Missing data estimation method for durability survey of reinforced concrete structures

Luis F Rincon, Bassel Habeeb, Elsa Eustaquio, Ameur El Amine Hamami, José Campos E Matos, Yina M Moscoso, Emilio Bastidas-Arteaga

Please cite this paper as: Rincon, L. F., Habeeb, B., Eustaquio, E., Hamami, A. E. A., Campos e Matos, J., Moscoso, Y. M., & Bastidas-Arteaga, E. (2025). Missing data estimation method for durability survey of reinforced concrete structures. *Structural Health Monitoring*, 14759217241303656. https://doi.org/10.1177/14759217241303656

Missing Data Estimation Method for Durability Survey of Reinforced Concrete Structures

Luis F. Rincon[1,2], Bassel Habeeb[2], Elsa Eustaquio[3], Ameur Hamami[2], José Campos e Matos[1], Yina M. Moscoso[1], Emilio Bastidas-Arteaga[2]

[1] University of Minho, ISISE, ARISE, Department of Civil Engineering, Guimarães, Portugal

[2] Laboratory of Engineering Sciences for the Environment (LaSIE - UMR CNRS 7356), La Rochelle University, La Rochelle, France

[3] Laboratório Nacional de Engenharia Civil, Lisboa, Portugal

Corresponding author: Luis F. Rincon, Email: luis.rincon_prada@univ-lr.fr

Abstract

Reinforced concrete structures are well-known for their high durability, however, they remain vulnerable to natural hazards and extreme events that can impact their performance over time. In aggressive environments, there is a high likelihood of increased maintenance, rehabilitation, and repair actions that constitute a significant portion of the total lifecycle spending. Monitoring systems have been implemented during the last decades to collect periodically or continuously essential data about the durability performance of the structures in real operation. However, the effectiveness of these systems is impacted by sensor efficacy, influenced in turn by environmental factors, sensor durability, and power outages, leading to intermittent or permanent data gaps. This study proposes a methodology to address the problem of missing data of a Structural Health Monitoring (SHM) system, specifically aiming to provide more accurate and continuous information from concrete resistivity and temperature sensors to support the early detection of corrosion. The proposed methodology was applied to a repaired reinforced concrete structure with over fourteen years of data, where significant gaps in the measurements were present. The approach combines several techniques to fill these gaps: deep machine learning for air temperature, generalized linear models for concrete temperature, and pattern recognition for concrete resistivity. To the best of the authors' knowledge, this is the first time a methodology has been proposed for imputing missing data from resistivity sensors in SHM systems, which are increasingly being implemented. This approach is innovative and offers potential benefits for SHM system managers, providing more information on long-term sensor data that could aid in early corrosion detection and maintenance planning. The application of the proposed methodology to a real case study indicated a successful imputation of 43.4% of missing data although some challenges persist for sensors located in areas characterized by high measurements variability. The code is available at https://github.com/LuisRinconP/Missing-Data-Estimation-Method-for-Durability-Survey-of-Reinforced-Concrete-Structures.

Keywords: Structural health monitoring; Sensors; Missing Data Estimation; Artificial neural network; Generalized linear models; Pattern recognition; Concrete resistivity.

1. Introduction

Reinforced concrete structures (RC) bridges play a crucial role in modern infrastructure, providing vital connections between communities and facilitating the continuous flow of people and goods. The long-term lifespan of these structures is essential to ensure the ongoing functionality of the transportation network. Therefore, the maintenance of these bridges relies on their ability to withstand operational and environmental challenges, thus providing reliable and safe service. Climate conditions and extreme events can affect the goal of maintaining and extending the lifespan of RC bridges[1].

Chloride ingress produced by climate conditions is the main corrosion mechanisms impacting durability of RC structures in coastal areas, where the exposure to saltwater accelerates the corrosion process[2]. Under natural exposure conditions, the rate of corrosion in reinforcing steel varies significantly due to several uncertainties including concrete properties[3]. Therefore, corrosion evolution is a complex phenomenon that initiates internally within the structure and can affect long-term structural safety and reliability without timely detection through inspections[1]. Hence, there has been a growing interest recently in the use of Structural Health Monitoring (SHM) systems in reinforcement concrete structures to gather information about current state of the materials and to detect early corrosion[4,5]. This is because sensors could provide real-time information about the condition of the structure, which can be crucial for making informed decisions about maintenance schedules and repairing techniques.

One of the challenges associated with long-term SHM is ensuring continuous measurements during the service life of the structure. However, some periods could not be monitored due to several factors, such as power outages, sensor malfunctions, data transmission issues, etc. In addition, certain data points also might be missing due to signal noise. Thus, missing data can occur in any experiment, and researchers typically address this issue by either recovering the information or imputing the missing data [6]. The effectiveness of data imputation methods is significantly influenced by the quality and quantity of the available data [7]. Various statistical imputation methods allow for the estimation of missing data, including mean imputation, spatial or temporal correlation, other statistical techniques, and machine learning algorithms [8–11].

Addressing the problem of missing data has been a subject of investigation in various research domains and has recently gained traction in the field of SHM [8,12,13]. Liu et al.[14] worked with accelerometers and presented a multivariate time-series analysis method for infrastructure damage detection, using a state-space embedding approach and singular value decomposition. The proposed approach demonstrates computational efficiency and successful damage identification in validation tests on a linear spring-mass system and a benchmark experimental structure. Wan and Ni [15] presented a methodology for SHM data recovery on temperature and accelerometers sensors using Bayesian multi-task learning with a multi-dimensional Gaussian process prior, efficiently modeling multiple tasks and their interrelations. The proposed approach demonstrates superior performance in reconstructing SHM data compared to traditional Bayesian single-task learning, with a focus on the impact of covariance function selection. Li et al. [8] address the issue of missing time series data in SHM systems, focusing on the calculation of cable force by constructing a matrix of correlations between days and within one day, and employing a probabilistic principal component analysis (PPCA) method to improve data imputation. The results show that fully capturing temporal correlations from measured values enhances imputation accuracy, with PPCA outperforming PCA, particularly in scenarios with continuous missing data, highlighting the potential for improved imputation by considering temporal correlations across dimensions. Niu et al.[16] also focused on cable force data and proposed a spatiotemporal graph attention network for restoring missing data in structural health monitoring systems, focusing on the spatial and temporal dependencies within the sensor network. Jiang et al.[17] proposed a novel data-driven generative adversarial network (GAN) to impute missing strain response

data from wireless sensors in structural health monitoring systems. The method was verified on a real concrete bridge and demonstrated superior imputation accuracy and efficiency by leveraging spatial-temporal relationships among strain sensors without needing a complete dataset during training. More recently, Gao et al [12] presented a slim generative adversarial imputation network (SGAIN) for recovering missing deflection data of SHM systems in a highway-railway dual-purpose bridge. The model used slim neural networks with a generator-discriminator architecture to efficiently impute missing data caused by sensor malfunctions or communication outages. The SGAIN network presented superior performance and execution speed when compared to the conventional GAIN model.

Other types of sensors have been also analyzed. Tang et al.[18] developed a convolutional neural network for recovering multi-channel SHM data with group sparsity awareness, effectively addressing segments of continuous missing data. The method demonstrated strong recovery performance on synthetic, field-test, and seismic response monitoring data. More recently, Luo et al.[19] analyzed the quantification and prediction of pitting corrosion of steel structures in using one-dimensional convolutional neural networks (1D CNN) in conjunction with electromechanical impedance (EMI) sensors. By using an EMI-instrumented circular piezoelectric-metal transducer, it was possible to detect corrosion-induced mass loss. The results showed high accuracy in predicting the extent of pitting corrosion, laying a technical foundation for real-time and quantitative monitoring of corrosion in steel structures.

Table 1. Publications on data imputation methods used in SHM systems.

Research	Type of Sensor/Feature Measured	Imputation Method
Liu et al. (2014)[14]	Acceleration data for Infrastructure damage detection	Multivariate time-series analysis, state-space embedding, Singular Value Decomposition (SVD)
Wan and Ni (2019)[15]	Temperature and acceleration data from Canton Tower	Bayesian multi-task learning with Gaussian process prior
Li et al. (2020)[8]	Cable force data	Probabilistic Principal Component Analysis (PPCA)
Niu et al. (2022)[16]	Cable force data	Spatiotemporal Graph Attention Network
Jiang et al. (2022)[17]	Strain response data from wireless sensors	Generative Adversarial Network (GAN)
Gao et al. (2022)[12]	Deflection data (highway-railway dual-purpose bridge)	Slim Generative Adversarial Imputation Network (SGAIN)
Tang et al. (2021)[18]	Multi-channel SHM data (seismic and synthetic data)	Convolutional Neural Network (CNN) with group sparsity awareness
Luo et al. (2023)[19]	Corrosion detection in steel structures (pitting corrosion) using Electromechanical Impedance (EMI) sensors	One-dimensional Convolutional Neural Networks (1D CNN)

The studies mentioned above have significantly contributed to deal with missing data estimation in SHM systems for different type of sensors (See Table 1). However, to the best of the author's knowledge, no research has been published regarding the imputation or filling of missing data for SHM durability sensors, in particularly, concrete resistivity sensors on reinforcement concrete structures.

Concrete resistivity sensors have proved to be useful for collecting information on chloride contamination and to be durable for long-term monitoring, which is particularly important, resulting in regular installations of sensors for SHM systems [20]. However, several external factors may affect the electrical resistivity of concrete [21,22]. Given the significant influence of temperature on resistivity, the installation of concrete temperature sensors is common when considering concrete electrical resistivity sensors, to account for temperature variations in data analysis [21].

In this study, we introduce a novel approach that focuses specifically on filling missing data for SHM durability sensors, particularly concrete resistivity sensors, which has not been addressed in previous research. The novelty of this research lies in the development of a comprehensive methodology that integrates multiple techniques for imputing missing sensor data, enhancing the reliability of long-term corrosion monitoring and is tested on a SHM system in a reinforced concrete bridge for over ten years period. The article presents a methodology to fill missing data found within the use of resistivity and temperature sensors on SHM system. The proposed methodology uses an external input, which is air temperature, to improve the estimation of missing data. However, the external input also had missing values that needed to be adjusted. To address this, first deep learning, specifically a Feed-Forward Neural Network, was implemented. Due to the high correlation between air and concrete temperature, generalized linear models were applied to estimate the missing concrete temperature values. Finally, the missing data for the resistivity sensor was estimated using pattern recognition and the inverse relationship between temperature and electrical resistivity. The results suggest that the proposed methodology can serve as a valuable tool to enhance the quality of sensor data and improve the effectiveness of monitoring systems in the analysis for early detection of corrosion. The paper is structured as follows. Section 2 describes the case study, Section 3 presents the methodology employed, Section 4 provides the results and discussion, and finally, the research conclusions are presented in Section 5.

2. Case study description

2.1. Test bed description

The bridge, inaugurated in the 1980s, is located in central Portugal. It features a main span of over 200 m and a total length of more than 900 m, supported by 85 m high piles in the tallest section. The analyzed bridge is located less than 5 km from the sea and serves to connect two regions of one of Portugal's major cities.

A detailed inspection revealed several issues: low execution quality with concreting defects, poor-quality painting of steel structures, reinforcement corrosion, alkali-silica reactions, sulphate attack (primarily in the foundations of the bridge), and frequent cracking in prestressed girders. These factors, along with updated design codes, dictated a rehabilitation of the structure in the 2000s. Additional information about the structure cannot be disclosed due to confidentiality concerns.

Concrete electrical resistivity and temperature data were collected from five repair zones on the bridge. The objective of the SHM system is to obtain information about the progress of the despassivation front in the concrete. Data collection occurred daily from July 2006 to November 2020.

2.2. Sensors and measurements

The air temperature data was obtained from the Instituto Português do Mar e da Atmosfera (IPMA), a public institute under the indirect administration of the state. The data comes from an automated weather station located less than 5 km from the analyzed bridge. The station is situated 4 m above sea level, and the daily average temperature, measured at a height of 1.5 m, was used.

Sensors were installed in five repaired zones of the structure, referred to as Location 1 (L1) through Location 5 (L5). Concrete electrical resistivity was measured using a two-graphite electrode resistivity sensor. Installation involved removing the concrete cover, placing the electrodes at depths of 15mm and 30mm, and then replacing the cover. Eight resistivity sensors were installed and will be referred to as L1-R1 if the sensor is located in Location 1 at a depth of 15mm, and L1-R2 if it is in Location 1 at a depth of 30mm. The concrete temperature was measured using a PT100 thermometer embedded in concrete installed at the same time. The temperature sensors will be named L1-T if located in Location 1, and similarly for the other locations. Data acquisition was performed automatically daily at midnight using a Datataker 500. The two-graphite electrode resistivity sensors measure daily concrete electrical resistivity of the bridge (Figure 1a). The concrete temperature was measured in Celsius using the same daily frequency (Figure 1b). After more than fourteen years of measurements, several data are loss due to problems with the data acquisition system and the power supply unit of the data acquisition system. A total of 27,032 electrical resistivity data points and 19,205 concrete temperature data points were collected, from eight resistivity sensors and five temperature sensors. Figure 2 presents the missing data for each of the sensors considered in this study, highlighting significant gaps in the concrete resistivity sensors.

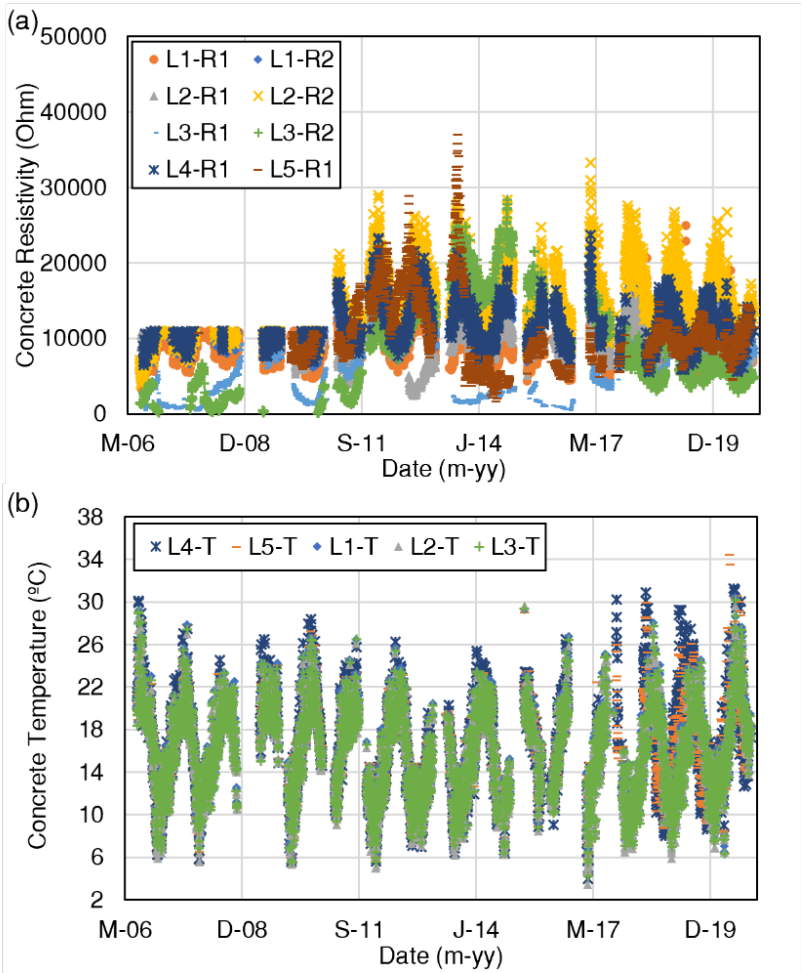

Figure 1. (a) Concrete electric resistivity and (b) Concrete temperature data obtained between 2006 and 2020.

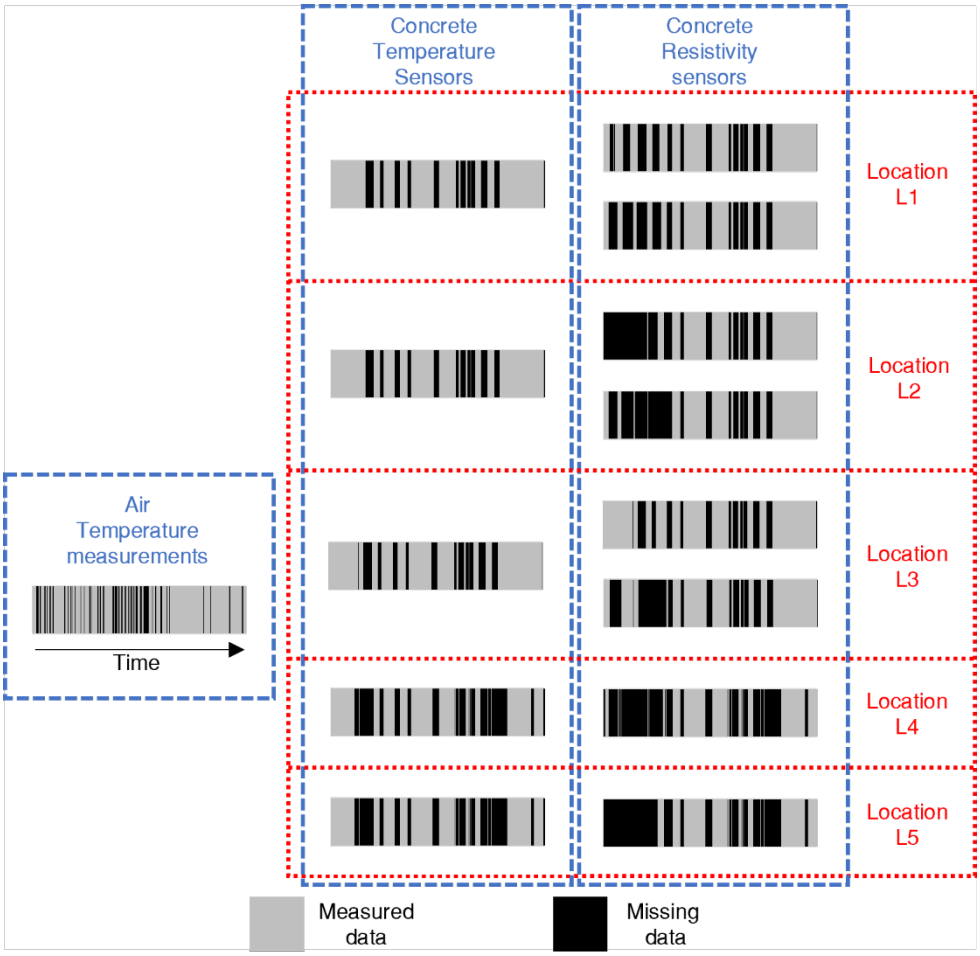

Figure 2. Missing data of each sensor consider in this research.

3. Proposed methodology

The methodology proposed for addressing missing data in this SHM system encompasses four key stages. This approach begins with assessing the sizes of data gaps (Stage A), followed by procedures to fill gaps in air and concrete temperature data (Stage B and C), and concludes with the implementation of pattern recognition techniques for missing resistivity data (Stage D). Each phase aims to systematically tackle the absence of information in the sensor datasets, ensuring a comprehensive approach to data completion. Figure 3 presents a diagram of the methodology used in this paper to fill in missing data from the concrete resistivity and temperature sensors, where the key stages are highlighted. Stage A presents a recommendation for data imputation based on the size of the data gap. Stage B introduces the methodology using Artificial Neural Networks to fill the missing data from the air temperature sensor (explained in Section 3.1). Stages C and D detail the methods for imputing missing data from concrete temperature and electrical resistivity sensors, which are described in Sections 3.2 and 3.3, respectively.

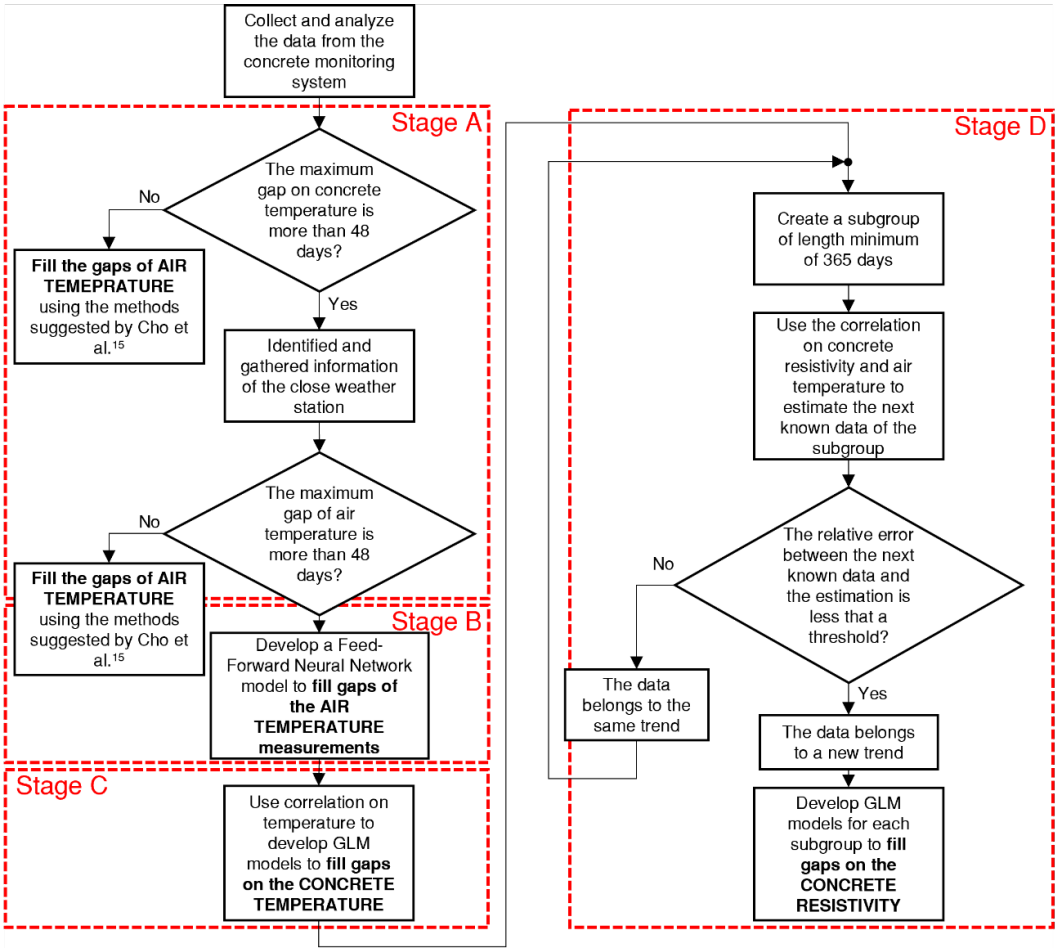

Figure 3. Flowchart of the methodology proposed to fill the missing data of the concrete resistivity sensors.

The methodology starts with the first part of stage A (See Figure 3), which consist in analyzing the maximum size of the gap to be filled. Table 2 presents the maximum sizes obtained for missing gaps. It is observed that the concrete temperature and resistivity present gaps of more than one year of lost information. Cho et al. [9] presented an extensive study to establish the best data imputation methods. In their study, three levels of gaps are established and numerical methods for filling are suggested (Table 3). The methodology proposed in this paper suggest following these recommendations. Therefore, for the missing concrete temperature and resistivity data require more intensive computational methods.

Table 2. Maximum gap size per type of sensor

Measurement	Maximum gap size (days)
Air temperature	59
Concrete temperature	336
Concrete resistivity	786

Table 3. The method suggested for data filling. Adapted from [9].

Gap classification	Maximum gap size	Method suggested
Small	1-8	Linear interpolation
Larger	9-48	K-nearest neighbors
Even larger	>48	More computational intensive

Lo Presti et al. [6] presented a methodology for estimating missing data, primarily applied to rainfall data in Italy. The methodology is divided into two stages. Firstly, they identify a similar weather station to the one being analyzed to determine suitable similarity coefficients. Secondly, a regression method is applied to estimate the missing data. A similar methodology is employed for filling the concrete temperature data gaps. Therefore, air temperature data is collected from a weather station located 3.1km away from the analyzed structure. However, this station also has missing data, with a maximum gap size of 59 days, which is smaller than those in the structure sensors but still significant according to Cho et al.[9]. In this initial step, a deep machine learning technique, specifically a Feed-Forward Neural Network, is used to compute the missing values of air temperature (stages A and B, section 3.1). Then, the proposed methodology considers the high correlation between air temperature and concrete temperature, to impute the missing data in the concrete temperature sensors (stage C, section 3.2).

Temperature is a crucial factor in the resistivity of concrete. However, it is important to recognize that it is not the only factor. According to the literature, concrete resistivity is affected by several factors such as pore structure, ion composition in pore water, cement content, and the degree of saturation, among others [23]. Temperature impacts resistivity by altering ion mobility, ion-ion and ion-solid interactions, and ion concentration in the pore solution [23,24]. Typically, as the temperature of concrete increases, its electrical resistivity decreases [25].

The relationship between electrical resistivity and electrical conductivity is commonly expressed as an inverse linear correlation [25,26]. Although temperature is not the sole influencing factor, it was chosen for this study due to its significant impact and the availability of temperature data from the sensors. Since comprehensive information on all factors influencing resistivity was not available, methods were employed to learn from the temperature-resistivity relationship in real conditions and attempt to extrapolate this relationship (stage D, section 3.3). Although this represents a limitation of the study, it is decided to fill the resistivity gaps using an intensive computational method that associates these parameters, considering the missing data on temperature and resistivity as missing at random [27].

3.1. Feed-Forward Neural Network method

As mentioned in the previous section, when there are gaps of more than 48 consecutive data points, more intensive computational methods must be used for data imputation. This section presents the Artificial Neural Networks method used to fill the missing data for the air temperature sensor, corresponding to stage B in Figure 3.

Artificial neural network models comprise a collection of neurons processing information individually and simultaneously, mirroring the functioning of the human brain [28], significantly enhancing the predictive accuracy by effectively capturing complex patterns in the data. In the context of time series forecasting, the Multilayer Feed-Forward Neural Network Autoregressive (FFNN-AR) model [29], stands out since it considers the evolution of time series data by integrating an autoregressive process of order p with a non-linear function to implement the complex dynamic behavior of the data instead of depending linearly on the previous values.

In this context, the temperature-lagged time series estimates are the inputs x to the model and are given by

$$x = x_{t-1}, x_{t-2}, x_{t-3}, \dots, x_{t-p} \quad (1)$$

The number of neurons n in the input layer corresponds to the autoregressive order p which is determined using the partial autocorrelation function. This model processes the input of lagged-time series temperature values (Eq. (1)) through a hidden layer in a one-direction flow and applies activation functions to the hidden and output layers (See Figure 4).

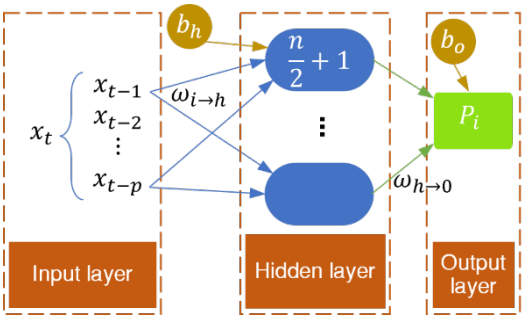

Figure 4. Structure of the FFNN-AR model. Adapted from [28].

The choice of the activation function in the layers corresponds to the type of the problem being solved and the nature of the input and output of the layers, and is determined through the loss function, i.e., Root Mean Square Error (RMSE) indicator (Eq. (2)) which provides a measure of accuracy by measuring the average magnitude of the differences between the predicted and the actual values to minimize the difference. In this study, the activation function for the hidden layer is a non-linear sigmoid activation function (Eq. (3)). A linear activation function (Eq. (4)) is applied for the output layer since the predictions in the output layer are the weighted sum of the resulting weights and biases from the hidden layer, making it directly proportional relation.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(Y_i - X_i)^2}{n}} \tag{2}$$

$$f_h(x) = \frac{1}{1+e^{-x}} \tag{3}$$

$$f_o(x) = \sum_{i=1}^{\frac{n}{2}+1} w_{h\to o} + b_h \tag{4}$$

The FFNN-AR model (Eq. (5)) considers the dynamic behavior of time series by implementing non-linearity within the hidden layer to represent nonlinearly the autoregressive process through a non-linear activation function (Eq. (3)) to the weighted sum of the inputs x_t(Eq. (1)), in which the weights ω and biases b are optimized using backpropagation to minimize the prediction error through a loss function (Eq. (2)). The prediction value P_i within the output layer applies the linear activation function (Eq. (4)) to the resulting weights and bias from the hidden layer, and is given by

$$P_i = \mathrm{f_o}\left[\sum_{i=1}^{\frac{n}{2}+1} \omega_{h\to o} f_h \left[\sum_{i=1}^{p} \omega_{i\to h} y_{t-l} + b_h\right]\right] + b_o \tag{5}$$

The FFNN-AR model was trained on temperature measurements for the interval from 04/07/2006 to 16/09/2006 and validated by predicting the temperature measurements from 17/09/2006 to 04/10/2006. The architecture of the FFNN-AR model consists of 15 nodes in the input layer, a hidden layer with 8 nodes using a sigmoid activation function, and a single output node. Training was conducted over 100 epochs using the Adam optimizer with a learning rate of 0.001. The loss function used for training was the Root Mean Square Error (RMSE), with a final training RMSE of 0.0003°C.

Figure 5 presents the comparison between the predicted and actual temperature measurements from 17/09/2006 to 04/10/2006, demonstrating good agreement between the two. Table 4 provides the validation error metrics: Mean Error (ME), Mean Absolute Error (MAE), and RMSE, with values of -

0.098°C, 0.99%, and 1.34°C, respectively. The model's performance was stable, as indicated by the low RMSE and MAE values.

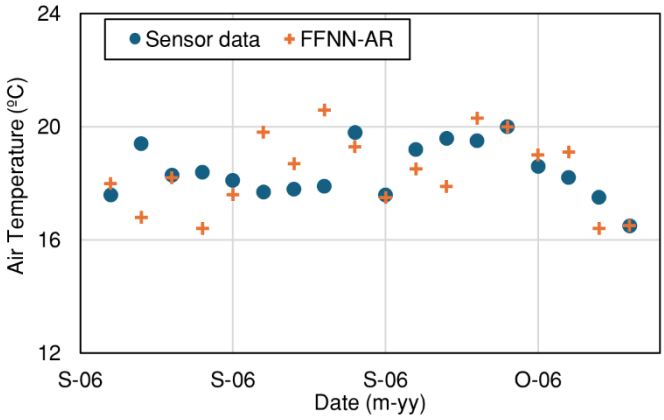

Figure 5. FFNN-AR model validation

Table 4. Validation error indicators.

Error indicators	FFNN-AR model
ME (C°)	-0.098
MAE (%)	0.99
RMSE (C°)	1.34

3.2. Generalized linear models.

Another intensive computational method used in the proposed methodology is Generalized Linear Models (GLMs). These are employed when a variable significantly influences the results of another variable. In this case, this model is part of stages C and D in Figure 3, where the predictor variable, air temperature, is used to estimate the values of concrete temperature and concrete electrical resistivity.

Generalized Linear Models (GLMs) constitute a statistical framework that extends classical linear regression models. These models have found widespread use in civil engineering due to their ability to provide greater flexibility in data distribution and the relationship between the dependent variable and the independent variables [30–32].

In a GLM, the relationship between the response variable Y and the predictor variables is modeled through a linking function g as follows:

$$g(\mu) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \quad (6)$$

where g is the linking function, μ is the expected value of Y, β_i are the estimated coefficients, and x_i are the predictor variables. Different linking functions can be considered in these models, including linear, quadratic, compound, growth, exponential, cubic, inverse, among others. In the present research, the identity link function was used since the relationship between air temperature and the dependent variables (concrete temperature and resistivity) was assumed to be linear, therefore the relationship is represented as a weighted sum of the predictor variables. In the present methodology, the predictor variable, x, was considered to be the air temperature, which was used to obtain the expected value, μ, corresponding to the concrete temperature and electrical resistivity.

Gaussian family distribution was applied for the error function due to the suitable for this problem according to the main key metrics used, including pseudo R^2, AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), Mean Squared Error (MSE) Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). This distribution is commonly used for modeling continuous outcomes like temperature and resistivity, as it assumes that the residuals (errors) are normally distributed.

3.3. Group pattern recognition.

To achieve better results with the GLMs, a group pattern recognition algorithm was implemented to identify variations in the readings of the concrete resistivity sensors. These variations could be due to degradation processes or changes in other factors influencing the sensors that were not considered in this study. This step corresponds to Stage D in Figure 3.

Pattern recognition for subgroup creation focuses on identifying and understanding patterns and structures within datasets involving multiple distinct groups or classes. The group-based approach aims to identify similarities and differences among datasets that can be divided into distinct groups or categories, aiding in better prediction of missing data. Various techniques exist to address group pattern recognition. The two primary methods focus on clustering algorithms and classification techniques to assign data to different classes or groups based on their characteristics [33]. Clustering algorithms were used in the group pattern recognition for this paper. Correlation was used as the variable to separate different subgroups.

4. Results and discussion

This section presents the results of the proposed methodology for the case study described in Section 2. In section 4.1, the missing data in the air temperature measurements are estimated, while in section 4.2 the results obtained in filling the missing data for the concrete temperature sensors are presented. Section 4.3 outlines the final part of the methodology, focusing on filling the missing data for the resistivity sensors.

4.1. Filling missing air temperature data

The first step in Figure 3 is to compute the missing air temperature data. This was achieved using a Feed-Forward Neural Network (FFNN), as explained in Section 3.1. The recorded air temperature data were used to train and validate the model. Table 5 presents the amount of known and missing data, and the largest continuous gap of missing data of the air temperature sensor.

Table 5. Information about the missing data in air temperature database.

Measured data	Missing data	Maximum continue gap
4975	277	59

Figure 6 presents the results of the FFNN method for filling missing air temperature data. It is observed that the calculated values align adequate with the temperature variations produced by the sensors. To estimate the approximation accuracy of the FFNN, five artificial gaps of 1, 5, 10, 25, and 60 days were created. Table 6 presents the main error metrics obtained between all the artificial gaps and the values measured by the meteorological station. The MAE of 3.51 indicates that, on average, the values are off by 3.51ºC, which is an acceptable value for the study. The R^2 value of 0.78 implies that 78% of the variability in the air temperature can be explained by the model, which is generally considered a strong

result according to Insukindro [34]. Therefore, it can be concluded that the results present an adequate fit of the proposed model, indicating that this step of the methodology functions properly.

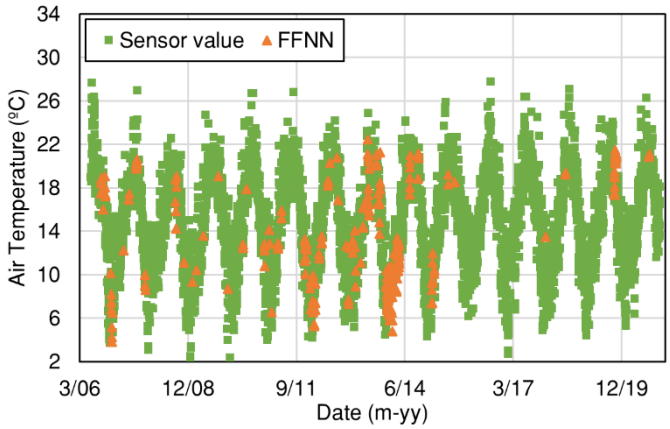

Figure 6. Data filling of air temperature using the FFNN method.

Table 6. Principal error indicators for the FFNN method.

Temperature sensor	Value	Units
Mean Absolute Error (MAE)	3.51	ºC
Mean Squared Error (MSE)	18.62	$ºC^2$
Root Mean Squared Error (RMSE)	4.32	ºC
Coefficient of Determination (R^2)	0.78	-

4.2. Filling missing concrete temperature data.

Once a gap-free air temperature database is obtained, the methodology is applied to the concrete temperature sensors. Table 7 displays the Pearson correlation coefficients between air temperature and the five concrete temperature sensors installed within the structure. Schober et al.[35] suggest that a Pearson coefficient between 0.7 and 0.89 can be considered a strong correlation. Therefore, with correlation coefficients greater than 0.8, the methodology used GLM to estimate missing data from the concrete temperature sensors. Table 7 also presents the amount of known and missing data, and the largest continuous gap of missing data of the five concrete temperature sensors. Two sensors (L4-T and L5-T) acquired less data and presented maximum continuum gap of 336 days.

Table 7. Information about the missing data in concrete temperature sensors.

Concrete Temperature sensor	Measured data	Missing data	Maximum continue gap	Pearson Correlation
L1-T	4106	1146	197	0.91
L2-T	4106	1146	197	0.92
L3-T	4106	1146	198	0.92
L4-T	3444	1808	336	0.83
L5-T	3444	1808	336	0.84

Figure 7 illustrates the results of filling missing data for the concrete temperature sensors. It is noteworthy that the estimations from the GLM adequately fill the gaps in the data for the five concrete temperature sensors. Additionally, a discernible seasonal trend is observed throughout the analyzed

period that is also well represented by the filled data. Table 8 presents the primary error indicators for the series of the GLM model adjusted for each sensor. The L2-T and L5-T sensors show the best MAE and RMSE values, indicating greater precision when filling concrete temperature. However, the average MAE is 1.296ºC, with low variability (standard deviation of 0.061), indicates that most sensors have similar precision in terms of mean absolute error. The average R^2 is 0.786, with a standard deviation of 0.072. This indicates that, on average, the sensors explain 78.6% of the variability in temperature measurements, although some sensors (such as the L4-T and L5-T) have lower R^2 values. Overall, the results demonstrate the high applicability of this methodology, even in cases where there is a high correlation among the data despite gaps of more than 48 days.

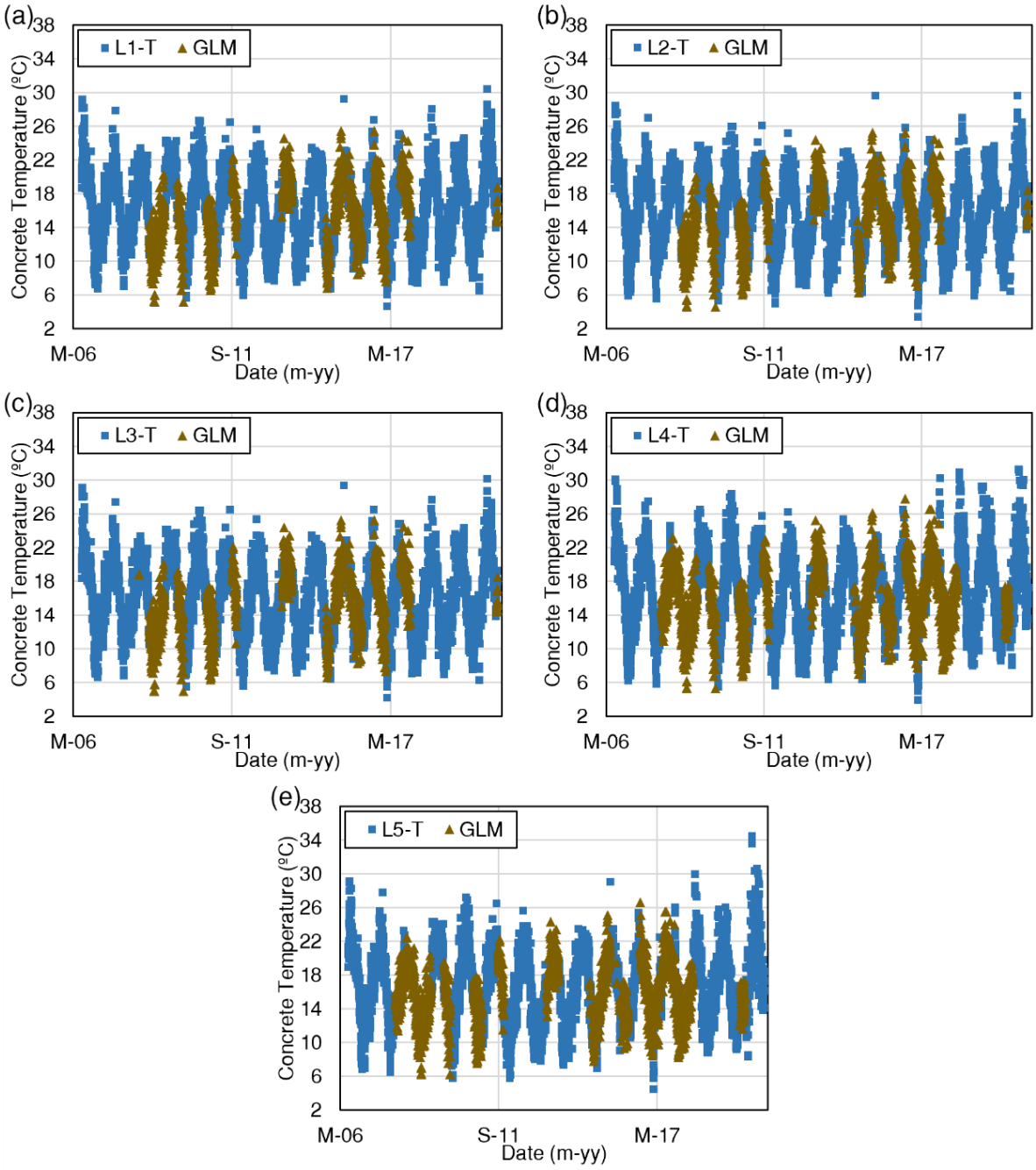

Figure 7. Data filling of the concrete temperature sensors using GLM.

Table 8. Error indicators for the GLM model.

Concrete temperature sensor	MAE (ºC)	MSE ($ºC^2$)	RMSE (ºC)	R^2
L1-T	1.29	2.98	1.73	0.84
L2-T	1.25	2.72	1.65	0.85
L3-T	1.29	2.98	1.73	0.84
L4-T	1.40	3.52	1.88	0.70
L5-T	1.25	2.80	1.67	0.70

4.3. Filling concrete resistivity data

The final step of the methodology is based on the premise of a correlation between temperature and electrical resistivity. This relationship is evident in Figure 8(a) and Table 9 where a negative correlation is observed for almost all cases. However, some sensors do not exhibit a clear correlation (Figure 8(b)). GLMs are consider on the estimation of the missing data in this section. However, the direct application of GLM is not feasible without first identifying patterns in the data. Therefore, pattern recognition techniques are employed to identify subgroups within the dataset that exhibit consistent trends, which then allows for the application of GLM for predictive purposes.

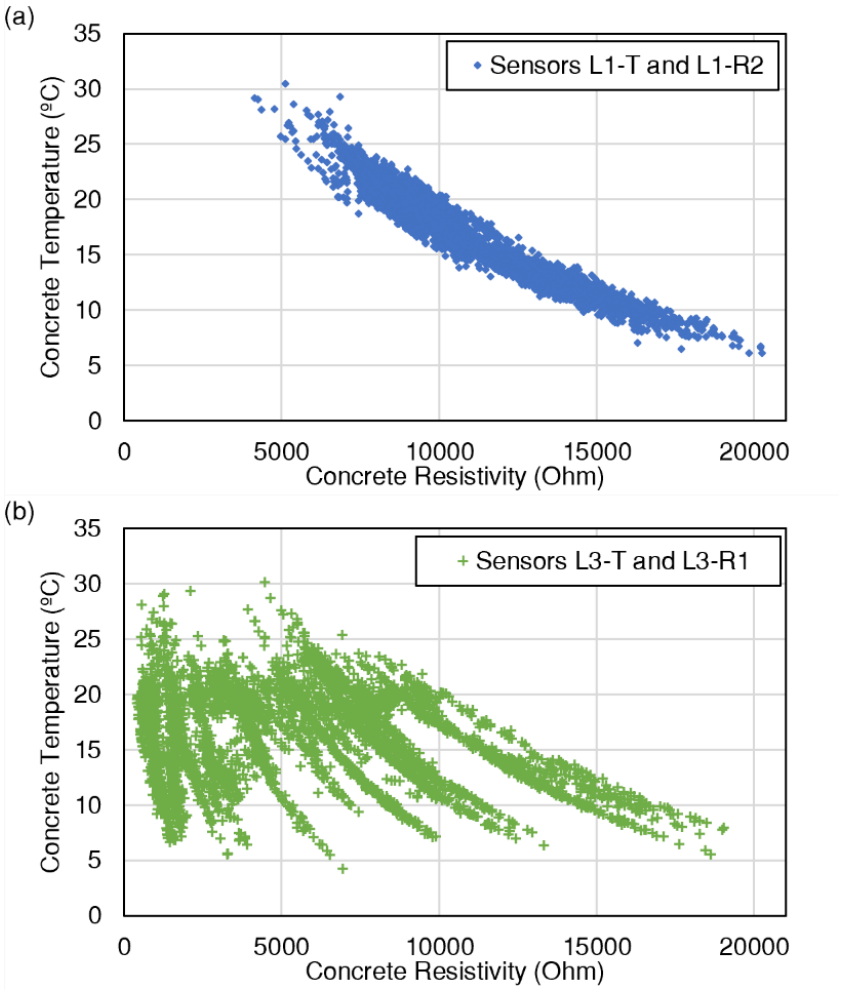

Figure 8. Relation between concrete temperature and resistivity sensors.

Table 9. Pearson correlation between concrete resistivity and concrete temperature.

Name of the sensors		Correlation
Resistivity sensor	Temperature sensor	
L1-R1	L1-T	-0.841
L1-R2	L1-T	-0.961
L2-R1	L2-T	-0.472
L2-R2	L2-T	-0.953
L3-R1	L3-T	-0.274
L3-R2	L3-T	-0.360
L4-R1	L4-T	-0.952
L5-R1	L5-T	-0.329

Figure 9 presents a 3D representation of Figure 8(b), highlighting the potential changes in correlation over time. These variations may be associated with fluctuations in concrete conditions. While the correlation varies, its association with temperature appears consistent. Hence, it is proposed to employ a group pattern recognition for resistivity data (Stage D in Figure 3, section 3.3).

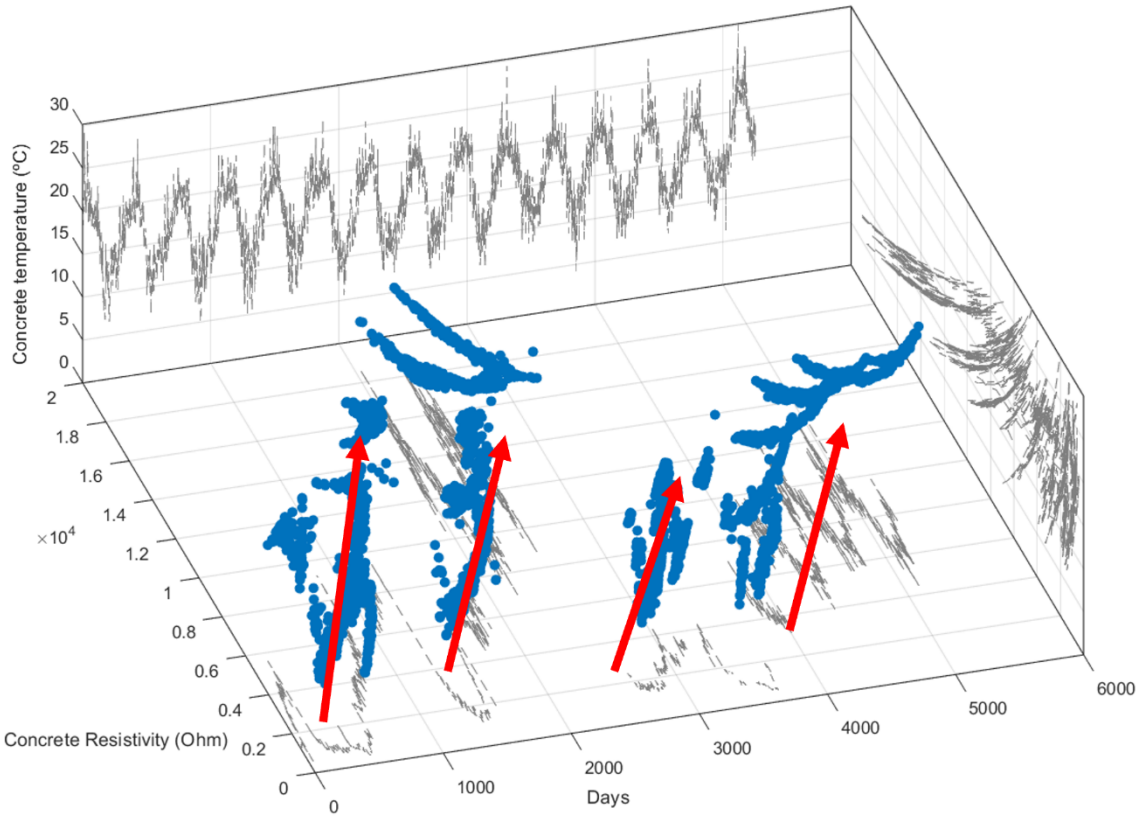

Figure 9. Relation of concrete resistivity, concrete temperature, and days of the sensors in location 3 (L3-T and L3-R1). In red, are possible different trends.

Group pattern recognition is used to identify subgroups among the analyzed sensors where data exhibit a consistent trend. The minimum subgroup size of 365 days was selected to reflect the seasonal cycles that influence concrete resistivity. This period was chosen because it aligns with typical climatic patterns, ensuring that the subgroups capture the variations that occur due to temperature and environmental changes over a complete year.

Starting from this minimum group size, a correlation is calculated, and the most recent data point is compared to the prediction using GLM. To minimize user-induced bias, the segmentation process was designed to be as objective as possible. We implemented a method that compares the relative error between the measured point and the prediction, triggering a new subgroup when this error exceeds a predefined threshold. This automatic procedure reduces manual intervention in subgroup creation, ensuring that the segmentation is based on statistical consistency rather than subjective visual interpretation. A relative error threshold of 0.8 was assumed based on engineering judgment to ensure good separation of subgroups during the pattern recognition process. When the error between the next measured point and the calculated prediction surpasses the threshold, a new subgroup is initiated.

Figure 10 displays resistivity values segmented by subgroups, while Table 10 presents the Pearson correlation for each subgroup. The methodology did not identify more than one subgroup for the L1-R1, L1-R2, and L4-R1 sensors, suggesting that the separation into subgroups it is not necessary because a strong correlation was estimated for the considered data indicating consistency in the measurements (See Figure 10(a, b, and g)). For the other sensors, the separation of the data into subgroups reveals a variability that is not consistent with the original correlations – i.e., positive correlations. This suggests that temporal patterns and trends may vary significantly over time, highlighting the importance of considering subgroups in the analysis to capture more complex dynamics. This pattern recognition also is useful to identify some subgroups with low or positive correlations for which the available data cannot be accurately used for filling purposes.

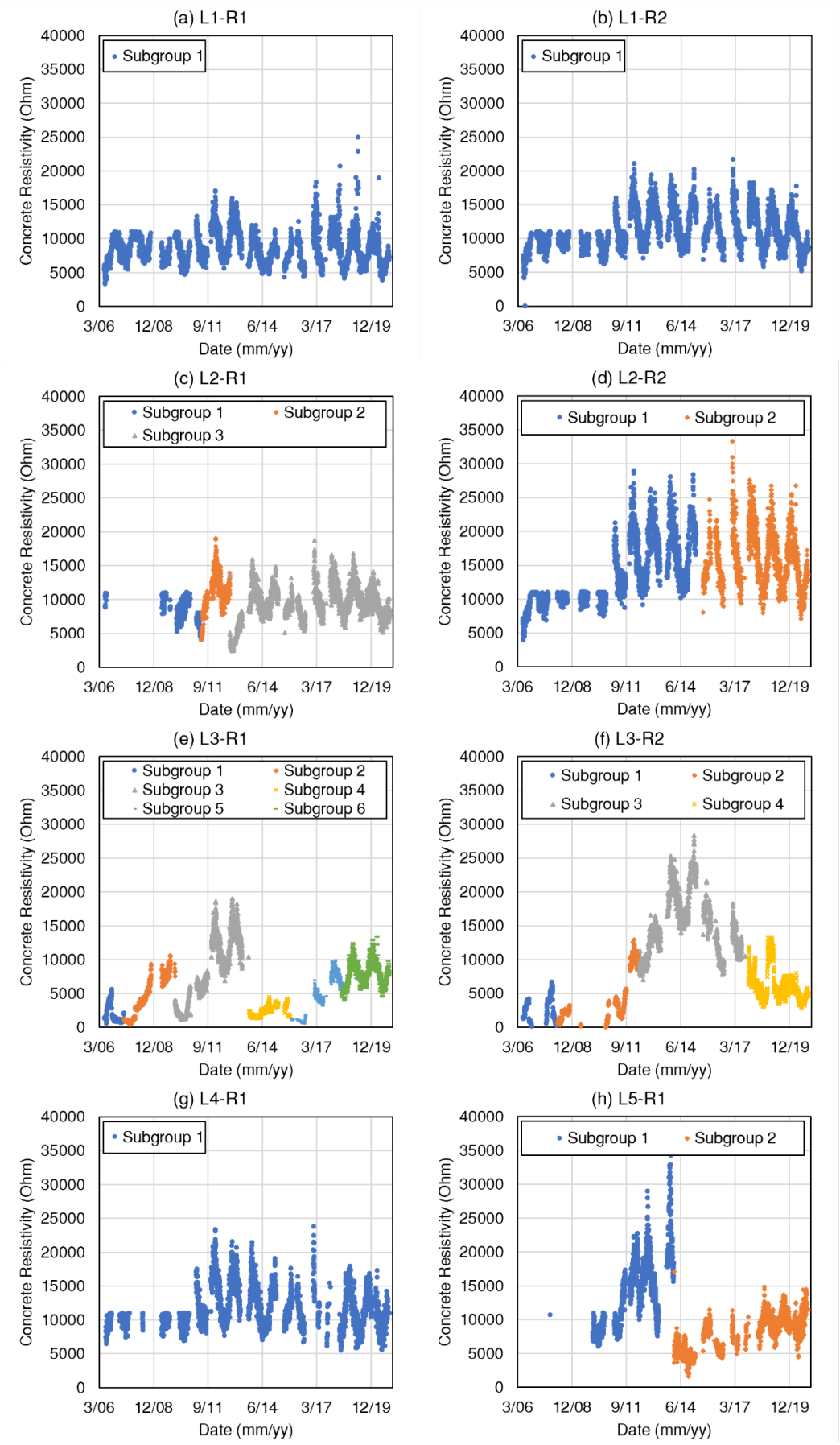

Figure 10. Subgroups identified using the proposed methodology in concrete resistivity sensors.

Table 10. Fundamental information of each subgroup of concrete resistivity sensors.

Concrete Resistivity sensor	Subgroup	Measured data	Missing data	Maximum continue gap	R^2	Subgroup correlation	Sensor correlation
L1-R1	1	3888	1364	200	0.60	-0.78	-0.78
L1-R2	1	3620	1632	210	0.80	-0.89	-0.89
L2-R1	1	364	1412	1012	0.12	0.35	-0.44
	2	459	64	64	0.60	-0.77	
	3	2253	700	151	0.23	-0.48	
L2-R2	1	1884	1427	219	0.76	-0.87	-0.87
	2	1533	408	151	0.84	-0.91	
L3-R1	1	365	0	0	-	0.29	-0.22
	2	654	277	197	0.10	0.31	
	3	1065	297	114	0.27	-0.52	
	4	560	206	120	0.13	0.35	
	5	586	358	151	0.28	-0.53	
	6	874	10	10	0.79	-0.89	
L3-R2	1	364	270	268	0.05	-0.22	-0.33
	2	673	831	455	0.48	-0.69	
	3	1318	690	151	0.28	-0.53	
	4	1096	10	10	0.41	-0.64	
L4-R1	1	2882	2370	339	0.62	-0.78	-0.78
L5-R1	1	1071	1709	786	0.12	-0.35	-0.27
	2	1523	949	165	0.00	0.03	

Once the subgroups for each sensor have been identified, the final part of stage D (Figure 3) is carried out. This part involves estimating the missing data for the subgroups using GLM models to complete the information for the resistivity sensors. For sensors where no subgroups were identified, the entire database of the sensor and the air temperature was used for the estimation of missing data. Table 10 also presents the subgroups obtained from the methodology, the amount of known and missing data, the largest continuous gap of missing data.

Table 10 also provides the R^2, the correlation for each subgroup and the original correlation of each sensor. It is observed that in most of the created subgroups, the correlation values are more consistent compared to the expected correlation, indicating that the methodology is capable of extracting specific trends from the analyzed data. The predictions for sensors L1-R1, L1-R2, L3-R2, and L4-R1 align remarkably well, with correlations below -0.7 and R^2 values exceeding 0.6. The authors consider an adequate estimation of the missing data occurs when the correlations are below -0.7 and R^2 values above 0.6. In that scenario, subgroup 2 of L2-R1 and subgroup 6 of L3-R1 should also be included among the successful predictions. Figure 11 showcases the data filled by the methodology for all concrete resistivity sensors. It is observed that the proposed approach provides a consistent data filling for the groups for which correlations are below -0.7 and R^2 values above 0.6, as it is characteristically shown in Figure 11. However, it is also noted (e.g., Figure 11(h), subgroup 2) that data filling provided by this method is not suitable when the above-mentioned conditions are not satisfied.

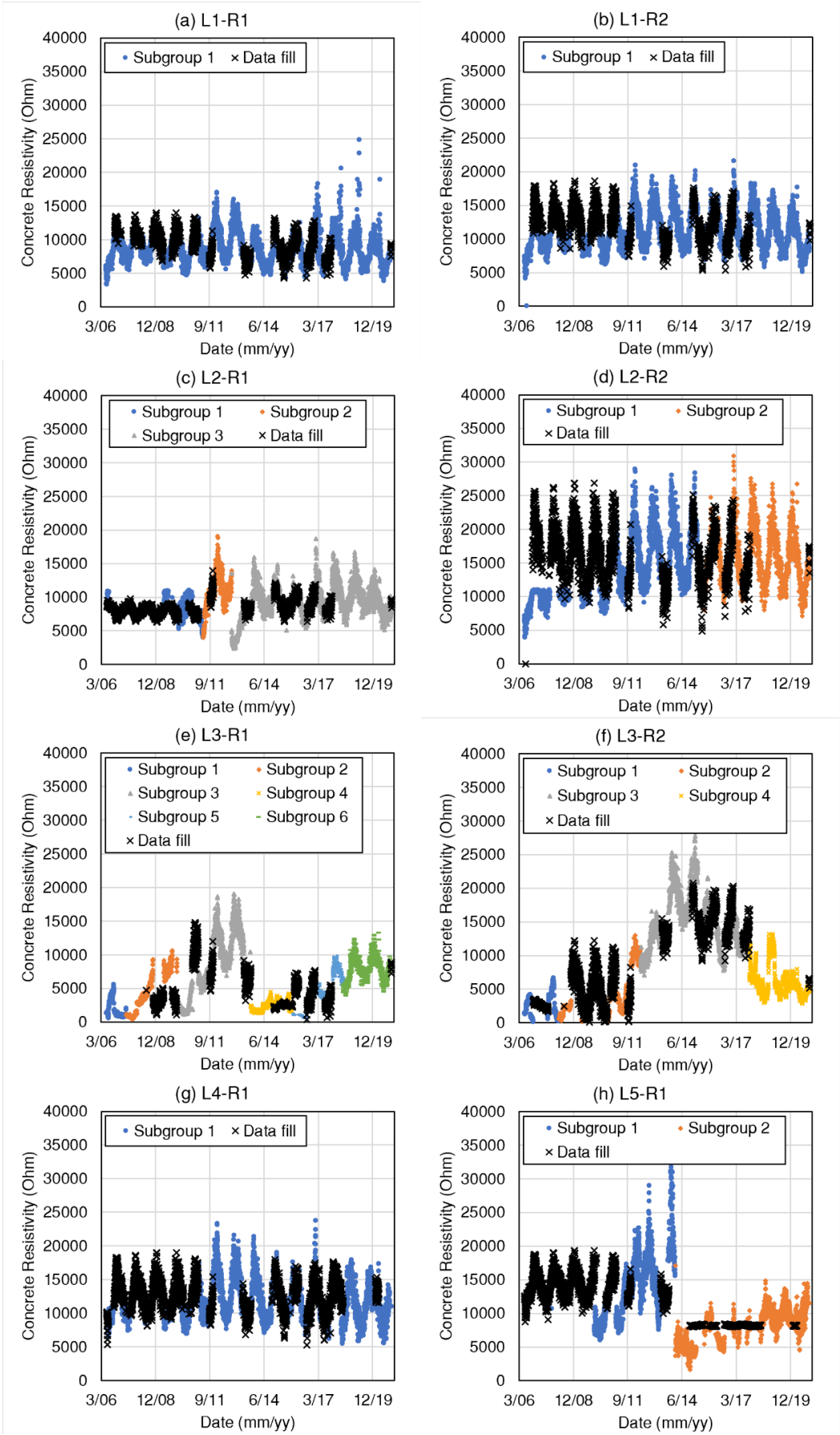

Figure 11. Data filling of the concrete resistivity sensors using pattern identification and GLMs.

4.4. Sensitivity analysis of error propagation

A sensitivity analysis was conducted to evaluate the impact of the MAE results of Table 6 in the imputation of air temperature on the subsequent predictions of concrete temperature and electrical resistivity. Since air temperature is a key input in the imputation process, it was necessary to assess how uncertainties in its estimation propagate through the model and affect the predicted values for other variables.

Three scenarios were defined to simulate the potential impact of the MAE on the imputed air temperature: Scenario 1: The imputations of air temperature were increased by the MAE value of 3.51°C. Scenario 2: The original imputed air temperature was used as a baseline. Scenario 3: The air temperature imputation were decreased by the MAE value of 3.51°C. For each scenario, the Generalized Linear Model (GLM) was applied to predict the missing values of concrete temperature and resistivity. The differences between the predictions in Scenario 1 and Scenario 3 were compared to those in Scenario 2 (baseline) to quantify the impact of perturbations in air temperature on the predicted variables.

The sensitivity analysis demonstrates that the MAE of 3.51°C in the imputed air temperature has a minimal effect on the predictions of concrete temperature and resistivity, with average impacts of less than 1.5%. Table 11 presents the impact on the imputation of concrete temperature and concrete resistivity, calculated as the relative error with respect to Scenario 2 (baseline). This low sensitivity indicates that the model is resilient to uncertainties in air temperature imputations, supporting the validity and robustness of the proposed methodology. The slight asymmetry observed in the results suggests that further investigation could be conducted to better understand the model's sensitivity to variations in lower temperature ranges, but these findings do not compromise the overall reliability of the results.

Table 11. Impact of Air Temperature Imputation Errors on Concrete Temperature and Resistivity Predictions.

Scenario	Impact on Concrete Temperature (%)	Direction (Concrete Temperature)	Impact on Concrete Resistivity (%)	Direction (Concrete Resistivity)
1. Air Temperature Imputation +3.51ºC	-0.66	Decrease	0.75	Increase
2. Baseline	-	-	-	-
3. Air Temperature Imputation -3.51ºC	1.48	Increase	-0.61	Decrease

5. Conclusions

The present research addresses the challenge of missing data in the durability performance monitoring of reinforced concrete structures using SHM systems. The proposed methodology employs a combination of feed-forward neural networks, generalized linear models, and pattern recognition techniques to impute missing data in air temperature measurements as well as in concrete resistivity and temperature sensors.

The scientific value of this study lies in the significant expansion and improvement of a preliminary methodology presented by the authors [10]. While the previous work was limited to a single year of data for a single resistivity and temperature sensor and focused on filling small data gaps of up to 61 days,

this paper introduces a more robust approach. By incorporating over fourteen years of sensor data and integrating air temperature as an additional input for data imputation, the present study demonstrates a more comprehensive and accurate methodology capable of filling longer data gaps. This is the first study, to the best of the authors' knowledge, to apply such an imputation methodology for missing data in resistivity sensors within SHM systems.

The results demonstrate that the methodology is particularly effective for sensors with strong correlations between temperature and resistivity (absolute Pearson correlation value greater than 0.7) and high R^2 values (above 0.62). Consequently, 43.4% of the missing data was estimated adequately. The integration of air temperature as an input improves the overall accuracy of the imputation process, particularly in long-term sensor data analysis, and offers practical benefits for SHM system managers in the context of the interpretation of the data for early corrosion detection and maintenance planning.

Despite these advances, some limitations remain. The methodology relies heavily on the correlation between temperature and resistivity, which can be problematic in scenarios of concrete deterioration, where this relationship may break down. This limits the effectiveness of the approach in certain deteriorated conditions. Furthermore, the use of GLMs proved to be effective in subgroups with high correlation, but less so in groups with lower correlation values, suggesting the need for future research into more advanced computational models that do not depend solely on correlation.

Pattern recognition allowed the identification of subgroups with similar behaviors, improving Pearson correlation and missing data estimation. However, with only 43.4% of the estimated data achieving a strong correlation with the measured data, there is still room for improvement. Future studies could explore the application of more sophisticated machine learning models or hybrid approaches that can address data variability more effectively, particularly in regions with high uncertainty.

Acknowledgments

This work is financed by national funds through FCT - Foundation for Science and Technology, under grant agreement 2021.05862.BD attributed to the first author. Doi: https://doi.org/10.54499/2021.05862.BD

This work was partly financed by FCT / MCTES through national funds (PIDDAC) under the R&D Unit Institute for Sustainability and Innovation in Structural Engineering (ISISE), under reference UIDB / 04029/2020 (doi.org/10.54499/UIDB/04029/2020), and under the Associate Laboratory Advanced Production and Intelligent Systems ARISE under reference LA/P/0112/2020.

The meteorological information used in the present research was provided by the Instituto Português de Mar e da Atmosfera (IPMA) - www.ipma.pt

Statements and Declarations

Ethical considerations: Not applicable

Consent to participate: Not applicable. This article does not contain any studies with human or animal participants.

Consent for publication: Not applicable

Declaration of conflicting interest: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article

Funding statements: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work is financed by national funds through FCT - Foundation for Science and Technology, under grant agreement 2021.05862.BD attributed to the first author. Doi: https://doi.org/10.54499/2021.05862.BD.

This work was partly financed by FCT / MCTES through national funds (PIDDAC) under the R&D Unit Institute for Sustainability and Innovation in Structural Engineering (ISISE), under reference UIDB / 04029/2020 (doi.org/10.54499/UIDB/04029/2020), and under the Associate Laboratory Advanced Production and Intelligent Systems ARISE under reference LA/P/0112/2020.

r2 in time series linear regression analysis

References

1. Verstrynge E, Van Steen C, Vandecruys E, et al. Steel corrosion damage monitoring in reinforced concrete structures with the acoustic emission technique: A review. *Constr Build Mater* 2022; 349: 128732.

2. Bastidas-Arteaga E, Stewart MG. Economic assessment of climate adaptation strategies for existing reinforced concrete structures subjected to chloride-induced corrosion. *Struct Infrastruct Eng* 2016; 12: 432–449.

3. Marsh PS, Frangopol DM. Reinforced concrete bridge deck reliability model incorporating temporal and spatial variations of probabilistic corrosion rate sensor data. *Reliab Eng Syst Saf* 2008; 93: 394–409.

4. Shevtsov D, Cao NL, Nguyen VC, et al. Progress in Sensors for Monitoring Reinforcement Corrosion in Reinforced Concrete Structures—A Review. *Sensors* 2022; 22: 3421.

5. Llorens M, Serrrano Á, Valcuende M. Sensores para la Determinación de la Durabilidad de Construcciones de Hormigón Armado. *Rev Ing Constr* 2019; 34: 81–98.

6. Lo Presti R, Barca E, Passarella G. A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy). *Environ Monit Assess* 2010; 160: 1–22.

7. Habeeb B, Bastidas-Arteaga E, Gervásio H, et al. Stochastic Carbon Dioxide Forecasting Model for Concrete Durability Applications. In: *18th International Probabilistic Workshop, Lecture Notes in Civil Engineering 153*. 2021, pp. 753–765.

8. Li L, Liu H, Zhou H, et al. Missing data estimation method for time series data in structure health monitoring systems by probability principal component analysis. *Adv Eng Softw* 2020; 149: 102901.

9. Cho B, Dayrit T, Gao Y, et al. Effective Missing Value Imputation Methods for Building Monitoring Data. In: *2020 IEEE International Conference on Big Data (Big Data)*. Atlanta, GA, USA: IEEE, pp. 2866–2875.

10. Rincon LF, Habeeb BA, Bastidas-Arteaga E, et al. Time series analysis for database completion and forecast of sensors measurements: application to concrete structures. *Acad J Civ Eng* 2023; 94–103.

11. Cho B, Dayrit T, Gao Y, et al. Effective Missing Value Imputation Methods for Building Monitoring Data. In: *2020 IEEE International Conference on Big Data (Big Data)*. Atlanta, GA, USA: IEEE, pp. 2866–2875.

12. Gao S, Zhao W, Wan C, et al. Missing data imputation framework for bridge structural health monitoring based on slim generative adversarial networks. *Measurement* 2022; 204: 112095.

13. Cui X, Gu H, Gu C, et al. A Novel Imputation Model for Missing Concrete Dam Monitoring Data. *Mathematics* 2023; 11: 2178.

14. Liu G, Mao Z, Todd M, et al. Damage assessment with state–space embedding strategy and singular value decomposition under stochastic excitation. *Struct Health Monit* 2014; 13: 131–142.

15. Wan H-P, Ni Y-Q. Bayesian multi-task learning methodology for reconstruction of structural health monitoring data. *Struct Health Monit* 2019; 18: 1282–1309.

16. Niu J, Li S, Li Z. Restoration of missing structural health monitoring data using spatiotemporal graph attention networks. *Struct Health Monit* 2022; 21: 2408–2419.

17. Jiang H, Wan C, Yang K, et al. Continuous missing data imputation with incomplete dataset by generative adversarial networks–based unsupervised learning for long-term bridge health monitoring. *Struct Health Monit* 2022; 21: 1093–1109.

18. Tang Z, Bao Y, Li H. Group sparsity-aware convolutional neural network for continuous missing data recovery of structural health monitoring. *Struct Health Monit* 2021; 20: 1738–1759.

19. Luo W, Liu T, Li W, et al. Pitting corrosion prediction based on electromechanical impedance and convolutional neural networks. *Struct Health Monit* 2023; 22: 1647–1664.

20. Figueira R. Electrochemical Sensors for Monitoring the Corrosion Conditions of Reinforced Concrete Structures: A Review. *Appl Sci* 2017; 7: 1157.

21. Azarsa P, Gupta R. Electrical Resistivity of Concrete for Durability Evaluation: A Review. *Adv Mater Sci Eng* 2017; 2017: 1–30.

22. Fan L, Shi X. Techniques of corrosion monitoring of steel rebar in reinforced concrete structures: A review. *Struct Health Monit* 2022; 21: 1879–1905.

23. Polder RB. Test methods for on site measurement of resistivity of concrete — a RILEM TC-154 technical recommendation. *Constr Build Mater* 2001; 15: 125–131.

24. Villagrán Zaccardi YA, Fullea García J, Huélamo P, et al. Influence of temperature and humidity on Portland cement mortar resistivity monitored with inner sensors. *Mater Corros* 2009; 60: 294–299.

25. Presuel F, Liu Y. Temperature Effect on Electrical Resistivity Measurement on Mature Saturated Concrete. In: *NACE - International Corrosion Conference Series*. 2012.

26. Pereira E, Figueira R, Salta MM, et al. A Galvanic Sensor for Monitoring the Corrosion Condition of the Concrete Reinforcing Steel: Relationship Between the Galvanic and the Corrosion Currents. *Sensors* 2009; 9: 8391–8398.

27. Schafer JL. *Analysis of incomplete multivariate data*. London: Chapman & Hall.

28. Rincon LF, Moscoso YM, Hamami AEA, et al. Degradation Models and Maintenance Strategies for Reinforced Concrete Structures in Coastal Environments under Climate Change: A Review. *Buildings* 2024; 14: 562.

29. Hyndman R, Athanasopoulos G, Bergmeir C, et al. Forecasting functions for time series and linear models. 2019. *URL Httppkg Robjhyndman Comforecast R Package Version*; 8.

30. Khedher MBB, Yun D. Generalized Linear Models to Identify the Impact of Road Geometric Design Features on Crash Frequency in Rural Roads. *KSCE J Civ Eng* 2022; 26: 1388–1395.

31. Chou J-S. Generalized linear model-based expert system for estimating the cost of transportation projects. *Expert Syst Appl* 2009; 36: 4253–4267.

32. Esmaeili B, Hallowell MR, Rajagopalan B. Attribute-Based Safety Risk Assessment. II: Predicting Safety Outcomes Using Generalized Linear Models. *J Constr Eng Manag* 2015; 141: 04015022.

33. Burgos DAT, Vejar MA, Pozo F. *Pattern Recognition Applications in Engineering*. IGI Global, 2019.

34. Insukindro Insukindro. Sindrum r2 dalam analisis regresi linier runtun waktu. *J Indones Econ Bus* 1998; 13: 1–11.

35. Schober P, Boer C, Schwarte LA. Correlation Coefficients: Appropriate Use and Interpretation. *Anesth Analg* 2018; 126: 1763–1768.