Effects of environmental conditions on historic buildings: interpretable versus accurate exploratory data analysis

Marco Parola[1][a], Hajar Dirrhami[2][b], Mario G.C.A. Cimino[1][c] and Nunziante Squeglia[2][d]
[1]*Dept of Information Engineering, University of Pisa, 56122 Pisa, Italy*
[2]*Dept of Civil and Industrial Engineering, University of Pisa, 56122 Pisa, Italy*
{*marco.parola, mario.cimino, nunziante.squeglia*}*@ing.unipi.it*

Keywords: Structural Health Monitoring, Leaning Tower of Pisa, Regression Analysis, Deep Learning, Interpretability

Abstract: The goal of structural health monitoring is to continuously assess the structural integrity and performance of a building or structure over time. This is achieved by collecting data on various structural parameters and using this data to identify potential areas of concern or damage. A critical challenge involves some properties being severely damaged by recurrent variations of external factors. These variations in environmental and operational conditions (such as humidity, temperature, and traffic) can deflect the variability in structural behavior caused by structural damage and make it difficult to identify the damage of interest. In this paper, we present a study on how regression analysis and deep learning can be used to measure the influence of environmental factors on the structural behavior of the Leaning Tower of Pisa. Transparent linear regressors offer the benefit of being simple to understand and interpret. They can provide insights about the relationship between input and target variables, as well as the relative importance of each input in forecasting the outcome. On the other hand, deep learning models are capable of learning nonlinear relationships between input and target variables. Definitively, in this work the accuracy-interpretability trade-off for structural health monitoring is discussed.

1 INTRODUCTION

In order to diagnose and assess the stability condition of historic monuments, Structural Health Monitoring (SHM) is crucial. Indeed, such structures are constantly exposed to the action of environmental effects, such as sun's radiation, temperature variations or wind motion, that eventually have a tendency to lose their structural integrity. It is therefore essential to ensure precautions and perform proper analyses to prevent any deterioration of these monuments and preserve culture.

One of the most common analysis consists of detecting whether the structure under examination is affected by damage. This information is fundamental and useful for structural engineers to undertake restoration measures and avoid unintended consequences. However, in SHM, this information is often not enough and other tasks are available to provide more specific additional information, such as the region of the structure where the damage is present by addressing a damage location task, or measuring damage degree by performing a damage quantification task (Parola. et al., 2022).

All of these tasks enable for the information collecting needed to evaluate the scenario severity to which a structure may be subjected, attempting to assess any changes in the sensitive features indicative of damage. Unfortunately, there are additional variation sources, such as variations brought on by environmental factors. If these impacts are not considered, they can result in incorrect damage diagnosis or less accurate injury detection. Differentiating between the two sources of variance in static or dynamic characteristics is crucial (Kullaa., 2014).

Effective monument maintenance, which distinguishes symptomatic changes of damage from environmental ones, relies on a framework composed of two main aspects: (i) the gathering of data and (ii) the processing of them to obtain useful information about the structure. The collection of data describing the structure exploits an SHM system involving the use of sensors installed directly at the building to measure its properties. SHM systems are divided into two groups based on the kind of parameters they

[a] https://orcid.org/0000-0003-4871-4902
[b] https://orcid.org/0000-0000-0000-0000
[c] https://orcid.org/0000-0002-1031-1959
[d] https://orcid.org/0000-0001-8104-503X

acquire: (i) Static systems, which track the temporal development of variables that change gradually over time (such as wall slopes or crack widths) by periodically sampling sensor devices; (ii) dynamic systems, which track vibrational variables like speeds or accelerations in order to collect information on broader dynamic features like natural frequencies intrinsic to the structure or modal forms (Zonzini et al., 2020).

About how the data are processed, Machine Learning (ML) and later Deep Learning (DL) were progressively introduced to enhance the analysis instruments adopted to solve these problems (Sujith et al., 2022), as they exploit a data-driven approach. These data-driven methods can be relied upon a supervised or unsupervised learning. The former makes use of labelled input-output pairs, where the input is structural response, while the value of the target structural parameters are the corresponding outputs. On the other side, instead of requiring an associated output label, unsupervised algorithms are frequently used to find damage-sensitive patterns or similarities in initial data (Cimino. et al., 2022). Specifically, transparent machine learning models have the advantage of being simple to understand. As an example, modular neural architectures refer to a design approach based on a collection of small neural (Cimino et al., 2009). However, such constrained architectures are limited in capturing input-output complex patterns.

Regression analysis is considered a supervised learning problem in the field of machine learning because it involves training a model on labeled data, where the true values of the output variable are known. In the SHM domain, regression analysis can be adopted to model the problem of measuring the influence of environmental conditions on the health and the structural behavior of a building (Farreras-Alcover et al., 2015) (Dervilis et al., 2015).

The case study investigated in this paper is the Leaning Tower of Pisa located in the Miracle Square in Tuscany. The SHM system installed on the Leaning Tower operates as a static system whereby, upon activation, it records a single value for each sensor. The system is programmed to activate on an hourly basis. However, since the monitoring system has been installed in 1993, the sampling frequency has varied by the service staff. To overcome this problem, time series resampling techniques can be adopted. The locations where the sensors have been installed on the tower are illustrated in Figure 1 and Table 1.

The novel contribution of this work is to explore the influence of the environmental conditions on a specific historical monument: the Leaning Tower of Pisa. Such analysis is performed through two different regressive methods; more specifically, we go deeper by measuring the impact of these conditions on the individual sensors.

The paper is organized as follows: Section 2 outlines the methods and methodologies we used to analyze the sensor data, while Section 3 provides an overview of the data discovery process. Section 4 describes the conducted experiments and displays the results achieved by the two compared models. Finally, Section 6 summarizes the work and addresses the conclusions.

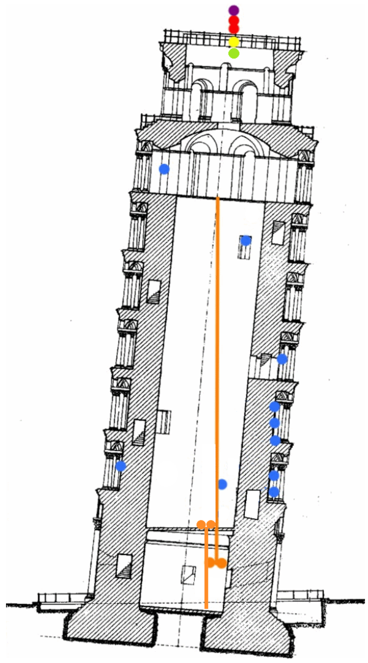

Figure 1: Sensor locations on the tower, as indicated by the color legend in Table 1

Table 1: Sensor system installed on the leaning tower

Sensor	#	Leg	Thresholds
Deformometer (D)	10	● (blue)	[-0.5,0.5] mm
Telecoordinom. (T)	4	● (orange)	[-2.1,1.8]* ”
Termometer (TM)	1	● (yellow)	[-10,42] °C
Wind speed (WS)	1	● (red)	[0,45] m/s
Wind dir (WD)	1	● (red)	[0,360] deg
Pressure (P)	1	● (light green)	[1, 1.04]*hPa
Solar radiation (SR)	1	● (purple)	[0,1]* W/m^2

$^* \cdot 10^3$

2 METHODS

Our methodology is structured in two main parts: data discovery and regression analysis. In the data discovery phase, we explore the data identifying aspects. This includes study the linear dependencies between environmental and operational factors through a correlation analysis. Indeed, by introducing the correlation matrix, correlation degree between pairs of variables in a dataset can be computed, allowing the identification of which ones are positively or negatively correlated, and how strongly they are related to each other.

During the regression analysis phase, we use regressive techniques to model the relationship between the dependent variable and one or more independent variables. This enables to identify key factors influencing the dependent variable and make predictions about future outcomes. Combining these two strategies provides a comprehensive and rigorous approach to investigating our research purpose and generating interesting insights.

A linear regression problem aims to find the line that best fits the data, which is expressed through the equation.:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{1}$$

Where $x1, x2, ..., xn$ are the independent variables and the goal is to find the best coefficients values (beta1, beta2, ..., beta_n) that minimize the sum of the squared differences between the predicted values of y and the actual values of y.

In non-linear regression, the relationship between the independent and dependent variables is modeled using a non-linear function, such as a polynomial or exponential function.

Overall, regression analysis is a powerful tool for understanding and predicting the relationship between variables. It can be used for a wide range of applications, such as the measurement of the influence of environmental factors on a structure's behavior (Farreras-Alcover et al., 2015) (Dervilis et al., 2015).

The goal of this strategy is to understand how environmental factors, such as temperature, humidity, wind, and precipitation, can affect the structural integrity of a building over time.

Gathering information on the environmental variables and the structural behavior of a building, such as displacement, strain, or vibration measurements, is one technique to employ regression analysis in this context. The operational quantities can then be predicted using a regression model built using this data and the environmental parameters.

Against the traditional statistical models, Neural Networks (NNs) have proven to be an effective solution for regression problems in many contexts. They offer several advantages for regression problems, such as the ability to learn complex non-linear relationships and the ability to handle large and noisy datasets.

The methodology entails resolving a regression problem and subsequently evaluating the performance of two distinct models using a chosen metric.

We introduce performance metrics to evaluate the performance of our regression model: mean squared error *mse* and coefficient of determination R^2. The *mse* measures the average squared difference between the predicted and actual values of the response variable as shown in Equation 2.

$$mse = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

where n is the number of observations, y_i is the actual value of the response variable for the ith observation, and $\hat{y}_i$ is the predicted value of the response variable for the ith observation.

On the other hand, the R^2 value measures the proportion of variation in the response variable explained by the regression model. The equation for R^2 is shown below:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{3}$$

where SS_{res} is the residual sum of squares, which is the sum of the squared differences between the actual and predicted values of the response variable, and SS_{tot} is the total sum of squares, which is the sum of the squared differences between the actual values of the response variable and the mean value of the response variable.

3 DATA PREPROCESSING AND MINING

In this section, we present the data preprocessing and mining techniques we applied to a sensor device dataset. The goal of this process is to extract valuable insights and knowledge from the raw data, which would then be used for further analysis in the next section.

3.1 Data preprocessing

This section presents the dataset collected from a static monitoring system. It hourly records samples by enabling continuous monitoring and tracking of environmental changes. The dataset covers a period of two months, from May to June 2020, during which the monitoring system captured the 19 parameters listed in Table 1. The collected data underwent a series of preprocessing steps to ensure its quality and consistency. The data preprocessing pipeline is composed of the following steps:

- *detection of out-of-scale samples*, based on upper and lower thresholds;
- *normalization*, to mitigate the phenomena of the curse of dimensionality by applying the z-score scaling;
- *statistical anomalies detection*, where observations with an actual value farther than ± 2.5 from the data distribution of the current moving average are dropped;
- *missing samples recovery*, short sequences of successive missing samples are reconstructed for isolated missing samples, by using a linear interpolation between the closest neighbors for the outliers that have already been identified (at most 4 samples). Long missing value sequences, such as those caused by device failure, are not recreated.
- *temporal data resampling*, since the sampling rate was changed over time during the entire monitoring period, the entire time series was resampled to 1-hour frequency.

3.2 Data mining

In order to avoid redundancy and be more concise, we present a subset of deformometers given the common trends among them. Figure 2 displays the partial correlation matrix between deformometers 4, 5, 11, 12, the four telecoordinometers with the five environmental factors; for space reasons we have not included the full matrix.

The highest correlation (absolute) values are related to the temperature, exept for north-south direction of both the telecoordinometers (*T1* and *T3*).

We can observe a reduction in the value measured by all the deformometers and the two east-west direction telecoordinometers, when the temperature increases. For the deformometer, an increase in temperature causes an expansion of the Tower, which will then tend to solicit the man-made cracks in which the sensors are installed, so the measured value decreases.

Regarding the telecoordinometer, the correlation is due to the fact that temperature causes not only an horizontal expansion, but also in the vertical direction; this expansion-contraction cycles causes movement in the Tower, which is then perceived by the sensor that measures the inclination in the east-west direction. The correlation related to solar radiation is also negative and strongly linked to the previous observations, considering the relationship between radiation and temperature.

For wind speed and direction, the correlation values are extremely close to zero, suggesting a low impact of the wind condition on the structure.

Finally, pressure is the only factor showing a positive correlation with the operational factors; in this case, as the pressure increases, the values measured by the deformometer and telecoordinometer sensors increase, although with a low correlation value.

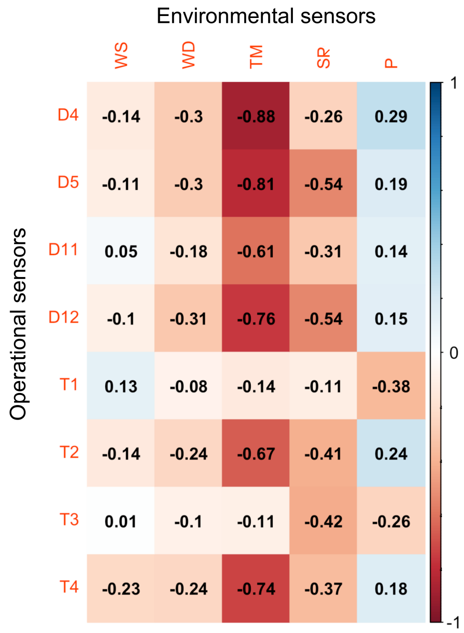

Figure 2: Partial correlation matrix - environmental sensors against operational sensors

To further explore the data discovery phase, we conduct a regression analysis over the entire monitoring period considered in this work. Equations 5, 6, 7 and 8 describe linear models of the relationship between environmental factors and deformometers *D4*, *D5*, *D11* and *D12*, respectively; while equations 9, 10, 11 and 12 similarly describe the dependence of telecoordinometers *T1*, *T2*, *T3* and *T4*, respectively.

This analysis provides valuable insights into our data, allowing to better understand how different sensors interact with each other and how they are affected by environmental factors. By examining these relationships, we can identify patterns and correlations in the data that may not be obvious from the visual inspection.

Correlation and linear regression analyses performed previously show a relevant dependence of op-

$$D1 = (4.43\text{e}{-4})\ WS - (1.39\text{e}{-6})\ WD - (2.81\text{e}{-3}) TM + (2.54\text{e}{-6})\ SR + (2.40\text{e}{-4})\ P \tag{4}$$

$$D4 = (4.43\text{e}{-4})\ WS - (1.39\text{e}{-6})\ WD - (2.81\text{e}{-3}) TM + (2.54\text{e}{-6})\ SR + (2.40\text{e}{-4})\ P \tag{5}$$

$$D5 = (3.43\text{e}{-4})\ WS - (9.45\text{e}{-7})\ WD - (1.39\text{e}{-5}) TM + (1.25\text{e}{-5})\ SR + (7.46\text{e}{-5})\ P \tag{6}$$

$$D11 = (7.67\text{e}{-4})\ WS + (9.7\text{e}{-7})\ WD - (1.23\text{e}{-3}) TM - (9.51\text{e}{-6})\ SR + (1.49\text{e}{-4})\ P \tag{7}$$

$$D12 = (2.75\text{e}{-4})\ WS - (3.91\text{e}{-6})\ WD - (1.01\text{e}{-3}) TM - (1.09\text{e}{-5})\ SR + (3.39\text{e}{-5})\ P \tag{8}$$

$$T1 = (4.10\text{e}{-2})\ WS - (1.19\text{e}{-3})\ WD - (3.17\text{e}{-2}) TM - (8.59\text{e}{-4})\ SR - (4.37\text{e}{-2})\ P \tag{9}$$

$$T2 = -(2.74\text{e}{-2})\ WS - (3.83\text{e}{-5})\ WD - (1.30\text{e}{-1}) TM - (2.17\text{e}{-3})\ SR + (1.46\text{e}{-2})\ P \tag{10}$$

$$T3 = (5.80\text{e}{-3})\ WS - (4.27\text{e}{-4})\ WD + (3.09\text{e}{-2}) TM - (2.18\text{e}{-3})\ SR - (2.36\text{e}{-2})\ P \tag{11}$$

$$T4 = -(2.88\text{e}{-2})\ WS + (6.03\text{e}{-4})\ WD - (2.18\text{e}{-1}) TM - (9.92\text{e}{-4})\ SR + (6.59\text{e}{-3})\ P \tag{12}$$

erational sensors on environmental factors. However, several studies show that the relationship between environmental conditions and structural parameters can introduce some nonlinear components (Hsu and Loh, 2010) (Shi et al., 2016). Therefore, in the experiments section we also explore a DL regression architecture to model the nonlinear components of the relationship between environmental and operational sensors.

4 EXPERIMENTS AND RESULTS

The whole machine learning pipeline has been implemented on Google Colaboratory (Colab)(Bisong, 2019), an open web platform design on top of the open-source Jupyter project. The virtual machines on which Colab is run are powered by a NVIDIA Tesla K80 GPU cards. The source code has been publicly released and can be accessed at (Parola, 2023).

This section shows the results achieved by the two regressive models previously described by presenting the value of evaluation metrics.

A single neuron with a linear activation function implemented the linear regressor (Jolivet et al., 2008), while the second model is a deep NN with a hidden layer containing 16 neurons. Both models have five neurons in the input layer, as the number of environmental sensors. The training phase has been run for 150 epochs where the Adam algorithm has been used as the optimizer and an early stopping condition was set to prevent overfitting with a patience value equal to 10.

Tables 2 and 3 show the evaluation metrics values of *mse* and R^2 obtained from the linear regressor model and NN model for the deformometer and telecoordinometer sensors, respectively.

Table 2: Deformometer mse and R^2

	LR		NN	
sensor	mse	R^2	mse	R^2
D1	.277	.719	.206	.791
D2	.242	.754	.188	.809
D3	.258	.737	.186	.810
D4	.202	.794	.168	.829
D5	.303	.551	.222	.774
D6	.241	.754	.190	.806
D8	.201	.797	.152	.847
D9	.431	.549	.237	.752
D11	.509	.482	.374	.619
D12	.299	.694	.239	.761
Mean	**.287**	**.679**	**.208**	**.838**

Table 3: Telecoordinometer mse and R^2

	LR		NN	
sensor	mse	R^2	mse	R^2
T1	.705	.290	.489	.507
T2	.568	.424	.479	.514
T3	.721	.280	.436	.565
T4	.716	.281	.542	.455
Mean	**.677**	**.318**	**.486**	**.510**

From above tables it is evident how the deep learning based strategy is more effective in estimating the influence of environmental effects on the structural behavior of the tower of Pisa; indeed, we can observe for each table row the *mse* value is higher for the linear regressor column compared with the DL model.

Specifically, both models are able to accurately forecast the deformometer sensor group behave, with a R^2 mean values for the linear regression being 67.9% and the NNs being 83.8%. The *mse* mean value of the deformometers is 28.7% using linear regressor and 20.8% using NN. As a result, we find that the deep learning architecture outperforms the linear models by a factor of 27%, computed as the error of the first method minus the error of the second one over the error committed by the first.

Telecoordinometer sensor modeling does not exhibit the same effectiveness, as the NN R^2 has a low value of 51.0%; while the performance of the linear regression is significantly poorer with an R^2 value of 31.8%.

5 DISCUSSION

Empirical results clearly denote how NNs outperform a linear regression approach in modeling operational sensors depending on environmental factors. However, the linear regression strategy may be preferred to NNs due to the lack of explainability of DL, which is considered a black-box approach (Guidotti et al., 2018).

By analyzing the linear regression coefficients, we can identify the environmental factors having the most significant impact on sensor measurements can be identify and a quantitative indication of them can be measured. This information can then be used to develop correction factors that take environmental influence into account and improve the accuracy of the monitoring system by calculating the corrected features adjusted from environmental effects (Roberts et al., 2023).

6 CONCLUSIONS

In this work, two regressive techniques to estimate the influence of environmental condition on structural behavior have been designed and compared, after a data mining phase to explore the time series data. The sensor network data of the leaning Tower of Pisa have been chosen as case study to implement the methodology.

In conclusion, transparent regression models may not be able to detect complex patterns in the data but have the benefit of being easy to understand and requiring less computing capabilities. Although deep learning models may capture complicated patterns, they can be challenging to interpret and need a lot of computational resources and training data. The choice between transparent regression models and deep learning models ultimately depends on various specific challenges of the problem and historical building to monitor: ranging from logical models to scoring systems. In any exploratory data analysis different models co-exist. Future research efforts aimed at establishing interconnections between different models could be founded on model-centric explanations derived from ontologies, which serve as standardized representations. This approach has the potential to help both system designers and users make systematic connections between explanations and their respective data sets and models.

ACKNOWLEDGEMENTS

This work has been partially carried out in the framework of the PRA_2022_101 project "Decision Support Systems for territorial networks for managing ecosystem services", funded by the University of Pisa. This work has been partially supported by the Tuscany Region in the framework of the "SecureB2C" project, POR FESR 2014-2020, Law Decree 7429 31.05.2017. Work partially supported by the Italian Ministry of Education and Research (MIUR) in the framework of the FoReLab project (Departments of Excellence).

REFERENCES

Bisong, E. (2019). *Building machine learning and deep learning models on Google cloud platform*. Springer.

Cimino, M. G., Pedrycz, W., Lazzerini, B., and Marcelloni, F. (2009). Using multilayer perceptrons as receptive fields in the design of neural networks. *Neurocomputing*, 72(10):2536–2548.

Cimino., M. G. C. A., Galatolo., F. A., Parola., M., Perilli., N., and Squeglia., N. (2022). Deep learning of structural changes in historical buildings: The case study of the pisa tower. In *Proceedings of the 14th International Joint Conference on Computational Intelligence - NCTA,*, pages 396–403. INSTICC, SciTePress.

Dervilis, N., Worden, K., and Cross, E. (2015). On robust regression analysis as a means of exploring environmental and operational conditions for shm data. *Journal of Sound and Vibration*, 347:279–296.

Farreras-Alcover, I., Chryssanthopoulos, M. K., and Andersen, J. E. (2015). Regression models for structural health monitoring of welded bridge joints based on temperature, traffic and strain measurements. *Structural Health Monitoring*, 14(6):648–662.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42.

Hsu, T.-Y. and Loh, C.-H. (2010). Damage detection accommodating nonlinear environmental effects by nonlinear principal component analysis. *Structural Control and Health Monitoring: The Official Journal of the International Association for Structural Control and Monitoring and of the European Association for the Control of Structures*, 17(3):338–354.

Jolivet, R., Schürmann, F., Berger, T. K., Naud, R., Gerstner, W., and Roth, A. (2008). The quantitative single-neuron modeling competition. *Biological cybernetics*, 99:417–426.

Kullaa., J. (2014). Structural health monitoring under nonlinear environmental or operational influences. *Shock and Vibration*, 2014.

Parola, M. (2023). Github data-2023 repository, https://github.com/marcoparola/torre-clima.

Parola., M., Galatolo., F. A., Torzoni., M., Cimino., M. G. C. A., and Vaglini., G. (2022). Structural damage localization via deep learning and iot enabled digital twin. In *Proceedings of the 3rd International Conference on Deep Learning Theory and Applications - DeLTA,*, pages 199–206. INSTICC, SciTePress.

Roberts, C., Cava, D. G., and Avendaño-Valencia, L. D. (2023). Addressing practicalities in multivariate nonlinear regression for mitigating environmental and operational variations. *Structural Health Monitoring*, 22(2):1237–1255.

Shi, H., Worden, K., and Cross, E. (2016). A nonlinear cointegration approach with applications to structural health monitoring. In *Journal of Physics: Conference Series*, volume 744, page 012025. IOP Publishing.

Sujith, A., Sajja, G. S., Mahalakshmi, V., Nuhmani, S., and Prasanalakshmi, B. (2022). Systematic review of smart health monitoring using deep learning and artificial intelligence. *Neuroscience Informatics*, 2(3):100028.

Zonzini, F., Malatesta, M. M., Bogomolov, D., Testoni, N., Marzani, A., and De Marchi, L. (2020). Vibration-based shm with upscalable and low-cost sensor networks. *IEEE Transactions on Instrumentation and Measurement*, 69(10):7990–7998.