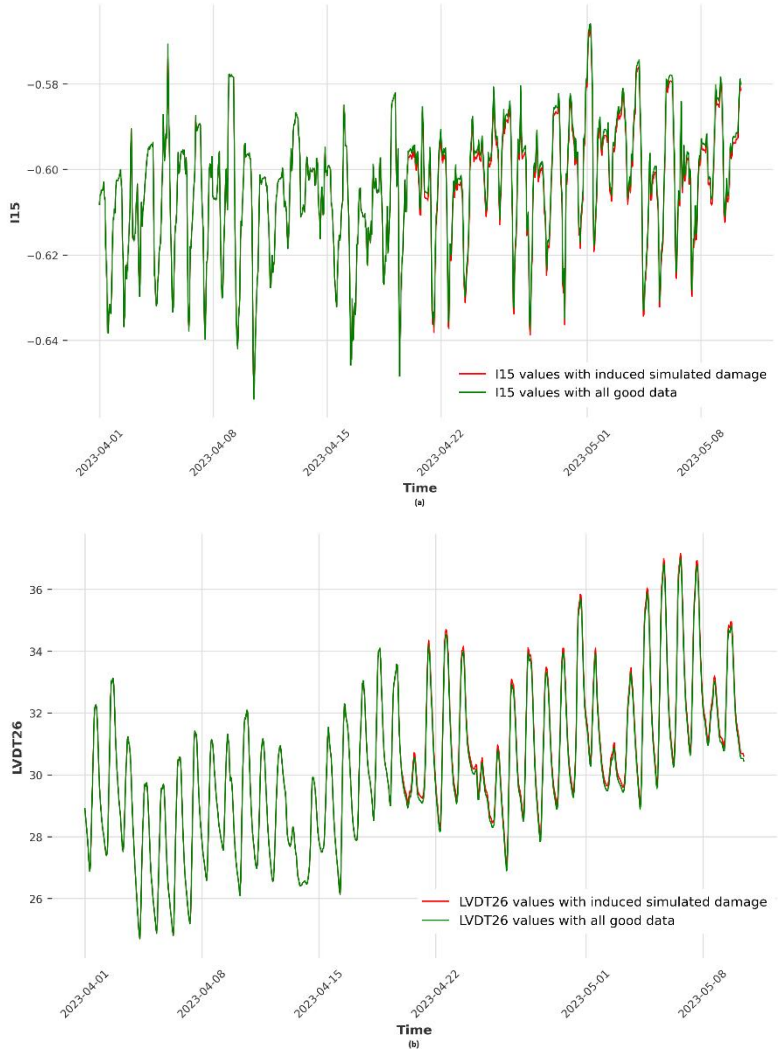


Figure 3.20: Unprocessed I15 and LVDT26 sensor data time series plot.

Whereas, after processing the data through the PCA+Regression model described in previous section, the damage induced can be clearly identified as shown in the Figure 3.21.

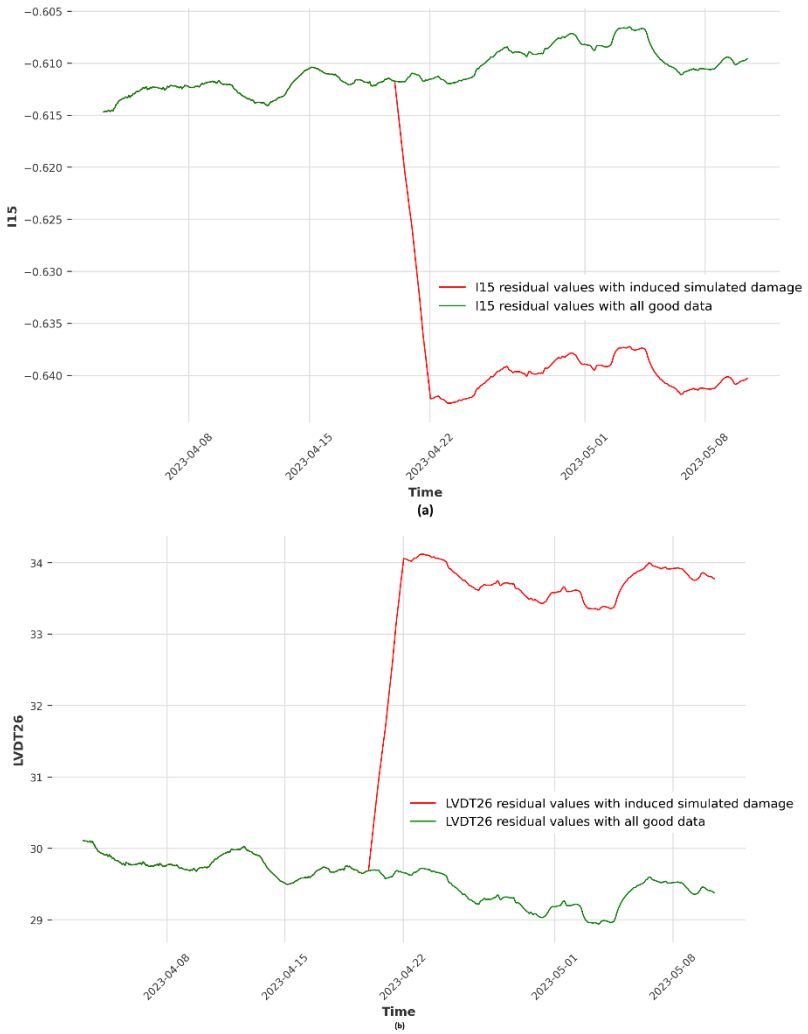


Figure 3.21: Processed I15 and LVDT26 sensor data time series plot

The ability of the refined residuals to primarily capture the variations due to induced damage while effectively removing temperature variations would demonstrate the success of this work. This outcome validates that the proposed combination of PCA and regression modeling has successfully removed the confounding effects of temperature and retained the variations linked to structural damage. The detection of damage patterns within the damaged residuals further confirms the accuracy and reliability of this method in damage identification and assessment.

4. Conclusions and Future Developments

This thesis work sought to address the major objective of reducing the effects of environmental and operational variability on structural health monitoring (SHM) systems, applying proposed methods to the real-time full-scale data of the Livenza Railway Bridge, and validate the performance of model using a damaged simulated data.

The methodology presented in Section.3.6 has been applied during this thesis. Initially, the Livenza Railway Bridge SHM system was described, and its susceptibility to environmental factors like temperature and operational factors like train loading was established.

Even though the focus was on a specific bridge case study, the methodology generated to validate the machine learning models aimed to be applicable to SHM systems in general.

Concerning the application of machine learning models to reduce environmental and operational variability:

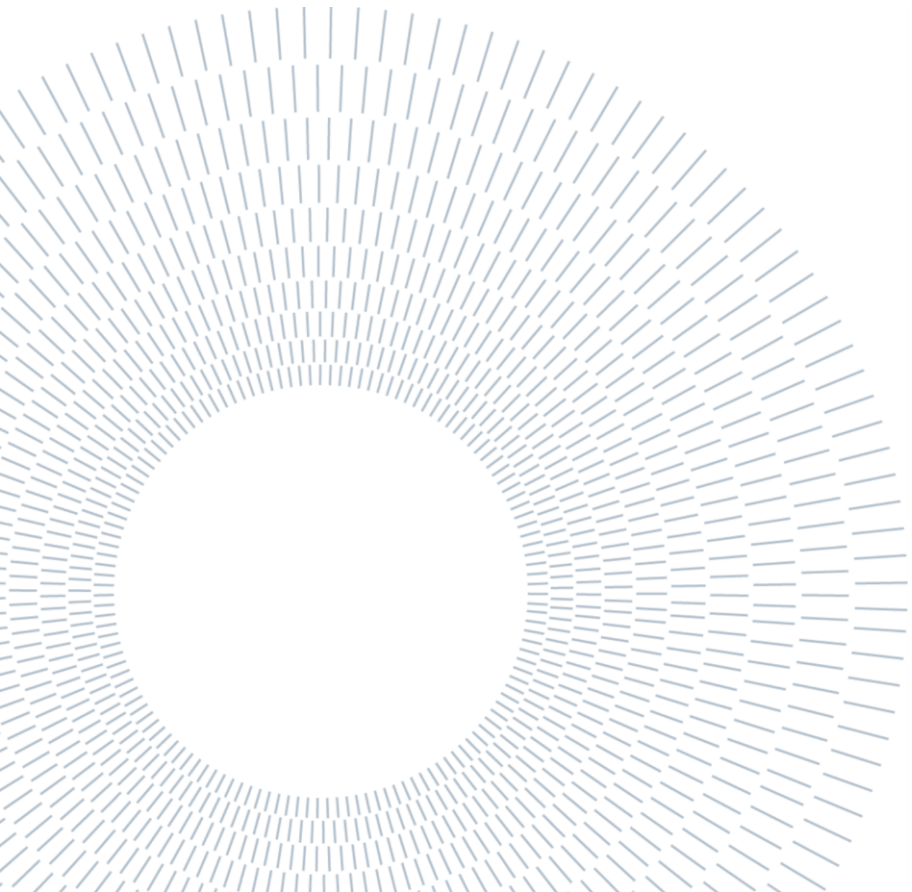
- Various visualization and statistical methods confirmed strong correlations between environmental factors like temperature and bridge response sensor data. Also, these techniques confirmed presence of operational effects. This highlighted the need for normalizing the data before applying damage detection algorithms.
- The influence of train loading on bridge response sensors were effectively reduced by utilizing an outlier removal method described in Section3.4.2.
- Different machine learning models were explored including static regression, dynamic regression, principal component analysis (PCA), and combinations thereof. Performance metrics showed dynamic regression with a 24-hour time lag to be optimal in predicting and removing temperature effects.

- Applying PCA on the regression residuals further filtered out unmeasured variations not accounted by the regression model. This integrated PCA-regression approach reliably isolated indicators of structural damage from environmental impacts.
- Simulated damage scenarios validated the effectiveness of the proposed methodology. While raw damaged data showed no visible indicators, the integrated model successfully extracted residuals reflecting the simulated damage patterns.

The computational framework presented enables distinguishing between changes in bridge response due to external factors and underlying structural degradation. This allows more accurate SHM analysis by minimizing environmental variability. While demonstrated on bridge data, the machine learning methods have the potential to be extended to other SHM applications where environmental or operational conditions may mask damage.

Several promising avenues exist to build upon the work presented in this thesis. Optimizing the LSTM model hyperparameters like number of layers, nodes, and regularization could further improve its performance. Additionally, more advanced system identification techniques like SINDy could derive an enhanced physics-based mathematical model tailored specifically to the dynamics of this bridge structure. Hybrid models that integrate machine learning techniques with physics-based principles also present a promising direction for further improving and isolating the environmental effects.

For detecting damage from the normalized data, unsupervised learning methods like autoencoders could be investigated.



Bibliography

[1] H. Zhang, J. Guo, X. Xie, R. Bie, and Y. Sun, "Environmental effect removal based structural health monitoring in the internet of things," *Proceedings - 7th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, IMIS 2013*, pp. 512–517, 2013, doi: 10.1109/IMIS.2013.91.

[2] J. C. A. Jauregui Correa and A. A. Lozano Guzman, "Condition monitoring," in *Mechanical Vibrations and Condition Monitoring*, Elsevier, 2020, pp. 147–168. doi: 10.1016/B978-0-12-819796-7.00008-1.

[3] TWI ltd, "Non-Destructive Testing | TWI." https://www.twi-global.com/technical-knowledge/faqs/what-is-non-destructive-testing#:~:text=Non%2Ddestructive%20testing%20is%20a,particle%20testing%20and%20penetrant%20testing. (accessed Sep. 03, 2023).

[4] C. R. Farrar and N. A. J. Lieven, "Damage prognosis: the future of structural health monitoring," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 365, no. 1851, pp. 623–632, Feb. 2007, doi: 10.1098/rsta.2006.1927.

[5] M. L. FUGATE, H. SOHN, and C. R. FARRAR, "VIBRATION-BASED DAMAGE DETECTION USING STATISTICAL PROCESS CONTROL," *Mech Syst Signal Process*, vol. 15, no. 4, pp. 707–721, Jul. 2001, doi: 10.1006/mssp.2000.1323.

[6] M. Martinez-Luengo and M. Shafiee, "Guidelines and Cost-Benefit Analysis of the Structural Health Monitoring Implementation in Offshore Wind Turbine Support Structures," *Energies (Basel)*, vol. 12, no. 6, p. 1176, Mar. 2019, doi: 10.3390/en12061176.

[7] E. Figueiredo, G. Park, C. R. Farrar, K. Worden, and J. Figueiras, "Machine learning algorithms for damage detection under operational and environmental variability," *Struct Health Monit*, vol. 10, no. 6, pp. 559–572, Nov. 2011, doi: 10.1177/1475921710388971.

[8] H. Sohn, "Effects of environmental and operational variability on structural health monitoring," *Philosophical Transactions of the Royal Society A:*

Mathematical, Physical and Engineering Sciences, vol. 365, no. 1851, pp. 539–560, Feb. 2007, doi: 10.1098/rsta.2006.1935.

[9] E. J. Cross, G. Manson, K. Worden, and S. G. Pierce, "Features for damage detection with insensitivity to environmental and operational variations," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 468, no. 2148, pp. 4098–4122, Dec. 2012, doi: 10.1098/rspa.2012.0031.

[10] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis (4th ed.)*. Wiley & Sons, 2006.

[11] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*, 3rd ed. Melbourne, Australia, 2021. Accessed: Sep. 03, 2023. [Online]. Available: OTexts.com/fpp3

[12] "Q-Q Plot (Quantile to Quantile Plot)," in *The Concise Encyclopedia of Statistics*, New York, NY: Springer New York, pp. 437–439. doi: 10.1007/978-0-387-32833-1_331.

[13] S. S. Shapiro and M. B. Wilk, "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika*, vol. 52, no. 3/4, p. 591, Dec. 1965, doi: 10.2307/2333709.

[14] IBM, "k-nearest neighbors | IBM," Feb. 2022. https://www.ibm.com/topics/knn (accessed Sep. 03, 2023).

[15] S. Sharma, "K-Nearest Neighbour: The Distance-Based Machine Learning Algorithm.," *Analytics Vidhya*, May 15, 2021. https://www.analyticsvidhya.com/blog/2021/05/knn-the-distance-based-machine-learning-algorithm/ (accessed Sep. 16, 2023).

[16] C. Aldrich, "Process Variable Importance Analysis by Use of Random Forests in a Shapley Regression Framework," *Minerals*, vol. 10, no. 5, p. 420, May 2020, doi: 10.3390/min10050420.

[17] A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network," *Physica D*, vol. 404, p. 132306, Mar. 2020, doi: 10.1016/j.physd.2019.132306.

[18] T. Bikmukhametov, "How to reshape data and do regression for time series using LSTM," *Towards data science*, Apr. 12, 2020. https://towardsdatascience.com/how-to-reshape-data-and-do-regression-for-time-series-using-lstm-133dad96cd00 (accessed Sep. 04, 2023).

[19] P. Van Overschee and B. De Moor, "N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems," *Automatica*, vol. 30, no. 1, pp. 75–93, Jan. 1994, doi: 10.1016/0005-1098(94)90230-5.

[20] R. Sadli, "Object Tracking: Simple Implementation of Kalman Filter in Python," *Machine Learning Space*, Feb. 15, 2020. https://machinelearningspace.com/object-tracking-python/ (accessed Sep. 16, 2023).

[21] A. M. Yan, G. Kerschen, P. De Boe, and J. C. Golinval, "Structural damage diagnosis under varying environmental conditions - Part I: A linear analysis," *Mech Syst Signal Process*, vol. 19, no. 4, pp. 847–864, Jul. 2005, doi: 10.1016/j.ymssp.2004.12.002.

[22] Milos Gajdos, "Principal Component Analysis - Part 2," *Cybernetist*, Apr. 20, 2016. https://cybernetist.com/2016/04/12/principal-component-analysis-part-2/ (accessed Sep. 04, 2023).

[23] J. Herzen *et al.*, "Darts: User-Friendly Modern Machine Learning for Time Series," *Journal of Machine Learning Research*, vol. 23, no. 124, pp. 1–6, 2022, [Online]. Available: http://jmlr.org/papers/v23/21-1177.html

[24] S. ALAMPALLI, "EFFECTS OF TESTING, ANALYSIS, DAMAGE, AND ENVIRONMENT ON MODAL PARAMETERS," *Mech Syst Signal Process*, vol. 14, no. 1, pp. 63–74, Jan. 2000, doi: 10.1006/mssp.1999.1271.

[25] K. WORDEN, G. MANSON, and N. R. J. FIELLER, "DAMAGE DETECTION USING OUTLIER ANALYSIS," *J Sound Vib*, vol. 229, no. 3, pp. 647–667, Jan. 2000, doi: 10.1006/jsvi.1999.2514.

[26] B. Peeters, J. Maeck, and G. De Roeck, "Vibration-based damage detection in civil engineering: excitation sources and temperature effects," *Smart Mater Struct*, vol. 10, no. 3, pp. 518–527, Jun. 2001, doi: 10.1088/0964-1726/10/3/314.

[27] B. Peeters and G. De Roeck, "One-year monitoring of the Z24-bridge: Environmental effectsversus damage events," *Earthq Eng Struct Dyn*, vol. 30, no. 2, pp. 149–171, 2001, doi: 10.1002/1096-9845(200102)30:2<149::AID-EQE1>3.0.CO;2-Z.

[28] A. Deraemaeker and K. Worden, "A comparison of linear approaches to filter out environmental effects in structural health monitoring," *Mech Syst Signal Process*, vol. 105, pp. 1–15, May 2018, doi: 10.1016/j.ymssp.2017.11.045.

[29] H. Sohn, M. Dzwonczyk, E. G. Straser, A. S. Kiremidjian, K. H. Law, and T. Meng, "An experimental study of temperature effect on modal parameters of the Alamosa Canyon Bridge," *Earthq Eng Struct Dyn*, vol. 28, no. 7–8, pp. 879–897, 1999, doi: 10.1002/(sici)1096-9845(199908)28:8<879::aid-eqe845>3.0.co;2-v.

[30] G. Comanducci, F. Magalhães, F. Ubertini, and Á. Cunha, "On vibration-based damage detection by multivariate statistical techniques: Application to a long-span arch bridge," *Struct Health Monit*, vol. 15, no. 5, pp. 505–524, Sep. 2016, doi: 10.1177/1475921716650630.

[31] F. Magalhães, A. Cunha, and E. Caetano, "Vibration based structural health monitoring of an arch bridge: From automated OMA to damage detection," *Mech Syst Signal Process*, vol. 28, pp. 212–228, Apr. 2012, doi: 10.1016/j.ymssp.2011.06.011.

[32] F. M. Bono *et al.*, "Low-frequency Structural Health Monitoring analysis combining numerical simulations with experimental measurements from continuous monitoring of a steel truss railway bridge."

[33] S. Seo, "A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets," 2006. [Online]. Available: https://api.semanticscholar.org/CorpusID:33756189

[34] "Winsorizing," *Wikipedia*. https://en.wikipedia.org/wiki/Winsorizing#:~:text=Winsorizing%20or%20winsorization%20is%20the,as%20clipping%20in%20signal%20processing. (accessed Sep. 04, 2023).

[35] T. Lumley, P. Diehr, S. Emerson, and L. Chen, "The Importance of the Normality Assumption in Large Public Health Data Sets," *Annu Rev Public Health*, vol. 23, no. 1, pp. 151–169, May 2002, doi: 10.1146/annurev.publhealth.23.100901.140546.

List of Figures

List of Tables