EWSHM 2024
11th European Workshop on Structural Health Monitoring

A systematic correlation analysis for regression model selection: Application to bridge response prediction using contact and remote sensor systems

Alireza ENTEZAMI[1], Bahareh BEHKAMAL[1], Carlo DE MICHELE[1], Stefano MARIANI[1]
[1]Department of Civil and Environmental Engineering, Politecnico di Milano, Piazza L. da Vinci 32, 20133 Milano, Italy, alireza.entezami@polimi.it, bahareh.behkamal@polimi.it, carlo.demichele@polimi.it, stefano.mariani@polimi.it

Abstract. Correlation analysis is a crucial step before undertaking any regression modeling for data prediction because it helps reveal the relationships between predictors and responses, especially in terms of linearity and nonlinearity. This analysis is often essential for selecting the most appropriate regression model. A major challenge is that linear correlation measures are suitable only for linear relationships, and there are limited measures for assessing nonlinearity. Moreover, a significant issue arises from the influence of unknown predictor data, which can lead to unrealistic and inaccurate outputs from both linear and nonlinear correlation measures. To address these challenges, this paper proposes a systematic correlation analysis that first assesses the impact of unknown predictors and then selects the most suitable regressor for modeling and forecasting. The proposed method utilizes a linear measure known as canonical correlation analysis and a nonlinear measure called maximal information criterion. Based on the correlation values obtained from these measures, one can suggest low, moderate, and high correlation levels. The effectiveness of the proposed method is demonstrated using measured data related to long-span bridge structures. This data includes temperature records, serving as a single predictor, and bridge displacement responses obtained from synthetic aperture radar images as products of remote sensing technology. Results confirm that the proposed method is highly effective and applicable for selecting the best regression model for prediction.

Keywords: Prediction, Structural Displacement, Long-Span Bridge, Regression Model Selection, Correlation Analysis, Remote Sensing Technology.

1. Introduction

Health assessment of critical civil structures is of paramount importance to their owners and stakeholders [1-3]. Recent advancements in sensor technologies have enabled civil engineers to capture a variety of structural responses. Consequently, the increased field measurement

Media and Publishing Partner
https://doi.org/10.58286/29843

of civil structures using diverse sensing systems has received considerable attention. In particular, next-generation sensing technologies, such as those based on smartphones [4], digital cameras [5], spaceborne remote sensing [6] significantly enhance the monitoring complex and huge civil structures.

Although field measurement is an important and practical component of structural health monitoring (SHM) in civil structures, several limitations can hinder the full utilization of this strategy. Firstly, it may not always be feasible to equip civil structures with all possible sensors due to difficult access caused by their geographic conditions. Secondly, harsh weather conditions can interrupt the recording of some structural responses. Thirdly, some next-generation sensors are sensitive to lighting conditions, which can lead to erroneous outputs. Fourthly, budget constraints may prevent civil engineers and researchers from completing field measurements, and costs can deter owners and stakeholders of civil structures from investing in SHM, especially for projects that require long-term monitoring. Some of the aforementioned limitations can be mitigated by benefiting spaceborne remote sensing. Notably, the use of synthetic aperture radar (SAR) images retrieved from certain satellites can overcome the limitations of harsh environments, lighting conditions, and weather fluctuations. For these reasons, SAR-aided SHM has emerged as a practical strategy for health assessment of various civil structures [7-13]. However, this strategy has its own weaknesses. Firstly, it is challenging to provide real-time data (SAR images) for urgent events (i.e., strong earthquakes, floods, typhoons, etc.). Secondly, the large file sizes of remote sensing products, often measured in gigabytes, require considerable storage space, which can be problematic for long-term monitoring programs. Third, the outputs of SAR-based SHM consist of a limited set of displacement responses, which may not encompass all structural properties and fully reflect all external loadings [9].

The remedy for these limitations is to take advantage of artificial intelligence and machine learning algorithms and models for predicting structural responses. In this regard, supervised regression modeling is an effective and reliable approach to data prediction [14]. Generally, the regression-based prediction process involves measuring and collecting all potential predictors (i.e., independent data) and responses (i.e., dependent data), followed by training a regression model. Subsequently, the regressor uses new predictors to forecast unseen response samples. A successful prediction process relies upon selecting a reliable regression model [15]. This selection is determined by the nature of the relationship between the predictors and responses, specifically whether it is linear or nonlinear.

On the other hand, it may not be possible to measure all potential predictors. In SHM, the main predictors (i.e., independent parameters), which affect structural responses include external loadings, environmental factors such as air temperature, humidity, rainfall, and wind. Some external loadings, despite their significant impacts on structural responses, cannot be measured. Additionally, sensor malfunction caused by aging and harsh weather may disrupt reliable measurements of some key environmental parameters leading to potential data missing. The other important issue is the variability in how measured response data are influenced across different locations of a civil structure. At some locations, the response data may be significantly affected by measured environmental and/or operational conditions, while at other locations, the data may be unaffected by these conditions or predominantly influenced by unmeasured factors [14]. Under such circumstances, accurate regression modeling may be challenging and an unsuitable regressor may result erroneous outputs. Therefore, an appropriate regressor selection is not only important for choosing the best model based on the relationship between the measured predictors and responses but also critical for recognizing whether the measured predictors are sufficient or other unmeasured factors affect responses.

The main objective of this paper is to propose a systematic correlation analysis for initially determining the relationship between the measured predictors and responses and

subsequently declaring whether the measurements are sufficient. The basis of the proposed method lies in two correlation coefficient measures; that is, canonical correlation analysis (CCA) and maximal information criterion (MIC). Computing the correlation coefficients, the proposed method defines three correlation cases. To validate the application and effectiveness of this method, measured temperature data of some contact-based temperature sensors and structural displacement responses of large-scale bridge structures are incorporated. Within the framework of long-term SHM plans, responses were derived from some SAR images concerning the remote sensing system. Results demonstrate that the proposed correlation analysis method significantly helps to make an accurate decision on response prediction and implementation of further field measurements.

2. Correlation Metrics

2.1 Canonical Correlation Analysis

In statistics, a correlation analysis presents a useful way for measuring the relationship between at least two variables and indicating their dependency. This approach is capable of identifying how changes in one variable influence variations in another. On this basis, a high correlation rate means that there is a strong relationship or dependency on the two variables. In contrast, a low correlation rate refers to some conditions such as a weak relationship and the effect of other variables that are not considered in the analysis.

Among correlation coefficient measures, the CCA is based on computing linear relationships between two multidimensional variables. This method can be considered as the problem of finding basis vectors for two sets of variables such that the correlation between the projections of the variables onto these basis vectors are mutually maximized [17]. In this case, the canonical correlation vectors are determined by a joint covariance analysis of the two variables [18: Chapter 16]. An important characteristic of the CCA is the feasibility of computing the correlation between two univariate datasets. Given the predictor and response data $\mathbf{x}=\{x_1,\ldots,x_n\}$ and $\mathbf{y}=\{y_1,\ldots,y_n\}$, the main objective of the CCA is to determine canonical scores of these datasets and attempt to describe the possible link between **x** and **y**. On this basis, these scores (i.e., **u** and **v**) can be written as follows:

$$\mathbf{u} = \mathbf{a}^T\mathbf{x} \tag{1}$$

$$\mathbf{v} = \mathbf{b}^T\mathbf{y} \tag{2}$$

where **a** and **b** are the canonical coefficients of the vectors **x** and **y**. The CCA seeks for vectors **a** and **b** such that the relation of the two indices $\mathbf{a}^{\mathrm{T}}\mathbf{x}$ and $\mathbf{b}^{\mathrm{T}}\mathbf{y}$ is quantified in some interpretable way. In this regard, the vectors **a** and **b** are obtained by maximizing the Pearson correlation coefficient between **u** and **v**, which can be expressed as:

$$\rho(\mathbf{u}\,\mathbf{v}) = \frac{\sum_{i=1}^{n}(u_i-\bar{u})(v_i-\bar{v})}{(\sum_{i=1}^{n}(u_i-\bar{u})^2\sum_{i=1}^{n}(v_i-\bar{v})^2)^{\frac{1}{2}}} \tag{3}$$

The canonical variables of **x** and **y** are the linear combinations of the canonical coefficients in **a** and **b,** respectively. Therefore, the canonical correlation between the predictor (**x**) and response (**y**) data is the positive value of the Pearson correlation coefficient between **u** and **v**; that is, CCA(**x**,**y**) = ρ(**u**,**v**). Accordingly, the CCA varies between zero and one so that CCA(**x**,**y**) = 1 implies a high linear correlation between **x** and **y**, whereas CCA(**x**,**y**) = 0 means that no correlation is available.

2.2 Maximal Information Criterion

Linear correlation measures are not able to represent nonlinear relationships between variables. In such cases, the best solution is to benefit nonlinear correlation measures. The MIC is one of the effective nonlinear correlation measures developed by Reshef et al. [19]. It is a statistical approach used to measure the strength in data (i.e., predictors and responses) and their linear or nonlinear correlation degree.

To understand the mathematical underpinning of MIC, it is important to initially understand the concept of mutual information, which is the basis of this statistical approach. In this regard, mutual information is a measure derived from information theory, quantifying the amount of information obtained about one random variable through observing another random variable. Given two random variables **x** (predictor) and **y** (response) with *n* samples, one can express their mutual information (I_{xy}) in the following form:

$$I_{xy} = \sum_{i=1}^{n} p(x_i\, y_i) \log \frac{p(x_i\, y_i)}{p(x_i)p(y_i)} \tag{4}$$

where $p(x,y)$ is the joint probability distribution function of x and y, and $p(x)$ and $p(y)$ are the marginal probability distribution functions of x and y, respectively. MIC benefits this concept by evaluating the mutual information across various grid partitions of the data. This process can be outlined as follows:

1) **Grid Construction**: For each pair of variables, a series of grids of varying sizes are laid over the scatter plot of their joint distribution.
2) **Mutual Information Calculation**: For each grid, one needs to calculate the mutual information of the variables (i.e., the predictor **x** and response **y**) when binned based on the grid.
3) **Normalization**: In this step, it is necessary to normalize the mutual information value by the logarithm of the smaller dimension (i.e., either the number of bins or the total number of samples) to account for the size of the grid. This normalized value is termed as:

$$MIC_g = \frac{I_{xy}}{\log\big(\min(k\, l)\big)} \tag{5}$$

where k and l represent the numbers of bins along each axis of the grid.

4) **Maximization Over Grids**: MIC is then defined as the maximum MIC_g obtained over all possible grids as MIC = max(MIC_g).

For assessing the correlation between the predictor and response data, MIC ranges from 0 to 1. In this regard, a MIC value equal to one suggests a robust linear or nonlinear correlation, while a value close to zero implies an absence of correlation between the predictor and response data.

3. Proposed systematic correlation analysis

The main framework for choosing the best supervised regression model is based on three correlation cases (i.e., *Case I*, *Case II*, and *Case III*) by analyzing the correlation coefficients obtained from the CCA and MIC under three correlation labels; that is, *High* (▲), *Moderate* (■), and *Low* (▼). For these labels, three correlation criteria are defined as follows:

- If the correlation coefficient is larger than 0.8, the correlation label is **High**.
- If the correlation coefficient is smaller than 0.8 and larger than 0.6, the correlation label is **Moderate**.

- If the correlation coefficient is smaller than 0.6, the correlation label is **Low**.

Based on these labels, one can define three correlation cases.

3.1 Case I: High Correlation Coefficients for CCA and MIC

This is the simplest condition in regression modeling for data prediction, for which there exists a high linear correlation between the predictor and response data. Thus, it can be inferred that since the environmental and/or operational conditions affect the response data linearly, linear regression models are well-suited for Case I.

3.2 Case II: Low and High Correlation Coefficients for CCA and MIC

In some conditions, the measured environmental and/or operational factors have nonlinear correlations with the response data, for which linear correlation measures such as CCA fail in indicating the accurate relationship. Under such circumstances, one should evaluate the outputs of nonlinear correlation measures such as MIC. For this case, one can ensure that the linear regression models are not effective and one should apply nonlinear regression models.

3.3 Case III: Different Correlation Coefficients for CCA and MIC

For this case, one can derive three potential conditions:

1) **When both CCA and MIC yield low correlation coefficients (Case III-1)**: This is the most obvious condition for the third case, which can ensure that unmeasured factors certainly impact on the response data. For this case, it is essential to consider new and additional sensing systems for providing further predictors. Moreover, nor linear and nonlinear regressors are suitable for Case III-1.
2) **When CCA and MIC yield low and moderate correlation coefficients (Case III-2)**: In this scenario, we can ensure that there is no linear correlation between the measured environmental and/or operational factors and the measured response data but we hesitate about the nonlinear correlation. Conservatively, it is better to prepare further predictors by using new sensing systems for new measurements.'
3) **When CCA and MIC yield moderate correlation coefficients (Case III-3)**: This is an uncertain and conservative condition of the regression-based prediction problem. For this issue, one can consider two scenarios: (i) benefiting rigorous and robust regressors with nonlinear capabilities, and (ii) conducing new measurements for further predictor preparation. It is worth remarking that rigorous regressors are more complicated models developed from advanced machine learning algorithms. In most cases, hybrid regression models fall in this category that residuals between the measured and predicted responses include information about unmeasured predictors and such residual samples serve as new predictor points.

4. Case Studies

To demonstrate the application of the proposed correlation analysis method, the measured data belonging to two large-scale bridges is considered. Fig. 1 shows the pictures of these bridges. The first case study is a steel arch bridge called the Lupu Bridge, see Fig. 1(b), where is located in Shanghai, China [7]. The second case study known as the Rainbow Bridge, see Fig. 1(c), is a steel arch bridge located in Tianjin, China [7]. For these bridge structures, limited SAR images captured from some satellites were used to extract displacement responses as different areas of these structures. For the Lupu Bridge, 38 SAR images belonging to TerraSAR-X were utilized to obtain the bridge displacement responses at the main girder and the bridge arch between April 16, 2013 to September 10, 2016. In relation

to the Rainbow Bridge, 53 SAR images from Sentinel-A1 were incorporated to determine the displacements at four piers and three main girders since 2015-2017. In both bridge structures, thermocouples were installed to record air temperature based on the paradigm of the contact-based sensing system. Fig. 2 displays the displacement responses as well as the recorded temperature data belonging to the Lupu Bridge. Moreover, Fig. 3 illustrates the same measured data regarding the Rainbow Bridge.

Fig. 1. Case studies: (a) the Lupu Bridge, (b) the Rainbow Bridge

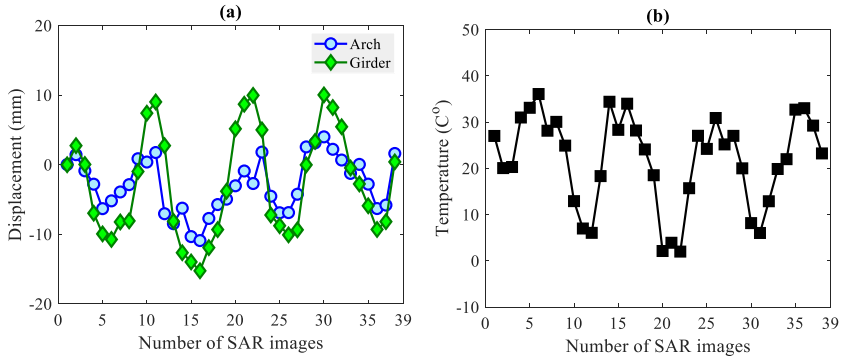

Fig. 2. The Lupu Bridge: (a) SAR-extracted displacement responses at the bridge arch and girder, (b) recorded temperature values

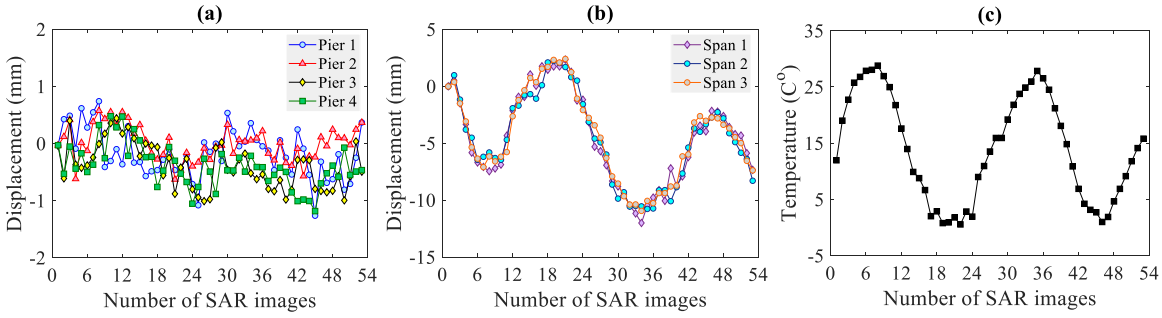

Fig. 3. The Rainbow Bridge: (a)-(b) SAR-extracted displacement responses at the bridge piers and spans, (c) recorded temperature values

5. Results

Using the proposed correlation analysis method, the relationship between each response and predictor data is examined to initially determine whether the ambient temperature is the sole significant environmental factor influencing the response data. Subsequently, it is evaluated to determine how the recorded temperature data influences displacement responses (i.e., in terms of linearity and nonlinearity concerning Case I and Case II) and make sure whether other unmeasured environmental and operational factors are dominant (Case III). At the first

step, the correlation coefficients via the CCA and MIC are computed as shown in Figs. 4-5 regarding the Lupu and the Rainbow Bridges, respectively. In Fig. 4, one can discern that both the CCA and MIC regarding the bridge arch are smaller than 0.6, while these metrics at the bridge girder indicate high correlations (i.e., coefficient larger than 0.8) between the temperature and displacement. From Fig. 5, it is seen that no correlation coefficients are over 0.8 so that some of them fall in 0.6-0.8 and the others are smaller than 0.6.

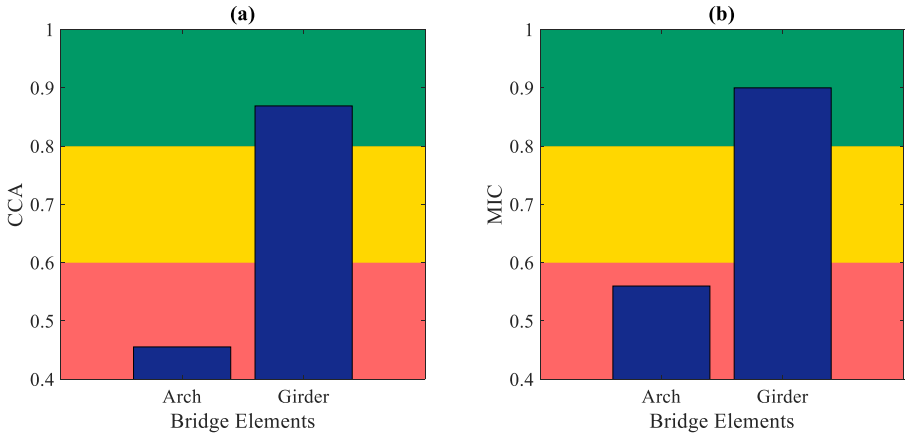

Fig. 4. Correlation analysis for the Lupu Bridge: (a) CCA, (b) MIC

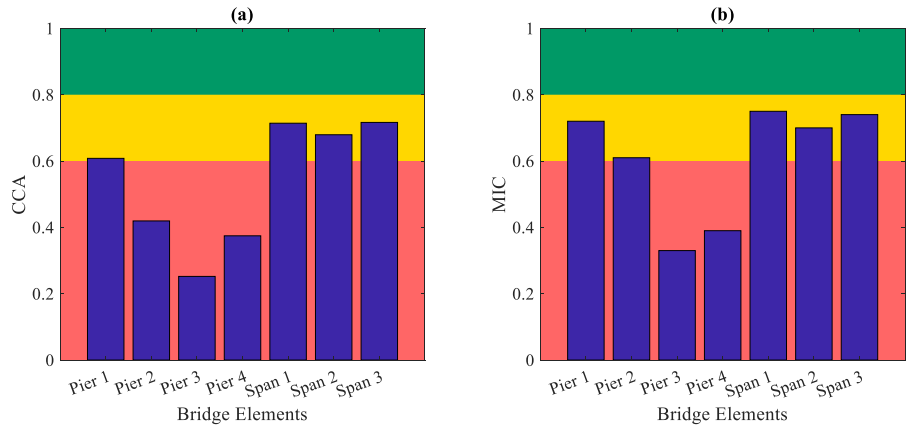

Fig. 5. Correlation analysis for the Rainbow Bridge: (a) CCA, (b) MIC

For more details, Tables 1 and 2 present the outputs of the proposed correlation analysis method for the Lupu and Rainbow Bridges, respectively. Based on Table 1, the correlation level at the arch of the Lupu Bridge is Case III-1 (i.e., the first condition of Case III), which means that unmeasured environmental and operational factors impact on the displacement data; hence, the bridge needs some additional sensors. In other words, the only temperature is not the most influential predictor, in which case nor linear neither nonlinear state-of-the-art regression models can yield reliable prediction outputs. In contrast, it can be ensured that air temperature is the main reason for changes in the displacement response of the bridge girder. This means that it is feasible to apply linear regression models for predicting future displacement responses by inputting new temperature data.

Table 1. The outputs of the proposed correlation analysis method in the Lupu Bridge

Element	Correlation labels		Decision	Suggestion
	CCA	MIC		
Arch	Low (▼)	Low (▼)	Case III-1	New sensors/field measurements
Girder	High (▲)	High (▲)	Case I	Linear regressors

In relation to the Rainbow Bridge, as the outputs in Table 2 appear, the third cases are the final results. For Piers 3 and 4, one should add new sensing systems to measure further environmental and operational factors. Moreover, additional inspections are welcomed to make sure that the bridge did not suffer from any damage. For Piers 1 and 2 as well as Spans 1-3, although it is possible to exploit or develop rigorous and robust regression models, it is better to add sensors for providing further predictor data.

Table 2. The outputs of the proposed correlation analysis method in the Rainbow Bridge

Element	Correlation labels		Decision	Suggestion
	CCA	MIC		
Pier 1	Moderate (■)	Moderate (■)	Case III-3	New sensors/field measurements or rigorous regressors
Pier 2	Low (▼)	Moderate (■)	Case III-2	
Pier 3	Low (▼)	Low (▼)	Case III-1	New sensors/field measurements
Pier 4	Low (▼)	Low (▼)	Case III-1	
Span 1	Moderate (■)	Moderate (■)	Case III-3	New sensors/field measurements or rigorous regressors
Span 2	Moderate (■)	Moderate (■)	Case III-3	
Span 3	Moderate (■)	Moderate (■)	Case III-3	

6. Conclusions

In this paper, a systematic correlation analysis method has been proposed to initially determine the relationship between predictors and responses and then suggest solutions to regression-based prediction. The proposed method has exploited linear and nonlinear correlation metrics, i.e., CCA and MIC, and has defined three cases of correlations. Measured temperature data and SAR-extracted displacement responses of two long-span bridges have been considered as the single predictor and response sets. The results of this paper have indicated that the use of the proposed method can significantly help to make a correct decision on response prediction. Moreover, it is possible to suggest the best solution based on the outputs.

Funding

This research was partially funded by the European Space Agency (ESA) under ESA Contract No. 4000132658/20/NL/MH/ac.

References

[1] Katam, R., Pasupuleti, V. D. K., and Kalapatapu, P., A review on structural health monitoring: past to present. Innovative Infrastructure Solutions 8(9), 248 (2023).

[2] Entezami, A., Sarmadi, H., Behkamal, B., and Mariani, S., Health monitoring of large-scale civil structures: An approach based on data partitioning and classical multidimensional scaling. Sensors 21(5), 1646 (2021).

[3] Daneshvar, M. H. and Sarmadi, H., Unsupervised learning-based damage assessment of full-scale civil structures under long-term and short-term monitoring. Engineering Structures 256, 114059 (2022).

[4] Sarmadi, H., Entezami, A., Yuen, K.-V., and Behkamal, B., Review on smartphone sensing technology for structural health monitoring. Measurement 223, 113716 (2023).

[5] Spencer, B. F., Hoskere, V., and Narazaki, Y., Advances in Computer Vision-Based Civil Infrastructure Inspection and Monitoring. Engineering 5(2), 199-222 (2019).

[6] Ge, P., Gokon, H., and Meguro, K., A review on synthetic aperture radar-based building damage assessment in disasters. Remote Sensing of Environment 240, 111693 (2020).

[7] Qin, X., Zhang, L., Yang, M., Luo, H., Liao, M., and Ding, X., Mapping surface deformation and thermal dilation of arch bridges by structure-driven multi-temporal DInSAR analysis. Remote Sensing of Environment 216, 71-90 (2018).
[8] Behkamal, B., Entezami, A., De Michele, C., and Arslan, A. N., Elimination of thermal effects from limited structural displacements based on remote sensing by machine learning techniques. Remote Sensing 15(12), 3095 (2023).
[9] Entezami, A., De Michele, C., Arslan, A. N., and Behkamal, B., Detection of partially structural collapse using long-term small displacement data from satellite images. Sensors 22(13), 4964 (2022).
[10] Milillo, P., Giardina, G., DeJong, M. J., Perissin, D., and Milillo, G., Multi-Temporal InSAR Structural Damage Assessment: The London Crossrail Case Study. Remote Sensing 10(2), 287 (2018).
[11] Entezami, A., Behkamal, B., and De Michele, C., Advanced ML Methods: Bridging SAR Images and Structural Health Monitoring, in *Long-Term Structural Health Monitoring by Remote Sensing and Advanced Machine Learning: A Practical Strategy via Structural Displacements from Synthetic Aperture Radar Images* (pp. 29-68). Cham: Springer Nature Switzerland (2024).
[12] Di Carlo, F., Miano, A., Giannetti, I., Mele, A., Bonano, M., Lanari, R., Meda, A., and Prota, A., On the integration of multi-temporal synthetic aperture radar interferometry products and historical surveys data for buildings structural monitoring. Journal of Civil Structural Health Monitoring 11(5), 1429-1447 (2021).
[13] Entezami, A., Behkamal, B., and De Michele, C., Pioneering Remote Sensing in Structural Health Monitoring, in *Long-Term Structural Health Monitoring by Remote Sensing and Advanced Machine Learning: A Practical Strategy via Structural Displacements from Synthetic Aperture Radar Images* (pp. 1-27). Cham: Springer Nature Switzerland (2024).
[14] Behkamal, B., Entezami, A., De Michele, C., and Arslan, A. N., Investigation of temperature effects into long-span bridges via hybrid sensing and supervised regression models. Remote Sensing 15(14), 3503 (2023).
[15] Sarmadi, H., Behkamal, B., and Entezami, A., Prediction of long-term dynamic responses of a heritage masonry building under thermal effects by automated kernel-based regression modeling, in *Artificial Intelligence Applications for Sustainable Construction*, M.L. Nehdi, et al., Editors, Woodhead Publishing. p. 257-283. 2024.
[16] Entezami, A., Sarmadi, H., and Behkamal, B., Short-term damage alarming with limited vibration data in bridge structures: A fully non-parametric machine learning technique. Measurement (2024), in press.
[17] Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J., Canonical Correlation Analysis: An Overview with Application to Learning Methods. Neural Computation 16(12), 2639-2664 (2004).
[18] Härdle, W. K. and Simar, L., Applied Multivariate Statistical Analysis. Springer International Publishing (2019).
[19] Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., and Sabeti, P. C., Detecting novel associations in large data sets. Science 334(6062), 1518-1524 (2011).