Ph.D. Dissertation

A Statistical Methodology for Classifying Time Series in the Context of Climatic Data

Author
Sandra Milena Ramírez Buelvas

Ph.D. Supervisors
Manuel Zarzo Castelló
Fernando Juan García-Diego

A doctoral thesis submitted to Department of Applied Statistics, Operational Research, and Quality

Universitat Politècnica de València

Valencia, December, 2021

To my God

Acknowledgements

I want to show my deep gratitude to professors Manuel Zarzo, Fernando García and Ángel Perles for their support and advice in the process of the development of the thesis. To Ramiro, my family and my friends for being there when I needed them. I would like to give special mention to Miguel Dawson for checking my English writing and helping me to improve my English. I gratefully acknowledge the financial support of Pontificia Universidad Javeriana Cali – PUJ and Instituto Colombiano de Crédito Educativo y Estudios Técnicos en el Exterior – ICETEX who awarded me the scholarships '*Convenio de Capacitación para Docentes O. J. 086/17*' and '*Programa Crédito Pasaporte a la Ciencia ID 3595089 foco-reto salud*' respectively. The scholarships were essential for obtaining the Ph.D. Also, I gratefully acknowledge the financial support of the European Union's Horizon 2020 research and innovation programme under grant agreement No. 814624. Finally, to the people who work at PUJ or ICETEX that made decisions to ensure I had the conditions to work on my thesis.

Abstract

According to different European Standards and several studies, it is necessary to monitor and analyze the microclimatic conditions in museums and similar buildings, with the goal of preserving artworks. With the aim of offering tools to monitor the climatic conditions, a new statistical methodology for classifying time series of different climatic parameters, such as relative humidity and temperature, is proposed in this dissertation.

The methodology consists of applying a classification method using variables that are computed from time series. The two first classification methods are versions of known sparse methods which have not been applied to time dependent data. The third method is a new proposal that uses two known algorithms. These classification methods are based on different versions of sparse partial least squares discriminant analysis PLS (sPLS-DA, SPLSDA, and sPLS) and Linear Discriminant Analysis (LDA). The variables that are computed from time series, correspond to parameter estimates from functions, methods, or models commonly found in the area of time series, e.g., seasonal ARIMA model, seasonal ARIMA-TGARCH model, seasonal Holt-Winters method, spectral density function, autocorrelation function (ACF), partial autocorrelation function (PACF), moving range (MR), among others functions. Also, some variables employed in the field of astronomy (for classifying stars) were proposed.

The methodology proposed consists of two parts. Firstly, different variables are computed applying the methods, models or functions mentioned above, to time series. Next, once the variables are calculated, they are used as input for a classification method like sPLS-DA, SPLSDA, or SPLS with LDA (new proposal). When there was no information about the clusters of the different time series, the first two components from principal component analysis (PCA) were used as input for k-means method for identifying possible clusters of time series. In addition, results from random forest algorithm were compared with results from sPLS-DA.

This study analyzed three sets of time series of relative humidity or temperate, recorded in different buildings (Valencia's Cathedral, the archaeological site of L'Almoina, and the baroque church of Saint Thomas and Saint Philip Neri) in Valencia, Spain. The clusters of the time series were analyzed according to different zones or different levels of the sensor heights, for monitoring the climatic conditions in these buildings.

Random forest algorithm and different versions of sparse PLS helped identifying the main variables for classifying the time series. When comparing the results from sPLS-DA and random forest, they were very similar for variables from seasonal Holt-Winters method and functions which were applied to the time series. The results from sPLS-DA were easier to interpret than results from random forest. When the different versions of sparse PLS used variables from seasonal Holt-Winters method as input, the clusters of the time series were identified effectively.

The variables from seasonal Holt-Winters helped to obtain the best, or the second best results, according to the classification error rate. Among the different versions of sparse PLS proposed, sPLS with LDA helped to classify time series using a fewer number of variables with the lowest classification error rate.

We propose using a version of sparse PLS (sPLS-DA, or sPLS with LDA) with variables computed from time series for classifying time series. For the different data sets studied, the methodology helped to produce parsimonious models with few variables, it achieved satisfactory discrimination of the different clusters of the time series which are easily interpreted. This methodology can be useful for characterizing and monitoring microclimatic conditions in museums, or similar buildings, for preventing problems with artwork.

Resumen

De acuerdo con las regulaciones europeas y muchos estudios científicos, es necesario monitorear y analizar las condiciones microclimáticas en museos o edificios, para preservar las obras de arte que se exponen en ellos. Con el objetivo de ofrecer herramientas para el monitoreo de las condiciones climáticas en este tipo de edificios, en esta tesis doctoral se propone una nueva metodología estadística para clasificar series temporales de parámetros climáticos como la temperatura y humedad relativa.

La metodología consiste en aplicar un método de clasificación usando variables que se computan a partir de las series de tiempos. Los dos primeros métodos de clasificación son versiones conocidas de métodos sparse PLS que no se habían aplicado a datos correlacionados en el tiempo. El tercer método es una nueva propuesta que usa dos algoritmos conocidos. Los métodos de clasificación se basan en diferentes versiones de un método sparse de análisis discriminante de mínimos cuadrados parciales PLS (sPLS-DA, SPLSDA y sPLS) y análisis discriminante lineal (LDA). Las variables que los métodos de clasificación usan como input, corresponden a parámetros estimados a partir de distintos modelos, métodos y funciones del área de las series de tiempo, por ejemplo, modelo ARIMA estacional, modelo ARIMA-TGARCH estacional, método estacional Holt-Winters, función de densidad espectral, función de autocorrelación (ACF), función de autocorrelación parcial (PACF), rango móvil (MR), entre otras funciones. También fueron utilizadas algunas variables que se utilizan en el campo de la astronomía para clasificar estrellas.

La metodología propuesta consta de dos partes. La primera consiste en calcular variables a partir de las series de tiempos, usando los métodos, modelos y funciones mencionadas anteriormente. La segunda parte consiste en usar las variables calculadas en el primer paso para ajustar alguno de los siguientes modelos: sPLS-DA, SPLSDA, o sPLS con LDA (nueva propuesta). En los casos que a priori no hubo información de los clusters de las series de tiempos, las dos primeras componentes de un análisis de componentes principales (PCA) fueron utilizadas por el algoritmo k-means para identificar posibles clusters de las series de tiempo. Adicionalmente, los resultados del método sPLS-DA fueron comparados con los del algoritmo random forest.

Tres bases de datos de series de tiempos de humedad relativa o de temperatura fueron analizadas. Estas series de tiempos se registraron en diferentes edificios (la Catedral de Valencia, el yacimiento arqueológico de L'Almoina y la iglesia barroca de Santo Tomás y San Felipe Neri) en Valencia, España. Los clusters de las series de tiempos se analizaron de acuerdo a diferentes zonas o diferentes niveles de alturas donde fueron instalados sensores para el monitoreo de las condiciones climáticas en los edificios.

El algoritmo random forest y las diferentes versiones del método sparse PLS fueron útiles para identificar las variables más importantes en la clasificación de

las series de tiempos. Los resultados de sPLS-DA y random forest fueron muy similares cuando se usaron como variables de entrada las calculadas a partir del método Holt-Winters o a partir de funciones aplicadas a las series de tiempo. Aunque los resultados del método random forest fueron levemente mejores que los encontrados por sPLS-DA en cuanto a las tasas de error de clasificación, los resultados de sPLS-DA fueron más fáciles de interpretar.

Cuando las diferentes versiones del método sparse PLS utilizaron variables re sultantes del método Holt-Winters, los clusters de las series de tiempo fueron mejor discriminados. Entre las diferentes versiones del método sparse PLS, la versión sPLS con LDA obtuvo la mejor discriminación de las series de tiempo, con un menor valor de la tasa de error de clasificación, y utilizando el menor o segundo menor número de variables.

En esta tesis doctoral se propone usar una versión sparse de PLS (sPLS-DA, o sPLS con LDA) con variables calculadas a partir de series de tiempo para la clasificación de éstas. Al aplicar la metodología a las distintas bases de datos estudiadas, se encontraron modelos parsimoniosos, con pocas variables, y se obtuvo una discriminación satisfactoria de los diferentes clusters de las series de tiempo con fácil interpretación. La metodología propuesta puede ser útil para caracterizar las distintas zonas o alturas en museos o edificios históricos de acuerdo con sus condiciones climáticas, con el objetivo de prevenir problemas de conservación con las obras de arte.

Resum

D'acord amb les regulacions europees i molts estudis científics, és necessari monitorar i analitzar les condiciones microclimàtiques en museus i en edificis similars, per a preservar les obres d'art que s'exposen en ells. Amb l'objectiu d'oferir eines per al monitoratge de les condicions climàtiques en aquesta mena d'edificis, en aquesta tesi es proposa una nova metodologia estadística per a classificar sèries temporals de paràmetres climàtics com la temperatura i humitat relativa.

La metodologia consisteix a aplicar un mètode de classificació usant variables que es computen a partir de les sèries de temps. Els dos primers mètodes de classificació són versions conegudes de mètodes sparse PLS que no s'havien aplicat a dades correlacionades en el temps. El tercer mètode és una nova proposta que usa dos algorismes coneguts. Els mètodes de classificació es basen en diferents versions d'un mètode sparse d'anàlisi discriminant de mínims quadrats parcials PLS (sPLS-DA, SPLSDA i sPLS) i anàlisi discriminant lineal (LDA). Les variables que els mètodes de classificació usen com a input, corresponen a paràmetres estimats a partir de diferents models, mètodes i funcions de l'àrea de les sèries de temps, per exemple, model ARIMA estacional, model ARIMA-TGARCH estacional, mètode estacional Holt-Winters, funció de densitat espectral, funció d'autocorrelació (ACF), funció d'autocorrelació parcial (PACF), rang mòbil (MR), entre altres funcions. També van ser utilitzades algunes variables que s'utilitzen en el camp de l'astronomia per a classificar estreles.

La metodologia proposada consta de dues parts. La primera consisteix a calcular variables a partir de les sèries de temps, usant els mètodes, models i funcions esmentades anteriorment. La segona part consisteix a usar les variables calculades en el primer pas per a ajustar algun dels següents models: sPLS-DA, SPLSDA, o sPLS amb LDA (nova proposta). En els casos que a priori no va haverhi informació dels clústers de les sèries de temps, les dues primeres components d'una anàlisi de components principals (PCA) van ser utilitzades per l'algorisme k-means per a identificar possibles clústers de les sèries de temps. Addicionalment, els resultats del mètode sPLS-DA van ser comparats amb els de l'algorisme random forest.

Tres bases de dades de sèries de temps d'humitat relativa o de temperatura varen ser analitzades. Aquestes sèries de temps es van registrar en diferents edificis (la Catedral de València, el jaciment arqueològic de L'Almoina i l'església barroca de Sant Tomàs i Sant Felip Neri) a València, Espanya. Els clústers de les sèries de temps es van analitzar d'acord a diferents zones o diferents nivells d'altures on van ser instal·lats sensors per al monitoratge de les condicions climàtiques en els edificis.

L'algorisme random forest i les diferents versions del mètode sparse PLS van ser útils per a identificar les variables més importants en la classificació de les sèries de temps. Els resultats de sPLS-DA i random forest van ser molt similars quan es van usar com a variables d'entrada les calculades a partir del mètode Holt-winters o a partir de funcions aplicades a les sèries de temps. Encara que els resultats

del mètode random forest van ser lleument millors que els trobats per sPLS-DA quant a les taxes d'error de classificació, els resultats de sPLS-DA van ser més fàcils d'interpretar.

Quan les diferents versions del mètode sparse PLS van utilitzar variables resultants del mètode Holt-Winters, els clústers de les sèries de temps van ser més ben discriminats. Entre les diferents versions del mètode sparse PLS, la versió sPLS amb LDA va obtindre la millor discriminació de les sèries de temps, amb un menor valor de la taxa d'error de classificació, i utilitzant el menor o segon menor nombre de variables.

En aquesta tesi proposem usar una versió sparse de PLS (sPLS-DA, o sPLS amb LDA) amb variables calculades a partir de sèries de temps per a classificar sèries de temps. En aplicar la metodologia a les diferents bases de dades estudiades, es van trobar models parsimoniosos, amb poques variables, i varem obtindre una discriminació satisfactòria dels diferents clústers de les sèries de temps amb fàcil interpretació. La metodologia proposada pot ser útil per a caracteritzar les diferents zones o altures en museus o edificis similars d'acord amb les seues condicions climàtiques, amb l'objectiu de previndre problemes amb les obres d'art.

Contents

List of Figures

List of Tables

1

Introduction and Objectives

1.1 | Introduction

1.1.1 | Justification

Time series clustering is a current area of research with applications in several fields such as astronomy, neuroscience, medicine, smart buildings, engineering, economics, and finance, among others. In particular, in art conservation, clustering of times series of climatic parameters (e.g., temperature, relative humidity, among others) is important because this can help to characterize the microclimatic conditions in different zones or heights in museums and similar buildings. In this field, two common problems are moisture and dust concentration on the artworks. These problems could be caused by high humidity and different changes of temperature among others. Classifying time series of temperature (T) and relative humidity (RH) can help to monitor and analyze microclimatic data, in order to preserve artworks.

Many dissimilarity measures have been proposed, using conventional clustering algorithms to evaluate dissimilarity between two time series [1]. Results from different studies have concluded that if the underlying clusters are very close to each other, the time series clustering performance might diminish significantly [2]. One problem when examining time series for art conservation is that time series of RH and T are very similar in distinct positions of the same building. Furthermore, in this area, in addition to classifying time series it is necessary that the results can be easily interpreted. A statistical methodology that uses both features that explain different aspects of time series and a clustering algorithm that determines the optimal features is required. It can help to improve the classification of time series with easy interpretation, for art conservation.

Artworks in Valencia's Cathedral, the archaeological site of L'Almoina, and the baro que church of Saint Thomas and Saint Philip Neri (Valencia, Spain) are being monitored in order to help preserve artefacts. In these three sites, the design of the building generates microclimatic conditions that need to be evaluated at least every year. Time series from sensors located in the mentioned sites can allow different statistical methodologies with real scenarios to be assessed. Also, it can help to determine relevant zones or levels for each site, in order to make a plan to preserve artworks, as well as to determine main features to explain the classification of the series. Furthermore, it could help to propose a set of steps for selecting relevant sensors in the buildings, in order to classify time series.

The following sections present the main background of time series, clustering time series, and cultural heritage related to clustering time series.

1.1.2 | Background

1.1.2.1 | Cultural heritage

Cultural heritage is a source of wealth because it promotes tourism and native culture. Artworks undergo certain degradation over time, caused for example by changes of cli matic conditions. With the goal to avoid damages it is necessary to maintain stable and control climatic scenarios [3]. Many studies have researched microclimate conditions of historical buildings analyzing time series of temperature (T) or rela tive humidity (RH) in order to improve indoor air conditions for preserving the cultural heritage, e.g., [4; 5; 6; 7; 8; 9; 10]. Among these, several studies have analyzed graphs of trajectories of time series to determine possible events that caused changes in the trajectories of series as well as differences among series according to positions and heights in the buildings. Another study also analyzed differences between series using analysis of variance (ANOVA) [11]. In the same way, principal component analysis (PCA) was employed for classify ing sensors in order to characterize different zones in the buildings [6; 7].

1.1.2.2 | Time series

A time series is a set of observations taken sequentially in time, which is denoted by $\{y_t\}$ with $t \in \mathbb{Z}$, where t indicates the time at which the observation was taken [12]. In this document, $\boldsymbol{y} = (y_1, \ldots, y_n)$, with $n \in \mathbb{Z}$ denotes an observed time series, namely, a time series of a finite sequence of length n.

1. Basic concepts

- *Strict Stationarity*:

  Strict stationarity of a time series $\{y_t\}$ is defined by the condition that $\{y_t\}$ is invariant to a translation in h times, i.e., $(y_1, \dots, y_n)$ and $(y_{1+h}, \dots, y_{n+h})$ have the same joint distributions for all integers h and $n > 0$, as in Equation 1.1 [13].

  $$F(y_{1+h}, y_{2+h}, \dots, y_{n+h}) = F(y_1, y_2, \dots, y_n) \tag{1.1}$$

- *Weak Stationarity*:

  A stochastic process $\{y_t\}$ is weakly stationary if,

  - It has a constant mean, i.e., $\mathbb{E}[y_t] = \mu < \infty \; \forall t \in \tau$.
  - It has a finite and constant second moment, i.e., $\mathbb{V}[y_t] = \sigma^2 < \infty \; \forall t \in \tau$.
  - There is a function $\gamma(\cdot)$ such that $\gamma(h) = Cov(y_t, y_{t+|h|}) \; \forall t, h \in \tau$ [12].

- *Autocovariance Function*:

  If a stochastic process $\{y_t\}$ is weakly stationary, the autocovariance function of $\{y_t\}$ at lag h is given by $\gamma(h) = Cov(y_t, y_{t+h})$ [12].

- *Autocorrelation Function*:

  The autocorrelation function (ACF) is a standardized measure of the dependence of observations y_t and y_{t+h}. ACF at lag h is given by $\rho(h) = \frac{\gamma(h)}{\gamma(0)}$ [12].

- *White Noise*:

  A weakly stationary process $\{\varepsilon_t\}$ is a white noise (WN) process if it is a sequence of uncorrelated random variables [12]. It can be denoted by $\{\varepsilon_t\} \sim WN(\mu, \sigma^2)$.

2. ARMA Model

   A time series $\{y_t\}$ follows an autoregressive moving average process with parameters p and q, ARMA(p, q), if $\{y_t\}$ can be written as in Equation 1.2, where B is the backshift operator (i.e., $By_t = y_{t-1}$, for $t > 1$ and $B^p y_t = y_{t-p}$, for $t > p$), $\boldsymbol{\phi_p}(B)$ is the autoregressive polynomial on the backshift operator B and $\boldsymbol{\theta_q}(B)$ is the moving average polynomial, with roots distinct from those of $\boldsymbol{\phi_p}(B)$ [12].

   $$\begin{aligned} \boldsymbol{\phi_p}(B)y_t &= \boldsymbol{\theta_q}(B)\varepsilon_t, \\ \boldsymbol{\phi_p}(B) &= 1 - \phi_1 B - \cdots - \phi_p B^p, \\ \boldsymbol{\theta_q}(B) &= 1 + \theta_1 B + \cdots + \theta_q B^q, \\ \text{where} \quad \{\varepsilon_t\} &\sim WN(0, \sigma^2). \end{aligned} \tag{1.2}$$

   If all the solutions of the equation $0 = 1 - \phi_1 x - \ldots - \phi_p x^p$ which is associated with the autoregressive polynomial are outside the unit circle, then $\{y_t\}$ is a statio

-nary process. In particular, an ARMA(p,0) model corresponds to an autoregressive AR(p) process and an ARMA(0,q) model corresponds to a moving average MA(q) process [12].

3. Wold Decomposition

   Any stationary process $\{y_t\}$ can be written as a unique decomposition that consists of the sum of a non-deterministic and deterministic process. Deterministic processes are of interest in fields such as electrical engineering where signal can be described by a random amplitude and a particular frequency [12]. A stationary process can be decomposed as in Equation 1.3, where $\{\eta_t\}$ is the deterministic part and ε_t is the limit of linear combinations of y_s, where $s \leq t$.

$$\begin{aligned} y_t &= \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} + \eta_t, \text{where,} \\ \{\varepsilon_t\} &\sim WN(0, \sigma_\varepsilon^2), \\ \psi_0 &= 1, \sum_{j=1}^{\infty} \psi_j^2 < \infty, \\ E(\eta_t \varepsilon_s) &= 0 \quad \forall s, t \in \mathbb{Z}. \end{aligned} \tag{1.3}$$

   If the time series is purely non-deterministic ($\eta_t = 0$), then it can be written as an MA(∞) representation, i.e., $y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$, where $\psi_0 = 1$, $\sum_{j=1}^{\infty} \psi_j^2 < \infty$ and $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$.

4. SARIMA Model

   A seasonal ARIMA model $(p, d, q)(P, D, Q)_S$ with period S has a non-seasonal component (p, d, q) and a seasonal component $(P, D, Q)_S$. A time series $\{y_t\}$ follows a SARIMA$(p, d, q)(P, D, Q)_S$ model with one seasonal component, if it can be written as in Equation 1.4,

$$\begin{aligned} \boldsymbol{\Phi_P}(B^S)\boldsymbol{\phi_p}(B)\nabla_S^D \nabla^d y_t &= \boldsymbol{\Theta_Q}(B^S)\boldsymbol{\theta_q}(B)\varepsilon_t, \text{where,} \\ \boldsymbol{\phi_p}(B) &= 1 - \phi_1 B - \cdots - \phi_p B^p, \\ \boldsymbol{\theta_q}(B) &= 1 + \theta_1 B + \cdots + \theta_q B^q, \\ \boldsymbol{\Phi_P}(B^S) &= 1 - \Phi_1 B^S - \Phi_P B^{PS}, \\ \boldsymbol{\Theta_Q}(B^S) &= 1 + \Theta_1 B^S + \cdots + \Theta_Q B^{QS}, \\ \{\varepsilon_t\} &\sim WN(0, \sigma^2), \end{aligned} \tag{1.4}$$

   where $\boldsymbol{\phi_p(B)}$ is the regular auto regressive (AR) operator of order p, $\boldsymbol{\theta_q(B)}$ is the regular moving average (MA) operator of order q, $\boldsymbol{\Phi_P(B^S)}$ is the seasonal AR

(SAR) operator of order P, $\mathbf{\Theta}_Q(\boldsymbol{B}^S)$ is the seasonal AR (SMA) operator of order Q and $\{\varepsilon_t\}$ is a WN sequence with zero-mean and variance σ^2. Also, $\nabla_S^D = (1 - B^S)^D$ represents the seasonal differences and $\nabla^d = (1 - B)^d$ represents the regular differences [12].

5. TGARCH-Student Model

   An ARMA model is not able to capture volatility or variance clustering that can be present in some time series. These patterns can be captured using GARCH family models. The most important models are the ARCH and generalized ARCH (GARCH) models developed by Engle [14] and later extended by Bollerslev [15]. The family of GARCH models includes models such as the asymmetric power ARCH, the threshold GARCH (TGARCH) and GJR-GARCH, among other models [16]. For GARCH models, error terms can be assumed from the Student distribution [17].

   Different papers and books have proposed various specifications for σ_t for the *TGARCH* process. This document will present the particular conditionally heteroskedastic processes employed in `rugarch` package of `R software` [16].

   The innovations $\{\varepsilon_t\}$ follow a conditionally heteroskedastic process, if it can be written as $\varepsilon_t = \sigma_t \epsilon_t$, where the conditional mean (μ_t) and the conditional variance of the process ε_t are given by $\mu_t = E(\varepsilon_t | \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots)$ and $\sigma_t^2 = E(\varepsilon_t^2 | \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots)$, respectively. Errors $\{\epsilon_t\}$ are an independent and identically distributed (i.i.d.) with mean 0 and variance 1.

   The innovations $\{\varepsilon_t\}$ follow a process of the family *GARCH* model (*fGARCH*) if it can be written as Equation 1.5, where the conditional mean and variance are used to scale the residuals $z_t = \frac{\varepsilon_t - \mu_t}{\sigma_t}$ [16].

$$\begin{aligned} \varepsilon_t &= \sigma_t \epsilon_t, \\ \sigma_t^\lambda &= (\omega + \Sigma_{j=1}^N \varsigma_j V_{jt}) \\ &\quad + \Sigma_{j=1}^q \alpha_j \sigma_{t-j}^\lambda (|z_{t-j} - \eta_{2j}| - \eta_{1j}(z_{t-j} - \eta_{2j}))^\delta \\ &\quad + \Sigma_{j=1}^p \beta_j \sigma_{t-j}^\lambda. \end{aligned} \tag{1.5}$$

   Equation 1.5 is a Box-Cox transformation for σ_t, where λ determines shape, δ transforms the absolute value function, which subjects it to rotations and shifts through the η_{1j} and η_{2j} respectively. Also, N refers to the number of external regressors V_j, which are passed pre-lagged [16]. The innovations $\{\varepsilon_t\}$ follow a *TGARCH* process with parameters s and m, the process is denoted by TGARCH(s, m), when

$\lambda = \delta = 1$, $\eta_{2j} = 0$, $|\eta_{1j}| \leq 1$ [16]. In this study, ϵ_t follows a Student distribution with parameter v and external regressors were not employed. Thus, the previous equation can be written as equation 1.6.

$$\begin{aligned} \varepsilon_t &= \sigma_t \epsilon_t, \\ \sigma_t &= \omega + \Sigma_{j=1}^{m} \alpha_j \sigma_{t-j}(|z_{t-j}| - \eta_{1j} z_{t-j}) + \Sigma_{j=1}^{s} \beta_j \sigma_{t-j}, \\ \{\epsilon_t\} &\sim t - Student(v). \end{aligned} \tag{1.6}$$

6. Spectral Density Function

The spectral approach to second-order properties helps to separate seasonal effects and short term [18]. A description of spectral density according to Venables and Ripley [18] is presented below.

Given a covariance-stationary process $\{y_t\}$ with mean μ and j-th autocovariance γ_j, the population spectrum of $\{y_t\}$ at frequency $\omega \in \mathbb{R}$ is given by Equation 1.7. This function is well defined, on condition that the sequence $\{\gamma_h : h \in \mathbb{Z}\}$ is absolutely summable. The population spectrum is symmetric around 0 and periodic with period π. Also, $\int_{-\pi}^{\pi} s(\omega) e^{iwk} d\omega = \gamma_k$ and $s(\omega) \geq 0$ with $\omega \in [-\pi, \pi]$.

$$s(\omega) = \frac{1}{2\pi} \sum_{j=1}^{\infty} \gamma_j e^{-i\omega j} \tag{1.7}$$

The population spectrum function and the autocovariance function contain the same information about process $\{y_t\}$. In particular, $\gamma_0 = V[y_t]$ can be calculated as in Equation 1.8.

$$\gamma_0 = \int_{-\pi}^{\pi} s(\omega) d\omega. \tag{1.8}$$

According to the spectral representation theorem, any covariance-stationary process $\{y_t\}$ with absolutely summable autocovariances can be represented as in Equation 1.9.

$$y_t = \mu + \int_{0}^{\pi} \{\alpha(\omega)\cos(\omega t) + \delta(\omega)\sin(\omega t)\} d\omega, \tag{1.9}$$

where $\alpha(.)$ and $\delta(.)$ have zero means. With words, y_t can be decomposed in terms of frequencies.

An estimator of the population spectrum $s(.)$ is the periodogram that is presented in Equation 1.10.

$$\tilde{s}(\omega) = \frac{1}{2\pi} \sum_{j=-T+1}^{T-1} \hat{\gamma}_j e^{-i\omega j}, \tag{1.10}$$

where T is the sample size and $\hat{\gamma}_j$ is the jth sample autocovariance.

The estimator $\tilde{s}(\omega)$ of $s(\omega)$ is unbiased but noisy. However, if it is assumed that s is smooth, the values of this naive estimator can be averaged over frequencies near ω to obtain a much more precise estimator of $s(\omega)$ which is denoted by $\hat{s}(\omega_j)$ and is presented in Equation 1.11.

$$\hat{s}(\omega_j) = \sum_{m=-l}^{l} \mathcal{W}_T(\omega_m)\tilde{s}(\omega_j - \omega_m), \tag{1.11}$$

where $\omega_j = 2\pi j/T$, and l indicates how many different frequencies can be considered close to ω_j, and $\mathcal{W}_T(.)$ is a weighting function that must have the properties presented in Equation 1.12.

$$\begin{aligned} \lim_{T\to\infty} \mathcal{W}_T(\omega_j)^2 &= 0, \\ \sum_{j=-l}^{l} \mathcal{W}_T(\omega_j) &= 1, \\ \mathcal{W}_T(\omega_j) &= \mathcal{W}_T(-\omega_j). \end{aligned} \tag{1.12}$$

7. Holt-Winters Method

   Holt's method, Holt-Winters method, and seasonal Holt-Winters method (SH-W) are extensions of simple exponential smoothing for computing the forecasting of data [19; 20]. The SH-W method captures the level, trend, and seasonality of a data set. The SH-W method employed a smoothing equation for the level component at time t (a_t), a smoothing equation for the trend or slope component at time t (b_t), and a smoothing equation for the seasonality components at time t (s_t). The additive SH-W prediction function for an observed time series $\mathbf{y}$ with period length p is given by Equation 1.13, where k is the integer part of $(l-1)/p$, and $\hat{y}_{t+l|t}$ is forecast of $\hat{y}_{t+l}$, based on all the data up to time t [21].

$$\begin{aligned} \hat{y}_{t+l|t} &= a_t + lb_t + s_{t+l-p(k+1)}, \text{where} \\ a_t &= \alpha(t_t - s_{t-p}) + (1-\alpha)(a_{t-1} + b_{t-1}) \\ b_t &= \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1} \\ s_t &= \gamma(t_t - a_{t-1} - b_{t-1}) + (1-\gamma)s_{t-p}, \\ \text{where} \quad 0 \le \alpha \le 1, &\quad 0 \le \beta \le 1, \quad 0 \le \gamma \le 1, \text{and} \quad t > s \end{aligned} \tag{1.13}$$

   When presenting outcomes from the algorithm, a, b and $S_1, S_2, \ldots, S_p$ corresponds to the last level of a_t, b_t, and s_t $(t = 1, \ldots, p)$, respectively.

1.1.2.3 | Clustering Time Series

Time series clustering is useful because it can help to discover interesting patterns in data sets and to understand their structure, anomalies, and other regularities in datasets, among other things [22; 23].

A crucial point in cluster analysis is establishing a suitable dissimilarity measure between two objects. In the context of time series, the dynamic character of the series makes it complex to determine a dissimilarity measure. The dissimilarities that are usually used in conventional clustering could not work appropriately with time dependent data, due to the fact that they pass over the interdependent relationship between values [24]. The dissimilarity measures are often adapted according to characteristic of the problem that it is of interest to solve, emphasizing properties of the time series that are of interest for the specific situation. For example, there are contexts where the main interest of the clustering is based on the properties of the predictions, others on profiles of series and features of models, among other things [24]. Different approaches for establishing the dissimilarity between time series have been proposed. Some approaches are presented below.

For comparing profiles of time series, some distances have been computed using raw data, e.g., each pair of sequences of data have been evaluated using a one-to-one mapping, in other cases, depending on the domain, the selection of dissimilarity measures must comply with properties of invariance to specific distortions of the time series [1]. Batista et al. [25] present a review of dissimilarity measures which were created to be invariant to features such as uniform scaling, amplitude scaling, phase and complexity, among others.

For representing the dynamic structure of each time series by a feature vector of lower dimension, dissimilarity measures can be determined by comparing sequences of serial features, computed from the original time series as spectral features, autocorrelations and wavelet coefficients, among other features [26; 27; 28; 29; 30].

For comparing levels of the complexity of time series algorithms based on data compression have been used [31; 32; 33; 34] as well as differences between permutation distributions [35].

For comparing future forecasts, Alonso et al. [36] and Vilar et al. [37] considered that two time series are in the same cluster if their forecasts for a specific future time are proximate.

Another approach consists of assuming specific underlying models and comparing fitted models [38; 39; 40]. The most common criterion is to assume that the time series are generated by ARIMA models. Piccolo [38] introduced the Euclidean distance between the corresponding autoregressive expansions from time series as a dissimilarity measure. The distance matrix between pairs of time series models was processed by a complete linkage clustering algorithm in order to construct the dendogram. For ARMA models, Maharaj [41] developed an agglomerative hierarchical clustering procedure that is based on the p-value of a test of hypothesis applied to every pair of given stationary time series. Kalpakis et al. [42] proposed using k-medoids algorithm and the Euclidean distance between the linear predictive coding cepstrum of two time series when a time series follows an ARIMA process. Furthermore, other methods such as PCA, DFT and DWT of the autocorrelation of time series were employed. Xiong and Yeung [43] derived an expectation maximization algorithm for learning the mixing coefficients from mixtures of ARMA models, as well as the parameters of the component models. One problem with the method was that its clustering performance can degrade significantly if the underlying clusters were very close to each other. Researchers in speech recognition and machine learning have also adopted alternative models such as Markov chains [44] or hidden Markov models [45; 46]. Model-based approaches have scalability problems [47], and performance reduces when the clusters are close to each other [2].

For comparing distributions, Khaleghi et al. [48] formulated a metric to quantify the distance between series, according to their distributions, and proved the consistency of k-means for clustering processes.

Finally, in the field of astronomy, an automated procedure for classifying time series (stars) where it was necessary to capture the peaks of time series, was developed [49; 50]. They proposed using certain features from the field of astronomy and two features that they designed as input for different classification methods, such as logistic regression, CART algorithm, boosting, random forest, support vector machine, artificial neural network, and Lasso regression. Some features were extracted from raw data and others after fitting a harmonic model.

1.1.2.4 | Classification Methods

When studying a classification problem like time series clustering, it is necessary to establish if there is a prior knowledge of the cluster that will be predicted in the data, or if there is not one. If the cluster is known, it is possible to carry out a supervised classification. A short introduction of three sparse versions of PLS (supervised) and random forest algorithm (supervised) are presented below.

1. Partial Least Squares

Partial least squares (PLS) regression [51] has been used as an alternative approach to ordinary least squares (OLS) regression in ill-conditioned linear regression models [52]. This method maximizes the covariance between components which correspond to linear combinations of original variables from two data sets, a regressor matrix and a response matrix. PLS is computationally fast and its results can be easily graphed and interpreted. For these reasons, PLS has acquired a lot of importance in high dimensional classification problems of computational biology [53].

Although PLS was not originally designed for classification, it has had an effective performance when it has been employed for that purpose [53]. In respect to the adaption of PLS to classification for high dimensional data, some approaches have been studied, e.g., PLS discriminant analysis (PLSDA) and generalized PLS (GPLS). The first, PLSDA [54; 55; 56] consists of treating the response as a continuous variable and employing PLS to compute latent components. Next, an off-the-shelf classification method such as logistic regression (LOG), linear discriminant analysis (LDA), and quadratic discriminant analysis (QDA) is employed. If the response is multicategorical, it is necessary to substitute the original categorical response into a numerical response matrix using dummy coding [53]. The second, GPLS [57; 58; 59; 60] incorporates PLS into a generalized linear model (GLM) framework. This method generalizes the weighted least squares problem arising within the Newton-Raphson algorithm to maximize the log-likelihood with PLS [53]. Both PLSDA and GPLS often employ variable filtering [54; 55; 56] as a pre-processing step before the PLS fit [53].

Even though in most cases PLS works very well when the number of predictors is greater than sample sizes, PLSDA and GPLS, often use variable filtering as a pre-processing step before fitting PLS. Although preselection approaches often improve the performance of PLS classification, the selection of predictors is often arbitrary [53]. Furthermore, variable filtering approaches that are commonly used are all univariate and ignore correlations among variables. Chung and Keles [53] proved that the existence of a high number of irrelevant variables leads to the inconsistency of estimates of parameters of the linear regression setting. As a solution, they proposed sparse partial least squares (SPLS) regression, which selects predictors, while reducing dimension. Thus, components depend only on a subset of the original set of predictors. Penalties such as Lasso and Ridge are used in PLS for variable selection. Also, Lê Cao et al. [61] and Waaijenborg et al. [62] proposed computational approaches introducing sparsity in PLS.

Chung and Keles [53] proposed two new methods extending SPLS [52] to classification problems: SPLS discriminant analysis (SPLSDA:
SPLSDA-LDA and SPLSDA-LOG) and sparse generalized PLS (SGPLS). Lê Cao et al. [63] proposed a natural extension to the sPLS [61; 64] by coding the response matrix with dummy variables: sPLS-discriminant analysis (sPLS-DA). The variable selection and the classification for SPLSDA is performed in two stages, whereas, for sPLS-DA the results are directly obtained from the by products of the sPLS. SGPLS is also performed in one stage. The three methods aim to improve the PLS classification approaches by using dimension reduction and variable selection simultaneously. The three different formulations of PLS are based on the formulation in Equation 1.14.

PLS models regressor matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ and response matrix $\mathbf{Y} \in \mathbb{R}^{n\times q}$ as $\mathbf{X} = \Xi\mathbf{C} + \mathbf{E}_1$ and $\mathbf{Y} = \Xi\mathbf{D} + \mathbf{E}_2 = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}_2$, where $\boldsymbol{\beta} \in \mathbb{R}^{n\times p}$ is the matrix of regression coefficients, $\mathbf{E}_1 \in \mathbb{R}^{n\times p}$ and $\mathbf{E}_2 \in \mathbb{R}^{n\times q}$ are random errors. $\Xi \in \mathbb{R}^{n\times H}$ is the latent component matrix, where $\Xi = \mathbf{XU}$, with $\mathbf{U} \in \mathbb{R}^{p\times H}$ as H direction vectors, with $1 \leq H \leq min\{n,p\}$ and $\mathbf{U} = (\boldsymbol{u}_1,\ldots,\boldsymbol{u}_H)$. The h-th direction vector $\widehat{\boldsymbol{u}}_h$ is obtained by solving successive optimization problems according to Equation 1.14 for $j = 1,\ldots h-1$, subject to $\|\boldsymbol{u}\|_2 = 1$ and $\boldsymbol{u}^\top S_{XX}\widehat{\boldsymbol{u}}_j = 0$, where S_{XX} is the sample covariance matrix of the predictors.

$$\max_{\boldsymbol{u}}\{\boldsymbol{u}^\top\mathbf{M}\boldsymbol{u}\}, \quad \text{where} \quad \mathbf{M} = \mathbf{X}^\top\mathbf{Y}\mathbf{Y}^\top\mathbf{X} \tag{1.14}$$

A description of three formulations of PLS are presented below.

- SPLSDA

  The optimization problem of SPLS corresponds to a new reformulation of the objective function in the Equation 1.14 with Lasso penalty function, Ridge penalty function and tuning parameter κ. This formulation corresponds to a convex problem and it is sufficiently sparse [52].

  The following formulation corresponds to multicategory classification, because in the context of art conservation, this is more probable.

  Consider the matrix $\mathbf{Y}^* \in \mathbb{R}^{n\times G}$, whose elements are given by

  $$y^*_{i,(g+1)} = I(y_i = g)$$

  for $i = 1,\ldots,n$, and $g = 0,1,\ldots,G-1$, where 0 is the 'baseline' class, n is the number of time series and $I(A)$ is an indicator function of event A. The resulting matrix $\mathbf{Y}^*$ is column centered before fitting SPLS.

SPLSDA modeled regressor matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ and response matrix $\mathbf{Y}^* \in \mathbb{R}^{n\times G}$ as $\mathbf{X} = \Xi\mathbf{C} + \mathbf{E}_1$ and $\mathbf{Y}^* = \Xi\mathbf{D} + \mathbf{E}_2 = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}_2$, where $\boldsymbol{\beta} \in \mathbb{R}^{n\times p}$ is the matrix of regression coefficients, $\mathbf{E}_1 \in \mathbb{R}^{n\times p}$ and $\mathbf{E}_2 \in \mathbb{R}^{n\times G}$ are random errors and $\Xi \in \mathbb{R}^{n\times H}$ is the latent component matrix, where $\Xi = \mathbf{XU}$, with $\mathbf{U} \in \mathbb{R}^{p\times H}$ as H direction vectors, with $1 \leq H \leq min\{n, p\}$ and $\mathbf{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_H)$. The h-th direction vector $\hat{\boldsymbol{u}}_h$ is obtained by solving successive optimization problems according to Equation 1.15 for $j = 1, \ldots h-1$, subject to $\|\boldsymbol{u}\|_2 = 1$.

$$\begin{aligned}
\min_{\boldsymbol{u},\boldsymbol{c}}\{-\kappa\boldsymbol{u}^\top\mathbf{M}\boldsymbol{u} + (1-\kappa)(\boldsymbol{c}-\boldsymbol{u})^\top\mathbf{M}(\boldsymbol{c}-\boldsymbol{u}) + P_{\lambda_1}(\boldsymbol{c}) + P_{\lambda_2}(\boldsymbol{c})\}, &\quad \text{where} \\
\mathbf{M} = \mathbf{X}^\top\mathbf{Y}^*\mathbf{Y}^{*\top}\mathbf{X}, & \\
P_{\lambda_1}(\boldsymbol{c}) = \lambda_1\|\boldsymbol{c}\|_1 \quad \text{(Lasso penalty), and} & \\
P_{\lambda_2}(\boldsymbol{c}) = \lambda_2\|\boldsymbol{c}\|_2 \quad \text{(Ridge penalty)} & \\
0 \leq \eta \leq 1 \quad \text{and justify setting} & \\
0 < \kappa \leq 0.5 \quad \text{and} & \\
\lambda_2 = \infty. &
\end{aligned} \tag{1.15}$$

This formulation has four tuning parameters (κ, λ_1, λ_2, and H). Furthermore, it promotes exact zero property by imposing $P_{\lambda_1}(\boldsymbol{c})$ onto $\boldsymbol{c}$ while keeping $\boldsymbol{u}$ and $\boldsymbol{c}$ close to each other. Also, $P_{\lambda_2}(\boldsymbol{c})$ takes care of the potential singularity of $\mathbf{M}$ when solving for $\boldsymbol{c}$ [53].

The first direction vector of SPLS is computed by Equation 1.16 [53].

$$\begin{aligned}
\hat{\boldsymbol{c}} = \arg\max_{\boldsymbol{c}} \sum_{g=0}^{G}\left(\frac{n_g n_{-g}}{n}\right)^2\left(\sum_{j=1}^{p} c_j(\widehat{\mu}_{j,g} - \widehat{\mu}_{j,-g})\right)^2 - P_{\lambda_1}(\boldsymbol{c}), \text{where} & \\
\widehat{\mu}_{j,g} \quad \text{is the sample mean of the j-th predictor in class } g & \\
n_g \quad \text{is the sample size in class } g & \\
n = n_1 + n_2 + \cdots + n_G & \\
n_{-g} = n - n_g & \\
\widehat{\mu}_{j,-g} \quad \text{is the sample mean of the j-th predictor across all but the class } g & \\
c_j \quad \text{is the element j-th of the direction vector } \boldsymbol{c} &
\end{aligned} \tag{1.16}$$

Chung and Keles [53] studied this solution and some conclusions were as follows: the contribution of each class to the construction of direction vectors is affected at the same time by both the class sample size $\left(\frac{n_g n_{-g}}{n}\right)$ and the difference between the out-of-class sample mean and within-class sample mean across predictors. Thus, if the effect sizes of the predictors across the class are comparable, it is probable that the first direction vector will be most affected by the class with a larger sample size. They also studied the solution for the binary classification case, this can be found in [53].

Once, the latent components were computed, a linear classifier, either LDA or LOG, was used [53]. The estimates of parameters for the original predictors $\left(\widehat{\boldsymbol{\beta}}\right)$ were computed using estimates of parameters for latent components from a linear classifier $\left(\widehat{\boldsymbol{\beta}}^{LC}\right)$ as $\widehat{\boldsymbol{\beta}} = \mathbf{U}\widehat{\boldsymbol{\beta}}^{LC}$ [53]. Chung and Keles [53] suggests using a linear classifier due to the fact that it might be better, from an interpretation point of view.

- SGPLS

  Consider $\mathbf{X}$ as a regressor matrix and $\boldsymbol{Y}$ as response vector. Also, the multinomial model in Equation 1.17 and its log-likelihood in Equation 1.18

$$\begin{aligned} \log\left(\frac{p_{ig}}{p_{i0}}\right) &= \boldsymbol{x}_i^{\top}\boldsymbol{\beta}_g, \text{where} \\ g &= 1, 2, \ldots, G \\ p_{ig} &= P\left(y_i = g|\boldsymbol{x}_i\right) \\ \boldsymbol{x}_i & \quad \text{is the} \quad i\text{-th row vector of} \quad \mathbf{X} \end{aligned} \tag{1.17}$$

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left\{\sum_{g=1}^{G} y_{ig}\boldsymbol{x}_i^{\top}\boldsymbol{\beta}_g - \log\left(1 + \sum_{g=1}^{G} exp(\boldsymbol{x}_i^{\top}\boldsymbol{\beta}_g)\right)\right\} \tag{1.18}$$

  Maximizing the log-likelihood in Equation 1.18 using the Newton-Raphson algorithm, is equivalent to solving an iteratively re-weighted least squares in Equation 1.19, for $\boldsymbol{\beta}_g$ while $l \neq g$, for $g, l = 1, \ldots, G$, where $\widetilde{\boldsymbol{\beta}}$ and g are the current estimates and class, respectively.

$$\begin{aligned} &\min_{\boldsymbol{\beta}_g}\sum_{i=1}^{n} v_{ig}\left(z_{ig} - \boldsymbol{x}_i^{\top}\boldsymbol{\beta}_g\right)^2, \quad \text{where} \\ &\widetilde{p}_{ig} = exp\left(\boldsymbol{x}_i^{\top}\widetilde{\boldsymbol{\beta}}_g\right) / \left(1 + \sum_{g=1}^{G} exp\left(\boldsymbol{x}_i^{\top}\widetilde{\boldsymbol{\beta}}_g\right)\right) \\ &v_{ig} = \widetilde{p}_{ig}(1 - \widetilde{p}_{ig}) \\ &z_{ig} = \boldsymbol{x}_i^{\top}\widetilde{\boldsymbol{\beta}}_g + (y_{ig} - \widetilde{p}_{ig})/v_{ig} \end{aligned} \tag{1.19}$$

Based on the problem in Equation 1.19, with the goal of incorporating variable selection into the logistic regression model, the minimization problem in Equation 1.15 with $\mathbf{M} = \mathbf{X}^\top \mathbf{V}_g \boldsymbol{z}_g \boldsymbol{z}_g^\top \mathbf{V}_g \mathbf{X}$, and $\mathbf{V}_g$ as a diagonal matrix with entries v_{ig} and $\boldsymbol{z}_g = (z_{1g}, \ldots, z_{ng})$, is solved.

In order to implement SGPLS, Chung and Keles [53] developed a computationally faster approximation to Firth's procedure. The details of the algorithm can be found in [53].

Computational experiments carried out by Chung and Keles [53] displayed that SPLSDA and SGPLS had comparable performance in terms of variable selection and classification accuracy, despite their structural differences. They concluded that if the classes are highly unbalanced in terms of their sample sizes, SGPLS has higher sensitivity than SPLSDA, in variable selection. However, the variable selection performance of SPLSDA improves when the sample sizes increase. SPLSDA has two main advantages over SGPLS. Firstly, it is computationally faster. Secondly, due to the fact SPLSDA treats dimension reduction and classification in two independent steps, there is a wide choice of classifiers that can be used for its second stage.

- sPLS-DA

Consider a regressor matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ with p variables and n sensors, response vector $\boldsymbol{Y} \in \mathbb{R}^{n \times 1}$ whose elements are the positions of the sensors. The vector $\boldsymbol{Y}$ was converted into a dummy matrix $\mathbf{Y}^* \in \mathbb{R}^{n \times K}$ given by

$$y^*_{i,k} = I(y_i = k),$$

where $k = 1, \ldots, K$, K classes, and $I(A)$ is an indicator function of event A.

sPLS-DA modeled $\mathbf{X}$ and $\mathbf{Y}^*$ as a linear regression, where $\mathbf{X} = \Xi \mathbf{C} + \mathbf{E}_1$ and $\mathbf{Y}^* = \Xi \mathbf{D} + \mathbf{E}_2 = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}_2$, where $\boldsymbol{\beta} \in \mathbb{R}^{n \times p}$ is the matrix of regression coefficients, $\mathbf{E}_1 \in \mathbb{R}^{n \times p}$ and $\mathbf{E}_2 \in \mathbb{R}^{n \times K}$ are random errors, $\Xi \in \mathbb{R}^{n \times H}$ is the latent component matrix, where $\Xi = \mathbf{X}\mathbf{U}$, with $\mathbf{U} \in \mathbb{R}^{p \times H}$ as H direction vectors, with $1 \leq H \leq min\{n, p\}$ and $\mathbf{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_H)$. Furthermore, $(\boldsymbol{u}_h, \boldsymbol{v}_h)$ is the solution the optimization problem according to Equation 1.20 for $j = 1, \ldots, h-1$, subject to $\|\boldsymbol{u}\|_2 = 1$.

$$\min_{\boldsymbol{u},\boldsymbol{v}} \{ \|\mathbf{M} - \boldsymbol{u}\boldsymbol{v}^\top\|_F^2 + P_\lambda(\boldsymbol{u}) \} \tag{1.20}$$

The optimization problem minimizes the Frobenius norm $\|\mathbf{M} - \boldsymbol{u}\boldsymbol{v}^\top\|_F^2$ which is computed as $\sum_{i=1}^{n} \sum_{j=1}^{p} (m_{ij} - u_i v_j)^2$, where $\mathbf{M} = \mathbf{X}^\top \mathbf{Y}^*$, $\boldsymbol{u}$ and $\boldsymbol{v}$ are the loading vectors, where $\mathbf{V} = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_H)$. Furthermore, $P_\lambda(\boldsymbol{u})$ is the Lasso penalty function, where $P_\lambda(\boldsymbol{u}) = \lambda\|\boldsymbol{u}\|_1$ [63; 64].

This optimization problem is solved based on the PLS algorithm [65] and Singular Value Decomposition (SVD) [66] of a matrix $\tilde{\mathbf{M}}_h$ per dimension h. The SVD decomposition of matrix $\tilde{\mathbf{M}}_h$ is subsequently deflated per iteration h. This matrix is computed as $\mathbf{U}\Delta\mathbf{V}^\top$, where $\mathbf{U}$ and $\mathbf{V}$ are orthonormal matrices, and Δ is a diagonal matrix whose diagonal elements are called the singular values. During the deflation step of PLS, $\mathbf{M}_h \neq \mathbf{X}_h^\top \mathbf{Y}_h^*$, due to the fact that $\mathbf{X}_h$ and $\mathbf{Y}_h^*$ are calculated separately, and the new matrix is called $\tilde{\mathbf{M}}_h$. At each step, a new matrix $\tilde{\mathbf{M}}_h = \mathbf{X}_h^\top \mathbf{Y}_h^*$ is calculated and decomposed by SVD. Furthermore, in the sPLS algorithm, the soft-thresholding function $g(\boldsymbol{u}) = (|\boldsymbol{u}| - \lambda)_+ sign(\boldsymbol{u})$, with $(x)_+ = max(0, x)$, was used in penalizing loading vectors $\boldsymbol{u}$ to perform variable selection in the regressor matrix, thus $\boldsymbol{u}_{new} = g_\lambda(\tilde{\mathbf{M}}_{h-1}\boldsymbol{v}_{old})$ [61].

The `mixOmics` package [67] offers different functions for carrying out multivariate analysis of data sets [68]. It proposes different functions for sPLS-DA. Rohart et al. [68] employs an algorithm for sPLS-DA that instead of using the soft-thresholding function $g(\boldsymbol{u})$ to perform variable selection, uses the function in Equation 1.21. Keep in mind that controlling η instead of the direction vector specific sparsity parameters λ_h with $h = 1, \dots, H$, evades combinatorial tuning of the set of sparsity parameters and supplies a bounded range for the sparsity parameter, i.e., $0 \leq \eta \leq 1$ [53].

$$\begin{aligned} g(\boldsymbol{u}) &= (|\boldsymbol{u}| - \eta max_{1\leq j\leq p}|\ u_j\ |)_+ sign(\boldsymbol{u}), \text{where} \\ 0 &\leq \eta \leq 1 \\ (x)_+ &= max(0, x) \end{aligned} \tag{1.21}$$

Lê Cao et al. [63] concluded that there are two parameters to tune in sPLS-DA: the number of latent components and the number of variables to select on each component. They displayed that, for most cases, the user could set $H = K - 1$, while the number of variables to select is more challenging and is still an open question. In their opinion, the number of variables can be orientated through the estimation of the generalisation classification error and a stability analysis [69; 70]. Nevertheless, when the sample size is small, these analyses might be seriously limited.

Lê Cao et al. [63] compared the classification performance of sPLS-DA against SGPLS and SPLSDA, among other methods such as LDA and the random forest algorithm, on five data sets. They concluded that sPLS-DA is competitive in terms of computational efficiency and superior in terms of interpretability of the results via graphical outputs. They did not intend to address the specific problem of unbalanced classes. While for this method Rohart et al. [68] suggested using the balanced error rate (BER) when the classes are unbalanced, because BER is less biased towards clusters with more elements during the performance assessment. Also, Lê Cao et al. [71] showed that sPLS-DA obtained relevant results in a microarray cancer data set.

While for running sPLS-DA, the number of variables to select for each latent component is required as an input parameter, SPLSDA-LOG, SPLSDA-LDA, and SGPLS require a tuned η parameter that varies between 0 and 1. The closer to 1, the smaller variable selection size.

2. Random Forest

Random forest is one of the algorithms which follows the ensemble classifier methodology (e.g., Bagging [72] and Boosting [73]) that determines a prediction model by combining the strengths of a collection of simpler base models [74]. Figure 1.1 (from [75]) illustrates the methodology of ensemble methods. In particular, for classification problems, the aggregation is performed employing the majority vote. Aggregation is the method employed for forming an aggregate which is the result of ensemble learning.

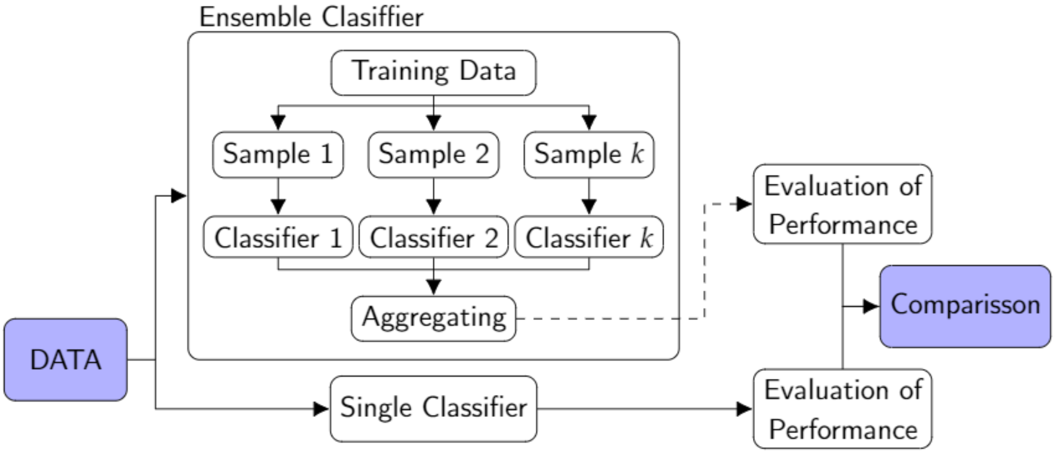


Figure 1.1: Workflow displays the methodology of ensemble methods.

Random forest [74] is a modification of the bagging method and it improves the variance reduction of bagging, by reducing the correlation between the trees. This is achieved by random selection of the input variables. Previous to each split, the algorithm selects a set of m ($m \leq p$) variables at random, as candidates for splitting. This algorithm is sensitive to the quality of input variables, because these are selected using a process of sampling.

1.1.3 | Proposed Solution

In the context of art conservation, the number of features computed from time series could be greater than the number of time series, because most of the cases, museum and similar buildings have restriction with the number of sensors that it is possible to install. This scenario might lead to severely ill-conditioned problems. Considering the previous idea and with the aim of identifying the main features that determine the clusters of time series, two known versions of sparse PLS, sPLS-DA [63] and SPLSDA [53], and a new proposal that combines sPLS [61] and LDA, are proposed. Furthermore, some features are proposed as input for these methods.

Features proposed correspond to estimates of parameters from time series models (e.g., seasonal ARIMA or seasonal ARIMA-TGARCH), estimates of components from seasonal Holt-Winters method, some coefficients of Wold decomposition, and features (e.g., maximum, mean, median, and variance) computed using function values of time series (e.g., sample ACF and PACF) or periodogram, as well as, some variables based on quantiles that are used in the field of astronomy. These features were organized according to the process used. Method 1, if variables were computed using functions applied to original time series; method 2, if variables were from seasonal Holt-Winters; method 3, if variables were from seasonal ARIMA or seasonal ARIMA-TGARCH; method 4, if variables were from Wold decomposition. Features from the same method were used as input in a classification method.

The interest of this work is to propose one statistical methodology for classifying time series, using features computed from them and employing a sparse method that helps to select the main features for determining clusters with easy interpretations.

Finally, the approach proposed for classifying time series is new in the context of both, clustering of time series and cultural heritage. For classifying time series, the dissimilarity measures computed for both sparse algorithms were applied to different features, computed according to at least two of the approaches mentioned before (i.e., profiles of time series, dynamic structure of series, assuming specific underlying models, future forecasts, among others). In this case, some measures were computed using a linear combination of a set of variables. These variables can correspond to different approaches, e.g., assuming specific underlying models and future forecasts, profiles of time series and the dynamic structure of series. In some cases, only one approach was used, e.g., profiles of time series. Also, this is probably the first time that sPLS-DA and SPLSDA are employed in the context of clustering of time series applied to microclimate monitoring, as well as, being the first time that the combination of the algorithms sPLS with LDA are used for classifying in this context. On the other hand, for art conservation, the classification of time series has been rarely explored and it has been basically analyzed using PCA.

1.2 | Objectives

1.2.1 | General Objective

Provide a statistical methodology with a robust framework to classify time series in the context of microclimate monitoring for the preventive conservation of artworks.

1.2.2 | Specific Objectives

1. To put forward different methods to extract features from time series that could be used as input for classification algorithms.

2. To propose at least two supervised methods for classifying time series that select the main features computed from series.

3. To compare the features selected for at least two supervised methods proposed.

4. To propose a methodology in order to select a subset of time series using the results from one of the supervised methods.

5. To characterize microclimatic conditions in distinct zones or levels in Valencia's Cathedral, the archaeological site of L'Almoina, and the baroque church of Saint Thomas and Saint Philip Neri (Valencia, Spain) using one of the methodologies proposed.

1.3 | Contributions

The following is a list of all the contributions during the progress of this Ph.D. thesis:

Publications:

1. Ramírez, Sandra.; Zarzo, M.; Perles, A.; Garcia-Diego, F.J. *Methodology for Discriminant Time Series Analysis Applied to Microclimate Monitoring of Fresco Paintings*. Sensors 2021, 21(2):436. doi:10.3390/s21020436.

2. Ramírez, Sandra.; Zarzo, M.; Garcia-Diego, F.J. *Multivariate Time Series Analysis of Temperatures in the Archaeological Museum of L'Almoina (Valencia, Spain)*. Sensors 2021, 21(13):4377. doi:10.3390/s21134377.

3. Ramírez, Sandra.; Zarzo, M.; Perles, A.; Garcia-Diego, F.J. *Characterization of Temperature Gradients According to Height in a Baroque Church by Means of Wireless Sensors*. Sensors 2021, 21(20): 6921. doi:10.3390/s21206921

Communications in conferences:

1. Ramírez, Sandra.; Zarzo, M.; Perles, A.; Garcia-Diego, F.J. *sPLS-DA to discriminate time series*. Joint Statistical Meetings JSM2020: 2-6 August 2020.

2. Ramírez, S.; Zarzo, M.; Perles, A.; Garcia-Diego, F.J. Sparse PLS-DA: A statistical methodology to characterize the relative humidity and temperature of an archaeological site. ICSA Symposium (Virtual Conference) 14-15 December 2020.

3. Ramírez, S.; Zarzo, M.; Perles, A.; Garcia-Diego, F.J. Sparse PLS-DA: Clustering time series for art conservation. 13th International Conference of the ERCIM WG on Computational and Methodological Statistics 14th International Conference on Computational and Financial Econometrics Virtual Conference 19-21 December 2020.

Proceeding:

1. Ramírez, S.; Zarzo, M.; Perles, A.; Garcia-Diego, F.J. sPLS-DA to discriminate time series. In JSM proceedings, 1221 statistics and the environment; JSM2020: Alexandria, VA: American Statistical Association., 2021; pp. 107–135. ISBN 978-1-7342235-2-1.

This PhD dissertation is composed of three scientific publications that were published in the Sensors journal, which is indexed in the Journal Citations Report (JCR®). The papers were properly formatted according to the requirements of this dissertation. Each article corresponds to one chapter of this document (i.e., Chapters 2, 3, and 4). Chapter 2 presents a new methodology to classify time series that employs sPLS-DA and variables extracted from seasonal ARIMA-TGarch models, seasonal H-W methods, and features computed using values of sample ACF, sample PACF, periodogram of time series, among others. The methodology was applied to time series of RH from sensors located in Valencia's Cathedral. Chapter 3 extends the methodology proposed in Chapter 2. For seasonal H-W method, values of predictions were used as input for sPLS-DA. Also, coefficients of Wold decomposition were employed as classification variables. Furthermore, time series were classified using Random Forests algorithm and the selected variables were compared with results from sPLS-DA. Additionally, PCA with k-means were used in order to determine possible clusters of the time series. Its Results were interpreted according to the technical knowledge of the microclimate of the site. The methodologies were applied to time series of RH from sensors located in L'Almoina museum. Chapter 4 describes the application of two versions of sparse PLS (sPLS-DA and sPLSDA) using features from methods employed in Chapters 2 and 3 as input. In addition, some features that were used in the context of astronomy for classifying time series were also employed. The methodologies were applied to time series of T from sensors located in the baroque church of Saint Thomas and Saint Philip Neri. Chapter 5 refers to general discussion and chapter 6 to general conclusions.

2

A Methodology for Discriminant Time Series Analysis Applied to Microclimate Monitoring of Fresco Paintings

Ramírez, Sandra.; Zarzo, M.; Perles, A.; Garcia-Diego, F.J. *Methodology for Discriminant Time Series Analysis Applied to Microclimate Monitoring of Fresco Paintings*. Sensors 2021, 21(2): 436. doi:10.3390/s21020436.

This chapter corresponds to the publication mentioned above, after changing the positions of some graphs and tables.

2.1 | Abstract

The famous Renaissance frescoes in Valencia's Cathedral (Spain) have been kept under confined temperature and relative humidity (RH) conditions for about 300 years, until the removal of the baroque vault covering them in 2006. In the interest of longer-term preservation and in order to maintain these frescoes in good condition, a unique monitoring system was implemented to record both air temperature and RH. Sensors were installed at different points at the vault of the apse during the restoration process. The present study proposes a statistical methodology for analyzing a subset of RH data recorded by the sensors in 2008 and 2010. This methodology is based on fitting different functions and models to the time series, in order to classify the different sensors. The methodology proposed, computes classification variables and applies a discriminant technique to them. The classification variables correspond to estimates of model parameters of and features such as mean and maximum, among others. These features are computed using values of functions such as spectral density, sample autocorrelation (sample ACF), sample partial autocorrelation (sample PACF), and moving range (MR). The classification variables computed were structured as a matrix. Next, sparse partial least squares discriminant analysis (sPLS-DA) was applied in order to discriminate sensors according to their position in the vault. It was found that the classification of sensors derived from Seasonal ARIMA-TGARCH showed the best performance (i.e., lowest classification error rate). Based on these results, the methodology applied here could be useful for characterizing the differences in RH, measured at different positions in a historical building.

Keywords: ARIMA; art conservation; Holt–Winters ; sensor diagnosis; sPLS-DA; TGARCH.

2.2 | Introduction

Over the past 300 years, the famed Renaissance frescoes in Valencia's cathedral were kept under confined conditions because they were covered by a baroque vault. However, this vault was removed in 2006 [7]. In the interest of longer-term preservation and in order to maintain these frescoes in good condition, a monitoring system was implemented to record both air temperature and relative humidity (RH). Sensors were located at different points in the apse vault. The approximate location of each sensor can be seen in Figure 2.1. The positions are: cornice $\mathcal{C}$, ribs $\mathcal{R}$, walls $\mathcal{W}$, and frescoes $\mathcal{F}$. Some sensors were inserted on the painting's surface itself. It is a unique system because sensors are rarely placed inside the frame or on the canvas of paintings. Details about the installation of probes in the frescoes can be seen in figures from [6; 7]. A perspective of the upper part of the apse and terrace above the frescoes can be found in [7].

The system was intended to detect water entering from the roof at specific points, or excessive general humidity in the vault itself. Any indication of high levels of thermo-hygrometric conditions would instigate corrective measures [6]. The data analysis carried out by Zarzo et al. [7] showed the advantages of using humidity sensors for the monitoring of frescoes, so as to maximize their protection and prevent deterioration. The microclimatic requirements for churches and cathedrals are similar to those of museums, which also contain valuable works of art [6]. As in the case of museums, the indoor thermo-hygrometric conditions should be maintained at optimal levels in order to conserve the artefacts. The risks constituted by ventilation systems, air-conditioning, central heating, and the presence of visitors should be assessed in order to prevent or slow down the process of deterioration. Ideally, the temperature of walls and their surfaces should be the same as the air in the immediate proximity because, otherwise, an airflow is generated along the wall surface that increments the aerodynamic deposition of airborne particles and wall soiling.

Cultural heritage sites are subjected to climatic changes that put them at risk, which has been widely discussed in [76; 77].

The internal environment should be appropriate [78] because changes in air temperature and RH can affect the conservation of fresco paintings [79; 80; 81]. Different studies [82; 83; 84; 85] have monitored thermo-hygrometric parameters inside museums in order to assess the potential risks related to temperature and humidity. Other authors, such as Camuffo et al. [81], have studied the interactions between the indoor atmosphere and walls supporting frescoes or mural paintings. Similar works have been carried out in churches [86; 87; 88; 89; 90]. Frasca et al. [86] performed a microclimatic monitoring of the historic church of Mogiła Abbey to analyze the impact of the environmental parameters on the works of art. Among their results, they found that vulnerable objects were at a high risk of mechanical damage approximately 15% of the time. The main cause of the vulnerability was the RH variability.

The problems of deterioration due to high humidity identified in the Renaissance frescoes at the cathedral of Valencia were studied by Zarzo et al. [7]. The researchers suggested that these problems could be caused by the infiltration of rainwater through the roof above the apse and, that maintenance or regular monitoring should therefore be conducted for the long-term preservation of the valuable frescoes. Bernardi et al. [91] studied the importance of waterproofing in the roof above frescoes in St. Stephan's church in Nessebar, Bulgaria.

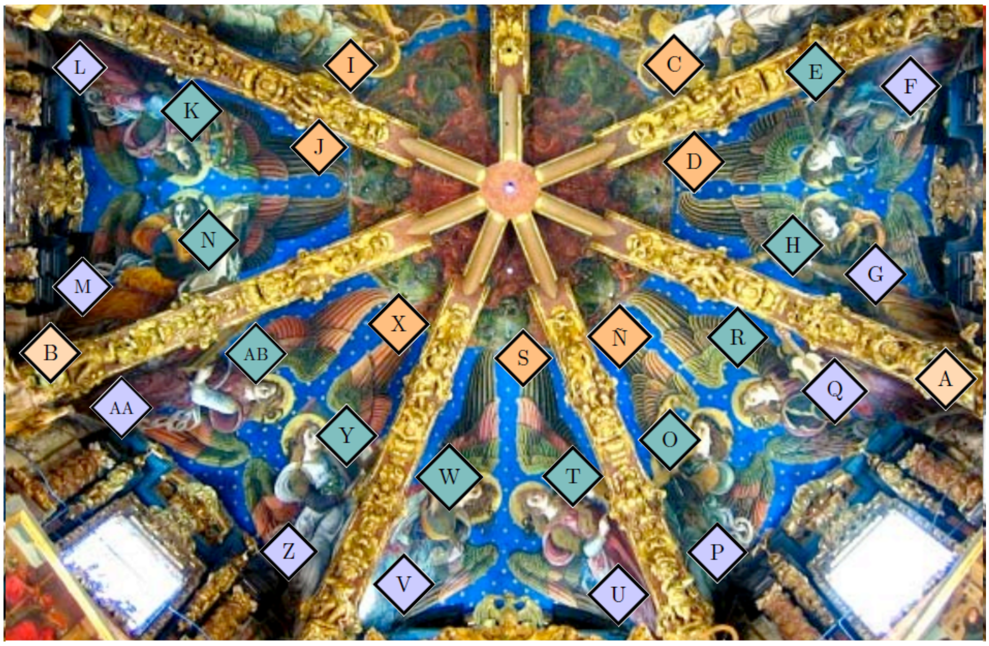


Figure 2.1: Approximate location of the probes and sensors at the apse vault of the cathedral of Valencia. Details of the installation of the probes and a scheme depicting the installation of probes in the frescoes can be seen in [7] and [6]. The image shows the position of the 29 probes for monitoring the relative humidity (RH) of the indoor atmosphere, displayed in different colors according to their position. Seven probes were located on the ribs (orange), two at the cornice (light orange), ten on the walls below the severies (purple), and ten probes on the frescoes (green).

In the same way, the European Standards [92; 93; 94; 95; 96; 97] summarized in [98] as well as Corgnati and Filippi [99] adopted the approach of the Italian Standard UNI 10829 (1999) for the monitoring, elaboration, and analysis of microclimatic data for the preservation of artefacts.

A big economic effort is being carried out by governments within the European Union to preserve artworks in museums. Several works have monitored the microclimate within museums to analyze its relationship with the degradation of materials from which works of art are made, for example, with the goal of preserving artwork and artefacts [100].

Concerning data analysis, García-Diego and Zarzo [6] used monthly principal components analysis (PCA) in their research for February, September, October, and November of 2007. Furthermore, Zarzo et al. [7] also fitted a PCA per year for the years 2007, 2008, and 2010. The resulting loading plots highlight the most relevant similarities and dissimilarities among sensors. Regarding RH recorded in 2007, researchers observed that the daily evolution versus time of the RH mean per hour ($\overline{RH}$) was rather parallel for all sensors. It was observed that sensors H, N, and R (inserted on the frescoes) recorded higher values of $\overline{RH}$ than those installed on the walls. Interestingly, sensors H and R were located in the zone where there was a moisture problem after their installation. In 2008 and 2010, the correlation between $\overline{RH}$ and the first principal component PC1 was very high (greater than 0.994) [7]. After computing the average of moving ranges with order 2 of RH (per hour HMV, day DMV and month MMV), it was shown that PC2 for 2008 could be predicted as: $PC2 = 232.88 - 2.69\overline{RH} - 32.39DMV$. Regarding 2010, the estimated regression model was: $PC2 = 297.03 - 3.87\overline{RH} - 32.12DMV$. Based on the results, researchers concluded that PC1 could be interpreted as the yearly RH average, while PC2 provided basic information about daily mean variations. Furthermore, researchers detected an abnormal performance in one sensor that might correspond to a failure of the monitoring system [101] or a change in the microclimatic conditions surrounding that particular sensor. They also concluded that the use of humidity sensors and the interpretation of the first two principal components can be very useful when discussing the microclimatic air conditions surrounding fresco paintings. Hence, PCA is a powerful statistical method for characterizing the different performance among sensors of the same type, located at different positions [7]. The advantage of PCA for sensor diagnosis has also been reported by Dunia et al. [101] and Zhu et al. [102].

The present study re-analyzes time series of RH from sensors located at the apse vault of Valencia cathedral. The data sets used here correspond to subsets of the database used in the study conducted by Zarzo et al. [7]. The present work focuses on RH measurements recorded from 23 sensors in 2008 and from 20 sensors in 2010. These time series of RH do not contain missing values.

This research aims to bring forward a methodology for discriminating sensors according to their position. For this purpose, the approach applied in this study consists of three stages: (1) The different time series were divided according to climatic condition and changes of the slope and level of the time series; (2) Three methods (M1, M2, and M3) were applied to obtain the classification variables per part of the time series identified in stage 1; (3) Sparse partial least squares discriminant analysis sPLS-DA was applied three times (one per method) as a discriminant technique in order to classify sensors, by using the set of classification variables as predictors.

The methodology proposed in this research is new in the context of time series clustering, as well as in sensor classification, when applied with the aim of conserving works of art. This methodology is unique because it uses a Seasonal ARIMA-TGARCH model to extract information from the time series, for discrimination purposes. It is also singular because it employs sPLS-DA in order to classify the time series.

According to the results of this study, sPLS-DA together with ARIMA-TGARCH-

Student has a high capability of classifying time series with very similar characteristics, which often occurs in museums or similar buildings. The proposed methodology is well-suited to monitoring the sensors in this type of building.

This approach can be very useful in defining how microclimatic measurements should be carried out for monitoring conditions in heritage buildings or similar sites. Furthermore, the methodology could be useful for reducing the number of sensors required to monitor the microclimate. In summary, this approach could help to better manage the preventive conservation of cultural heritage sites.

2.3 | Materials and Methods

2.3.1 | Materials: Description of the Data Sets

Regarding the frescoes in Valencia's cathedral, 29 probes were implemented to monitor the indoor air conditions. Each probe contains an integrated circuit model DS2438 (Maxim Integrated Products, Inc.) that incorporates an analogue-to-digital voltage converter. This converter measures the output voltage of a humidity sensor (HIH-4000, Honeywell International, Inc.) and a temperature sensor. The recorded values of RH have an accuracy of ±3.5%. Details of the probes, RH sensors, functions of calibration, and their installation in the apse vault are described elsewhere [6; 7]. Seven probes were placed on the ribs ($\mathcal{R}$), two at the cornice ($\mathcal{C}$), ten on the walls below the severies ($\mathcal{W}$), and ten on the frescoes ($\mathcal{F}$) (see Figure 2.1).

The data sets used here do not contain missing values and correspond to subsets of the database used by Zarzo et al. [7]. The electronics platform Arduino was used `https://www.arduino.cc/en/Guide/Introduction`. Such data sets correspond to the mean RH per hour or day ($\boldsymbol{RH_h}$ or $\boldsymbol{RH_d}$), where RH_{h_t} is the average of measurements per hour, while RH_{d_t} corresponds to the average of measurements per day.

The $\boldsymbol{RH}$ datasets correspond to those sensors located in the cornice $\mathcal{C}$, ribs $\mathcal{R}$, walls $\mathcal{W}$, and frescoes $\mathcal{F}$. As the statistical analysis was performed separately for each season, sensors `M` in $\mathcal{W}$ and `Ñ` at the $\mathcal{R}$ were discarded because it was necessary to deal with time series comprising a time period of at least 300 observations without missing values. Thus, time series are available in 2008 for 23 sensors: two on the cornice (`A` and `B`), five at the ribs (`C`, `D`, `I`, `J`, and `X`), nine on the frescoes (`E`, `H`, `K`, `O`, `R`, `T`, `W`, `Y`, and `AB`), and seven on the walls (`G`, `L`, `P`, `U`, `V`, `Z`, and `AA`). In 2010, information from sensors `H`, `Y`, `AB`, `G`, and `Z` was not available, but there were two additional sensors (`S` and `Q`) located at the $\mathcal{C}$ and $\mathcal{W}$, respectively. Hence, 20 sensors could be used for 2010: 8 at position $\mathcal{RC}$ ($\mathcal{R}$ or $\mathcal{C}$), 6 on the walls ($\mathcal{W}$), and 6 on the frescoes ($\mathcal{F}$) (see Figure 2.1).

For both years, the follow-up time spanned the seasons of winter Wr, spring Sp, and summer Sm. The data set $\boldsymbol{RH_h}$ for 2008 consists of 3851 observations: 1430 for Wr, 2099 for Sp, and 322 for Sm. Regarding 2010, 3414 observations are available: 636 for Wr, 2178 for Sp, and 600 for Sm.

Autumn was not considered because the number of observations was not high enough according to the established conditions of the study (i.e., at least 300 observations). Data sets correspond to the periods between 15 January and 4 July 2008, and from 22 February to 18 July 2010. Sensors from the selected times—both in 2008 and 2010—did not experience electronic malfunction. In 2008, there was no evidence of conservation problems in the frescoes where these sensors were located. By contrast, in 2010 there was evidence of salt efflorescence found in the same zones, before the first restoration works in the apse vault [7].

The periods corresponding to each season were defined as follows: spring was considered as being between 19 March and 20 June, summer between 20 June and 22 September, and winter between 22 September and 22 January.

2.3.2 | Statistical Methods

Three different methods (M) were applied to the RH data in order to extract estimates of parameters and features used subsequently as classification variables. These methods consist of fitting the time series to different statistical models or functions. M1 included functions such as spectral density, sample autocorrelation function (ACF), sample partial ACF (PACF), and moving range (MR) [18; 103; 104; 105]; M2 was the Additive Seasonal Holt–Winters method (Additive SH-W) [19; 20]; and M3 was a Seasonal ARIMA with threshold generalized autoregressive conditional heteroskedastic (TGARCH) model considering the Student distribution for residuals (Seasonal ARIMA-TGARCH-Student) [12; 16; 106; 107].

The three methods were carried out separately for various seasons of the year (Wr, Sp, and Sm) for both 2008 and 2010. Once the classification variables were computed from the three methods, sparse partial least square discriminant analysis (sPLS-DA) [63] was applied to classification variables three times—one time per method—using all classification variables per season. sPLS-DA was used to discriminate between sensors according to their three possible positions in the vault: $\mathcal{RC}$ ($\mathcal{R}$ or $\mathcal{C}$), $\mathcal{W}$, and $\mathcal{F}$. The classes $\mathcal{R}$ and $\mathcal{C}$ were joined as a new class called $\mathcal{RC}$ in order to ensure a similar number of sensors per group.

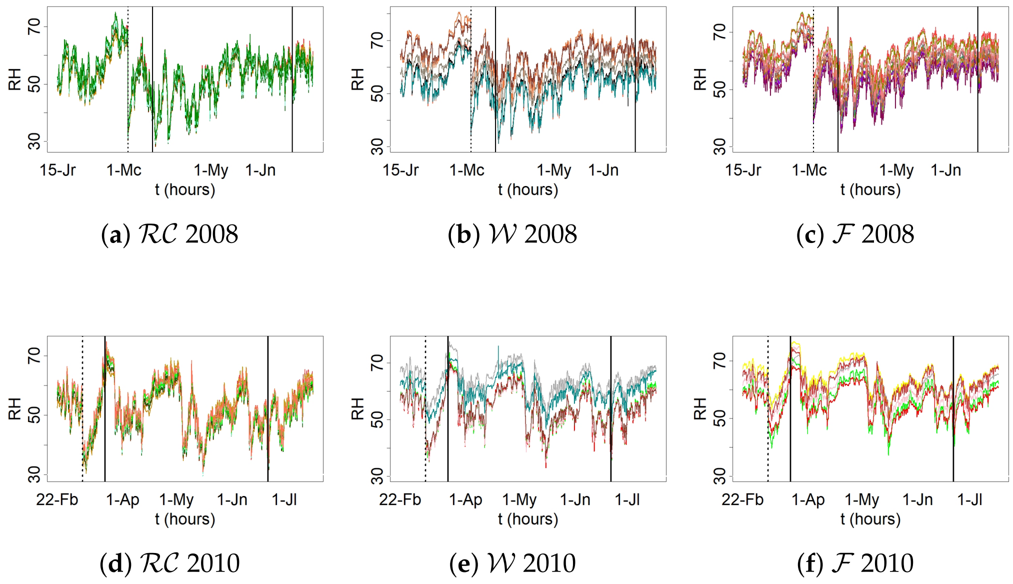
The statistical methodology applied consisted of different steps: Firstly, the identification of structural breaks in the time series (Section 2.3.2.1), which leads to the establishment of periods where the analyses were carried out. Secondly, calculation of classification variables using M1 (Section 2.3.2.2). Thirdly, the calculation of classification variables using M2 (Additive SH-W; Section 2.3.2.3). Fourthly, the calculation of classification variables using M3 (Seasonal ARIMA-TGARCH-Student; Section 2.3.2.4). Finally, sensor classification was done by means of sPLS-DA (Section 2.3.2.5).

`R` software [108] version 4.03 was used to carry out the analyses. The main `R` packages used were `aTSA` [109], `forecast` [110; 111], `mixOmics` [67; 68], `rugarch` [16; 112], `strucchange` [113], `tseries` [114], and `QuantTools` [105].

2.3.2.1 | Identification of Structural Breaks in the Time Series

Many time series models (e.g., ARMA [106], ARCH, and GARCH [16]) assume the lack of sudden changes due to external factors that might appear occasionally. However, when analyzing time series of real situations, it can be found that external factors produce dramatic shifts such as a change in the slope of a linear trend, which cannot be properly modeled. Such occasional events are known as structural breaks [12]. In order to detect such events, different tests can be applied from the generalized fluctuation test framework (e.g., CUSUM and MOSUM), which are based on empirical fluctuation processes [115]. Others like the Chow [116] test are based on checking sequences of F statistics [117; 118; 119], while the supF test [120] consists of applying the former at all possible structural breaks. The null hypothesis is "no structural change", versus the alternative: "the vector of coefficients varies over time" [120].

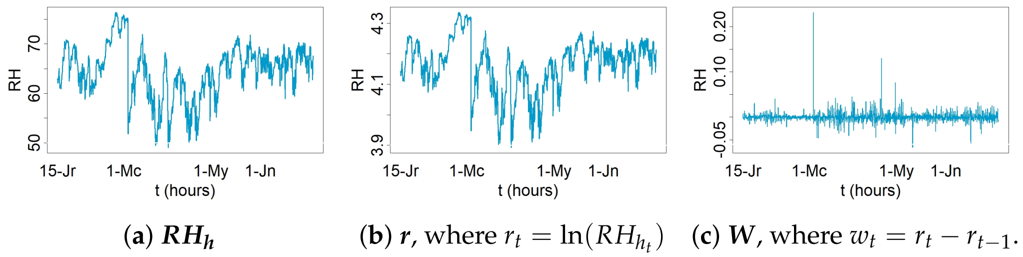
By visually inspecting the evolution versus time of $\boldsymbol{RH_h}$ for both 2008 and 2010 (see Figure 2.2), potential structural breaks were identified in at least two points. Their significance was assessed by means of the CUSUM and supF tests. Both tests were carried out with the logarithmic transformation [121], that is, $r_t = \ln(RH_{h_t})$, which has been used in other works to stabilize the data variance [12; 121; 122]. It was observed that most of the daily time series undergo seasonal trends, which makes it necessary to a-pply regular differentiation (i.e., $w_t = r_t - r_{t-1}$) in order to remove any trend [121], which is a common pretreatment in time series analysis. In this paper, $\boldsymbol{r}$ refers to transformed values using the logarithmic function, while $\boldsymbol{W}$ refers to data that was subjected to a logarithmic transformation and one regular differentiation. Figure 2.3 displays the plot of the time series of RH from sensor `Y` (2008). Additionally, this figure shows the plots of the time series of the logarithmic transformation of RH as well as one regular differentiation of the previous time series.



(**a**) $\mathcal{RC}$ 2008 (**b**) $\mathcal{W}$ 2008 (**c**) $\mathcal{F}$ 2008

(**d**) $\mathcal{RC}$ 2010 (**e**) $\mathcal{W}$ 2010 (**f**) $\mathcal{F}$ 2010

Figure 2.2: Evolution of $\boldsymbol{RH_h}$. Trajectories of sensors located at equivalent positions in the apse vault are depicted in the same chart (data recorded between January 15 and July 4 2008): cornice and ribs ($\mathcal{RC}$) (**a**), walls ($\mathcal{W}$) (**b**), and frescoes ($\mathcal{F}$) (**c**). Likewise for data collected between February 22 and July 18 2010: $\mathcal{RC}$ (**d**), $\mathcal{W}$ (**e**), and $\mathcal{F}$ (**f**). Separation by seasons (Wr, Sp and Sm) is indicated by means of vertical solid lines. Wr is divided into two periods (dashed line) because a structural break was identified according to the $supF$ and $CUSUM$ tests.



(**a**) $\boldsymbol{RH_h}$ (**b**) $\boldsymbol{r}$, where $r_t = \ln(RH_{h_t})$ (**c**) $\boldsymbol{W}$, where $w_t = r_t - r_{t-1}$.

Figure 2.3: (**a**) Observed time series of RH from sensor `Y` (2008), (**b**) logarithmic transformation of the time series of (a), (**c**) one regular differentiation of the time series of (b).

The supF and CUSUM tests were applied to six groups: Wr 2008 (group 1, $n = 1429$), Sp 2008 (group 2, $n = 2098$), Sm 2008 (group 3, $n = 321$), Wr 2010 (group 4, $n = 635$), Sp 2010 (group 5, $n = 2177$), and Sm 2010 (group 6, $n = 559$).

According to the supF test, a structural break was identified in Wr 2008, after the 1058th observation (March 27 at 7:00 AM, p-value = 0.01). Another break was found in 2010 at the 338th observation (March 8 at 1:00 PM, p-value = 0.04). The CUSUM chart identified a significant shift at the same instant of time (1058th value in 2008 and 338th observation in 2010). The main reason for structural breaks could have been the strong changes of RH that occur in Valencia.

Ignoring structural breaks can lead to negative implications such as inconsistency of the parameter estimates and forecast failures [123]. Accordingly, for each structural break, it wasdecided to fit one model before this event, and another one after the structural break. On the other hand, in congruence with the physical characteristics of the data, it might be convenient to split the statistical analysis per season and year. According to both considerations, the analysis was carried out separately in four periods, denoted as WrA, WrB, Sp, and Sm. WrA corresponds to the winter period before the structural break, while WrA refers to the following period (see Figure 2.2).

2.3.2.2 | Calculation of Classification Variables—Method M1

This method is based on features using estimates of ACF (ρ_l at lag l), PACF (α_l at lag l), as well as features using mean (μ) and moving range (MR). Furthermore, features from spectral density were used, which was estimated using the periodogram ($I(w)$) of signals w. These features can help to characterize and reveal interesting properties of the underlying stochastic process without using any specific parametric model. Figure 2.4 shows a summary of the steps of M1.

$\boldsymbol{RH}$ values were used for estimating PACF, mean, and MR. By contrast, logarithm transformation and regular differencing were applied before estimating ACF and spectral density in order to stabilize the variances and remove the trend (i.e., $\boldsymbol{W}$ was employed). The objective of using both the ACF and spectral density with $\boldsymbol{W}$ was to focus on the seasonal component of the time series. These functions are briefly explained below:

Firstly, the mean of $\boldsymbol{RH}_h$ was estimated for each period because this variable appeared as important for discrimination purposes in the preliminary study [6; 7].

Secondly, MR with order n correspond to range values over n past values [105]. This function was applied to $\boldsymbol{RH}_h$ and $\boldsymbol{RH}_d$. For each period, the mean and variance were computed for all MR values with order 2. These variables were calculated in order to estimate `HMV` and `DMV`, which were used in the preliminary investigations of this project [6; 7]. However, `MMV` (i.e., MR of order 2 for $\boldsymbol{RH}_m$, estimated with the average RH per month) could not be calculated in this research because the number of observations per month was too low. For $\boldsymbol{RH_h}$, the mean of MR ($\widehat{\mu}_{MR}$) corresponds to `HMV` and for $\boldsymbol{RH_d}$, the mean of MR is represented by `DMV`.

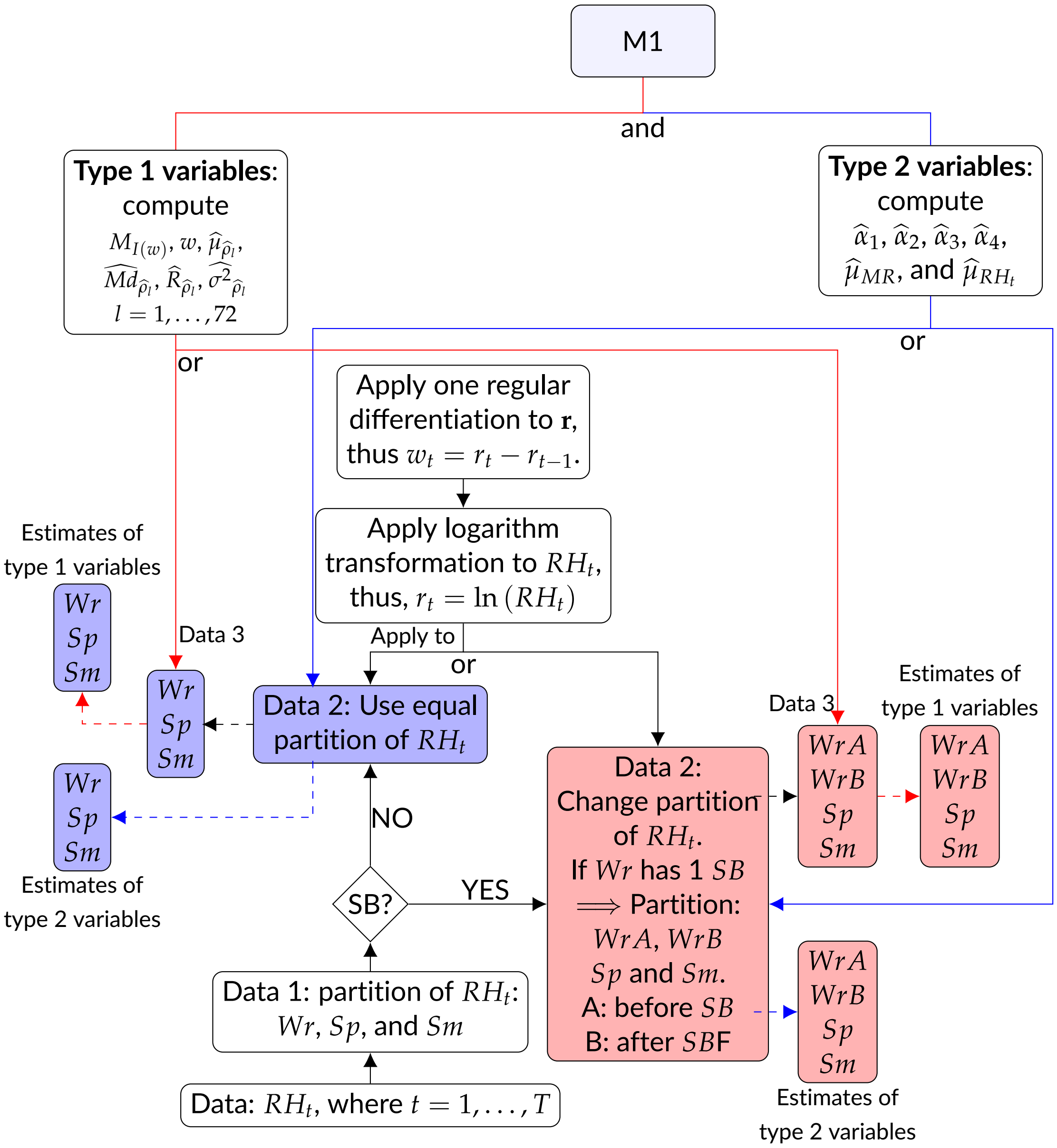

M1
and
Type 1 variables: compute
$M_{I(w)}$, w, $\widehat{\mu}_{\widehat{\rho}_l}$, $\widehat{Md}_{\widehat{\rho}_l}$, $\widehat{R}_{\widehat{\rho}_l}$, $\widehat{\sigma^2}_{\widehat{\rho}_l}$
$l = 1,\ldots,72$
Type 2 variables: compute
$\widehat{\alpha}_1$, $\widehat{\alpha}_2$, $\widehat{\alpha}_3$, $\widehat{\alpha}_4$, $\widehat{\mu}_{MR}$, and $\widehat{\mu}_{RH_t}$
or
or
Apply one regular differentiation to $\mathbf{r}$, thus $w_t = r_t - r_{t-1}$.
Apply logarithm transformation to RH_t, thus, $r_t = \ln(RH_t)$
Apply to
or
Estimates of type 1 variables
Wr
Sp
Sm
Data 3
Wr
Sp
Sm
Data 2: Use equal partition of RH_t
Wr
Sp
Sm
Estimates of type 2 variables
NO
SB?
YES
Data 1: partition of RH_t: Wr, Sp, and Sm
Data: RH_t, where $t = 1,\ldots,T$
Data 2: Change partition of RH_t. If Wr has 1 SB $\Longrightarrow$ Partition: WrA, WrB Sp and Sm. A: before SB B: after SBF
Data 3
WrA
WrB
Sp
Sm
Estimates of type 1 variables
WrA
WrB
Sp
Sm
WrA
WrB
Sp
Sm
Estimates of type 2 variables

Figure 2.4: Flow chart for the steps of method 1: Blue lines indicate type 2 variables. Red lines indicate type 1 variables. Solid lines indicate processes. Dashed lines indicate results.

Thirdly, spectral density was estimated by means of the periodogram, which was calculated on the log scale using a spectrum function [18]. The periodogram displays information about the strengths of the various frequencies for explaining the seasonal components of a time series. The maximum peaks of spectral density and their corresponding frequencies were identified [18]. These functions were applied to $\boldsymbol{RH}_h$.

Finally, an estimation of ACF at lag l is the correlation (quantified by means of Pearson's correlation coefficient) between the values of a given time series, with the lagged values of the same time series at l time steps (l refers to lags) [18]. $\boldsymbol{W}$ values were used for this calculation. The values of sample ACF for the lags from 1 to 72 were used as classification variables because they showed greater variations, while for further lags they displayed lower values close to zero, comprised between the limits of a 95% confidence interval in the ACF correlogram.

Regarding sample PACF, according to Cowpertwait and Metcalfe [124], "the partial autocorrelation at lag l is the correlation that results after removing the effect of any correlations due to the terms at shorter lags". Sample ACF and sample PACF plots are commonly used in time series analysis and forecasting (e.g., autoregressive moving average (ARMA) models and their particular cases such as autoregressive (AR) and moving average (MA) models [124]). These plots, also called correlograms, illustrate the strength of a relationship between the values observed at a certain instant of time with those recorded in previous moments (with lag l) in the same time series. If sample ACF values decline exponentially and there are spikes in the first or more lags of sample PACF values, the time series can be modeled as an AR process. If sample PACF values decline exponentially and there are spikes in the first or more lags of the sample ACF values, the time series can be modeled as an MA process. If both sample PACF and sample ACF values decline exponentially, the time series can be modeled as a mixed ARMA process [124]. Sample PACF was computed for $\boldsymbol{RH}_h$ values. The sample PACFs for the first four lags were calculated for each period and were regarded as classification variables because they are usually the most important ones for capturing the relevant information in time series.

The features computed using the values of $\boldsymbol{RH}$ were called type 1 variables and features calculated using values of $\boldsymbol{W}$ were referred to as type 2 variables. The list of type 1 variables resulting from M1 are the estimates of the following parameters:

- Mean of $\boldsymbol{RH_h}$ ($\widehat{\mu}_{RH}$).
- Mean of MR ($\widehat{\mu}_{MR}$) of order 2 for $\boldsymbol{RH_d}$ and $\boldsymbol{RH_h}$.
- Variance of MR ($\widehat{\sigma^2}_{MR}$) of order 2 for $\boldsymbol{RH_d}$ and $\boldsymbol{RH_h}$.
- PACF for the first four lags ($\widehat{\alpha}_1$, $\widehat{\alpha}_2$, $\widehat{\alpha}_3$, and $\widehat{\alpha}_4$).

The list of type 2 variables resulting from M1 are the estimates of the following parameters:

- Maximum of spectral density ($M_{I(w)}$) and frequency corresponding to the maximum (w).
- Mean ($\widehat{\mu}_{\widehat{\rho}_l}$), Median ($\widehat{Md}_{\widehat{\rho}_l}$), range ($\widehat{R}_{\widehat{\rho}_l}$), and variance of the sample ACF ($\widehat{\sigma^2}_{\widehat{\rho}_l}$) for the first 72 lags.

2.3.2.3 | Calculation of Classification Variables—Method M2: Additive SH-W

Winters [20] extended Holt's method [19] for capturing the seasonality of a time series. Hence, it was called Holt–Winters (H-W), which is a particular method of exponential smoothing aimed at forecasting [125].

The Seasonal H-W approach (SH-W) is based on three smoothing equations: for the level, trend, and seasonality. The parameter S denotes the number of values per season, while three additional parameters capture the information at time t: a_t denotes the time series level, b_t is the slope, and s_t is the seasonal component [125]. There are two different SH-W methods, depending on whether seasonality is modeled additively or multiplicatively [125]. The Seasonal H-W approach (SH-W) is based on three smoothing equations: for the level, trend, and for seasonality. The parameter S denotes the number of values per season, and three additional parameters capture the information at time t: a_t denotes the time series level, b_t is the slope, and s_t is the seasonal component [125].

In this research, the Additive SH-W method was fitted to both time series of RH and to their logarithmic transformations, but it turned out that the best outcomes were obtained with the transformed data. The period, the number of observations per season, was considered as $S = 24$ (i.e., 24 hourly values per day). Although this method does not require a residual analysis, one was carried out in an attempt to extract further information. Autocorrelation within the time series appeared in at least 10 out of the 22 lags for over 80% of the Ljung–Box Q (LBQ) tests [126] applied. Furthermore, the Kolmogorov–Smirnov normality (KS normality) [127] and Shapiro–Wilk (SW) tests [122; 128; 129] rejected the hypothesis of normality for at least 80% of the cases applied. The KS normality test compares the empirical distribution function with the cumulative distribution function. The test statistic is the maximum difference between the observed and theoretical values (normality). The statistic of the KS normality test was used as a classification variable in order to gather information about the distribution pattern of residuals and quantify departure from a normal distribution. The SW test detects deviations from normality due to either skewness or kurtosis, or both. The statistic of the SW test was also employed as a classification variable in order to identify lack of normality in the residuals according to skewness and kurtosis. Furthermore, given that data sets in this study are seasonal with a period of 24 hours, where 72 is the maximum of lags, this maximum value was also considered for estimation of the mean, median, range, and variance of the sample ACF for the residuals.

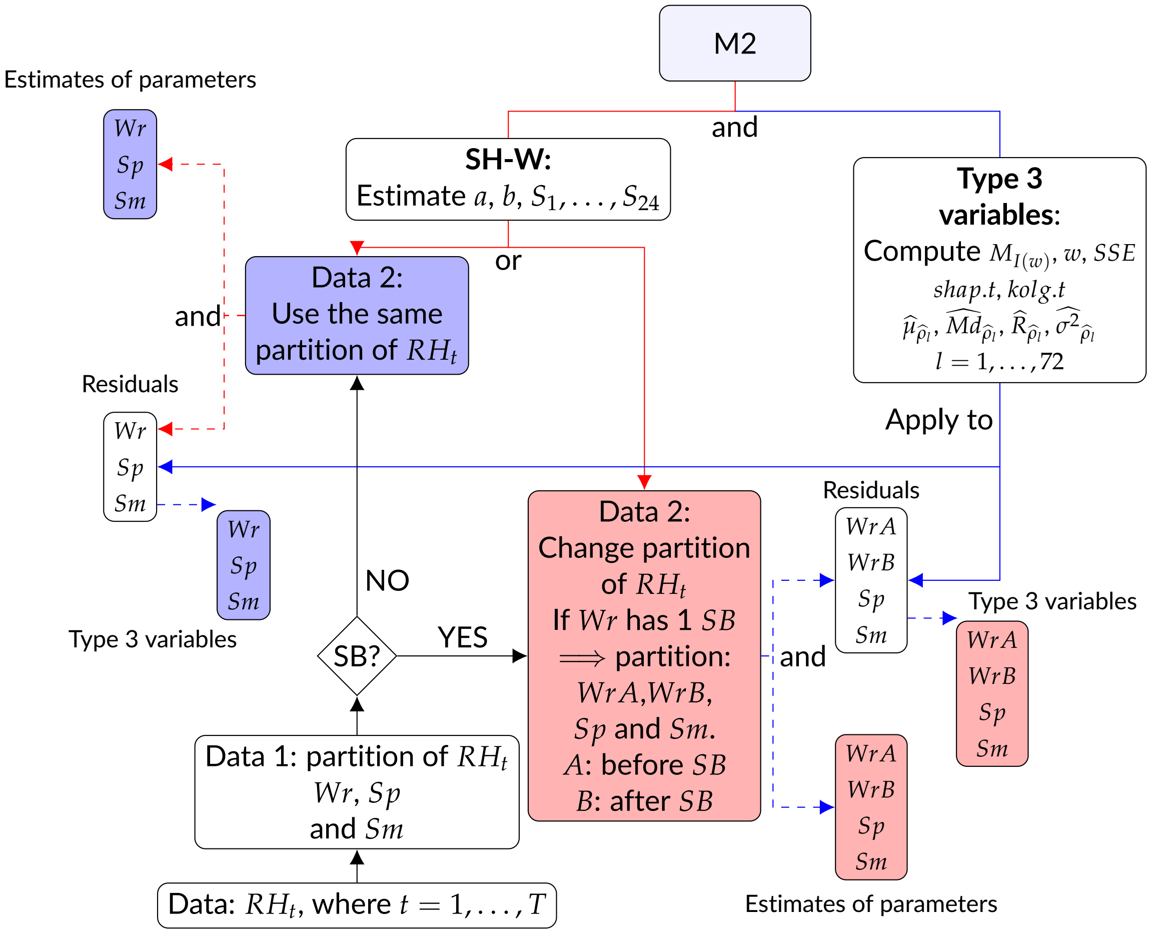

M2
and
Estimates of parameters
Wr
Sp
Sm
SH-W:
Estimate a, b, $S_1, \ldots, S_{24}$
Type 3 variables:
Compute $M_{I(w)}$, w, SSE
$shap.t$, $kolg.t$
$\widehat{\mu}_{\widehat{\rho}_l}$, $\widehat{Md}_{\widehat{\rho}_l}$, $\widehat{R}_{\widehat{\rho}_l}$, $\widehat{\sigma^2}_{\widehat{\rho}_l}$
$l = 1, \ldots, 72$
or
and
Data 2:
Use the same partition of RH_t
Residuals
Apply to
Type 3 variables
NO
YES
SB?
Data 2:
Change partition of RH_t
If Wr has 1 SB $\Longrightarrow$ partition:
WrA,WrB,
Sp and Sm.
A: before SB
B: after SB
Residuals
WrA
WrB
Sp
Sm
and
Type 3 variables
Data 1: partition of RH_t
Wr, Sp
and Sm
Data: RH_t, where $t = 1, \ldots, T$
Estimates of parameters

Figure 2.5: Flow chart for the steps of method 2: Red lines indicate estimated parameters. Blue lines indicate type 3 variables. Solid lines indicate processes. Dashed lines indicate results.

Figure 2.5 shows a summary of the steps of M2. The first step consists of dividing the different time series according to the climatic conditions: Wr, Sp, and Sm (Data 1). The second step consists of dividing the time series (Data 1) according to possible structural breaks (SBs) (Data 2). The third step applies a logarithm transformation and one regular differentiation to Data 2. The result is Data 3. The fourth step consists of applying the formulas of type 2 variables to Data 2. This is the first result. The fifth stage is carried out by applying the formula of type 1 variables to Data 3. The outcome produced is result 2. Different boxes contain symbols such as Wr, Sp, and Sm (or WrA, WrB, Sp, and Sm). They indicate that the results were computed for all different parts of the time series.

The features computed from residuals of the SH-W method were called type 3 variables. The other classification variables corresponded to the estimates of the method's parameters. The list of classification variables resulting from M2 are the following:

- Estimates of the parameters of the SH-W method: trend ($\widehat{b}$), level ($\widehat{a}$), and seasonal components ($\widehat{S}_1, \ldots, \widehat{S}_{24}$).
- type 3 variables: sum of squared estimate of errors (SSE), maximum of spectral density ($\widehat{M}_{I(w)}$), frequency corresponding to maximum of spectral density ($\widehat{w}$), and the mean ($\widehat{\mu}_{\widehat{\rho}_k}$), median ($\widehat{Md}_{\widehat{\rho}_k}$), range ($\widehat{R}_{\widehat{\rho}_k}$), and variance ($\widehat{\sigma^2}_{\widehat{\rho}_k}$) of sample ACF for 72 lags. The statistic of the SW test ($shap.t$), and the statistic of the KS normality test ($kolg.t$) are also included in this list.

2.3.2.4 | Calculation of Classification Variables—Method M3: Seasonal ARIMA-TGARCH-Student

ARMA models were popularized by Anderson [130], who developed a coherent three-step iterative cycle for time series estimation, verification, and forecasting. This method is also known as the Box–Jenkins approach. The ARMA model assumes that the time series is stationary; if this is not the case, differencing the time series one or more times is required, resulting in an ARIMA model. In the ARIMA(p, d, q) approach, p is the number of AR terms, d is the number of regular differences taken, and q is the number of the MA. Furthermore, ϕ_i $(i = 1, \ldots, p)$ are the parameters of the AR part of the model, θ_j $(j = 1, \ldots, q)$ are the parameters of the MA part, and the ε_t are error terms—generally assumed to be a white noise sequence [12].

Although ARIMA is flexible and powerful in forecasting, it is not able to properly handle continuously changing conditional variance or the non-linear characteristics of the variance that can be present in some time series [131]. This is often referred to as variance clustering or volatility [132; 133]. If it is assumed that a given time series follows an ARIMA process, the conditional variance of residuals is supposed to be constant versus time. When this condition is not fulfilled, it is known as a conditional variance process [121; 132; 133]. In such a case, data can also be affected by non-linear characteristics of the variance. These patterns can be studied using the GARCH family of models. Two of the most important ones for capturing such changing conditional variance are the ARCH and generalized ARCH (GARCH) models developed by Engle [14] and later extended by Bollerslev [15]. Engle and Bollerslev [134] were pioneers in the area of volatility modeling by introducing ARCH and, subsequently, GARCH models, which provide motion dynamics for the dependency in the conditional time variation of the distributional parameters of the mean and variance.

In recent years, different studies have applied hybrid forecasting models in various fields, and have shown a good performance for rainfall data [135], for the price of gold [136], for forecasting daily load patterns of energy [137], and for stock market prices [138].

According to Ghalanos [16], the family of GARCH models is broad, including the standard, integrated, and exponential models, as well as the GJR-GARCH, the asymmetric power ARCH, and the threshold GARCH (TGARCH) of Zakoian [139]. They capture the asymmetry of occasional impacts as well as abnormal distributions to account for the skewness and excess kurtosis. For GARCH models, error terms can sometimes be assumed from the Student distribution [17]. Bollerslev [140] described the GARCH-Student model as an alternative to the normal distribution for fitting the standardized time series. In particular, in the TGARCH-Student(s,r) model, s is the number of GARCH parameters β_i ($i = 1, \dots, s$), r is the number of ARCH and rotation parameters α_j and η_{1j}, respectively ($j = 1, \dots, r$), while ω is the variance intercept parameter. Error terms ϵ_t are assumed to be a white noise sequence following a Student distribution with degrees of freedom v [16].

Thus, instead of considering the standard ARIMA approach, whose focus is the conditional mean, it seems convenient to use here a hybrid approach based on ARIMA and GARCH models which can simultaneously deal with both the conditional mean and variance [12].

Given that data sets in this study are seasonal, it is necessary to use Seasonal ARIMA models, which are capable of modeling a wide range of seasonal data. A Seasonal ARIMA$(p,d,q)(P,D,Q)_S$ model is characterized by additional seasonal terms: P is the number of seasonal AR (SAR) terms, D is the number of differences taken, Q is the number of seasonal MA (SMA) terms, and S is the number of observations per period ($S = 24$ in this study). Furthermore, Φ_i ($i = 1, \dots, P$) are the parameters of the SAR part of the model, Θ_j ($j = 1, \dots, Q$) are the parameters of the SMA part, and the ε_t are error terms, which are assumed to be a white noise sequence [12]. In particular, in the Seasonal ARIMA$(p,d,q)(P,D,Q)_S$ -TGARCH(s,r) -Student model, the errors ε_t from Seasonal ARIMA$(p,d,q)(P,D,Q)_S$ follow a TGARCH(s,r) -Student process of orders s and r, so that their error terms ϵ_t are assumed to be a white noise sequence following a Student distribution, with degrees of freedom v.

Two steps were considered for the application of a hybrid approach based on Seasonal ARIMA and GARCH models, as briefly explained below. Firstly, the most successful Seasonal ARIMA (or ARIMA) model was selected and the residuals were computed.

Next, the most successful GARCH model was applied to fit these residuals. The following steps were carried out:

- The condition of stationarity was checked, that is, whether the statistical characteristics of the time series were preserved across the time period. The null hypothesis was that mean and variance do not depend on time t and the covariance between observations RH_t and RH_{t+l} does not depend on t [12]. To examine this null hypothesis, the augmented Dickey–Fuller (ADF) [141] and LBQ tests were applied for 48 lags. Furthermore, the sample ACF and sample PACF plots were also used.
- Transformation and differencing: the logarithmic transformation and regular differentiation were applied to $\boldsymbol{RH}_h$ data before fitting ARMA in order to transform nonstationary data into stationary data [124]. The criterion for determining the values of d (differencing) is explained in the next step. The logarithmic transformation was preferred over other transformations because the variability of a time series becomes more homogeneous using logarithmic transformation, which leads to better forecasts [142].
- Identification of the most appropriate values for (p, d, q) and (P, D, Q). Sample ACF and sample PACF plots were used to identify the appropriate values of (p, q). Furthermore, the *corrected Akaike information criterion* ($AICc$) [125] was useful for evaluating how well a model fits the data and determining the values of both (P, D, Q) and (p, d, q), taking into account the restriction that d and D should be 0 or 1. The most successful model for each time series was chosen according to the lowest $AICc$ value. The $AICc$ values were compared for models with the same orders of differencing, that is, equal values of d and D.

Secondly, the maximum likelihood estimation (MLE) method was used for estimating the parameters of the Seasonal ARIMA (or ARIMA) [12]. The models chosen were statistically examined in order to ensure that the resulting residuals do not contain useful information for forecasting. For this purpose, different tests were applied to determine whether all conditions and model assumptions were fulfilled. The analysis of residuals was carried out as follows:

- The condition of white noise was checked. Error terms can be regarded as white noise if their mean is zero and the sequence is not autocorrelated [12]. In order to check this issue, the ADF and LBQ tests were applied to the residuals and their squared values for 48 lags. Furthermore, the sample ACF plots were also used.
- To study the absence of Arch effects: for this purpose, the Lagrange multiplier test [14] and sample ACF plots [132] were applied to the residuals and their squared

values [143].

- To check the distribution of residuals: by means of the Q–Q normal scores plots as well as the SW and KS normality tests.

The analysis of residuals revealed that error terms follow a GARCH process in all the different ARIMA models that were fitted. Therefore, it was necessary to fit a GARCH model to these residuals. The estimated model parameters were checked to determine if they were statistically significant, and their residuals were evaluated as described above. Finally, a hybrid model was fitted for each sensor and each period using the combined Seasonal ARIMA (or ARIMA) and TGARCH approach. After repeating the steps iteratively in order to select the most successful model, the normality tests applied to their residuals rejected the hypothesis of normality in all cases. Furthermore, all Q–Q normal scores plots showed that residuals were not falling close to the line at both extremes. Thus, a Student distribution was used to fit the residuals of the different TGARCH models.

For each period, a common model was applied to the hourly data of each sensor (one day corresponds to a sequence of 24 values).

- WrA (2008): seasonal $\text{ARIMA}(1,1,0)(2,0,0)_{24}-$ TGARCH(1,1)-Student.
- WrA (2010): $\text{ARIMA}(1,1,2)-$ TGARCH(1,1)-Student.
- WrB (2008 and 2010): seasonal $\text{ARIMA}(1,1,1)(2,0,0)_{24}-$ TGARCH(1,1)-Student.
- Sp (2008 and 2010): seasonal $\text{ARIMA}(1,1,2)(0,0,2)_{24}-$ TGARCH(1,1)-Student.
- Sm (2008 and 2010): seasonal $\text{ARIMA}(1,1,1)(1,0,0)_{24}-$ TGARCH(1,1)-Student.

A seasonal model was not selected for WrA (2010) because the analysis of residuals of the selected model showed similar results to the best seasonal model, and the selected model was simpler.

When analyzing the residuals from Seasonal ARIMA-TGARCH-Student models for 2008, it turned out that in the case of WrA, the time series from 12 sensors out of the 23 available did not satisfy all expected conditions. The same happened for Sp: 14 out of the 20 models did not fulfill all requirements. Thus, in an attempt to extract further information not properly captured by these models, some features were calculated from the residuals.

Figure 2.6 shows a summary of the steps of M3. The first step divides the different time series according to the climatic conditions (Wr, Sp, and Sm) (Data 1). The second step organises the time series according to possible structural breaks (SBs) (Data 2). The third step applies the method to Data 2 in order to obtain the estimates of the method's parameters (first result) and then the residuals from the method. The fourth step consists of applying the formulas for type 3 variables to the residuals (second result). Different boxes display symbols Wr, Sp, and Sm (or WrA, WrB, Sp, and Sm). This indicates that the results correspond to all different parts of the time series.

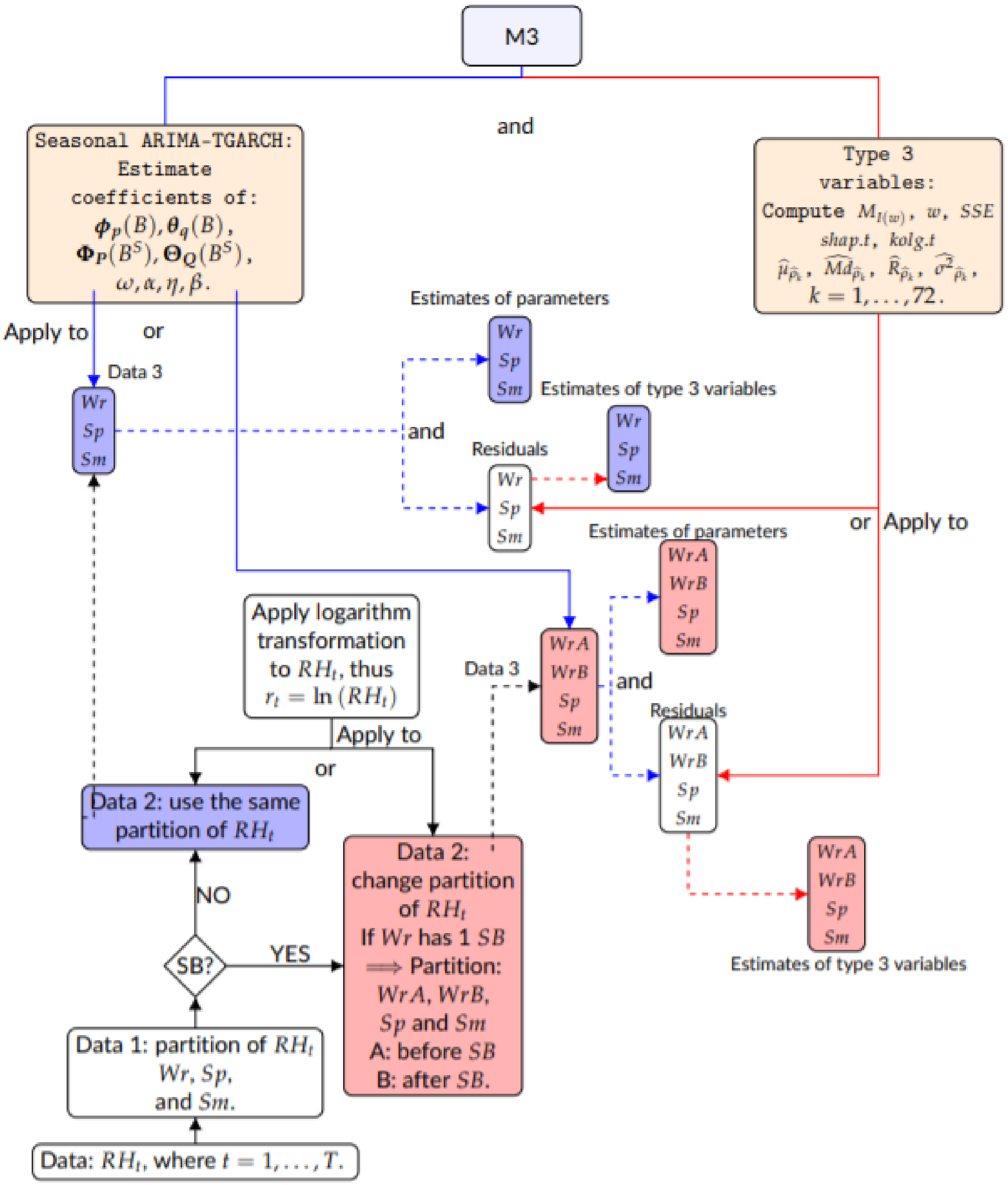

M3
and
Seasonal ARIMA-TGARCH: Estimate coefficients of: $\boldsymbol{\phi}_p(B), \boldsymbol{\theta}_q(B)$, $\boldsymbol{\Phi}_P(B^S), \boldsymbol{\Theta}_Q(B^S)$, $\omega, \alpha, \eta, \beta$.
Type 3 variables: Compute $M_{I(w)}$, w, SSE $shap.t$, $kolg.t$ $\hat{\mu}_{\hat{\rho}_k}$, $\widehat{Md}_{\hat{\rho}_k}$, $\hat{R}_{\hat{\rho}_k}$, $\widehat{\sigma^2}_{\hat{\rho}_k}$, $k = 1, \ldots, 72$.
Apply to
or
Data 3
Wr Sp Sm
Estimates of parameters
Wr Sp Sm
Estimates of type 3 variables
Wr Sp Sm
and
Residuals
Wr Sp Sm
or
Apply to
Estimates of parameters
WrA WrB Sp Sm
Data 3
WrA WrB Sp Sm
and
Residuals
WrA WrB Sp Sm
WrA WrB Sp Sm
Estimates of type 3 variables
Apply logarithm transformation to RH_t, thus $r_t = \ln(RH_t)$
Apply to
or
Data 2: use the same partition of RH_t
Data 2: change partition of RH_t If Wr has 1 SB $\Longrightarrow$ Partition: WrA, WrB, Sp and Sm A: before SB B: after SB.
NO
SB?
YES
Data 1: partition of RH_t Wr, Sp, and Sm.
Data: RH_t, where $t = 1, \ldots, T$.

Figure 2.6: Flow chart for the steps of method 3: Blue, red, solid, and dashed lines indicate estimated parameters, type 3 variables, processes, and results, respectively.

In all cases, residuals from the ARIMA-GARCH-Student models displayed evidence of stationarity for 48 lags. However, in some cases, there was evidence of autocorrelation as well as the presence of ARCH effect. For the tests applied to residuals, 0.03 was the maximum p-value found to reject the null hypothesis. Regarding 2008, the number of time series (from the 23 sensors) that satisfied all tests in the residual analysis, is the following: 12 in WrA, 22 in WrB, 22 in Sp, and 20 in Sm. In 2010, out of the 20 sensors available, the values are: 18 in WrA, 19 in WrB, 14 in Sp, and 15 in Sm.

The features computed from residuals of the models were called type 3 variables. The other classification variables corresponded to the estimates of the parameters of the selected models. The estimates of the parameters are as follows:

- Estimated parameters from ARIMA of: (1) the regular autoregressive operator ($\boldsymbol{\phi_p}(B)$) of order p and the regular moving average operator ($\boldsymbol{\theta_q}(B)$) of order q: $\widehat{\phi}_1$, $\widehat{\phi}_2$, $\widehat{\theta}_1$, $\widehat{\theta}_2$, etc.; (2) the seasonal autoregressive operator ($\boldsymbol{\Phi_P}(B^{24})$) of order P and the seasonal moving average operator $\boldsymbol{\Theta_Q}(B^{24})$ of order Q: $\widehat{\Phi}_1$, $\widehat{\Phi}_2$, $\widehat{\Theta}_1$, $\widehat{\Theta}_2$, etc.
- Estimated parameters from TGARCH (1,1) : α_1, η_{11}, β_1, ω, and v (for Student distribution).

The estimate of type 3 variables:

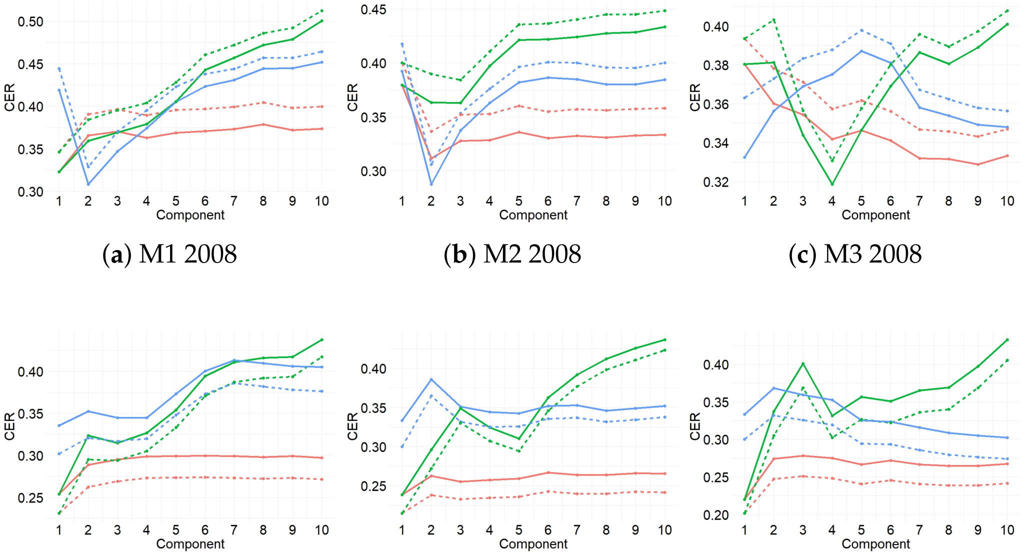
- Variance of the residuals ($\widehat{\sigma^2}$), maximum of spectral density of the residuals ($\widehat{M}_{I(w)}$), frequency corresponding to maximum of spectral density ($\widehat{w}$), mean ($\widehat{\mu}_{\widehat{\rho}_k}$), median ($\widehat{Md}_{\widehat{\rho}_k}$), range ($\widehat{R}_{\widehat{\rho}_k}$) , and variance ($\widehat{\sigma^2}_{\widehat{\rho}_k}$) of sample ACF for 72 lags. The statistic of the SW test ($shap.t$) and the statistic of the KS normality test ($kolg.t$) are also included.

2.3.2.5 | Sensor Classification by Means of sPLS-DA

Once all classification variables were computed as described above for the data from 2008, they were structured in three matrices, one per method (denoted as $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$, respectively), with 23 rows (sensors) where the classification variables are in columns. The total number of variables obtained from each method was 53 for $\mathbf{X}_1$, 141 for $\mathbf{X}_2$, and 59 for $\mathbf{X}_3$. Likewise, regarding 2010, classification variables were structured in three analogous matrices with 20 rows and with the same number of variables.

As the number of classification variables is much greater than the number of sensors, this scenario suggests a high degree of multicollinearity, and it might lead to severely ill-conditioned problems. Different approaches can be considered to deal with this problem. One solution is to perform variable selection, or to apply methods based on projection to latent structures like partial least squares discriminant analysis (PLS-DA).

One advantage of this multivariate tool is that it can handle many noisy and collinear classification variables, being computationally very efficient when the number of variables is much greater than the number of sensors. Even though PLS-DA is extremely efficient in a high-dimensional context, the interpretation of results can be complex in the case of a high number of variables. In such a case, sparse PLS-DA (sPLS-DA) has very satisfying predictive performance, and is able to select informative variables easily. Therefore, it was decided to apply sPLS-DA [63] here using the classification data sets mentioned above in order to identify a small subset of components and classification variables aimed at sensor clustering.



(**a**) M1 2008 (**b**) M2 2008 (**c**) M3 2008

(**d**) M1 2010 (**e**) M2 2010 (**f**) M3 2010

Figure 2.7: Classification error rate (CER) from PLS-DA for 10 components. The CER was computed for each prediction distance (maximum, centroid, and Mahalanobis) per method (M1, M2, and M3) for 2008 and 2010. Two types of error rate are indicated: balanced BER (dashed lines) or Overall (solid lines). Blue lines refer to maximum distance, red lines to centroid distance, and green lines to Mahalanobis distance. PLS-DA was carried out using repeated three-fold CV 1000 times. For 2010, BER and centroid distance showed the best performance achieved by one component. In 2008, M1 and M2 performed the best for maximum distance and Overall, while Mahalanobis distance performed the best in M3. The second-best distance for the three methods was the centroid distance.

The algorithm of sPLS-DA used here was the one proposed by Rohart et al. [68], which corresponds to a modified version developed by Lê Cao et al. [63]. This new version uses the penalty ℓ_1 (lasso) on the loading vector of the regressor matrix by shrinking to zero the coefficient of some variables according to Rohart et al. [68].

With the aim of sensor clustering, sPLS-DA was applied to the previously mentioned classification data sets. Three categories of positions were considered for the sensors: $\mathcal{RC}$, $\mathcal{W}$, and $\mathcal{F}$. This method was applied to the three matrices ($\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$) containing the classification variables, with dimension $n \times p$, where p is the number of classification variables and n is the number of sensors. Furthermore, $\boldsymbol{Y}$ is a vector of length n that indicates the class of each sensor, with values coded as 1 (for $\mathcal{RC}$), 2 ($\mathcal{F}$) and 3 ($\mathcal{W}$). This vector has to be converted into a dummy matrix ($\mathbf{Z}$, i.e., with values either 0 or 1) with dimension $n \times K$, where n is the number of sensors and $K = 3$ the number of classes or positions of sensors.

Before applying sPLS-DA, all anomalous values of each classification variable were removed and considered as missing data after being previously identified using normal probability plots and box plots for each variable. As a result, in 2008: 1.20% (M1), 1.04% (M2), and 1.06% (M3) were the percentages of missing values of the classification data sets. In 2010, the corresponding percentages were 0.40%, 1.39%, and 0.49%, respectively, for each method. These values are relatively low. Furthermore, all classification variables were normalized (i.e., centered and scaled to unitary variance). The package `mixOmics` [68] was used to perform sPLS-DA, which is able to handle missing values by using the NIPALS algorithm [68; 144].

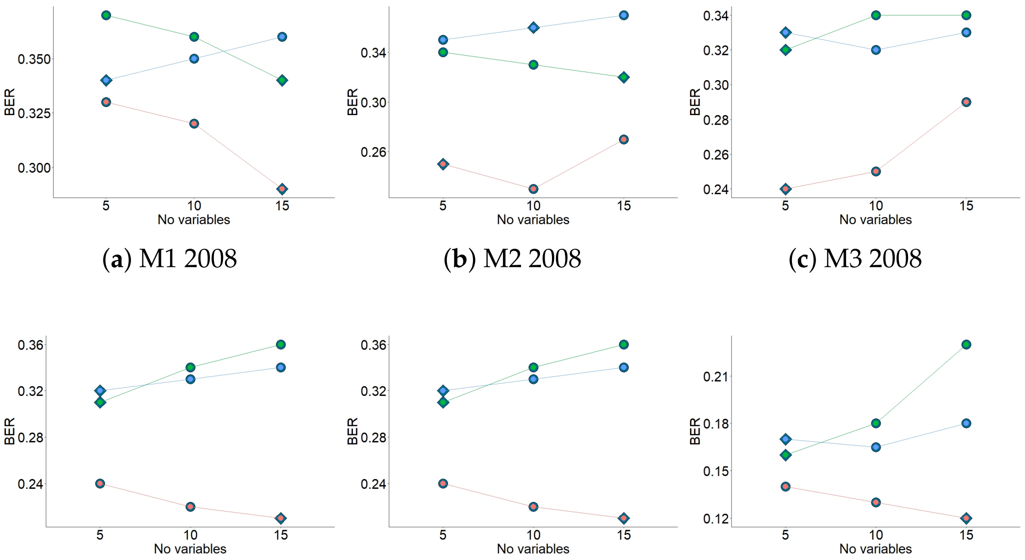
Three-fold cross-validation (three-fold CV, S1 supplementary information of [68]) was used to evaluate the performance (i.e., low classification error rate) of the PLS-DA. It was used to determine both the optimal number of components and the optimal number of variables. The three-fold CV was performed with stratified subsampling, where all positions ($\mathcal{RC}$, $\mathcal{F}$, and $\mathcal{W}$) are represented in each fold.

In order to select the optimal number of components, three-fold CV was applied for a maximum number of ten components, which was repeated 1000 times for each fold. With the objective of assessing the PLS-DA performance, the classification error rate (CER), the overall classification error rate (denoted as Overall), and the balanced classification error rate (BER) were computed [68]. Each BER value corresponds to the average proportion of wrongly classified sensors in each class, weighted by the number of sensors in each class. BER is less biased towards majority classes during the performance assessment when compared with the Overall criterion [68]. Thus, BER was considered instead of the latter.

The classification of sensors was determined according to different prediction distances (PD): maximum, centroid, and Mahalanobis [68]), which were computed for each sensor. Among the three distances calculated, it was found that the centroid one performed better in most cases for the classification, and hence it was selected. Regarding the centroid distance, the software computed the centroid (G) of the learning set of sensors (training data) belonging to the classes ($\mathcal{RC}$, $\mathcal{F}$, and $\mathcal{W}$). Each centroid G was based on the H latent components associated with $\mathbf{X}$. The distances were calculated from the components of the trained model. The position of the new sensor was assigned according to the minimum distance between the predicted score and the centroids G calculated for the three classes considered.

The optimal number of components H was achieved by determining the best performance, based on the BER criterion and prediction distances according to the centroid distance. Once the optimal number of components was determined, repeated three-fold CV was carried out to establish the optimal number of variables according to the criteria of centroid and BER. Finally, once the optimal number of components and variables was decided, the final PLS-DA model was computed.

Figure 2.7 displays the results from the first three-fold CV for the three methods and for both years. For 2010, the values of BER and centroid distance suggested that one component is enough to classify the time series, while for 2008 the results indicate that one or two components are necessary. From this step, the centroid distance and BER were selected in order to determine the number of components.



(**a**) M1 2008 (**b**) M2 2008 (**c**) M3 2008

(**d**) M1 2010 (**e**) M2 2010 (**f**) M3 2010

Figure 2.8: BER according to the number of variables (5, 10 or 15) and different number of components (1: orange dots, 2: blue dots, or 3: green dots) for each method, for 2008 and 2010. Three-fold CV was run 1000 times using centroid distance prediction. Diamonds indicate the optimal number of variables per component according to the lowest value of BER.

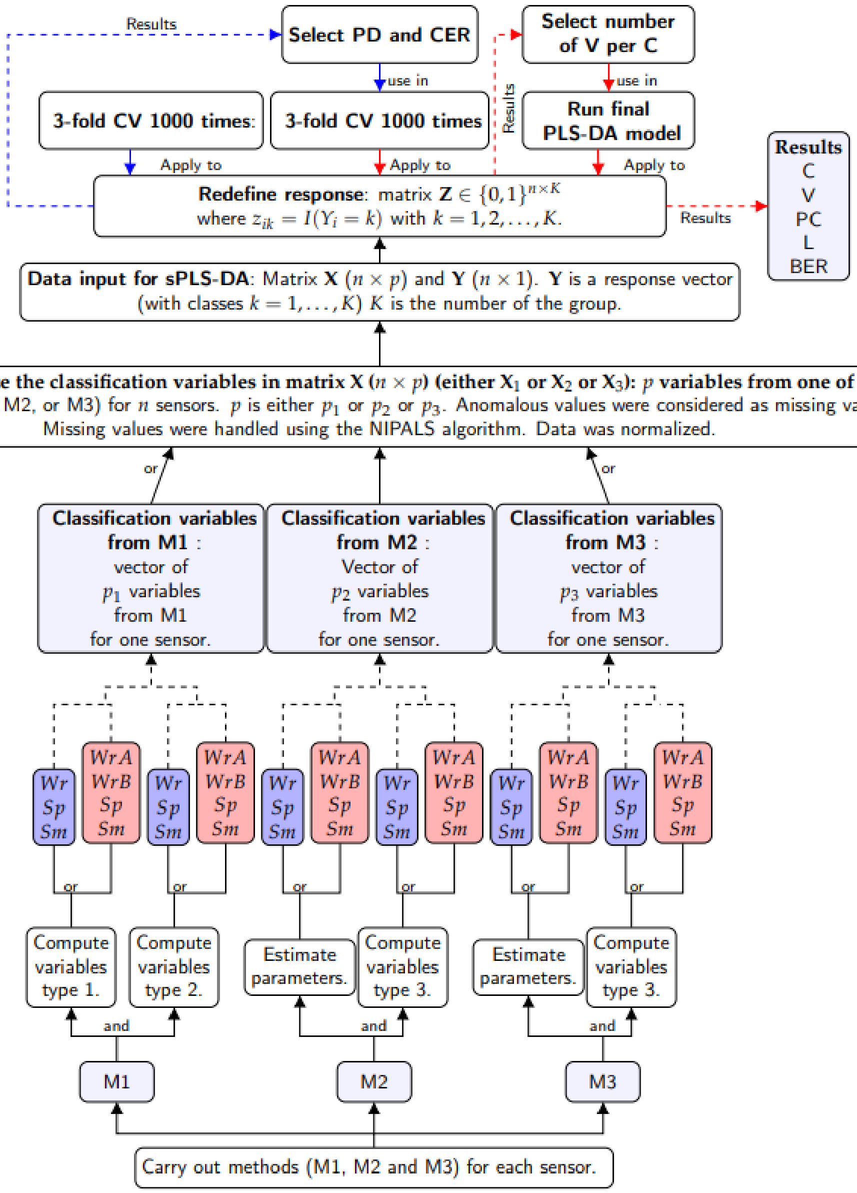

Results
Select PD and CER
Select number of V per C
use in
use in
Results
3-fold CV 1000 times:
3-fold CV 1000 times
Run final PLS-DA model
Apply to
Apply to
Apply to
Results
C
V
PC
L
BER
Redefine response: matrix $\mathbf{Z} \in \{0,1\}^{n \times K}$ where $z_{ik} = I(Y_i = k)$ with $k = 1, 2, \ldots, K$.
Results
Data input for sPLS-DA: Matrix $\mathbf{X}$ $(n \times p)$ and $\mathbf{Y}$ $(n \times 1)$. $\mathbf{Y}$ is a response vector (with classes $k = 1, \ldots, K$) K is the number of the group.
Data: Organize the classification variables in matrix $\mathbf{X}$ $(n \times p)$ (either $\mathbf{X}_1$ or $\mathbf{X}_2$ or $\mathbf{X}_3$): p variables from one of the methods (M1 or M2, or M3) for n sensors. p is either p_1 or p_2 or p_3. Anomalous values were considered as missing values. Missing values were handled using the NIPALS algorithm. Data was normalized.
or
or
Classification variables from M1 : vector of p_1 variables from M1 for one sensor.
Classification variables from M2 : Vector of p_2 variables from M2 for one sensor.
Classification variables from M3 : vector of p_3 variables from M3 for one sensor.
Wr Sp Sm
WrA WrB Sp Sm
or
and
Compute variables type 1.
Compute variables type 2.
Estimate parameters.
Compute variables type 3.
Estimate parameters.
Compute variables type 3.
M1
M2
M3
Carry out methods (M1, M2 and M3) for each sensor.

Figure 2.9: Flow chart for the stages used to apply sparse partial least squares discriminant analysis (sPLS-DA) to the results from the three methods. In the box titled "Data", the information corresponds to the variables from one of the three methods. If the information is from M_i then $\mathbf{X}$=$\mathbf{X}_i$, $i = 1, 2$ and 3. The values were computed for all sensors. Thus, a matrix $\mathbf{X}$ was obtained.

Figure 2.8 shows the results from the second three-fold CV for the three methods and for both years. For 2008, the results suggested that the number of variables per one component were 15 (M1), 10 (M2), and 5 (M3). For 2010, the results suggest the number of variables per one component was 15 for all methods. The information in Figures 2.7

and 2.8 (centroid.dist, BER, and number of variables per component) was used to apply the final PLS-DA. Figure 2.9 describes the steps used to apply sPLS-DA, using the results from the three methods in the study.

Figure 2.9 shows the summary of steps of the sPLS-DA. The values were treated before running the sPLS-DA algorithm. In the box titled "Data input for sPLS-DA", the information corresponds to the response vector converted into a dummy matrix $\mathbf{Z}$. In the following boxes the PLS-DA algorithm runs from left to right. The first three-fold CV was used to evaluate PLS-DA and the prediction distance PD, classification error rate (CER) with the optimal number of components selected. This information was used in the second three-fold CV to check PLS-DA in order to select the optimal number of variables V. The information obtained using both three-fold CVs was used to run the final PLS-DA.

The main outputs from the analysis are: (1) a set of components (C) associated with $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$ for the matrix $\mathbf{Z}$; (2) a set of loading (L) vectors containing the coefficients assigned to each variable that define each component; (3) a list of selected variables (V) from $\mathbf{X}_i$ $(i = 1, 2, 3)$ associated with each component; (4) the values of BER for each component; and (5) the predicted class (PC) for each sensor. Coefficients in a given loading vector indicate the importance of each variable.

2.4 | Results

Components from sPLS-DA are linear combinations of variables that might correspond to WrA, WrB, Sp, or Sm. By applying sPLS-DA to $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$, only one component appeared to be relevant in all cases. The variables selected (per component) by the sPLS-DA algorithm are indicated in the following paragraph. The final model which used the classification variables from M1 (2008) is based on 15 selected variables. The selected model from M2 (2008) consists of 10 variables, while just 5 variables were considered for M3 (2008). The final model for M1, M2, and M3 (2010) comprises 1 component and 15 selected variables from each model (see Table 2.1).

The BER values are indicated in Table 2.1a,b for the three methods, using data from 2008 and 2010, respectively. For both years, the classification variables which turned out to be the most important for the first component were ordered according to the absolute value of their loading weights, from highest to lowest. The notation of the results is $\widehat{M}_{I(w)}$ (`spec.mx`); for $\boldsymbol{RH_h}$: $\widehat{\mu}_{MR}$ (`rMh`), $\widehat{\sigma^2}_{MR}$ (`rVh`); for $\boldsymbol{RH_d}$: $\widehat{\mu}_{MR}$ (`rMd`), $\widehat{\sigma^2}_{MR}$ (`rVd`). Also, SSE (`sse`), $kolg.t$ (`kolg.t`), $\widehat{\sigma^2}$ (res.v), ω (`omega`), $\widehat{\alpha}_2$ (`pacf2`), S_1 (`s1`), S_{18} (`s18`), S_{20} (`s20`), S_{24} (`s24`), α (`alpha`), v (`shape`), $\widehat{\mu}_{\widehat{\rho}_l}$ (`acf.m`), $\widehat{Md}_{\widehat{\rho}_l}$ (`acf.md`). An explanation about the most important variables, regardless of period and year, is presented below.

Table 2.1: Results from *sPLS-DA* (2008 and 2010): variables selected per component (C) and per Method (M) for each period (WrA stands for winter-A, WrB for winter-B, Sp for spring and Sm for summer). For each component, variables are ordered according to the *absolute value* of their loading weights, from highest to lowest. Variables with negative weights are highlighted. The *Balanced classification Error Rate* (BER) is indicated for each component.

M	Variables	BER
1	$\boldsymbol{WrA_{spec.mx}}$, $\boldsymbol{WrA_{rMh}}$, $\boldsymbol{WrB_{rMh}}$, $\boldsymbol{Sp_{rMh}}$, $\boldsymbol{Sm_{rMh}}$, $\boldsymbol{WrA_{rVh}}$, $\boldsymbol{WrB_{rVh}}$, $\boldsymbol{Sp_{rVh}}$, $\boldsymbol{Sm_{rVh}}$, $\boldsymbol{WrA_{rMd}}$, $\boldsymbol{WrB_{rMd}}$, $\boldsymbol{Sp_{rMd}}$, $\boldsymbol{Sm_{rMd}}$, $\boldsymbol{WrA_{rVd}}$, $\boldsymbol{WrB_{rVd}}$	30.02%
2	WrA_{sse}, $WrA_{spec.mx}$, WrB_{sse}, Sp_{sse}, $\boldsymbol{WrA_{kolg.t}}$, $Sp_{spec.mx}$, $\boldsymbol{WrB_{kolg.t}}$, $\boldsymbol{Sm_{kolg.t}}$, Sm_{sse}, $Sm_{spec.mx}$	24.05%
3	$\boldsymbol{WrA_{spec.mx}}$, $\boldsymbol{WrA_{res.v}}$, $\boldsymbol{WrB_{spec.mx}}$, $\boldsymbol{WrB_{res.v}}$, $\boldsymbol{WrA_{omega}}$	22.60%

a Results from *sPLS-DA* (2008).

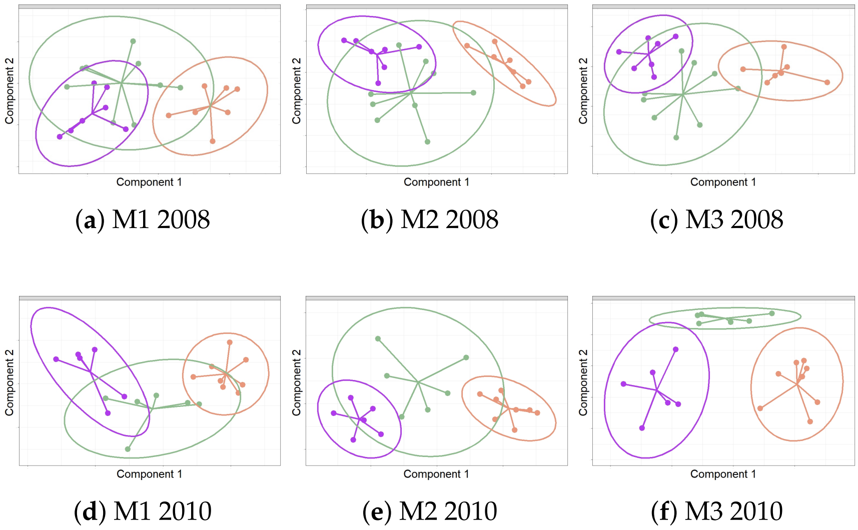
M	Variables	BER
1	$WrA_{spec.mx}$, WrA_{rMd}, WrB_{rMd}, Sp_{rMd}, Sm_{rMd}, WrA_{rMh}, WrB_{rMh}, Sp_{rMh}, Sm_{rMh}, WrA_{rVh}, WrB_{rVh}, Sp_{rVh}, Sm_{rVh}, $WrB_{spec.mx}$, $\boldsymbol{WrB_{pacf2}}$	24.08%
2	$WrB_{spec.mx}$, Sm_{sse}, $\boldsymbol{Sp_{kolg.t}}$, Sp_{sse}, $\boldsymbol{WrA_{kolg.t}}$, WrB_{s1}, $\boldsymbol{Sm_{kolg.t}}$, WrB_{sse}, WrB_{s24}, $Sm_{spec.mx}$, $\boldsymbol{Sm_{s19}}$, $\boldsymbol{Sm_{s18}}$, $\boldsymbol{WrA_{b}}$, $\boldsymbol{Sm_{s20}}$, Sp_{s24}	21.17%
3	$Sp_{res.v}$, $Sm_{res.v}$, $Sm_{spec.mx}$, $WrA_{res.v}$, $WrB_{res.v}$, $Sp_{spec.mx}$, Sp_{omega}, Sm_{omega}, $WrA_{spec.mx}$, WrB_{omega}, $\boldsymbol{WrB_{spec.mx}}$, $\boldsymbol{WrA_{alpha}}$, $\boldsymbol{WrA_{shape}}$, $\boldsymbol{Sp_{acf.md}}$, $\boldsymbol{Sp_{acf.m}}$	12.81%

b Results from *sPLS-DA* (2010).

- M1: `spec.mx`, `rMh`, `rMd`, `rVh`, `rVd`, and `pacf2` (see Table 2.1). The features `rMh` and `rMd` account for changes in the mean of the time series, while `rVh` and `rVd` are intended to explain changes in the variance. The rest of the features mentioned provide information about the dynamic structure of each time series. It was found that `rMh`, `rMd`, and `rVh` were important in the four periods considered, both in 2008 and 2010. `rMd` was relevant for WrA and WrB in 2008. The variable `spec.mx` was relevant in WrA and WrB for 2008 and 2010, as well as WrB. The variable `pacf2` was found in WrB 2010. Hence, consistent results were derived from the two years under study.
- M2: `sse`, `kolg.d`, and `spec.mx` (computed from the residuals), as well as `b`, `s1`, `s18`, `s19`, `s20`, and `s24` (from the models). From the residuals, `sse` accounts for the variance that is not explained by the models. This parameter appeared as important in all periods considered, except WrA 2010. `kolg.d` quantifies the deviation from normality for the residuals, and was relevant in all periods except Sp 2008 and WrB 2010. The third feature, `spec.mx`, which provides information about the dynamic structure of each time series, was relevant for all periods except WrB 2008, WrA 2010, and Sp 2010. Regarding the parameters computed from the mo-

dels, `b` is related to the trend component of the time series, which was important in WrA 2010. The other variables mentioned are related to the seasonal components of the time series, which were shown to be important in Sm 2010.

- M3: `res.v`, `shape`, `spec.mx`, `acf.m`, and `acf.md` (computed from the residuals), as well as `omega` and `alpha` (from the model). From the residuals, `res.v` is aimed to explain the variance not explained by the models. It was relevant in all periods except Sp and Sm 2008. The variable `shape` provides information about the distribution of residuals, but it was only relevant in WrA 2010. The other features (i.e., `spec.mx`, `acf.m`, and `acf.md`) are intended to describe the dynamic structure of each time series. `Spec.mx` was important in all periods except Sp and Sm 2008, while the last two only appeared in Sp 2010. Regarding the parameters from the models, `omega` explains the changes in the mean of the conditional variance, while `alpha` quantifies the impact of the rotation on the conditional variance. The variable `alpha` only appeared in WrA 2010. Again, the fact that most variables were common in the three periods and in both years suggests strong consistency in the underlying phenomena explaining the discrimination between sensors.



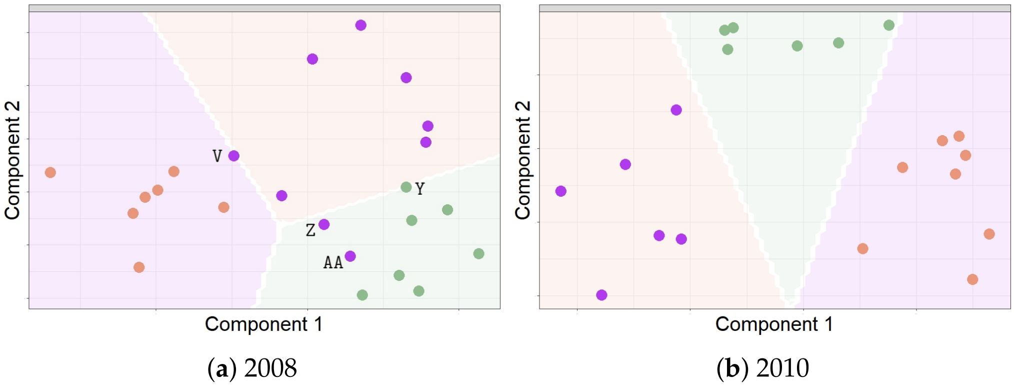
(**a**) M1 2008 (**b**) M2 2008 (**c**) M3 2008

(**d**) M1 2010 (**e**) M2 2010 (**f**) M3 2010

Figure 2.10: Discrimination of the time series of RH according to the position of sensors: Frescoes ($\mathcal{F}$), Cornice and Ribs ($\mathcal{RC}$), and Wall ($\mathcal{W}$). Color codes: $\mathcal{F}$ sensors are shown in green, $\mathcal{RC}$ in orange, and $\mathcal{W}$ in purple. Graphics correspond to the projection of sensors over the first two components from sPLS-DA. Each graph shows confidence ellipses for each class to highlight the strength of the discrimination at a confidence level of 95%.

In all cases, the classification variables corresponded to the different parts of the time series (WrA, WrB, Sp, and Sm), except for M3 in 2008 which only showed variables from winter (see Table 2.1).

The results shown in Figure 2.10 correspond to the score plots for the first two components from sPLS-DA applied to the classification of sensors. They depict their projection over the two principal latent structures that best discriminate sensors according to their position. In 2008, the first component for each method allowed a rather good discrimination of sensors at the $\mathcal{RC}$ position with respect to the rest, though a poor discrimination was achieved between $\mathcal{F}$ and $\mathcal{W}$ (see Figure 2.10a–c).



(**a**) 2008 (**b**) 2010

Figure 2.11: Prediction classes derived from M3: these plots display the predicted classes of each sensor located in different points in a simulated grid. The final classification prediction for the position of the sensors is displayed. The points in orange ($\mathcal{RC}$), purple ($\mathcal{W}$) and green ($\mathcal{F}$) represent the class prediction for the sensors in the study. (**a**) For 2008: according to the classification prediction, the wrongly classified sensors were as follows: `V`, `Y`, `Z`, and `AA`. (**b**) For 2010: all sensors were classified correctly.

In 2010, the first and second components for M3 displayed a clear discrimination between sensors located on the three positions. However, for M1 and M2, only $\mathcal{RC}$ sensors appear far apart from those on the walls, while the $\mathcal{F}$ group is located in between (see Figure 2.10d–f).

Regarding the performance of the three methods for achieving the classification of sensors, the best results were derived from M3 and the worst from M1. M3 yielded higher correct classification percentages: 77.40% in 2008 and 87.19% in 2010 (see Table 2.1b). For 2008, the final classification resulting from M3 variables displayed the following wrongly classified sensors: `Y`, `AA`, `Z`, and `V` (see Figure 2.11a). Three of them (`Y`, `Z`, and `AA`) were installed near the location where the salt efflorescence was found. For 2010, the final results from sPLS-DA for M3 showed that all sensors were classified correctly (see Figure 2.11b).

2.5 | Discussion

The methodology proposed here consists of using sPLS-DA to classify time series of RH according to classification variables that were computed from different functions (e.g., sample ACF, sample PACF, spectral density, and MR). Additionally, the Seasonal ARIMA-TGARCH-Student model and the Additive SH-W method were used. Furthermore, estimated parameters of the models, as well as the mean, variance, and maximum values of the functions (e.g., sample ACF, sample PACF, spectral density, and statistics of the KS normality test, among others) were applied to the residuals derived from the models. The centroid distance was applied to classify the sensors, and the lasso penalty was used to select the optimal variables that determine the relevant components. Additionally, the BER parameter was employed to evaluate the performance of the classification methodology.

We used sPLS-DA because the classification data in this study are characterized by more variables than the number of time series (sensors), and in the interest of easily interpreting the results. This technique leads to underlying latent variables (components) that summarize the relevant information from the data for the purpose of discrimination. It performs variable selection for each component, which is an advantage. The key issue in time series clustering is how to characterize the similarities and dissimilarities between time series. Various metrics for measuring such similarity have been proposed, based on: parameters from models [38; 39; 40; 145; 146], serial features extracted from the original time series [27; 28; 29; 30], the complexity of the time series [31; 32; 34; 35; 147; 148], the properties of the predictions [36; 37], and the comparison of raw data [25]. Regarding methods based on model parameters, the criterion most commonly considered is to assume that time series are properly explained by ARIMA processes. Piccolo [38] introduced the Euclidean distance between their corresponding AR expansion [12] as a metric and used a complete linkage clustering algorithm to construct a dendrogram. One problem of this metric is related to the numerical computations of AR coefficients. The same metric was also considered by Otranto [149] for dealing with GARCH processes. For ARMA models, Maharaj [41] developed an agglomerative hierarchical clustering procedure based on the p-value of a hypothesis test applied to every pair of stationary time series. Kalpakis et al. [42] studied the clustering of ARIMA time series by using the Euclidean distance between the linear predictive coding (LPC) cepstrum of two time series as their dissimilarity measure. Xiong and Yeung [43] classified univariate ARIMA time series by considering ARMA models. They derived an expectation maximization (EM) algorithm for estimating the coefficients and parameters of the models. However, if the underlying clusters are very close to each other, the clustering performance might diminish significantly. According to the review of the previously mentioned studies about the clustering of time series, it seems that the methodology applied here is rather unique because it uses a hybrid model comprising ARIMA and GARCH to calculate a distance for classifying time series. This is also probably the first using sPLS-DA in order to classify time series.

We found that the time series of RH, one per sensor, were very similar despite their different positions in the apse vault of the cathedral. When classifying the sensors, it turned out that few parameters appeared as relevant, most of which were features extracted from the residuals of models. This is most likely due to the similarity among the time series studied. As a consequence, the information that was not properly explained by the models was decisive for characterizing the differences between time series. The classification variables derived from the ARIMA-TGARCH-Student model yielded better performance than those from SH-W, which might suggest that the former model captures more information from the data than the latter. In fact, SH-W is an algorithm intended for producing point forecasts [125].

A comparison of the results from method M1 with those from the preliminary study of Zarzo et al. [7] indicates that a better classification was obtained here. The variables `HMV` and `DMV` (`rMh` and `rMd`) were relevant in both studies. Although the mean for the total observations was important in the preliminary project [7], this variable was not selected by the sPLS-DA. The classification variables selected per sPLS-DA explained the changes in the mean and variance of the time series with `rMh`, `rMd`, `rVh`, and `rVd`. Furthermore, the method obtains variables from sample PACF and spectral density which explain the autocorrelation of the time series. The research by Zarzo et al. [7] did not use variables related to the autocorrelation of the time series.

One disadvantage of sPLS-DA is the need to use the same number of classification variables for each sensor. As a consequence, a unique ARIMA-TGARCH-Student model was used for all sensors in the same part of the time series (WrA, WrB, Sp, and Sm). This means that a better fit might result, as it considers a different model for each time series. Another disadvantage is that it is necessary to know a priori the number of classes of the time series (sensors) for their classification. According to the previous idea, the limitations of the statistical methodology proposed in this study are: (1) sPLS-DA needs to know the number of classes before implementing the algorithm. (2) When applying a unique ARIMA-TGARCH parametric model to all sensors, it is unlikely that the best values for the classification variables will be found. This can affect the classification error rate of the sensors.

One advantage of using both sPLS-DA and ARIMA-TGARCH-Student is the capability of classifying time series with very similar characteristics. Additionally, the functions and models utilized here can be easily implemented because different packages are available in `R` software. One such example is the `mixOmics` package of `R`, which has different functions that allow sPLS-DA to be implemented simply and makes it easy to display the different results for interpretation. Furthermore, this package can handle missing values using the NIPALS approach. It takes advantage of the PLS algorithm which performs local regressions on the latent components. There are two main advantages of using PLS—it both handles missing values and calculates the components sequentially. In this study, the anomalous values of classification variables were considered as missing values in order to avoid possible problems with the classification of the sensors. The percentage of values that were used as missing were lower than 2%.

In relation to future studies, alternative classification variables could be considered depending on the different scenarios and according to the characteristics of the time series. In order to obtain classification variables that capture more information from the data, flexible models can be proposed. Some options for calculating the classification variables might be the following:

- Cepstral coefficients: Ioannou et al. [150] studied several clustering techniques in the context of the semiparametric model: spectral density ratio. They found that the cepstral- based techniques performed better than all the other spectral-domain-based methods, even for relatively small subsequences.

- Structural time series model: the flexibility required from this model can be achieved by letting the regression coefficients change over time [151].

- A nonparametric approach of the GARCH [152; 153].

Regarding classification techniques when there are fewer variables than time series, sPLS-DA can be extended by using the elastic net [154] as the penalization. Finally, a further study might be carried out in controlled scenarios, where time series can be computationally simulated by controlling different characteristics in order to identify the strengths and weaknesses of the proposed methodology. In alignment with the previous ideas for improving the methodology, future research will use sPLS-DA with two methods: a nonparametric Seasonal ARIMA-GARCH model and a structural time series model. Furthermore, several time series will be computationally simulated in controlled scenarios in order to evaluate the results when using sPLS-DA, together with one of the previously mentioned methods.

García-Diego and Zarzo [6] concluded that the environment surrounding the Renaissance frescoes was not the same at all points of the apse vault of the cathedral. Sensors located on the walls or on the paintings registered higher RH values than those in the vault ribs. Thus, the mean value of RH is related with the three previously mentioned classes. The ideal goal is obviously to achieve a correct classification of all sensors. However, a poor classification error rate might be caused by either the malfunctioning of some sensors or a poor performance of the classification technique, or if there is a problem related with the microclimate where the sensors are located. Those sensors incorrectly classified by the technique should be checked to identify possible moisture problems in the artworks. In this work, the main cause of sensor malfunction was the development of salt deposits around the probes as a consequence of fitting some of the probes inside the layer of plaster supporting the frescoes.

The results indicate that sPLS-DA could be implemented for the online monitoring of fresco paintings aimed at preventive conservation using the parameters and features previously extracted from the hybrid models based on GARCH and ARIMA as classification variables. This analysis might be carried out for each season of every year.

2.6 | Conclusion

The methodology proposed here is useful for understanding the differences in thermo-hygrometric conditions monitored inside large buildings or museums, which might provide a basis for better assessing the potential risks related to temperature and humidity on the artworks. Among the methods proposed, a hybrid approach based on ARIMA and GARCH models with sPLS-DA yielded the best performance. Parsimonious models with a small subset of components and classification variables were obtained using sPLS-DA, which offers satisfactory results with easy interpretation. Another advantage of sPLS-DA is that it can be implemented easily with `mixOmics`, which allows a focus on graphical representation in order to better understand the relationships between the different observations and variables. Furthermore, this package can deal with missing values. Finally, the use of a hybrid approach based on ARIMA and GARCH models as well as sPLS-DA is a novel proposal for classifying different time series.

In order to improve the methodology proposed in this research, future research will use sPLS-DA with two methods that are more flexible than those applied in this study. This will capture more information from the data. Furthermore, a computational simulation will be carried out in order to evaluate the new methodology in different possible scenarios.

3

Multivariate Time Series Analysis of Temperatures in the Archaeological Museum of L’Almoina (Valencia, Spain)

Ramírez, Sandra.; Zarzo, M.; Garcia-Diego, F.J. *Multivariate Time Series Analysis of Temperatures in the Archaeological Museum of L’Almoina (Valencia, Spain)*. Sensors 2021, 21 (13): 4377. doi:10.3390/s21134377.

This chapter corresponds to the publication mentioned above, after changing the positions of some graphs and tables.

3.1 | Abstract

An earlier study carried out in 2010 at the archaeological site of L'Almoina (Valencia, Spain) found marked daily fluctuations of temperature, especially in summer. Such pronounced gradient is due to the design of the museum, which includes a skylight as a ceiling, covering part of the remains in the museum. In this study, it was found that the thermal conditions are not homogeneous and vary at different points of the museum and along the year. According to the European Standard EN10829, it is necessary to define a plan for long-term monitoring, elaboration and study of the microclimatic data, in order to preserve the artifacts. With the aforementioned goal of extending the study and offering a tool to monitor the microclimate, a new statistical methodology is proposed. For this propose, during one year (October 2019–October 2020), a set of 27 data-loggers was installed, aimed at recording the temperature inside the museum. By applying principal component analysis and k-means, three different microclimates were established. In order to characterize the differences among the three zones, two statistical techniques were put forward. Firstly, Sparse Partial Least Squares Discriminant Analysis (sPLS-DA) was applied to a set of 671 variables extracted from the time series. The second approach consisted of using a random forest algorithm, based on the same functions and variables employed by the first methodology. Both approaches allowed the identification of the main variables that best explain the differences between zones. According to the results, it is possible to establish a representative subset of sensors recommended for the long-term monitoring of temperatures at the museum. The statistical approach proposed here is very effective for discriminant time series analysis and for explaining the differences in microclimate when a net of sensors is installed in historical buildings or museums.

Keywords: ARIMA; art conservation; Holt–Winters; k-means; random forest; sensor diagnosis; sPLS-DA

3.2 | Introduction

The environmental conditions of historical buildings, exhibition facilities and storage areas in museums have been shown to be the most crucial factor in the preservation of collections and artifacts. Temperature, humidity and lighting can potentially deteriorate or even destroy historical or cultural objects that are kept, protected and displayed in collections [155]. A continuous monitoring of the indoor environment can provide information about the microclimatic conditions affecting the works of art. Monitoring is an essential tool for developing a preventive control program aimed at maintaining the optimal microclimatic conditions for preservation. As a consequence, long-term monitoring has to be applied to prevent the deterioration of artworks [99]. Furthermore, it is necessary to find practical solutions and tools for the incorporation of climate change adaptation in the preservation and management of cultural heritage [77]. In particular, in archaeological sites, temperature differences between various minerals in block surfaces and alternative surfaces cause thermal stress. Humidity and thermal stresses are important causes of microfractures between the mineral grains of blocks [99].

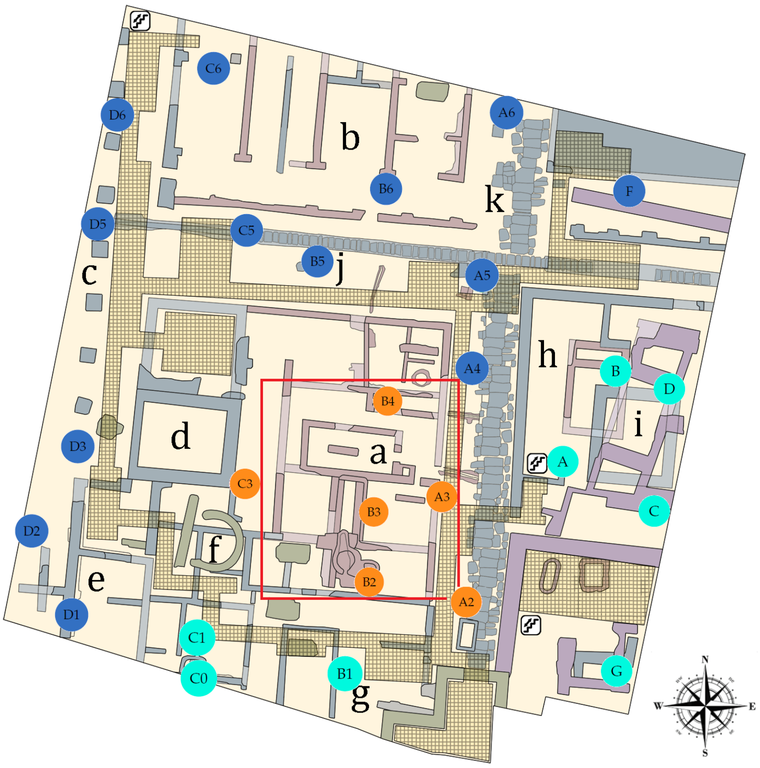


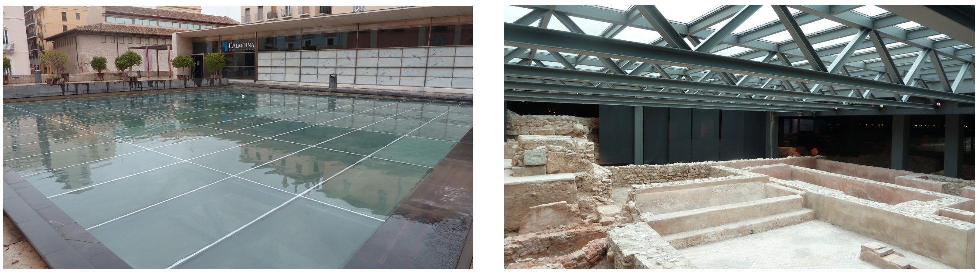
Figure 3.1: Plan of the L'Almoina archaeological site, indicating the position of 27 dataloggers for monitoring the air conditions inside the museum. Based on the multivariate analysis of temperatures (see Section 3.4), three zones were established: North West (NW, in blue), South East (SE, in green) and Skylight (Sk, in orange). The different observable structures and the construction phases in the museum are indicated: (a) Roman baths; (b) Imperial granary; (c) Portico of the imperial forum; (d) Imperial chapel; (e) Imperial basilica; (f) Byzantine apse and tombs; (g) Byzantine Cathedral Baptistery; (h) Republican and Imperial Asklepieion; (i) Alcázar Aldalusí; (j) Decumani; and (k) Cardus [156].

In L'Almoina museum (Valencia, Spain), a pronounced gradient of temperature was found [157] due to a skylight which was included in the architectonical design of the museum (Figure 3.1). A thermo-hygrometric monitoring study carried out in this museum in 2010 [157] discussed the significant effect of the skylight on the variations in T and RH. A pronounced greenhouse effect was noted, as a consequence of the skylight and the high temperatures reached in summer in Valencia. In 2011 the layer of water was removed due to a leak that had to be repaired. To replace the beneficial effect of the water layer, a provisional canvas cover was installed directly over the skylight, in order to avoid the overheating of the archaeological site below, by preventing direct sunlight. A second thermo-hygrometric monitoring study was performed in 2011 to assess the effect of different corrective measures and changes implemented in the museum [158]. The microclimatic data of RH and T recorded in 2010 before laying the canvas cover was compared with air conditions in 2013, after its installation. It was found that the presence of the canvas covering the skylight improved the T and RH conditions, so that the microclimate was in accordance with the international standards [159; 160].

Given the marked detrimental influence of the skylight, a long-term monitoring is required for the control of thermal conditions. Thus, it is necessary to find practical solutions and tools for the preservation and management of the ruins. For this purpose, a statistical methodology for classifying different time series of temperature that are very similar is of interest. Such methodology can help to characterize different zones in the museum and to provide guidelines for monitoring the thermal conditions.

Some studies have been carried out in order to propose a plan for monitoring either temperature (T) or relative humidity (RH) for art conservation. Three reported studies proposed a methodology for classifying different time series using either observations of time series, or features from time series. García-Diego and Zarzo [6] and Zarzo et al. [7] applied Principal Component Analysis (PCA) in order to study the values of RH from sensors installed in different positions at the apse vault of Valencia's Cathedral (Spain). Zarzo et al. [7] reported that the first and second principal components could be estimated according to a linear combination of the average RH values and the moving range of RH. Based on the two first components, the differences between the time series of RH in different positions in the apse were discussed. Ramírez, Sandra et al. [161] proposed a statistical methodology in order to classify different time series of RH, which pointed to those zones with moisture problems in the apse vault of the Cathedral of Valencia. Merello et al. [11] analyzed 26 different time series of RH and T, recorded at Ariadne's house (Pompeii, Italy), using graphical descriptive methods and Analysis of Variance (ANOVA), in order to assess the risks for long-term conservation of the mural paintings in the house. The work provided guidelines about the type, calibration, number and position of thermo-hygrometric sensors in outdoor or semi-confined environments.

The proposed methodology in this article has an adequate capability for discriminating time series with similar features. In addition, it can help to obtain parsimonious models with a small subset of variables leading to a satisfactory discrimination. As a consequence, its results can be easily interpreted and can help to select a subset of representative sensors for the long-term monitoring of indoor air conditions inside the museum. Finally, this methodology can be effective in order to establish the different zones in the archaeological site and to discriminate the microclimate of these areas.



(**a**) (**b**)

Figure 3.2: View of the skylight covering part of L'Almoina archaeological site: (**a**) external view from the pedestrian plaza; and (**b**) internal view.

Aimed at better understanding the differences of microclimate in L'Almoina archaeological site (Figure 3.2), a set of 27 autonomous data-loggers was installed at different points of the museum (Figure 3.1). The time period under study was of about one year, from 22 October 2019 to 20 October 2020. The main goal of this research was to identify different microclimates at the museum and to characterize the differences in temperature between such zones. The purpose was to classify the sensors according to features and variables extracted from the time series of T. Another target was to identify those variables that best discriminate the different time series per zone. For this purpose, a methodology was applied based on sparse partial least squares discriminant analysis (sPLS-DA) and random forest (RF) with models and functions of time series [161; 162]. Another target was to identify a subset of representative sensors for a long-term monitoring of thermic air conditions in the museum, as well as to determine the best location recommended for these sensors. The proposed methodology is rather new in the context of clustering of time series applied to cultural heritage. Furthermore, this methodology can be useful for defining different zones in the museum, according to features of the time series of T, as well as to achieve a correct classification of all sensors in such zones.

This article is structured as follows. In Section 3.3, a short background related to art conservation is presented. Characteristics of the dataset and the sensors, criteria for determining the stages of the time series of T, methods for calculating features of the time series and strategies for classifying time series are introduced in Section 3.4. The most notable results of the different analyses and their discussion are presented in Section 3.5. Finally, conclusions appear in Section 3.6.

3.3 | Background

3.3.1 | Studies for the Long-Term Preservation of Artworks

Many studies have been conducted in recent years which monitor the climatic parameters for the long-term preservation of cultural heritage. Frasca et al. [86] studied the impact of T, RH and carbon dioxide (CO_2) on the organic and hygroscopic artworks in the church of Mogiła Abbey. They found that artworks were at high risk of mechanical damage for approximately 15% of the time under study, due to an excessive variability of RH. Huijbregts et al. [163] proposed a method for evaluating the damage risk of long-term climate changes on artifacts in museums or historic buildings. This method was applied for two historic museums in the Netherlands and Belgium. For the examined case studies, they found that the expected climate change would significantly increase both indoor T and RH, with the increase of the latter having the highest impact on the damage potential for artifacts in museums. Zítek and Vyhlídal [164] proposed a novel air humidity control technique for preventing the moisture sensitive materials from varying the equilibrium of their moisture content, maintaining desirable environmental conditions for the preventive conservation of cultural heritage. Angelini et al. [4] designed and installed a wireless network of sensors for monitoring T and RH, aimed at establishing a correlation between the environmental conditions and the conservation state of artifacts. In addition, Lourenço et al. [5] studied air T and RH, among other parameters, in the historical city center of Bragaça (Portugal). Other researchers have studied the variation of some environmental parameters, in order to identify the main factors involved in the deterioration of certain remains, e.g., exposed and buried remains at the fourth century Roman villa of Chedworth in Gloucester, England [165].

3.3.2 | European Standards

Experts suggest that it is necessary to investigate the actual environmental dynamics in a museum before any structural intervention. Furthermore, it is important to define the compatibility between the climate control potentials and the preservation requirements [99]. Several European Standards [92; 93; 94; 95; 96; 97; 98] have been developed for the monitoring, elaboration and study of the microclimatic data, as supporting actions for the preservation of artifacts. Long-term monitoring is required, as well as an appropriate statistical approach for the data management.

A large economical investment is being provided by governments within the European Union to preserve artworks in museums. Different research projects have monitored the indoor microclimate within museums, in order to analyze the relationship between thermo-hygrometric conditions and the degradation of materials, from which works of art are made. For example, with the goal of preserving artwork and artifacts, the CollectionCare Project is currently working on the development of an innovative system for making decisions about the preventive conservation of artworks in small- to medium-sized museums and private collections [166].

3.3.3 | Characteristics of the L'Almoina Museum

The archaeological site of L'Almoina in Valencia (Spain) is an underground museum at about 3 m below the city floor level. It occupies an area of about 2500 m^2. The archaeological remains are covered by a concrete structure, which forms an elevated plaza above the ruins. This cover connects with walkways and steps at different heights along its perimeter. There is no vertical retaining wall inside the museum to isolate the remains and prevent water diffusion through capillarity from the surrounding areas. An external glass skylight (225 m^2) was adapted to the museum so that part of the ruins could be observed from the pedestrian plaza. Nowadays, the skylight is protected by a layer of water (see Figure 3.2) to prevent high temperatures of the glass.

3.4 | Materials and Methods

3.4.1 | Materials: Description of the Datasets

In total, 27 data-loggers were installed for the monitoring of T and RH at L'Almoina museum. Technical details are as follows: data-logger model WS-DS102-1, with a temperature range between -40 and $+60\,^{\circ}$C and an accuracy according to the MI-SOL manufacturer of $\pm 1\,^{\circ}$C, under 0–50 $^{\circ}$C [167]. The data acquisition rate was of one recording every five minutes, so that a manual download of data was necessary every two months. The monitoring experiment started on 22 October 2019 and ended on 20 October 2020. The initial number of available observations of T was approximately 104,832 per sensor (i.e., 364 days$\cdot$24 h/day$\cdot$12 values/h), which were arranged as a matrix containing 27 columns (i.e., temperatures recorded by each sensor) by 104,832 rows. Daily cycles are clearly marked, which implies a repetitive pattern every 288 values, but such a data sequence seems too long. Thus, it was decided to calculate the median of values recorded per hour, which leads to daily cycles every 24 values. This frequency seems more convenient for the use of seasonal methods of time series analysis. Thus, a new matrix was arranged comprising 8730 observations by 27 sensors. This dataset did not contain missing values.

3.4.2 | Data Calibration

All sensors were calibrated prior to their installation by means of an experiment carried out inside a climatic chamber, model alpha 990-40H from Design Environmental Ltd. (Gwent, UK). The temperature was maintained at three different levels: 5, 23 and 30 °C. For each stage, the RH was 75%, 50% and 30%, respectively. Each stage of temperature was maintained for 2 h, so that the total calibration experiment lasted for 6 h. The frequency of temperature recorded was one datum per five minutes from each sensor. Next, the median of T per hour was calculated by sensor. Linear regression was applied to obtain calibration functions, one per sensor, relating the measured median of T as a function of the real temperature inside the chamber. Finally, the temperature matrix (8730 × 27) was modified by correcting the bias of each sensor, according to the resulting calibration functions, which leads to a corrected matrix containing the median temperature per hour, after the calibration.

3.4.3 | Statistical Methods

The statistical methodology comprises the following steps: (1) Identify structural breaks in all time series. (2) Extract features directly from the time series. (3) Compute classification variables using the additive seasonal Holt–Winters approach. (4) Compute classification variables by means of seasonal ARIMA. (5) Compute classification variables according to the Wold decomposition. (6) Determine the optimum number of classes and establish the class for each sensor, using PCA and k-means algorithm. (7) Check the classification of sensors using sPLS-DA and identify the optimal subset of variables that best discriminate between classes, per method. (8) Check the classification of sensors by means of the RF algorithm and identify the optimal set of variables calculated from each method, that best discriminate between the classes. (9) Propose a methodology for selecting a subset of representative sensors for future long-term monitoring experiments in the museum.

The main R software packages [108] (version 4.3) used to carry out the statistical analyses were: mixOmics [67; 68], aTSA [109], forecast [110; 111], strucchange [113], tseries [114], moments [168], PerformanceAnalytics [169], NbClust [170] and QuantTools [105].

3.4.3.1 | Identification of Structural Breaks in the Time Series

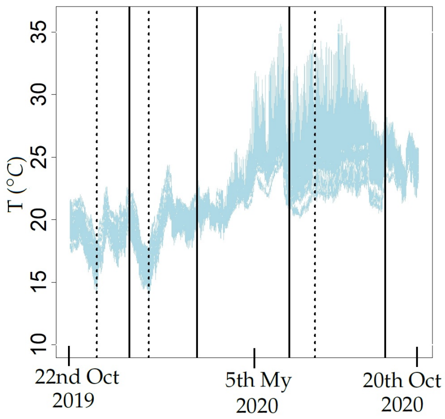


Figure 3.3: Trajectories of the different time series of T from the 27 sensors located in the museum. Values were recorded between 22 October 2019 and 20 October 2020. The separation of different stages (Wr1, Cd, Tr, Ht and Wr2) is indicated by means of solid vertical lines. Dashed vertical lines indicate the structural breaks identified within the stages Wr1, Cd and Ht.

In real conditions, time series can undergo sudden shifts or changes in the slope of a linear trend. Such events are known as structural breaks [12]. The CUSUM and supF tests, among others, can be used to detect structural breaks in a time series [115; 120]. By carefully inspecting the evolution of all observed time series of temperature (***T***) over time (Figure 3.3), certain potential structural breaks can be observed. Both the CUSUM and supF tests were applied after computing the logarithmic transformation and one regular differentiation to the time series. Such logarithmic transformation was intended to stabilize the variance, while the regular differentiation was used to remove the trend of the different time series [107]. The notation used throughout this paper is the following: ***r*** refers to the logarithmic transformation of ***T*** and ***W*** denotes one regular differentiation of the logarithmic transformation of ***T***. Thus, each value of ***W*** corresponds to $w_t = r_t - r_{t-1}$, being $r_t = \ln(T_t), t = 1, \ldots, t_{max}$. The tests were computed with functions Fstats and efp from the strucchange package [113]. Initially, 5 stages were tentatively established: warm 1 (Wr1, comprising n = 1490 observations), cold (Cd, n = 1703), transition (Tr, n = 2303), hot (Ht, n = 2327) and warm 2 (Wr2, n = 903) (see Figure 3.3). Wr1 corresponds to 22 October–23 December 2019, Cd to 24 December–3 March 2020, Tr to 4 March–7 June, Ht to 8 June–12 September and Wr2 to 13 September–20 October 2020.

3.4.3.2 | Calculation of Classification Variables—Method M1

This method consists of computing features such as mean, median and maximum, among others, from estimates of the Auto Correlation Function (ACF), Partial Auto Correlation Function (PACF), spectral density and Moving Range (MR) [18; 104]. The ACF and PACF correlograms of the observed time series are commonly used to fit Auto Regressive Moving Average (ARMA) models. Firstly, a set of variables denoted as Type 1 comprised the mean, MR and PACF, which were estimated for the 8 stages of $\boldsymbol{T}$ values. The average of $\boldsymbol{T}$ was calculated to capture the level or position in each stage of the time series. The mean of MR with order 2 (i.e., average range over 2 past values) was computed to identify sudden shifts or increases in the level of $\boldsymbol{T}$. For each stage of $\boldsymbol{T}$, the sample PACF parameters (α_l at lag l) were estimated for the first four lags (l = 1, 2, 3, 4), which are usually the most important ones for capturing the relevant information in time series.

Secondly, another set of variables called Type 2 comprised spectral density and ACF, which were estimated for $\boldsymbol{T}$ values after applying the logarithm transformation and regular differencing. The objective of using this transformation and differencing was to stabilize the variances and remove the trend of $\boldsymbol{T}$, so that the computed variables provide information about the seasonal component of the time series. Spectral density was estimated using the periodogram of observed time series $\boldsymbol{W}$ ($I(w)$ of signals w). The maximum peak of the periodogram and its frequency were identified. Values of ACF (ρ_l at lag l) were estimated to analyze the correlation between $\boldsymbol{W}$ values with the lagged values of the same observed time series at each lag, for the first 72 lags. This criterion was used because the values of the ACF correlogram for further lags were comprised within the limits of a 95% confidence interval in the correlogram.

The different steps involved in M1 are depicted in Figure 3.4. Firstly, the 27 time series were split according to the climatic stages observed: Wr1, Cd, Tr, Ht and Wr2 (Data 2). Secondly, some of the main stages (Data 2) were subdivided according to the structural breaks (SB) identified in Wr1, Cd and Ht (Data 3). In the third step (Data 4), the logarithm transformation was applied and, next, one regular differentiation (Data 5). The fifth step consists of applying the formulas of Type 2 variables to w_t: maximum of periodogram ($M_{I(w)}$) and its frequency (w), as well as mean, median, range and variance of the sample ACF for the first 72 lags ($\mu_{\widehat{\rho}_l}$, $Md_{\widehat{\rho}_l}$, $R_{\widehat{\rho}_l}$ and $\sigma^2_{\widehat{\rho}_l}$, with $l = 1, \ldots, 72$). Finally, the formulas of Type 1 variables were applied to $\boldsymbol{T}$ values (Data 3): mean of $\boldsymbol{T}$ (μ_T), mean of MR of order 2 (μ_{MR}) and PACF for the first four lags (α_1, α_2, α_3 and α_4).

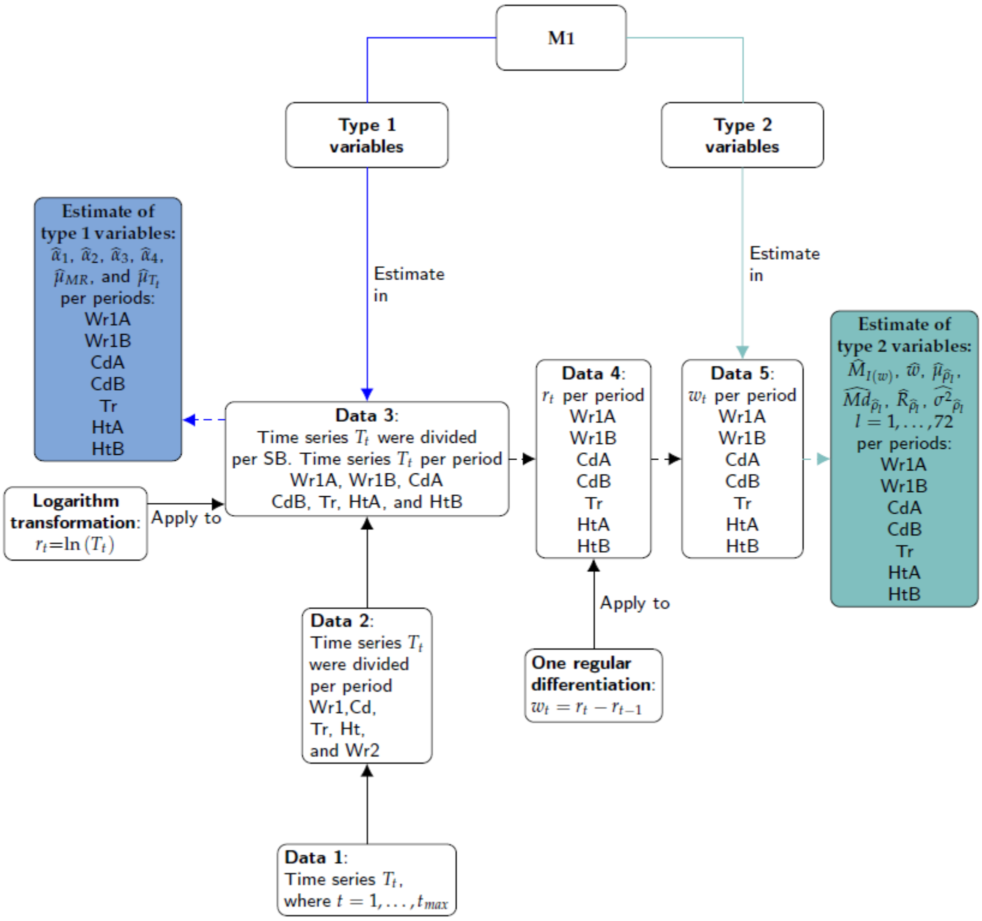

M1
Type 1 variables
Type 2 variables
Estimate in
Estimate in
Estimate of type 1 variables: $\widehat{\alpha}_1$, $\widehat{\alpha}_2$, $\widehat{\alpha}_3$, $\widehat{\alpha}_4$, $\widehat{\mu}_{MR}$, and $\widehat{\mu}_{T_t}$ per periods: Wr1A Wr1B CdA CdB Tr HtA HtB
Data 3: Time series T_t were divided per SB. Time series T_t per period Wr1A, Wr1B, CdA CdB, Tr, HtA, and HtB
Data 4: r_t per period Wr1A Wr1B CdA CdB Tr HtA HtB
Data 5: w_t per period Wr1A Wr1B CdA CdB Tr HtA HtB
Estimate of type 2 variables: $\widehat{M}_{I(w)}$, $\widehat{w}$, $\widehat{\mu}_{\widehat{\rho}_l}$, $\widehat{Md}_{\widehat{\rho}_l}$, $\widehat{R}_{\widehat{\rho}_l}$, $\widehat{\sigma^2}_{\widehat{\rho}_l}$ $l = 1, \ldots, 72$ per periods: Wr1A Wr1B CdA CdB Tr HtA HtB
Logarithm transformation: $r_t = \ln(T_t)$
Apply to
Apply to
One regular differentiation: $w_t = r_t - r_{t-1}$
Data 2: Time series T_t were divided per period Wr1, Cd, Tr, Ht, and Wr2
Data 1: Time series T_t, where $t = 1, \ldots, t_{max}$

Figure 3.4: Summary of steps involved in method M1: blue lines, Type 1 classification variables; green lines, Type 2 variables; solid lines, process; dashed line, results. Different boxes contain the name of the stages (i.e., Wr1A, Wr1B, CdA, CdB, Tr, HtA, HtB and Wr2) to indicate that the procedure was applied to all parts of the time series. A structural break was found in Wr1, Cd and Ht, so that the suffixes A and B denote the substages before and after the break, respectively.

The estimates of MR values were computed with the R software according to the function `rollrange` from the `QuantTools` package [105]. The sample ACF and sample PACF values were calculated with the function `acf` (`stats`) [171] and `pacf` (`tseries`) [114], respectively. The values of the periodogram and their frequencies were obtained with the function `spectrum` (`stats`).

3.4.3.3 | Calculation of Classification Variables—Method M2

The Holt–Winters method (H-W) [19] is a type of exponential smoothing that is used for forecasting in time series analysis. The seasonal H-W (SH-W) method uses a smoothing equation for each component of a given time series: the level, slope and seasonality, which are denoted at time t^* as a_{t^*}, b_{t^*} and s_{t^*}, respectively. The additive SH-W prediction function of a time series of T is given by Equation (3.1), where p denotes the number of observations per period and k is the integer part of $(l-1)/p$ [21; 125]. This equation was implemented with the conditions: $0 \leq \alpha \leq 1$, $0 \leq \beta \leq 1$, $0 \leq \gamma \leq 1$ and $t^* > s$. When the algorithm converges, a corresponds to the last level of a_{t^*}, b is the last slope of b_{t^*} and s_1–s_{24} are the last seasonal contributions of each s_{t^*}. A forecast of $\hat{t}_{t^*+l}$ based on all the data up to time t^* is denoted by $\hat{t}_{t^*+l|t^*}$; a simpler notation for $\hat{t}_{t^*+l|t^*}$ is $\hat{t}_{t^*+l}$.

$$\begin{aligned}
\hat{t}_{t^*+l|t^*} &= a_{t^*} + l b_{t^*} + s_{t^*+l-p(k+1)}, \text{where} \\
a_{t^*} &= \alpha(t_{t^*} - s_{t^*-p}) + (1-\alpha)(a_{t^*-1} + b_{t^*-1}) \\
b_{t^*} &= \beta(a_{t^*} - a_{t^*-1}) + (1-\beta) b_{t^*-1} \\
s_{t^*} &= \gamma(t_{t^*} - a_{t^*-1} - b_{t^*-1}) + (1-\gamma) s_{t^*-p}
\end{aligned} \tag{3.1}$$

The method M2 consists of fitting additive SH-W equations in two steps. Firstly, for each stage of the 27 time series of T, the classification variables are the last level of smoothing components of the additive SH-W method per sensor: level (a), slope (b) and seasonal components ($s_1, s_2, \ldots, s_{24}$). The method was fitted by considering $p = 24$ as the number of observations per day. These smoothing components were called as Type 3 variables. Secondly, by considering the complete time series, the first 24 predictions of $\boldsymbol{T}$ for each unique additive SH-W model per sensor were regarded as additional classification variables, which were denoted as Type 5 variables.

Although a residual analysis is not necessary when using SH-W method, estimates of the features from residuals were also computed per stage of the time series (Type 4 variables): sum of squared estimate of errors (SSE), maximum of periodogram ($M_{I(w)}$) and its frequency (w) and several parameters (mean, median, range and variance) of sample ACF for 72 lags ($\mu_{\widehat{\rho}_l}$, $Md_{\widehat{\rho}_l}$, $R_{\widehat{\rho}_l}$ and $\sigma^2_{\widehat{\rho}_l}$, with $l = 1, \ldots, 72$). Moreover, the Kolmogorov–Smirnov (KS) normality test [127] and Shapiro–Wilk test (SW) [122; 128; 129] were applied in order to extract further information from the different stages of the observed time series. The statistic of the KS test (denoted as Dn) was used to compare the empirical distribution function of the residuals with the cumulative distribution function of the normal model. Likewise, the statistic of the SW test (Wn) was employed to detect deviations from normality, due to skewness and/or kurtosis. The statistics of both tests were also used as classification variables because they provide information about deviation from normality for the residuals derived from the SH-W method.

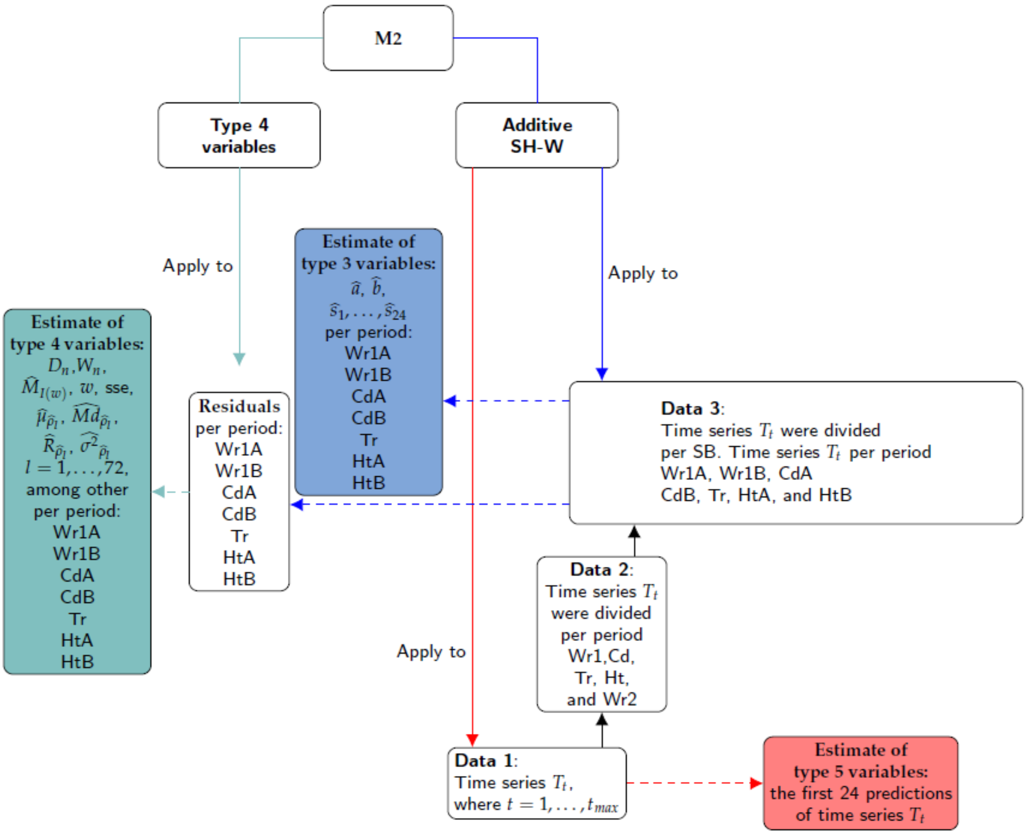

M2
Type 4 variables
Additive SH-W
Apply to
Apply to
Apply to
Estimate of type 3 variables: $\widehat{a}$, $\widehat{b}$, $\widehat{s}_1, \ldots, \widehat{s}_{24}$ per period: Wr1A Wr1B CdA CdB Tr HtA HtB
Estimate of type 4 variables: D_n, W_n, $\widehat{M}_{I(w)}$, w, sse, $\widehat{\mu}_{\widehat{\rho}_l}$, $\widehat{Md}_{\widehat{\rho}_l}$, $\widehat{R}_{\widehat{\rho}_l}$, $\widehat{\sigma^2}_{\widehat{\rho}_l}$ $l = 1, \ldots, 72$, among other per period: Wr1A Wr1B CdA CdB Tr HtA HtB
Residuals per period: Wr1A Wr1B CdA CdB Tr HtA HtB
Data 3: Time series T_t were divided per SB. Time series T_t per period Wr1A, Wr1B, CdA CdB, Tr, HtA, and HtB
Data 2: Time series T_t were divided per period Wr1,Cd, Tr, Ht, and Wr2
Data 1: Time series T_t, where $t = 1, \ldots, t_{max}$
Estimate of type 5 variables: the first 24 predictions of time series T_t

Figure 3.5: Summary of steps in method M2: blue lines, Type 3 classification variables; green lines, Type 4 variables; red lines, Type 5 variables; solid lines, process; dashed line, results.

The steps involved in M2 are depicted in Figure 3.5. Firstly, the different time series were split according to the climatic stages observed (Data 2) and the structural breaks (SB) identified (Data 3). Secondly, the seasonal H-W method was applied to Data 3 in order to obtain the last level of smoothing components (Type 3 variables) and then the model residuals. The third step consisted of applying the formulas of Type 4 variables to the residuals. Finally, the seasonal H-W method was applied to Data 1 in order to obtain the first 24 predictions of T (Type 5 variables).

The `HoltWinters` function (`stats`) was used to fit the Additive SH-W method. The `shapiro.test` (`stats`) and `ks.test` (`dgof`) [172] were used to apply the normality tests. Values of the sample ACF and sample PACF were computed with the functions acf (stats) and pacf (tseries), respectively. Values of the periodogram and their frequencies were calculated with the function spectrum (stats).

3.4.3.4 | Calculation of Classification Variables—Method M3

The ARMA model is also known as the Box–Jenkins approach, which focuses on the conditional mean of the time series and assumes that it is stationary [107]. By contrast, the ARIMA model can be used when a time series is not stationary. It employs regular differencing of the time series prior to fitting an ARMA model. ARIMA (p, d, q) models employ p Auto Regressive (AR) terms, d regular differences and q Moving Average (MA) terms. Parameters of the AR component are denoted as ϕ_i $(i = 1, \ldots, p)$ and parameters of the MA component as θ_j $(j = 1, \ldots, q)$. The error terms ϵ_t are assumed to be a sequence of data not autocorrelated with a null mean, which is called White Noise (WN) [107]. In addition, if a given time series is assumed to follow an ARIMA process, the conditional variance of residuals is supposed to be constant. If this is not the case, then it is assumed that an ARCH effect exists in the time series. Two of the most important models for capturing such changing conditional variance are the ARCH and Generalized ARCH (GARCH) models [12].

Seasonal ARIMA $(p, d, q)(P, D, Q)_S$ models are more appropriate in this case given the marked daily cycles. P refers to the number of seasonal AR (SAR) terms, D to the number of difference necessary to obtain a stationary time series, Q to the number of seasonal MA (SMA) terms and S to the number of observations per period ($S = 24$ in this case). Parameters of the SAR component are denoted as Φ_i $(i = 1, \ldots, P)$ and the SMA component as Θ_j $(j = 1, \ldots, Q)$. The error terms ϵ_t are assumed to be a WN sequence [107].

A seasonal ARIMA $(p, d, q)(P, D, Q)_S$ model is given by Equation (3.2), where the polynomial $\boldsymbol{\phi_p}(B)$ is the regular AR operator of order p, $\boldsymbol{\theta_q}(B)$ is the regular MA operator of order q, $\boldsymbol{\Phi_P}(B^S)$ is the seasonal AR operator (SAR) of order P, $\boldsymbol{\Theta_Q}(B^S)$ is the seasonal MA operator (SMA) of order Q and B is the backshift operator (i.e., $Bv_t = v_{t-1}$, for $t > 1$ and $B^{24}v_t = v_{t-24}$, for $t > 24$).

$$\begin{aligned}\boldsymbol{\Phi_P}(B^S)\boldsymbol{\phi_p}(B)v_t &= \boldsymbol{\Theta_Q}(B^S)\boldsymbol{\theta_q}(B)\varepsilon_t, \text{where,}\\ \boldsymbol{\phi_p}(B) &= 1 - \phi_1 B - \cdots - \phi_p B^p\\ \boldsymbol{\theta_q}(B) &= 1 + \theta_1 B + \cdots + \theta_q B^q\\ \boldsymbol{\Phi_P}(B^S) &= 1 - \Phi_1 B^S - \Phi_P B^{PS}\\ \boldsymbol{\Theta_Q}(B^S) &= 1 + \Theta_1 B^S + \cdots + \Theta_Q B^{QS}\end{aligned} \tag{3.2}$$

Furthermore, ∇_S^D represents the seasonal differences while ∇^d accounts for the regular differences, so that ∇_S^D is defined as $(1-B^S)^D$ and ∇^d as $(1-B)^d$ [12]. In this study, $v_t = \nabla_S^D \nabla^d r_t$, which was obtained by differentiating the series once regularly ($d = 1$) and once seasonally ($D = 1$). Thus, $v_t = \nabla_{24}^1 \nabla^1 r_t = \nabla_{24}^1 w_t = w_t - w_{t-24}$.

For each stage of the time series, a common seasonal ARIMA model was fitted for the 27 observed time series (see Table 3.1). Firstly, each observed stage of the time series $\boldsymbol{T}$ was checked to determine if it could be regarded as stationary, which implies that the mean and variance are constant over time t and the covariance between one observation and another (in lagged l steps) from the same time series does not depend on t [107]. The ACF and PACF correlograms were used to examine this condition. Furthermore, the Augmented Dickey–Fuller (ADF) test [141] was applied for checking the null hypothesis of non stationarity, as well as the Lagrange multiplier (LM) test [14] for examining the null hypothesis about the absence of ARCH effect. In addition, the autocorrelation Ljung–Box Q (LBQ) test [126] was applied for inspecting the null hypothesis of independence in a given time series. This LBQ test was carried out on the different lags from nt to $nt + 48$, where nt is the sum of the number of AR, MA, SAR and SMA terms of the seasonal ARIMA models. It was applied to the time series of the model residuals and the squared residuals.

The condition of stationarity is necessary when fitting an ARMA model. For this purpose, logarithmic transformation, one regular differentiation ($d = 1$) and one seasonal differentiation ($D = 1$) were applied to all time series of $\boldsymbol{T}$, in order to stabilize the variance and remove both the trend in mean and seasonal trend [124]. Seasonal differentiation was applied to the observed time series $\boldsymbol{W}$, and the results were denoted as $\boldsymbol{V}$ ($v_t = w_t - w_{t-24}$), being $w_t = r_t - r_{t-1}$.

In order to determine the appropriate values of (p, d, q) and (P, D, Q), the corrected Akaike's Information Criterion ($AICc$) [125] was used, which checks how well a model fits the time series using the restriction $d = 1$ and $D = 1$. The most successful model for each stage of the different observed time series $\boldsymbol{T}$ was chosen according to the lowest $AICc$ value. Next, the maximum likelihood estimation method was used to estimate the parameters of the seasonal ARIMA models [107]. Different tests were used to determine whether model assumptions were fulfilled. After computing the model residuals, ADF and LBQ [126] tests were applied to the residuals and their squared values, for 48 lags, in order to evaluate the condition of WN process. The ACF and PACF correlograms were also used. The next step was to evaluate the absence of Arch effects in the residuals. For this purpose, the LM test was applied to the residuals and their squared values [132; 143]. Although the normality of errors is not an assumption for fitting ARIMA models, the distribution of residuals derived from the fitted models were compared with the normal distribution by means of the Q-Q normal scores plots, as well as the SW and KS normality tests.

Table 3.1: The most successful models per stage of the different observed time series $\boldsymbol{r}$ are presented in the second column. Column 3 presents the percentages of the LBQ test on the different lags from nt to $nt + 48$ from the 27 sensors that fulfill the assumptions of independence. Column 4 presents the percentages of the LM test from the 27 sensors that fulfill the assumptions of the absence of Arch effect. The significance level used was 0.01.

Stage	Model	LBQ	LM
(a) For Method 3			
Wr1A	Seasonal ARIMA$(0,1,2)(2,1,0)_{24}$	77.00	92.52
Wr1B	Seasonal ARIMA$(0,1,0)(2,1,0)_{24}$	3.00	44.44
CdA	Seasonal ARIMA$(0,1,0)(2,1,0)_{24}$	25.00	77.77
CdB	Seasonal ARIMA$(0,1,2)(2,1,0)_{24}$	18.00	18.52
Tr	Seasonal ARIMA$(0,1,3)(2,1,0)_{24}$	11.00	3.70
HtA	Seasonal ARIMA$(1,1,3)(0,1,1)_{24}$	22.00	22.22
HtB	Seasonal ARIMA$(0,1,3)(2,1,0)_{24}$	0.00	37.04
Wr2	Seasonal ARIMA$(0,1,2)(2,1,0)_{24}$	18.00	37.04
(b) For Method 4			
Wr1A	A seasonal ARIMA per sensor	92.59	96.30
Wr1B	A seasonal ARIMA per sensor	51.85	59.26
CdA	A seasonal ARIMA per sensor	81.48	81.48
CdB	A seasonal ARIMA per sensor	48.15	25.93
Tr	A seasonal ARIMA per sensor	25.93	3.70
HtA	A seasonal ARIMA per sensor	59.26	29.63
HtB	A seasonal ARIMA per sensor	25.93	44.44
Wr2	A seasonal ARIMA per sensor	55.55	37.04

Given that the errors of all models cannot be regarded as WN in this case, it is possible that the model residuals contain useful information about the performance of the different time series. In order to extract further information from the residuals, some features were calculated using ACF, PACF and statistics of normality tests, among others. They were used as additional classification variables.

The steps involved in Method 3 are illustrated in Figure 3.6. Firstly, the different time series were split according to the climatic stages observed (Data 2) and the structural breaks (SB) identified (Data 3). Secondly, the logarithm transformation was applied to Data 3 (the result is denoted as Data 4). The third step consisted of applying the seasonal ARIMA model to Data 4 in order to obtain the estimates of model coefficients: $\boldsymbol{\phi_p}(B)$, $\boldsymbol{\theta_q}(B)$, $\boldsymbol{\Phi_P}(B^S)$ and $\boldsymbol{\Theta_Q}(B^S)$. These parameters were denoted as Type 6 variables. Next, different features (Type 7 variables) were computed from the residuals: variance (σ^2), maximum of periodogram ($M_{I(w)}$) and its frequency w. For the set of 72 lags, additional features were computed from sample ACF values: mean ($\mu_{\widehat{\rho}_l}$), median ($Md_{\widehat{\rho}_l}$), variance ($\sigma^2_{\widehat{\rho}_l}$) and range ($R_{\widehat{\rho}_l}$), with $l = 1, \ldots, 72$. Finally, the first four values of sample PACF (α_1, α_2, α_3 and α_4) were computed, as well as the statistics of the KS normality (Dn) and SW (Wn) tests.

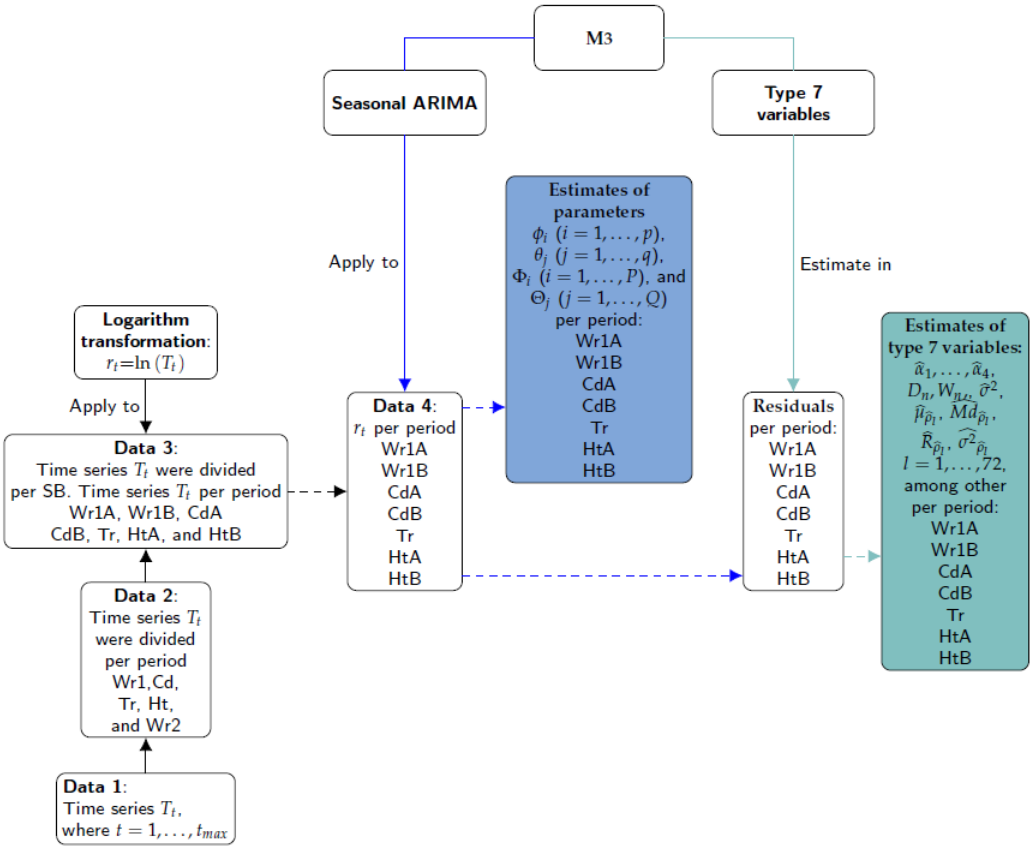

M3
Seasonal ARIMA
Type 7 variables
Apply to
Estimates of parameters ϕ_i $(i = 1, \ldots, p)$, θ_j $(j = 1, \ldots, q)$, Φ_i $(i = 1, \ldots, P)$, and Θ_j $(j = 1, \ldots, Q)$ per period: Wr1A Wr1B CdA CdB Tr HtA HtB
Estimate in
Logarithm transformation: $r_t = \ln(T_t)$
Apply to
Data 3: Time series T_t were divided per SB. Time series T_t per period Wr1A, Wr1B, CdA CdB, Tr, HtA, and HtB
Data 4: r_t per period Wr1A Wr1B CdA CdB Tr HtA HtB
Residuals per period: Wr1A Wr1B CdA CdB Tr HtA HtB
Estimates of type 7 variables: $\widehat{\alpha}_1, \ldots, \widehat{\alpha}_4$, $D_n, W_n, \widehat{\sigma}^2$, $\widehat{\mu}_{\widehat{\rho}_l}$, $\widehat{Md}_{\widehat{\rho}_l}$, $\widehat{R}_{\widehat{\rho}_l}$, $\widehat{\sigma^2}_{\widehat{\rho}_l}$ $l = 1, \ldots, 72$, among other per period: Wr1A Wr1B CdA CdB Tr HtA HtB
Data 2: Time series T_t were divided per period Wr1, Cd, Tr, Ht, and Wr2
Data 1: Time series T_t, where $t = 1, \ldots, t_{max}$

Figure 3.6: Summary of steps in method M3: blue lines, Type 6 classification variables; green lines, Type 7 variables; solid lines, process; dashed line, results.

In order to choose a seasonal ARIMA model and the estimations of the model parameters for each sensor, the `arima` (`stats`) and `auto.arima` (`forecast`) functions [110; 111] were used. The ADF test was computed using the `adf.test` (`aTSA`) [109]. The LBQ test was applied by means of the `Box.test` function (`stats`). The LM test was carried out using the `arch.test` function (`aTSA`). The SW and KS normality tests were applied using the `shapiro.test` (`stats`) and `ks.test` functions (`dgof`), respectively.

3.4.3.5 | Calculation of Classification Variables—Method M4

The Wold decomposition establishes that any covariance stationary process can be written as the sum of a non-deterministic and deterministic process. This decomposition, which is unique, is a linear combination of lags of a WN process and a second process whose future values can be predicted exactly by some linear functions of past observations. If the time series $\{v_t; t \in \mathbb{Z}\}$ is purely non-deterministic, then it can be written as a linear combination of lagged values of a WN process (MA(∞) representation), that is, $v_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$, where $\psi_0 = 1$, $\sum_{j=1}^{\infty} \psi_j^2 < \infty$ and ε_t is a WN [12]. Although the Wold decomposition depends on an infinite number of parameters, the values of coefficients of the decomposition decay rapidly to zero.

For Method 3, a unique model was fitted for the 27 sensors in the same stage of the time series because it is necessary to have the same number of classification variables per sensor, in order to apply later the sPLS-DA method. It is not possible to work with 'the best model' per sensor in each stage. In order to obtain the same number of variables using the 'best model' per sensor, the Wold decomposition was applied to each sensor. Hence, Method 4 consists of obtaining the Wold decomposition for the ARMA models. Firstly, different seasonal ARIMA models were fitted iteratively to time series r_t per sensor and stage of the time series, and the most successful model was determined.

As an illustration, consider that a time series r_t follows a
seasonal ARIMA $(2, 1, 1)(0, 1, 0)_{24}$ process. Now, consider that $w_t = r_t - r_{t-1}$ and $v_t = w_t - w_{t-24}$, with $t = 1, \ldots, t_{max}$. Then, the time series v_t follows an ARMA$(2, 1)$ process, which can be decomposed according to the Wold approach by obtaining the polynomials $\boldsymbol{\phi_p}(B)$ and $\boldsymbol{\theta_q}(B)$ that determine the best ARMA (p, q) model. In summary, for each seasonal ARIMA $(p, d, q)(P, D, Q)_S$, it was possible to find the best ARMA (p, q) model and its Wold decomposition.

The analysis of residuals of the different models fitted suggests that the condition of not autocorrelation is not fulfilled in all cases. Nonetheless, the Wold decomposition of each model was fitted independently, in order to have the 'best seasonal ARIMA model' per sensor and the same number of parameters per sensor. For each model, the first five coefficients of the Wold decomposition were calculated and used as classification variables. In all cases, the most successful seasonal ARIMA model per sensor used $D = 1$ and $d = 1$.

The steps involved in Method 4 are illustrated in Figure 3.7. Firstly, the different time series were split according to the climatic stages observed (Data 2) and the structural breaks (SB) identified (Data 3). Secondly, the logarithm transformation was applied to Data 3 (the result is denoted as Data 4). The third step consisted of applying the seasonal ARIMA model to Data 4 in order to obtain the estimates of parameters and their residuals. Next, the same formulas of Type 7 variables used in M3 were applied to the residuals. Finally, the Wold decomposition was determined using the estimates of parameters of seasonal ARIMA models. The first five coefficients of the MA weights (i.e., $\psi_1, \ldots, \psi_5$) were denoted as Type 8 variables.

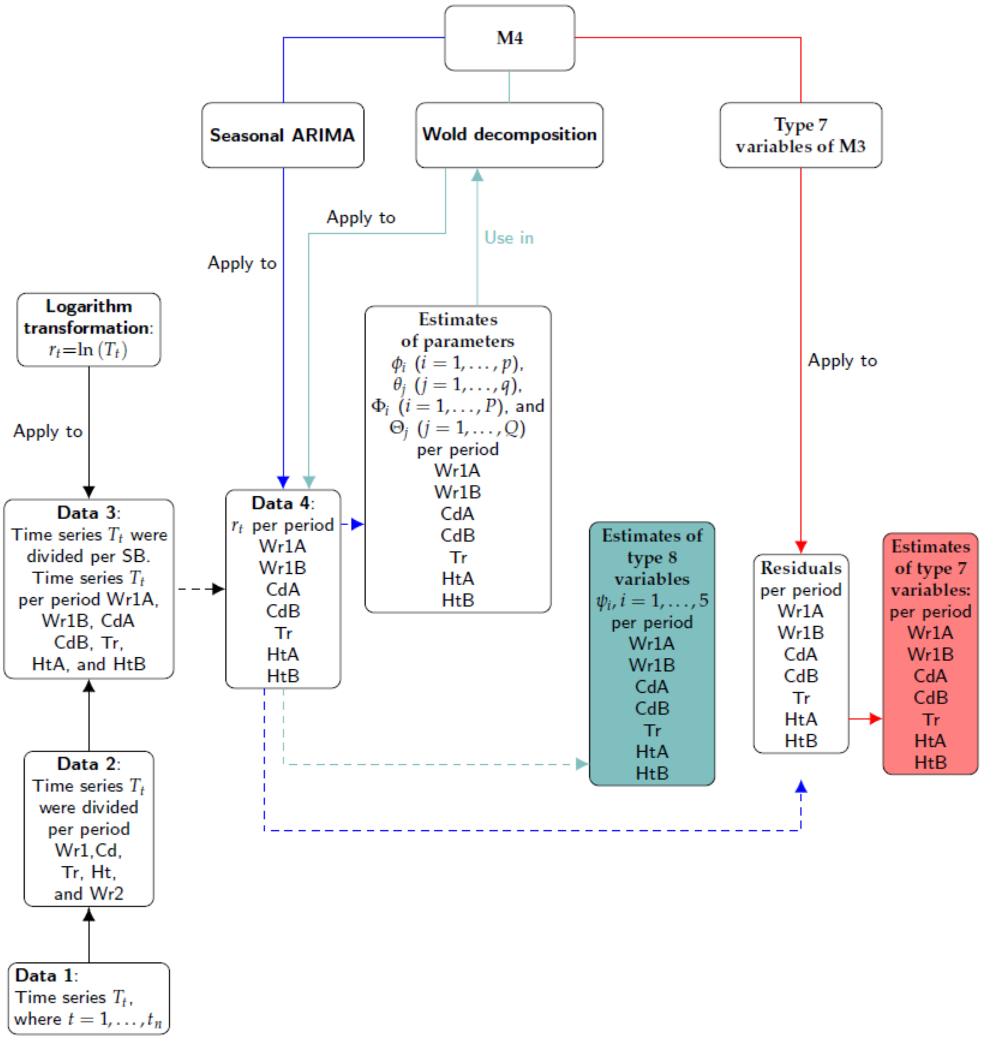

M4
Seasonal ARIMA
Wold decomposition
Type 7 variables of M3
Apply to
Apply to
Use in
Apply to
Logarithm transformation: r_t=ln (T_t)
Apply to
Data 3: Time series T_t were divided per SB. Time series T_t per period Wr1A, Wr1B, CdA CdB, Tr, HtA, and HtB
Data 4: r_t per period Wr1A Wr1B CdA CdB Tr HtA HtB
Estimates of parameters ϕ_i $(i = 1, \ldots, p)$, θ_j $(j = 1, \ldots, q)$, Φ_i $(i = 1, \ldots, P)$, and Θ_j $(j = 1, \ldots, Q)$ per period Wr1A Wr1B CdA CdB Tr HtA HtB
Estimates of type 8 variables $\psi_i, i = 1, \ldots, 5$ per period Wr1A Wr1B CdA CdB Tr HtA HtB
Residuals per period Wr1A Wr1B CdA CdB Tr HtA HtB
Estimates of type 7 variables: per period Wr1A Wr1B CdA CdB Tr HtA HtB
Data 2: Time series T_t were divided per period Wr1, Cd, Tr, Ht, and Wr2
Data 1: Time series T_t, where $t = 1, \ldots, t_n$

Figure 3.7: Summary of steps of method M4: red lines, Type 7 variables of M3; green lines, Type 8 variables; solid lines, process; dashed line, results.

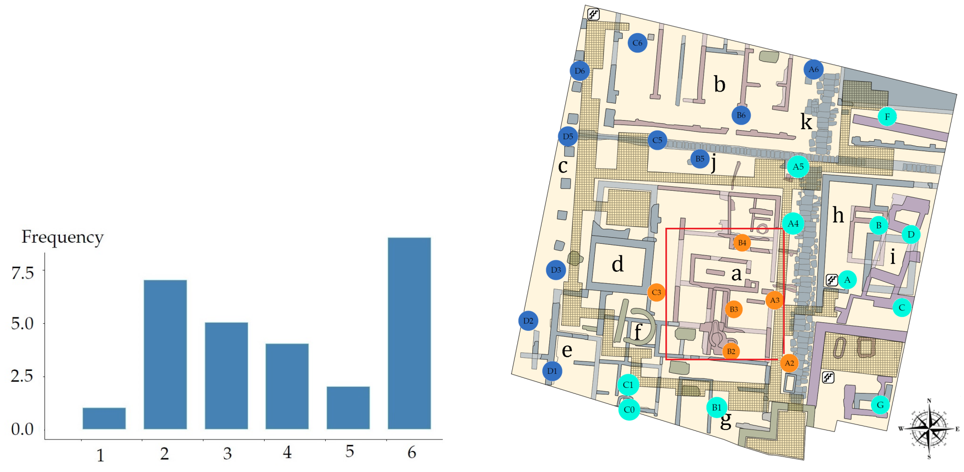
Apart from the same functions used for M3, this method employed `ARMAtoMA` (`stats`). One function was created for reducing a polynomial with AR and SAR components to a polynomial with just AR component. Likewise, another function converted a polynomial with MA and SMA components to one with MA component.

3.4.3.6 | Determination of Number of Classes and Class per Sensor Using PCA and K-Means Algorithm

All classification variables calculated as described above for each sensor were arranged as a matrix called 'total classification dataset' (TCD) with 27 rows (sensors) by 671 columns corresponding to the classification variables from the four methods. The total number of variables was 88, 296, 143 and 139 from M1, M2, M3 and M4, respectively. The multivariate analysis of the TCD matrix would allow the identification of different microclimates in the archaeological museum. It was checked that the statistical distribution of some classification variables was strongly skewed, which recommends to apply a data pretreatment prior to the multivariate analysis. For those variables with a strongly skewed distribution, different standard (simple) Box–Cox transformations [173] were applied with the goal of finding a simple transformation leading to a normal distribution. In particular, the Box–Cox transformations were used on those classification variables with a Fisher coefficient of kurtosis [174] or with a Fisher–Pearson standardized moment coefficient of skewness [174] outside the interval from -2.0 to 2.0. Before applying the Box–Cox transformations, an absolute value function was used for variables with a negative skewness that fulfilled one of the aforementioned conditions. The skewness statistic was computed for each variable in order to check the asymmetry of the probability distribution. The kurtosis parameter indicates which variables were heavy-tailed or light-tailed, relative to a normal distribution. Moreover, the estimates of kurtosis were useful measures for identifying outliers in the classification variables. The functions `kurtosis` and `skewness` (`PerformanceAnalytics`) [169] were used to compute the coefficients of kurtosis and skewness, while `boxcoxfit` (`geoR`) [175] was employed to apply different Box–Cox transformations. The function `prcomp` (`stats`) was used to carry out PCA.

Those values of a given classification variable that clearly departed from a straight line on the normal probability plot were removed and regarded as missing data. These were estimated using the NIPALS algorithm [144] implemented in the `mixOmics` package [67], which is able to cope with this drawback and returns accurate results [176]. After the data normalization, all variables were mean-centered and scaled to unit variance, which is the common pretreatment in PCA. Next, PCA was carried out to reduce the dimensionality of the TCD matrix. Each observation (sensor) was projected onto the first few principal components to obtain lower-dimensional data, while preserving as much of the data variation as possible.

Given that the two first components maximize the variance of the projected observations (TCD), only two components were employed to run the k-means clustering. This method is a centroid-based algorithm that computes the distance between each sensor and the different centroids, one per cluster or class. The algorithm determines *K* clusters so that the intra-cluster variation is as small as possible. However, prior to applying this method, the number of clusters *K* has to be determined, which depends on the type of clustering and previous knowledge about the time series of T. For this purpose, different criteria can be used [170; 177]. Such methods do not always agree exactly in their estimation of the optimal number of clusters, but they tend to narrow the range of possible values. The `NbClust` function of the `NbClust` package [170] incorporates 30 different indices for determining the number of clusters [177]. This function claims to use the best clustering scheme from the different results obtained, by varying all combinations of the number of clusters, distance measures and clustering methods. It allows the user to identify the value *K* in which more indices coincide, providing assurance that a good choice is being made.



(**a**)

(**b**)

Figure 3.8: Results associated with the k-means method. (**a**) Absolute frequency (number) of indices that indicates the best number of classes in the museum. For example, two classes are selected by seven indices. (**b**) Classification of sensors installed in the museum, according to the k-means method. Each color (blue, green and orange) corresponds to a different class.

In this clustering method, the measure used to determine the internal variance of each cluster was the sum of the squared Euclidean distances between each sensor and each centroid. The distances were used to assign each sensor to a cluster. For this purpose, the k-means algorithm of Hartigan and Wong [178] was applied by means of the function `kmeans` (`stats`). It performs better than the algorithms proposed by MacQueen [179], Lloyd [180] and Forgy [181]. However, when the algorithm of Hartigan and

Wong is carried out, it is often recommended to try several random starts. In the present study, 100 random starts were employed. This algorithm guarantees that, at each step, the total intra-variance of the clusters is reduced until reaching a local optimum. Results from the k-means algorithm depend on the initial random assignment. For this reason, the algorithm was run 100 times, each with a different initial assignment. The final result was the one leading to a classification with the lowest total variance value. By comparing the classification obtained with the position of sensors in the museum (Figure 3.8a), the three zones were denoted as North West (NW), South East (SE) and Skylight (Sk).

3.4.3.7 | Sensor Classification Using sPLS-DA

Partial Least Squares (PLS) regression [51] is a multivariate regression method which relates two data matrices (predictors and answer). PLS maximizes the covariance between latent components from these two datasets. A latent component is a linear combination of variables. The weight vectors used to calculate the linear combinations are called loading vectors.

Penalties such as Lasso and Ridge [182] have been applied to the weight vectors in PLS for variable selection in order to improve the interpretability when dealing with a large number of variables [52; 61; 64]. Chung and Keles [53] extended the sparse PLS [52] to classification problems (SPLSDA and SGPLS) and demonstrated that both SPLSDA and SGPLS improved classification accuracy compared to classical PLS [57; 58; 60]. Lê Cao et al. [63] introduced a sparse version of the PLS algorithm for discrimination purposes (sPLS-Discriminant Analysis, sPLS-DA) which is an extension of the sPLS proposed by Lê Cao et al. [61, 64]. They showed that sPLS-DA has very satisfying predictive performances and is able to select the most informative variables. Contrary to the two-stages approach (SPLSDA) proposed by Chung and Keles [53], sPLS-DA performs variable selection and classification in a single-step procedure. In order to classify the sensors and improve the interpretability of results, sPLS-DA was applied to the different classification datasets.

Since the original PLS algorithm proposed by Wold [51], many variants have arisen (e.g., PLS1, PLS2, PLS-A, PLS-SVD [183] and SIMPLS [184]), depending on how the regressor matrix ($\mathbf{X}$) and response matrix ($\mathbf{Z}$) are deflated. Alternatives exist whether $\mathbf{X}$ and $\mathbf{Z}$ are deflated separately or directly, using the cross product $\mathbf{M} = \mathbf{X}^{\top}\mathbf{Z}$ and the Singular Value Decomposition (SVD). A hybrid PLS with SVD is used in the version sPLS-DA [61]. For sPLS-DA, a regressor matrix is denoted as $\mathbf{X}$ ($\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$ or $\mathbf{X}_4$ in this case), with dimension $n \times p$. The number of rows (sensors) is $n = 27$ and the columns correspond to the classification variables p (p_1, p_2, p_3 or p_4). A response qualitative vector denoted as $\boldsymbol{Y}$ has length n and it indicates the class of each sensor, with values coded as 1 (for NW), 2 (SE) and 3 (Sk).

sPLS-DA was carried out using Lasso penalization of the loading vectors associated to $\mathbf{X}$ [66] using a hybrid PLS with SVD decomposition [185]. The penalty function is included in the objective function of PLS-DA, which corresponds to PLS carried out using a response matrix $\mathbf{Z}$ with values of either 0 or 1, created with the values of response vector $\boldsymbol{Y}$. Thus, this vector was converted into a dummy matrix $\mathbf{Z}$ with dimension $n \times K$, being $n = 27$ the number of sensors and $K = 3$ the number of sensor classes.

Regarding the optimization problem of sPLS-DA, in this case: $\mathbf{X} \in \mathbb{R}^{n\times p}$ is a matrix with p variables and n sensors, $\boldsymbol{Y} \in \mathbb{R}^{n\times 1}$ is a response vector (classes $k = 1,2,3$) and $\mathbf{Z} \in \{0,1\}^{n\times 3}$ is an indicator matrix, where $z_{ik} = I(Y_i = k)$, with $k = 1,2,3$. The sPLS-DA method modeled $\mathbf{Z}$ and $\mathbf{X}$ as $\mathbf{X} = \Xi\mathbf{C} + \mathbf{E}_1$ and $\mathbf{Z} = \Xi\mathbf{D} + \mathbf{E}_2$, where $\mathbf{C}$ and $\mathbf{D}$ are matrices that contain the regression coefficients of $\mathbf{X}$ and $\mathbf{Z}$ on the H latent components associated to $\mathbf{X}$, while $\mathbf{E}_1$ and $\mathbf{E}_2$ are random errors. Furthermore, each component of Ξ is a combination of selected variables, where $\Xi = [\xi_1, \ldots, \xi_H]$, and each vector ξ_h was computed sequentially as $\xi_h = \mathbf{X}_{h-1}\boldsymbol{u}_h$, where $\mathbf{X}_{h-1}$ is the orthogonal projection of $\mathbf{X}$ on subspace $span\{\xi_1, \ldots, \xi_{h-1}\}^{\perp}$ and $(\boldsymbol{u}_h, \boldsymbol{v}_h)$ is the solution the optimization problem according to Equation (3.3), subject to $\|\boldsymbol{u}_h\|_2 = 1$.

$$(\boldsymbol{u}_h, \boldsymbol{v}_h) = \underset{\boldsymbol{u}_h, \boldsymbol{v}_h}{\arg\min}\{\|\mathbf{M}_h - \boldsymbol{u}_h\boldsymbol{v}_h^{\top}\|_F^2 + P_{\lambda}(\boldsymbol{u}_h)\} \tag{3.3}$$

The optimization problem minimizes the Frobenius norm between the current cross product matrix ($\mathbf{M}_h$) and the loading vectors ($\boldsymbol{u}_h$ and $\boldsymbol{v}_h$), where $\mathbf{M}_h = \mathbf{X}_h^{\top}\mathbf{Z}_h$ and $\mathbf{Z}_{h-1}$ is the orthogonal projection of $\mathbf{Z}$ on subspace $span\{\xi_1, \ldots, \xi_{h-1}\}^{\perp}$. Furthermore, $\|\mathbf{M}_h - \boldsymbol{u}_h\boldsymbol{v}_h^{\top}\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{p}(m_{ij} - u_i v_j)^2$, and $P_{\lambda}(\boldsymbol{u}_h)$, defined as $\lambda\|\boldsymbol{u}_h\|_1$, is the Lasso penalty function [63; 64]. This optimization problem is solved iteratively based on the PLS algorithm [65]. The SVD decomposition of matrix $\mathbf{M}_h$ is subsequently deflated for each iteration h. This matrix is computed as $\mathbf{M}_h = \boldsymbol{u}\Delta\boldsymbol{v}^{\top}$, where $\boldsymbol{u}$ and $\boldsymbol{v}$ are orthonormal matrices and Δ is a diagonal matrix whose diagonal elements are called the singular values. During the deflation step of PLS, $\mathbf{M}_h \neq \mathbf{X}_h^{\top}\mathbf{Z}_h$, because $\mathbf{X}_h$ and $\mathbf{Z}_h$ are computed separately, and the new matrix is called $\tilde{\mathbf{M}}_h$. At each step, a new matrix $\tilde{\mathbf{M}}_h = \mathbf{X}_h^{\top}\mathbf{Z}_h$ is computed and decomposed by SVD [61]. Furthermore, the soft-thresholding function $g(\boldsymbol{u}) = (|\boldsymbol{u}| - \lambda)_+ sign(\boldsymbol{u})$, with $(x)_+ = max(0, x)$, was used in penalizing loading vectors $\boldsymbol{u}$ to perform variable selection in regressor matrix, thus $\boldsymbol{u}_{new} = g_{\lambda}(\tilde{\mathbf{M}}_{h-1}\boldsymbol{v}_{old})$ [61].

Rohart et al. [68] implemented an algorithm to solve the optimization problem in Equation (3.3), where the parameter λ needs to be tuned and the algorithm chooses λ

among a finite set of values. It is possible to find a value λ for any null elements in a loading vector. These authors implemented the algorithm using the number of non-zero elements of the loading vector as input, which corresponds to the number of selected variables for each component. They implemented sPLS-DA in the `R` package `mixOmics`, which provides functions such as `perf`, `tune.splsda` and `splsda`, in order to determine the number of components and elements different to zero in the loading vector before running the final model.

In order to compare the performance of models constructed with a different number of components, 1000 training and test datasets were simulated and the sPLS-DA method (for a maximum number of 10 components) was tuned by three-fold cross-validation (CV) for $\mathbf{X}$. The `perf` function outputs the optimal number of components that achieve the best performance based on both types of classification error rate (CER): Balanced Error Rate (BER) and the Overall classification error rate (Overall). BER is the average proportion of wrongly classified sensors in each class, weighted by the number of sensors. In most cases, the results from sPLS-DA were better (or very similar) using Overall than when using BER. However, BER was preferred to Overall because it is less biased towards majority classes during the performance assessment of sPLS-DA. In this step, three different prediction distances were used, maximum, centroid and Mahalanobis [68], in order to determine the predicted class per sensor for each of the test datasets. For each prediction distance, both Overall and BER were computed.

The maximum number of components found in the present work was three ($K = 3$) when BER was used instead of Overall. Furthermore, when sPLS-DA used classification variables from M2 and M3, two components led to the lowest BER values. Among the three prediction distances calculated (i.e., maximum, centroid and Mahalanobis), it was found that centroid performed better for the classification. Thus, this distance was used to determine the number of selected variables and to run the final model. Details of distances can be found in the Supplementary Materials of [68].

In order to compare the performance of diverse models with different penalties, 1000 training and test datasets were simulated and the sPLS-DA method was carried out by three-fold CV for $\mathbf{X}$. The performance was measured via BER, and it was assessed for each value of a grid (`keepX`) from Component 1 to H, one component at a time. The different grids of values of the number of variables were carefully chosen to achieve a trade-off between resolution and computational time. Firstly, a two coarse tuning grids were assessed before setting a finer grid. The algorithm used the same grids of `keepX` argument in `tune.splsda` function to tune each component.

Once the optimal parameters were chosen (i.e., number of components and number of variables to select), the final sPLS-DA model was run on the whole dataset $\mathbf{X}$. The performance of this model in terms of BER was estimated using repeated CV.

In summary, three-fold CV was carried out for a maximum of ten components, using the three distances and both types of CER. The optimal number of components was obtained by using the BER and centroid distance. Next, the optimal number of variables was identified by carrying out the second three-fold CV. It was run using BER, centroid distance and values of three grids with different number of variables. Next, when both optimal numbers were obtained, the final model was computed.

Regarding the most relevant variables for explaining the classification of sensors, there are many different criteria [65; 186; 187]. The first measure selected is the relative importance of each variable for each component and another is the accumulated importance of each variable from components. Both measures were employed in this research.

Lê Cao et al. [63] applied sPLS-DA and selected only those variables with a non-zero weight. The sparse loading vectors are orthogonal to each other, which leads to uniquely selected variables across all dimensions. Hence, one variable might be influential in one component, but not in the other. Considering the previous argument and that the maximum number of components was three (h = 1, 2, 3), Variable Importance in Projection (VIP) [65] was used to select the most important ones. It is defined using loading vectors and the correlations of all response variables, for each component. VIP_{hj} denotes the relative importance of variable $\mathbf{X}_j$ for component h in the prediction model. Variables with $VIP_{hj} > 1$ are the most relevant for explaining the classification of sensors. VIP_{hj} was calculated using the `vip` function (`mixOmics`). Although the assumption of the sparse loading vectors being orthogonal was considered, in practice, some selected variables were common in two components. Then, a second measure of VIP [187] was employed: VIP_j denotes the overall importance of variable $\mathbf{X}_j$ on all responses (one per class) cumulatively over all components. It is defined using the loading vectors and the sum of squares per component. Variables with $VIP_j > 1$ are the most relevant for explaining the classification of sensors. The selected variables were ranked according to both types of VIPs, which are discussed below for each stage of the time series.

3.4.3.8 | Sensor Classification Using Random Forest Algorithm

The RF algorithm [74] handles big datasets with high dimensionality. It consists of a large number of individual decision trees that were trained with a different sample of classification variables ($\mathbf{X}$) generated by bootstrapping. The overall prediction from the algorithm was determined according to the predictions from individual decision trees. The class which receives most of the votes was selected as the prediction from each sensor. In addition, it can be used for identifying the most important variables.

An advantage of using the bootstrap resampling is that random forests have an Out-Of-Bag (OOB) sample that provides a reasonable approximation of the test error, which allows a built-in validation set that does not require an external data subset for validation. The following steps were carried out to obtain the prediction (output) from the algorithm, using the different classification data. Firstly, from a classification dataset, B random samples with replacements (bootstrap samples) were chosen, as well as one sample of 17 sensors. Next, from each bootstrap sample, a decision tree was grown. At each node, m variables out of total p were randomly selected without replacement. Each tree used an optimal number of variables that was determined by comparing the OOB classification error of a model, based on the number of predictors evaluated. Each node was divided using the variable that provided the best split according to the value of a variable importance measure (Gini index). Each tree grew to its optimal number of nodes. The optimal value of this hyper-parameter was obtained by comparing the OOB classification error of a model, based on the minimum size of the terminal nodes. From each bootstrap sample, a decision tree came up with a set of rules for classifying the sensors. Finally, the predicted class for each sensor was determined using those trees which excluded the sensor from its bootstrap sample. Each sensor was assigned to the class that received the majority of votes.

If the number of trees is high enough, the OOB classification error is roughly equivalent to the leave-one-out cross-validation error. Furthermore, RF does not produce overfitting problems when increasing the number of trees created in the process. According to previous arguments, 1500 trees were created for all cases to run the algorithm and the OOB classification error was used as an estimate of the test error. Prior to running the final RF algorithm, the optimal values of the number of predictors and the terminal nodes were determined (i.e., those corresponding to a stable OOB error).

In order to select the most important variables, the methodology proposed by Han et al. [188] was used, which is based on two indices: Mean Decrease Accuracy (MDA) and Mean Decrease in Gini (MDG). The combined information is denoted as MDAMDG. Some reasons for using the methodology were: (1) The OOB error usually gives fair estimations compared to the usual alternative test set error, even if it is considered to be a little bit optimistic. (2) The use of both indices is more robust than considering any individual one [188]. MDA corresponds to the average of the differences between OOB error before permuting the values of the variable and OOB error after permuting the values of the variable for all trees. Because a random forest is an ensemble of individual decision trees, the expected error rate called Gini impurity [189] is used to calculate MDG. For classification, the node impurity is measured by the Gini index, and the

MDG index is based on this. The former is the sum of a variable's total decrease in node impurity, weighted by the proportion of samples reaching that node in each individual decision tree in the random forest [190]. The higher is the MDG, the greater is the contribution of a given variable in the classification of sensors. The functions `randomForest` and `importance` [191] were used to carried out the RF algorithm.

3.5 | Results and Discussion

The methodology employed consists of using sPLS-DA and RF in order to classify time series of T and to determine the optimal variables for discriminating the series. Both methods had input features calculated from different functions (e.g., sample ACF, sample PACF, spectral density and MR), as well as estimated parameters from the seasonal ARIMA model, Wold decomposition and the last level of smoothing components from the additive SH-W method. Additionally, other features were computed such as the mean, variance and maximum values of functions applied to residuals of the seasonal ARIMA or SH-W models (e.g., sample ACF, sample PACF, spectral density and statistics of the SW and KS normality test, among others). For sPLS-DA, centroid distance and BER were considered when running the final model. Two indicators of VIPs were used to rank the selected variables. The first one measures the relative importance of each relevant variable, per component in the prediction model, while the second indicator evaluates the overall importance of each variable over all components. Additionally, based on the results from sPLS-DA, a methodology was proposed to reduce the number of sensors required for a long-term microclimate monitoring. Regarding RF, the OOB classification error was used as an estimate of the test error and the parameters MDG and MDA were computed as indicators of the variable importance. Both sPLS-DA and RF methods were applied to select and compare the optimal variables that explain the classification of sensors according to values of T.

Before carrying out sPLS-DA and RF, the k-means algorithm with PCA was employed, in order to characterize the different zones in the museum, according to the indoor temperature. Once the three zones (NW, SE and Sk) were established, these classes were used as input for sPLS-DA and RF.

3.5.1 | Identification of Structural Breaks in the Time Series

The supF and CUSUM tests were applied to study the stages. In Wr1, the former test revealed a structural break after the 682nd observation (20 November at 7:00 a.m., p-value = 0.01). In Cd, this test suggests another break after the 1981st observation (13

January at 10:00 a.m., p-value = 0.02). Finally, a structural break was also found in the Ht stage at the 6133rd observation (4 July at 10:00 a.m., p-value = 0.03). Accordingly, the stages Wr1, Cd and Ht were split in two parts (i.e., before and after the structural break). Thus, the following eight stages of T were considered: Wr1A (warm 1 before the structural break, n = 682), Wr1B (warm 1 after the break, n = 808), CdA (cold period before the break, n = 491), CdB (cold stage after the break, n = 1212), Tr (transition, n = 2303), HtA (hot stage before the break, n = 637), HtB (hot stage after the structural break, n = 1690) and Wr2 (warm 2, n = 903).

The CUSUM chart identified a significant shift at the same instances for the stages Wr1, Cd and Ht. The main reason for these structural breaks could have been sudden changes of T outside the museum or possibly modifications in the air conditioning and heating systems. January is usually the coldest month of the year in Valencia, while July and August are the hottest ones. It is reasonable to assume that the configuration of the air conditioning system was modified in these months, in order to maintain an appropriate microclimate inside the archaeological site.

According to the structural breaks, all time series were split into eight stages, and the four methods explained in the next sections were applied to each stage. This step is necessary to avoid possible problems with the properties of the estimated parameters of the models applied [123]. The four methods (M1–M4) were carried out separately for each stage of the 27 time series of $\boldsymbol{T}$ or $\boldsymbol{W}$: Wr1A, Wr1B, CdA, CdB, Tr, HtA, HtB and Wr2. As an exception, in Method 2, apart from modeling each stage separately, the complete time series was also considered.

3.5.2 | Calculation of Classification Variables—M1–M4

For Method 3, for the most successful seasonal ARIMA models, both tests (KS and SW) rejected the normality hypothesis of the errors in 100% of cases. Furthermore, all Q-Q normal score plots displayed that residuals were not falling close to the line at both extremes. ADF test suggests that the errors are stationary in all cases. According to the LBQ test, the errors are independent (lags from nt to $nt + 48$) as maximum in 77.00% of the 27 sensors (in stage Wr1A) and as minimum 0.0% (HtB). The LM test suggests the absence of Arch effects as maximum 92.52% (Wr1A) and as minimum 3.70% (Wr1B) (see Table 3.1).

For Method 4, for the most successful seasonal ARIMA models (see Table 3.2), both tests of normality rejected the hypothesis of normal distribution for the errors in 100% of models. The ADF test suggests that errors are stationary in all cases. The analysis of residuals of the different fitted models indicates that the condition of not autocorrelation

(WN) is not fulfilled in all cases. In particular, according to the LBQ test, errors are autocorrelated (up to lag 24) at least in 25.93% of the models (for stage HtB) and the maximum was 92.59% (Wr1A). The LM test suggests absence of Arch effects, with a maximum of 96.30% (Wr1A) and a minimum of 3.70% (Tr) (see Table 3.1). In order to extract further information, the same features computed in M3 for the residuals (Type 7 variables) were also estimated in M4 and used as classification variables.

The results of Methods 1–4 were arranged as matrices, denoted as $\mathbf{X}_1$ (27 sensors × 88 variables), $\mathbf{X}_2$ (27 sensors × 296 variables), $\mathbf{X}_3$ (27 sensors × 143 variables) and $\mathbf{X}_4$ (27 sensors × 139 variables), respectively. The regressor matrices $\mathbf{X}_1$ to $\mathbf{X}_4$ contained the following percentages of missing values: 5.28%, 5.66%, 3.55% and 8.61%, respectively.

3.5.3 | Determination of Number of Classes and Class per Sensor Using PCA and K-Means Algorithm

The best option appeared to be $K = 6$ (coincidence of 8 out of the 27 indices), followed by $K = 2$ and $K = 3$ (coincidences of 7 and 5, respectively), and then $K = 4$ and $K = 5$ (Figure 3.8a). According to previous research, there is a pronounced temperature gradient at the museum, particularly in summer, caused by the greenhouse effect of the skylight. Then, each cluster or class was expected to be related to the distance of each sensor from the skylight, among other factors, such as the influence of the weather conditions outdoors and the effect of the air conditioning system. Furthermore, due to the large size of the museum (2500 m^2) and the temperature gradient that exists at the entrance and below the skylight, it seems better to consider three zones instead of just two. Hence, $K = 3$ was the number of clusters used for the k-means algorithm.

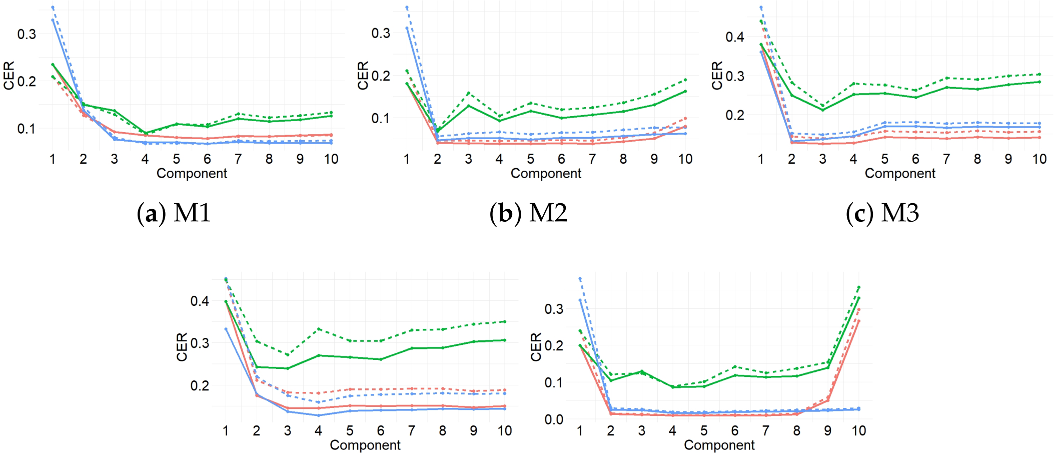
The k-means method classified sensors `C0`, `C1`, `A4`, `A5` and `F` in the SE zone. However, by checking their position on the map of the museum (Figure 3.8b), these sensors could be regarded in the boundary between the NW and SE zones. Hence, it should be discussed whether such classification is appropriate, or if they should be regarded within the NW zone. In order to study this issue, two classifications were analyzed using sPLS-DA: (1) by considering A4, A5 and F in the NW zone; and (2) by locating C0 and C1 in the NW zone. For both cases, the rate of misclassified sensors was computed by means of sPLS-DA. The error rates for the classification from the k-means algorithm were: 0.25 (M1), 0.30 (M2), 0.29 (M3) and 0.42 (M4). For Case (1), the classification error rates were: 0.15 (M1), 0.19 (M2), 0.31 (M3) and 0.35 (M4). For Case (2), the error rates were: 0.21 (M1), 0.23 (M2), 0.30 (M3) and 0.41 (M4). It turns out that the lowest error rates were found for Case (1). Thus, sensors A4, A5 and F were considered as part of the NW zone for the next sections. Finally, 13 sensors were classified in the NW zone

(A4, A5, A6, B5, B6, C5, C6, D1, D2, D3, D5, D6 and F), eight in the SE zone (A, B, B1, C, C0, C1, D and G) and six in the Sk zone (A2, A3, B2, B3, B4 and C3). The proposed classification of sensors is shown in Figure 3.1a, which depicts an association between T values and the three zones of the museum: NW, SE and Sk.

Table 3.2: The most successful seasonal ARIMA $(p, d, q)(P, D, Q)_S$ model (M) per sensor for different stages (Stg), with $S = 24$ and $D = d = 1$.

Stg	**M**	**D**	**B**	**A**	**C**	**G**	**B2**	**B3**	**B4**	**A3**	**A2**	**C3**	**B5**	**C5**	**C6**	**D6**	**D5**	**A4**	**A5**	**B6**	**A6**	**F**	**D1**	**D3**	**C0**	**C1**	**B1**	**D2**
Wr1A	p	0	0	0	0	0	1	0	0	0	0	0	0	1	1	1	0	0	0	0	1	1	0	1	0	2	2	0
	q	2	2	1	1	1	1	2	1	2	3	1	1	1	2	1	3	2	1	0	0	0	2	1	3	0	0	1
	P	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
	Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Wr1B	p	1	1	1	1	1	0	2	1	0	1	0	1	0	2	2	1	2	3	1	1	0	0	1	1	1	1	1
	q	1	1	1	1	2	1	0	2	2	2	0	1	1	0	0	1	1	0	1	2	2	1	1	1	2	2	2
	P	2	2	2	2	2	0	2	2	2	2	0	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
	Q	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CdA	p	1	0	0	0	1	2	0	2	1	1	2	2	1	0	0	1	3	0	0	2	3	0	1	0	1	0	1
	q	1	1	2	2	3	0	2	0	1	0	1	1	0	1	0	1	0	2	0	2	0	3	1	1	0	3	1
	P	2	2	2	2	0	2	0	2	2	2	0	2	0	2	0	2	0	2	2	0	2	0	1	2	2	2	2
	Q	0	0	1	1	1	1	1	1	0	0	1	0	2	2	1	0	2	0	0	1	0	2	2	1	2	0	1
CdB	p	0	0	0	0	3	0	1	1	1	1	0	1	0	2	0	0	1	0	0	0	1	3	3	1	3	1	3
	q	2	1	3	3	0	3	2	2	1	2	1	2	1	0	1	1	2	3	3	1	0	0	0	0	0	2	0
	P	2	2	2	2	2	2	0	0	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2
	Q	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HtA	p	2	1	2	1	3	0	0	3	0	2	3	2	0	1	0	0	0	0	1	1	1	0	0	2	0	3	1
	q	2	3	0	1	1	1	3	0	4	2	0	1	2	1	3	3	1	1	1	0	1	1	0	2	2	1	1
	P	0	0	2	2	0	2	1	2	0	0	2	2	2	2	2	2	2	2	2	2	0	2	0	0	2	0	2
	Q	1	1	1	1	1	2	0	0	1	1	0	0	0	0	0	0	0	0	0	2	1	0	1	1	1	1	0
HtB	p	2	3	0	0	1	1	3	0	0	0	1	1	0	3	0	0	0	1	0	0	1	0	2	0	2	0	0
	q	1	0	2	1	1	2	1	2	0	2	2	1	1	0	3	0	2	2	2	1	0	1	0	3	0	2	1
	P	0	2	2	2	2	0	0	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
	Q	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Tr	p	2	3	0	0	3	0	1	0	1	0	3	1	3	3	3	3	1	3	0	3	0	3	1	0	3	0	1
	q	0	0	1	3	0	1	3	3	1	2	0	2	0	0	0	0	2	0	0	0	3	0	1	3	0	3	2
	P	2	2	2	2	2	0	0	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
	Q	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Wr2	p	3	2	0	0	0	0	0	2	0	2	0	1	3	2	0	0	2	1	0	3	0	0	0	0	0	0	0
	q	0	1	3	0	2	3	3	1	3	0	2	2	0	1	3	3	0	2	2	0	0	3	2	2	3	3	2
	P	0	2	2	2	2	2	2	0	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
	Q	1	0	0	0	0	0	0	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

3.5.4 | Sensor Classification Using sPLS-DA



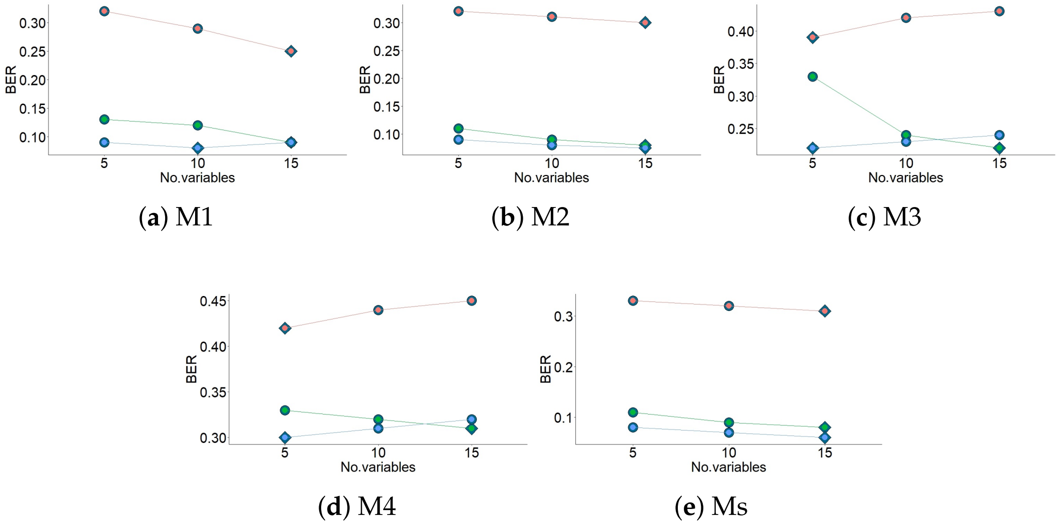
(**a**) M1 (**b**) M2 (**c**) M3

(**d**) M4 (**e**) Ms

Figure 3.9: Evaluation of the PLS-DA performance for the classification of sensors into three categories. Vertical axes indicate the classification error rate (CER) for each prediction distance as a function of the number of components (horizontal axis) for: M1 (**a**); M2 (**b**); M3 (**c**); M4 (**d**); and using the variables from all methods (**e**). Three types of prediction distances were considered: Mahalanobis (green lines), maximum (blue lines) and centroid (red lines). Two types of CER were computed: balanced error rate (dashed lines) and overall error rate (solid lines). PLS-DA was carried out using repeated three-fold CV 1000 times.

Figure 3.9 displays the results from the first three-fold CV for the four methods (M1–M4) and when using variables from all methods (Ms). According to the results, centroid distance performed the best for the first two components, in the case of M1, M4 and Ms. For M1 and Ms, Overall or BER performed the best for both classification error rates. For the other methods (M2, M3 and M4), Overall was the best. The centroid distance and BER from these results were selected as input in the next step of the method (see Figure 3.10). The values of BER and centroid distance suggested that three components are enough to classify the sensors. Figure 3.10 shows the results from the second three-fold CV for the final grid (5, 10 and 15 variables). The results suggest that the number of variables for the first component of each method were the following: 15 (M1 and M2), 5 (M3 and M4) and 15 for Ms. The information displayed in Figures 3.9 and 3.10 (centroid distance, BER and number of variables per component) was used to apply the final model.

The notation used in this study are the estimates of the following parameters: mean, range, median and variance of ACF values (`acf.m`, `acf.r`, `acf.md` and `acf.v`); PACF at lags 1–6 (`pacf1–pacf6`); mean of the time series (`M`); statistics of the KM test (`kolg.t`); statistics of the SW test (`shap.t`); maximum values of the periodogram (`spec.mx`); frequency of the maximum values of the periodogram (`freq`); variance of the residuals (`res.v`); SSE (`sse`); seasonal components (`s1–s24`); level (`a`); slope (`b`); coefficients of the Wold decomposition (`psi1–psi5`); the first AR term (`ar1`); the first MA term (`ma1`); the first SMA term (`sma1`); the first SAR term (`sar1`); and the 17th prediction of T, which was denoted as `pred.17`.



(**a**) M1 (**b**) M2 (**c**) M3

(**d**) M4 (**e**) Ms

Figure 3.10: Evaluation of the PLS-DA performance (considering three components) for the classification of sensors into three categories. Vertical axes indicate the Balance error rate (BER) per component (orange lines, Component 1; green lines, Component 2; and blue lines, Component 3). BER values were computed across all folds using 5, 10 or 15 variable (horizontal axes) for each method: M1 (**a**); M2 (**b**); M3 (**c**); M4 (**d**); and all methods (Ms) (**e**). The three-fold CV technique was run 1000 times, using maximum distance prediction. Diamonds highlight the optimal number of variables per component.

The most relevant variables per method are the estimates of the following parameters:

- Selected variables from M1: PACF at lags 1–4, MR and parameters of the sample ACF (mean, range, median and mean) (Table 3.3 (a)).
- Selected variables from M2: Level, slope, some seasonal components (5, 8, 10–12, 16–21 and 23), mean and median of sample ACF (residuals), SSE (residuals) and maximum of the periodogram (residuals) (Table 3.3 (b)).
- Selected variables from M3: Parameters of MA, SAR and SMA of seasonal ARIMA models, PACF at lags 1–5 (residuals), mean and median of sample ACF (residuals),

statistic of KM normality test (residuals) and frequency for the maximum of the periodogram (residuals) (Table 3.4 (a)).

- Selected variables from M4: The first four Wold coefficients, mean and median of sample ACF (residuals), statistics of the SW and KM normality tests (residuals), PACF at first five lags (residuals), variance (residuals) and maximum of the periodogram (residuals) (see Table 3.4 (b)).
- Selected variables when using the total set of variables from the four methods (some values are highlighted in bold and blue in Tables 3.3 and 3.4): Slope (M2), some seasonal components (2, 3, 5, 8, 11, 12, 14, 16–18 and 21–23) (M2); SSE (M2); Wold Coefficient 1 (M4); mean (M1–M4), median (M2), range (M1 and M2) and variance (M1 and M2) of sample ACF values; and maximum of the periodogram (M2–M4).

Based on the information in Tables 3.3 and 3.4, the two most influential stages of the time series are highlighted in bold in Table 3.5 per component and method, or just one stage when it had a value greater than 50%. It can be deduced from the results in Table 3.5 that the two most important stages for classifying sensors were HtA and Tr, which is intuitively appealing because important temperature fluctuations occur in summer due to the greenhouse effect caused by the skylight.

Figure 3.11 displays the estimations of the average BER over all components from the sPLS-DA. According to the results, M1 obtained the best performance, followed by M2.

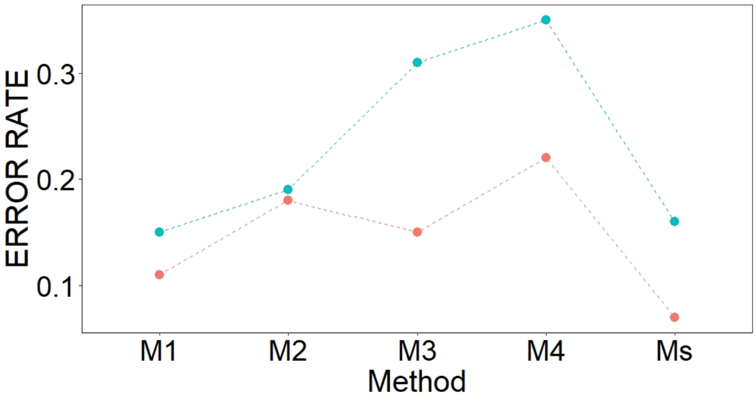


Figure 3.11: Error rates derived from the sPLS-DA and RF algorithms. Red points are OOB classification error rates per method, based on different sets of classification variables: M1, M2, M3, M4 and all variables (Ms). Blue points are mean values of BER for all the components, per method, for sPLS-DA.

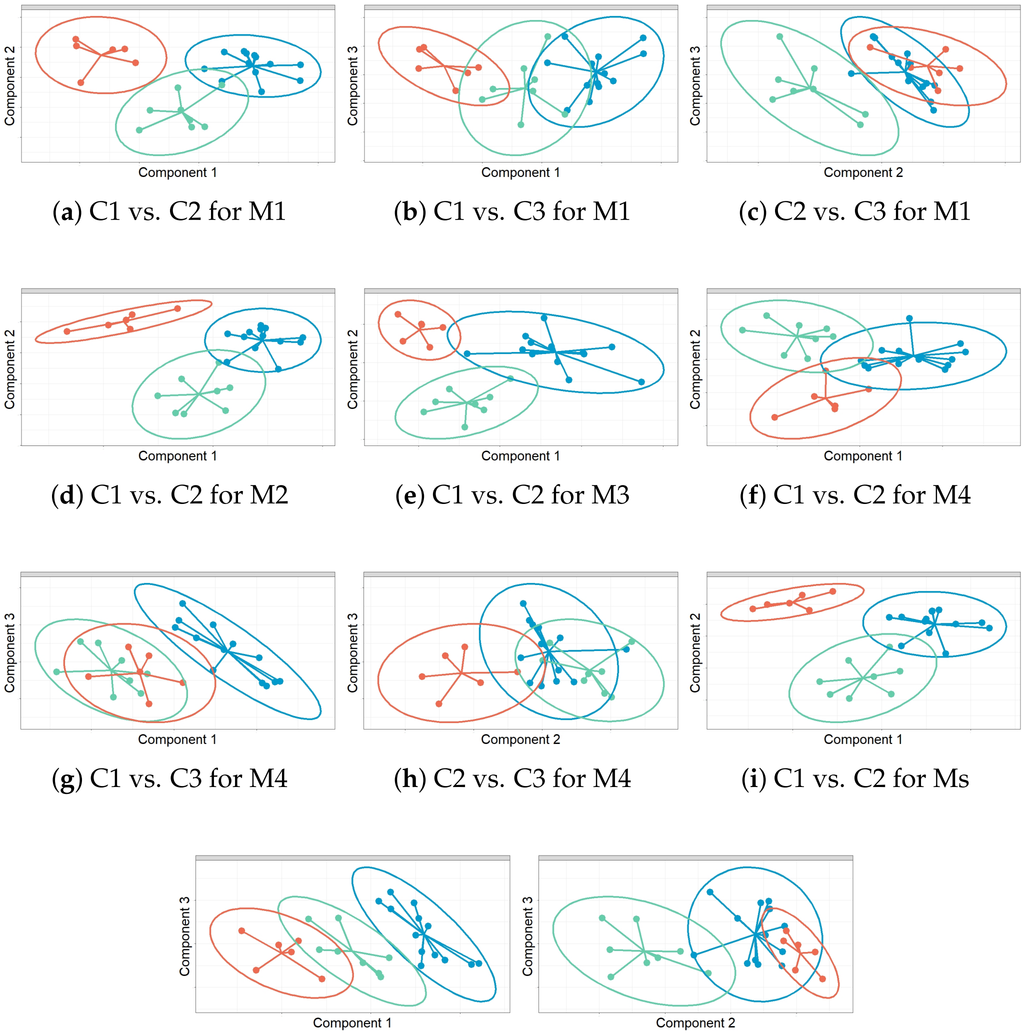
For Component 1 (C1) versus Component 2 (C2) from sPLS-DA: C1 displays a clear discrimination between sensors in the Sk class against NW and SE for both M1 and M2 (see Figure 3.12a,d). This fact is more clear for M2. For M2, C2 clearly separates Sk vs. SE, as well as Sk vs. NE (see Figure 3.12a). For M3, C1 discriminates Sk vs. NW satisfactorily, and C2 shows the same performance for this method (see Figure 3.12e). For M4, C1 properly separates SE vs. NW, while C2 discriminates between SE and Sk (see Figure 3.12d). When using all variables (Ms), C1 shows an adequate discrimination of Sk vs. NW and C2 between Sk vs. SE (see Figure 3.12i). For C1 vs. C3 from sPLS-DA: For M1, C1 clearly separates Sk against NW (see Figure 3.12b). For M4, C1 discriminates NW against the other classes, but Sk and SE appear overlaid (see Figure 3.12g). With C3, the classes could not be discriminated. For C2 vs. C3 from sPLS-DA: C2 properly separates SE against the other clusters in M1 and Ms, but the other classes appear overlaid (see Figure 3.12c,k). For M4, C2 shows a good discrimination of Sk against the other classes (see Figure 3.12h), but C3 did not yield any separation.

3.5.5 | Sensor Classification Using Random Forest Algorithm

Figure 3.11 displays the values of OOB classification error. The best results were achieved from Ms and the second best result from M1. Table 3.6 shows the selected variables in descending order, according to MDAMDG. The most relevant variables per method are the estimates of the following parameters:

- Selected variables from M1: Mean, range and variance of the sample ACF values, the first four values of sample PACF, the maximum value of the periodogram and the mean.
- Selected variables from M2: Level, slope and 18 seasonal components (2, 3, 7–12, 14 and 16–24) from SH-W method, SSE, maximum of the periodogram values, statistic of the KS normality test and 17th prediction of T.
- Selected variables from M3: First term of the AR, MA and SMA components of the seasonal ARIMA models, sample PACF at lags 1, 2 and 5, mean and median of the sample ACF values, residual variance, maximum of the periodogram values and statistic of the KS normality test.
- Selected variables from M4: Coefficients 1, 3, 4 and 5 from the Wold decomposition; the first five values of the sample PACF; the mean, median, range and variance of the sample ACF values; residual variance; maximum of the periodogram values of the residuals; and statistics of both the SH and KS normality tests.
- Selected variables when using all classification variables from the four methods (Ms): mean (M1–M3), median (M2 and M3), range (M2) and variance (M2) of the sample ACF values; sample PACF at lags 1 (M1 and M3), 2 (M1) and 3 (M1); first

coefficient of the Wold decomposition (M3); and statistic of the KS normality test (M2) (the results from Ms comprise 20% of variables from M1, 60% from M2, 9% from M3 and 11% from M4).



(**a**) C1 vs. C2 for M1 (**b**) C1 vs. C3 for M1 (**c**) C2 vs. C3 for M1

(**d**) C1 vs. C2 for M2 (**e**) C1 vs. C2 for M3 (**f**) C1 vs. C2 for M4

(**g**) C1 vs. C3 for M4 (**h**) C2 vs. C3 for M4 (**i**) C1 vs. C2 for Ms

(**j**) C1 vs. C3 for Ms (**k**) C2 vs. C3 for Ms

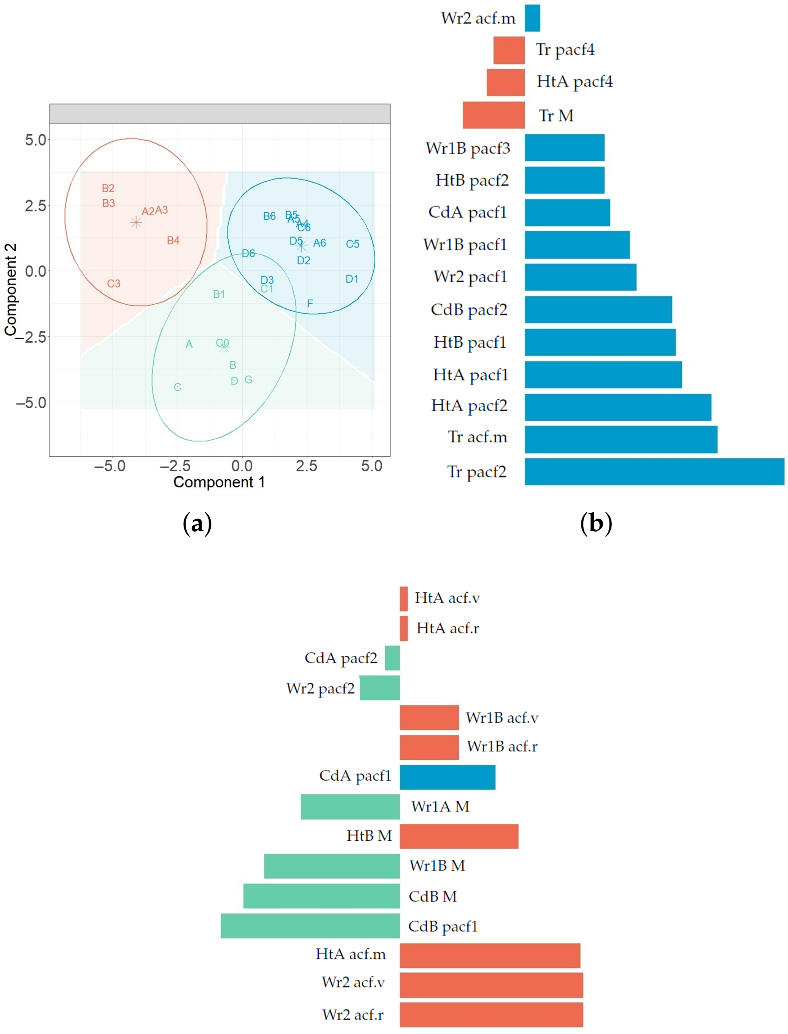
Figure 3.12: Projection of sensors over the three relevant components (C1–C3) from sPLS-DA, per method (M1–M4) or when using all variables (Ms). Graphs show discrimination of the sensors, according to three classes: North Western (NW), South Eastern (SE) and Skylight (Sk). Color codes: NW sensors in blue, Sk in orange and SE in green. Each graph displays a confidence ellipse for each class (at a confidence level of 95%), in order to highlight the strength of the discrimination.

Table 3.3: Selected variables (V) per component (C) from sPLS-DA. Variables highlighted in bold and gray correspond to selected variables for Ms.

	C1		**C2**		**C3**	
	Stage	**V**	**Stage**	**V**	**Stage**	**V**
(a) For Method 1						
1	Tr	**pacf2**	Wr2	**acf.r**	CdA	pacf3
2	Tr	**acf.m**	Wr2	**acf.v**	Wr1A	pacf4
3	HtA	**pacf2**	HtA	**acf.m**	CdA	pacf4
4	HtA	**pacf1**	CdB	pacf1	Wr1A	acf.r
5	HtB	**pacf1**	CdB	M	Wr1A	acf.v
6	CdB	pacf2	Wr1B	M	Wr1A	rMh
7	Wr2	pacf1	HtB	M	Wr1B	rMh
8	Wr1B	pacf1	Wr1A	M	CdA	rMh
9	CdA	pacf1	CdA	pacf1	CdB	rMh
10	HtB	pacf2	Wr1B	acf.r	Tr	rMh
11	Wr1B	pacf3	Wr1B	acf.v	HtA	rMh
12	Tr	M	Wr2	pacf2	HtB	rMh
13	HtA	pacf4	CdA	pacf2	Wr2	rMh
14	Tr	pacf4	HtA	acf.r		
15	Wr2	acf.m	HtA	acf.v		
(a) For Method 1						
1	HtA	**sse**	HtA	**acf.m**		
2	Tr	**sse**	HtA	**acf.md**		
3	Tr	**s18**	HtA	**s12**		
4	Tr	**b**	HtA	**s11**		
5	Tr	**s16**	HtA	**acf.r**		
6	HtB	**sse**	CdA	**acf.v**		
7	Tr	**s17**	HtA	**acf.v**		
8	HtB	**s21**	HtA	**s23**		
9	Tr	**spec.mx**	CdA	**s14**		
10	Wr2	s20	CdA	**s5**		
11	HtB	acf.m	Tr	**a**		
12	Wr1B	s12	CdA	s10		
13	Wr2	acf.md	HtA	**s8**		
14	HtB	s19	HtA	s10		
15	HtB	s23	Wr1A	acf.m		

Table 3.4: Selected variables (V) per component (C) from sPLS-DA. Variables highlighted in bold and gray correspond to selected variables for Ms.

	C1		**C2**		**C3**	
	Stage	**V**	**Stage**	**V**	**Stage**	**V**
(a) For Method 3						
1	Tr	**acf.m**	CdA	acf.m		
2	HtA	res.v	CdA	pacf4		
3	Tr	ma1	CdA	pacf5		
4	HtB	ma1	Wr2	pacf4		
5	Tr	kolg.t	Tr	sar2		
6			CdB	pacf2		
7			HtB	pacf1		
8			Wr2	sar2		
9			CdA	sar1		
10			HtA	sma1		
11			HtA	pacf3		
12			HtA	freq		
13			HtA	pacf1		
14			HtB	acf.md		
15			Wr1A	pacf5		
(b) For Method 4						
1	HtA	psi1	HtA	pacf5	Wr1B	**psi1**
2	HtA	spec.mx	HtB	pacf1	CdB	**acf.m**
3	HtA	res.v	Wr2	acf.v	CdB	**spec.mx**
4	Tr	psi1	HtA	shap.t	Wr2	acf.m
5	Wr1A	acf.m	HtA	acf.m	Wr2	shap.t
6			Wr2	acf.r		
7			Tr	pacf1		
8			Tr	acf.md		
9			Wr1B	psi4		
10			CdA	pacf5		
11			Tr	psi3		
12			CdA	acf.md		
13			Wr2	psi2		
14			Tr	kolg.t		
15			Wr2	kolg.t		



(**a**) (**b**)

(**c**)

Figure 3.13: (**a**) Projection of sensors over components C1 and C2 from sPLS-DA for M1. The graph displays a good discrimination of the sensors according to the classes. They are color coded according to the zone where the sensor is located: NW in blue, SE in green and Sk in orange. The most important variables for (**b**) C1 (**c**) and C2, according to the absolute value of their coefficients, are ordered from bottom to top. The color corresponds to the zone in which the variable yields the highest mean component value.

Table 3.5: Percentages of selected variables per stage of the time series for each component (C) and each method (M). Values were computed according to the information contained in Tables 3.3 and 3.4 (e.g., the value 60.0% for C2 of M2 means that 9 out of the 15 selected variables correspond to HtA, according to Table 3.3 (b)). The two highest values of each column are highlighted in bold and gray, but only one is selected in case of a percentage > 50%.

	M1			**M2**		**M3**		**M4**			**Ms**		
Stage	**C1**	**C2**	**C3**	**C1**	**C2**	**C1**	**C2**	**C1**	**C2**	**C3**	**C1**	**C2**	**C3**
Wr1A	0.00	6.70	**30.80**	0.00	6.70	0.00	6.70	20.00	0.00	0.00	0.00	0.00	0.00
Wr1B	13.30	**20.00**	7.70	6.70	0.00	0.00	0.00	0.00	6.70	20.00	0.00	0.00	33.30
CdA	6.70	13.30	**23.10**	0.00	26.70	0.00	**26.70**	0.00	13.30	0.00	0.00	20.00	6.70
CdB	6.70	13.30	7.70	0.00	0.00	0.00	6.70	0.00	0.00	**40.00**	0.00	0.00	**53.30**
Tr	**26.70**	0.00	7.70	**40.00**	6.70	**60.00**	6.70	20.00	**26.70**	0.00	**60.00**	6.70	0.00
HtA	**20.00**	**20.00**	7.70	6.70	**60.00**	20.00	**26.70**	**60.00**	20.00	0.00	20.00	**60.00**	6.70
HtB	13.30	6.70	7.70	**33.30**	0.00	20.00	13.30	0.00	6.70	0.00	20.00	0.00	0.00
Wr2	13.30	**20.00**	7.70	13.30	0.00	0.00	13.30	0.00	**26.70**	**40.00**	0.00	13.30	0.00

Table 3.6: Selected variables (V) from RF per method and when using the variables from the four methods.

	M1		M2		M3		M4		Ms		
	Stage	**V**	**Stage**	**V**	**Stage**	**V**	**Stage**	**V**	**M**	**Stage**	**V**
1	CdA	pacf1	HtA	sse	Tr	acf.m	HtA	psi1	M1	CdA	pacf1
2	HtA	pacf4	HtA	s24	HtA	sma1	CdB	psi1	M2	HtA	sse
3	HtA	pacf1	HtA	s23	Tr	acf.md	Tr	res.v	M3	Tr	acf.m
4	Tr	pacf2	HtA	s8	CdB	pacf2	HtA	pacf5	M2	HtA	s24
5	CdB	pacf2	HtA	s11	HtA	res.v	HtA	res.v	M4	HtA	psi1
6	HtA	acf.m	HtA	s12	Tr	kolg.t	Tr	acf.m	M1	HtA	pacf4
7	Tr	acf.m	HtB	acf.m	HtA	spec.mx	Tr	spec.mx	M2	HtA	s23
8	CdA	acf.m	CdA	acf.v	CdB	ma1	Tr	psi1	M3	HtA	sma1
9	Tr	pacf1	Tr	sse	Tr	ma1	Tr	kolg.t	M2	HtA	s8
10	HtB	pacf2	HtA	acf.r	HtA	pacf1	CdA	pacf1	M2	HtA	s11
11	HtB	M	HtA	acf.md	CdA	pacf5	HtA	spec.mx	M2	HtA	acf.r
12	Wr2	acf.m	Wr2	s23	HtA	ar1	Wr1A	acf.m	M2	HtA	s12
13	Wr2	pacf1	Wr1B	a	Wr1A	sar2	CdA	acf.md	M2	HtA	acf.m
14	CdB	pacf1	Tr	b	Wr1A	acf.md	CdA	pacf4	M2	CdA	acf.v
15	Tr	M	HtA	s22	Tr	spec.mx	CdA	acf.v	M2	Wr1B	a
16	Wr2	acf.v	Tr	s16	CdA	acf.m	HtB	res.v	M2	HtA	s22
17	Wr2	pacf2	Tr	a	Tr	res.v	CdA	kolg.t	M2	HtB	acf.m
18	HtA	pacf2	Tr	s17	HtB	ma1	Wr1A	psi1	M1	HtA	pacf1
19	Wr2	acf.r	CdA	s10	HtB	sar2	Wr1B	acf.m	M2	HtA	acf.md
20	CdA	M	CdA	a	HtB	pacf1	Tr	pacf1	M3	Tr	acf.md
21	Wr1B	M	HtA	acf.v	CdB	kolg.t	Tr	pacf5	M2	HtA	s9
22	HtB	pacf1	Wr2	s18	CdA	sar1	HtB	psi1	M2	Wr2	s23
23	Wr1B	pacf3	HtB	kolg.t	HtA	pacf3	Tr	acf.md	M2	Tr	s17
24	HtA	M	Wr2	s19	HtA	pacf5	Wr2	pacf4	M2	CdA	a
25	CdB	M	HtB	s21	Tr	freq	CdA	acf.m	M1	Tr	pacf2
26	Wr1A	pacf4	HtA	s10	Wr1B	acf.md	HtA	shap.t	M2	Tr	sse
27	HtB	pacf4	HtB	s9	Wr1A	ma1	HtA	acf.m	M2	HtA	acf.v
28	Wr1B	pacf2	Tr	spec.mx			Wr1B	psi4	M4	CdB	psi1
29	HtB	spec.mx	Tr	s18			Wr2	res.v	M2	HtA	s10
30	Wr1B	acf.m	CdA	s5			CdA	pacf5	M2	Tr	spec.mx
31	HtA	pacf3	Tr	s20			Wr2	acf.r	M3	CdB	pacf2
32	Wr1B	pacf1	CdA	s14			CdA	shap.t	M2	Tr	s16
33	CdA	pacf3	CdA	s1			Wr2	acf.v	M2	Tr	b
34	CdB	pacf4	HtA	acf.m			HtB	pacf1	M2	Wr2	s20
35	Wr1A	pacf2	HtA	s9			Wr2	spec.mx	M3	CdB	ma1
36	Wr1A	M	Wr2	s17			Wr1B	psi3	M2	CdA	s1
37	CdA	pacf2	All	pred.18			Wr2	pacf2	M2	CdA	s10
38	Tr	acf.v	CdA	s22			Wr1B	pacf5			
39	Wr2	pacf4	HtA	spec.mx			CdB	shap.t			
40	Tr	pacf4	Wr2	s3			HtA	psi5			
41	HtA	spec.mx	CdA	s23			Tr	shap.t			
42	CdB	acf.m	Wr2	s2			Wr1A	pacf5			
43	Tr	acf.r					Wr2	kolg.t			
44	HtB	acf.m					CdA	pacf3			

According to Table 3.7, the most important stages for classifying the time series were HtA and Tr. The same result was found from the sPLS-DA method. This result makes sense because there is a pronounced temperature gradient inside the museum, due to the skylight.

Table 3.7: Results from the random forest algorithm: percentages of selected variables per stage and method (M1–M4) or when using all variables from the four methods (Ms). M2 was the only method which used 'all observations' and the 'time series split per stage' for computing the prediction of the time series of T. For the first column (Stage), 'All' refers to 'all observations of a time series' and this category is only used for M2. The two largest percentages per method are highlighted in bold and gray.

Stage	M1	M2	M3	M4	Ms
Wr1A	6.80	0.00	11.10	8.80	0.00
Wr1B	11.40	2.40	3.70	10.50	2.70
CdA	11.40	**19.00**	11.10	15.80	13.50
CdB	11.40	0.00	11.10	7.00	8.10
Tr	**15.90**	**19.00**	**25.90**	**17.50**	**21.60**
HtA	**15.90**	**33.30**	**25.90**	**17.50**	**45.90**
HtB	13.60	9.50	11.10	7.00	2.70
Wr2	13.60	14.30	0.00	15.80	5.40
All	0.00	2.40	0.00	0.00	0.00

For the purpose of time series discriminant analysis, various metrics have been proposed in the literature for measuring such similarity, based on serial features extracted from the original time series [27; 28; 29; 30], parameters from models [38; 39; 40; 145; 146], complexity of the time series [31; 32; 34; 35; 147; 148], properties of the predictions [36; 37] and the comparison of raw data [25]. A description about time series clustering was reported by Vilar and Montero [1]. Furthermore, the sPLS-DA has been applied in a previous study for the classification of different time series of RH [161] using centroid distance. In addition, BER was used to compare the performance of various models. In this research, the variables computed as input for sPLS-DA were the following: estimates of parameters from seasonal ARIMA-GARCH, the last level of smoothing components from SH-W and features extracted from the original time series (ACF, PACF, spectral density, MR and mean). A common seasonal ARIMA-TGARCH was used for the same "stage" of the different time series because it was necessary to deal with the same number of variables as input for sPLS-DA. For the SH-W method, the last level of smoothing components was used instead of RH predictions because each series was split into different stages, according to the seasons and structural breaks identified per season. Regarding the results of the aforementioned study, sPLS-DA with

ARIMA-GARCH yielded the best results according to BER. The second best results were achieved using sPLS-DA with SH-W. With respect to sPLS-DA carried out with variables extracted from the original time series, the results of the aforementioned work were compared with previous studies using PCA directly applied to the time series [7].

In the present study, considering that SH-W is intended for producing point forecasts [125] and that two time series are similar if their forecasts are alike [37], predictions of T were included as additional variables. However, according to results from the sPLS-DA, these variables did not show up as important for the classification of sensors. As an exception, by applying the RF, one prediction of T was selected among the relevant variables. Furthermore, the first five coefficients of Wold decomposition were computed using the estimates of parameters for the best ARMA model per sensor. Previously, the ARIMA polynomials were determined according to the seasonal ARIMA that best fitted the time series. The 'best model' was obtained after one regular and one seasonal differentiation. This procedure made it possible to find the 'best model' and to compute the same number of classification variables for each sensor. However, results from the sPLS-DA and RF were better when using a common seasonal ARIMA, for all sensors in the same stage of the series. By contrast, the previous study [161] used a common ARIMA-GARCH model and did not compute forecasts for SH-W. Expanding the methods in this research, by using the Wold decomposition and predictions of T from SH-W, did not improve the results.

Few parameters of the seasonal ARIMA models and Wold decomposition were found as relevant for the classification of sensors. Regarding the SH-W, most of the important variables were the last level of smoothing components, but, in the other methods, most of the key variables were features extracted from the model residuals, as discussed below. Actually, the percentage of important variables selected from the SH-W that do not correspond to estimates from residuals was 60.0% and 74.0% for the sPLS-DA and RF methods, respectively. Such appropriate results could be explained by the flexibility of this method. The associated weight becomes higher for the most recent observation, which generates both reliable forecasts and smoothing components for a wide range of time series [125]. From sPLS-DA, the percentage of key variables selected that do not correspond to estimates from residuals was 25.0% and 24.0% for the seasonal ARIMA and Wold decomposition, respectively. From RF, these percentages were 33.3% and 21.1%, respectively. The main reason for these low percentages might be the similarity between the time series and the fact that the estimates of parameters were very similar. Given that residuals account for the variability not explained by the models, their information was relevant in this case for classifying the different time series.

According to the BER values from sPLS-DA, M1 performed better than M2. Moreover, the latter yielded better results than using M3, which in turn led to a better classification than with M4. Regarding the results from sPLS-DA, compared with those from the previous study [161], in both cases, the second best results were derived from the SH-W, possibly because it reliably generates the last level of the smoothing components for a wide range of time series. In the previous research, seasonal ARIMA-TGARCH yielded the best result, while in the present work it was achieved using different functions applied to the time series.

With respect to values of the OOB error from the RF, the use of functions applied to different time series (M1) performed better than when applying seasonal ARIMA (M3), which in turn achieved better results than when using SH-W (M2). The worst results were derived from the Wold decomposition (M4). When using all variables from the four methods, two of them were more important, depending on the number of variables selected per method, as follows. For sPLS-DA and RF, the most relevant method was M2, possibly due to the flexibility of the exponential smoothing. The second most important was M1 for sPLS-DA and M3 for RF, although the percentage of variables from M1 was just 3% less than M3. Results from sPLS-DA were consistent with the BER values, calculated separately for each method because, according to BER, M1 and M2 appeared as the most efficient. By contrast, different results were derived from the RF procedure because M1 and M3 appeared as the best methods, according to OOB error.

The most important variables from sPLS-DA and RF are the following. For M1, PACF at lag 2 and the mean values of ACF, which explain the autocorrelation of the different time series; for M2, seasonal Component 18, mean of ACF values that account for the autocorrelation of the residuals from the SH-W method and the residual variance (SSE); for M3, the first term of MA was a key variable, as well as two others explaining the autocorrelation according to the ARIMA model (PACF at lag 2 and mean ACF values) and the residual variance of the model; and for M4, the first coefficient of the Wold decomposition explaining the autocorrelation of T, PACF at lag 5 that accounts for the autocorrelation of residuals from the ARIMA model and the residual variance. In summary, for M1, the key variables for classifying the different sensors explain the dynamic dependency of the time series of T. In contrast, for M2–M4, some of the most relevant variables explain the dynamic dependency of time series and the residuals.

One disadvantage of sPLS-DA is that it is necessary to know a priori the number of clusters of the time series for their classification. Different indices were employed to determine the number of clusters, and the k-means was used to establish the class for each sensor. The results led to a classification of the series which was consistent with the areas of the museum and the knowledge of the microclimate in this site.

For the task of identifying the key variables that characterize each cluster of sensors, the centroid per class was proposed, according to the first two components. Another solution was to use the mean for each variable, in each class, to identify the class where the mean was the highest. This solution helped to provide a value for characterizing each zone, using the main variables.

By using sPLS-DA with estimates from the seasonal ARIMA, SH-W method or functions applied to the time series, one advantage is the capability for classifying time series with very similar characteristics. The best option among the different possible inputs depends on the characteristics of the time series. The procedure was based on a previous study using a similar methodology to classify time series of RH [161]. SH-W also turned out to be the second best approach according to BER, as was found in the present study, while the best method was a hybrid model with seasonal ARIMA and GARCH. In contrast, in this research, the best result of BER was found, using functions applied to the time series (M1). Another advantage is that the centroids per class, the distances between centroids and the projection of sensors onto the first two components might be helpful in order to select a subset of representative sensors.

3.5.6 | Methodology to Select a Subset of Sensors for Future Monitoring Experiments in the Museum

One drawback of the data-loggers used in the monitoring experiment is that they required a manual download of data every two months. It could be much more efficient to use wireless sensors that transmit the readings to a server or a cloud. Commercial wireless sensors are able to instantly transmit the recordings of indoor air conditions, such as temperature and humidity, among others. In addition, these devices can alert the user via e-mail and/or SMS when the recorded values are outside the range of values established as appropriate. Such limits are defined by users according to either the European standards or the requirements of materials or type of artwork which needs to be preserved. However, wireless sensors are more expensive (about 300–400 euros) than the autonomous type of data-loggers used in the present research (approximate price 30 euros) [167]. Moreover, the former can suffer signal transmission problems in some cases. The number of data-loggers used in the present experiment is too big for the long-term monitoring of indoor conditions in this museum. Certain equilibrium has to be reached between the accuracy required and other factors such as sensor price, maintenance and time required to download the data. Thus, for the long-term monitoring of microclimate conditions, it is necessary to decide the optimal number of sensors. For this purpose, one option is to select a subset of sensors per class, according to the results from sPLS-DA. A methodology is proposed in this research for this purpose.

It was assumed that the minimum number of sensors for three clusters should be 15, because three sensors per class are the minimum for applying methodologies such as sPLS-DA, and it is necessary to add a few extra sensors in the case of failure or malfunctioning. Based on this criterion, the recommended position for the representative subset of sensors was decided.

In this study, 27 data-loggers were used because the different zones with similar microclimates were not known a priori, but this amount of sensors seems excessive for long-term monitoring. Employing so many data-loggers on a routine basis is not the best option for monitoring the microclimate in the long term, because it takes a long time to download the data, and nearby sensors offer redundant information. One solution is to make a selection of sensors that capture the relevant information. Thus, it is important to determine how many sensors would be necessary and to establish their location. With the goal of selecting an optimal number of sensors per class and the 'best' option among the 27 sensors, the following methodology was applied. The minimum number of training sensors was established as 15, because each class should have at least three sensors in order to calculate the variance of any variable per class or to apply methods such as sPLS-DA. The idea was to select a set of sensors, based on the first two components from sPLS-DA, centroids of the three classes and the distances between each centroid and the position of sensors in the multivariate space (i.e., pair of coordinates using C1 for x-axis and C2 for y-axis for each sensor and the centroid of its class).

The first step consists of deciding the optimal number of sensors per class, which can be computed using the variance of the centroid distances of sensors in the same class. The class with the highest variance should select more sensors, while the opposite applies to the class with lowest variance. Secondly, sensors in the same class were split into three subsets, which were created according to the distances to the centroid. The idea is to draw three concentric circles per class, with three different values of the radius (R1–R3) and the same center (O), the centroid. Each area between circumferences makes it possible to identify the sensors per subset. Thus, there are three groups per class: the first (G1) is determined by the inner circle, the second (G2) by the area between the circumferences with R1 and R2 radius and the third group (G3) by the area between the circumferences with R2 and R3 radius. Thirdly, the optimal number of sensors was selected per subset. Such optimal value can be computed using the variance of the distances of sensors, within the same subset. Finally, in order to guarantee representation for each subset, a sample of sensors per subset was randomly selected, according to M1, which is the method with the lowest value of mean BER values.

In respect to the selection of a subset of sensors using the outcomes from sPLS-DA for M1, the percentages of the sum of variances of the centroid distances of the classes NW, SE and Sk were 35%, 40% and 25%, respectively. A proposal based on the results from sPLS-DA with M1 are the following:

- For NW, the number of representative sensors was $15 \times 0.35 = 5.25 \approx 6$. The 13 sensors in this zone were classified in each of the three concentric circles, as follows: D2, D5, A6 and C6 for the first area (G1); B6, B5, A5 and A4 for the second area (G2); and D6, D3, F, D1 and C5 for the third area (G3). The number of representative sensors selected per group (G1–G3) in this NW class was decided according to the variance of the distances: one sensor in G1, three in G2 and two in G3. The proposed subset of representative sensors is the following: D2 (for G1); B6, A5 and A4 (for G2); and D6 and F (for G3).
- For SE, the number of representative sensors was $15 \times 0.4 = 6$. The eight sensors in this zone were classified according to the concentric circles as: C0 (G1); A and B (G2); and C, C1, B1, D and G (G3). The number of SE sensors selected per group was determined based on the variance of the distances: one sensor in G1, one in G2 and four in G3. The proposed subset of representative sensors is as follows: C0 (for G1); A (for G2); and C, B1, D and G (for G3).
- For Sk, the number of representative sensors was established as $15 \times 0.25 = 3.75 \approx 4$. The six sensors regarded as Sk were classified as: A2 and A3 (G1); B2, B3 and B4 (G2); and C3 (G3). The number of sensors chosen per group was decided according to the variance of the distances: one in G1, one in G3 and two in G2. The proposed subset of representative sensors was: A3 (for G1); B2 and B4 (for G2); and C3 (for G3).

HtA and Tr appeared as the most relevant stages for the sensor discrimination. This result is consistent with a previous research [158], which reported a pronounced temperature gradient at the museum, particularly in summer, caused by the greenhouse effect of the skylight. The identification of the key stages of the time series for discriminating the sensors might help to select and enhance the criteria of adequate sampling intervals in automated systems for microclimate monitoring.

Based on the results, the methodology proposed seems effective for characterizing the indoor air conditions in a heritage building, aimed at preventive conservation. This approach can use variables previously calculated by means of either functions applied to different time series or fitting the SH-W method. Furthermore, sPLS-DA might be useful to select a subset of representative sensors, in order to decide the best location for autonomous data-loggers or wireless sensors.

3.6 | Conclusions

The temperature in the L'Almoina museum varies according to the location inside the museum and along the year. In order to define a plan for the long-term monitoring for preserving the artifacts, a statistical methodology was proposed. Some of the most important results found in this research are the following.

Both the sPLS-DA and RF methods were useful for identifying the most important variables that explain the differences among the three zones in the museum. For M1, the relevant variables explain the dynamic dependency of the time series. By contrast, with the other methods, the most important variables explain the dynamic dependency of both, the time series of T and the residuals from either SH-W or ARIMA, although most key variables were computed from residuals. Both approaches showed a good capability for discriminating time series. It was possible to obtain parsimonious models with a small subset of variables leading to satisfactory discrimination. Results from sPLS-DA can be easily interpreted. PCA and k-means with sPLS-DA and RF were effective in establishing the different zones in the site and to discriminate the microclimate of these areas. Furthermore, the stages HtA and Tr were the most relevant ones in order to discriminate the different sensors.

The best method for determining the input of variables for sPLS-DA depends on the characteristics of the time series. The SH-W approach appeared to be more flexible for modeling the different time series and obtaining low values of the classification error rate. By applying SH-W, the percentage of selected variables that do not correspond to residuals was higher than when using seasonal ARIMA and Wold decomposition.

For establishing the most important variables for each zone, the centroids of the two first components from sPLS-DA were used. Another option was to identify the class where the mean value of the selected variable was the highest. Thus, the variable with the highest mean for a class could characterize that class.

The methodology proposed might be useful for characterizing different zones in a building, according to the values of T and RH. Furthermore, it might be helpful to establish the optimal number of sensors, in order to manage resources and to monitor the microclimate according to the European Standards.

Regarding future studies: (1) Other versions of sparse PLS DA could be considered, such as SPLSDA and SGPLS [53], in order to compare the capability of classifying time series. Another unsupervised method could also be used to establish the classes before applying sPLS-DA. For example, Guha et al. [192] recently proposed a novel Bayesian-nonparametric strategy for setting the number of clusters and their labels. (2) Further studies about the indoor air conditions at this museum should focus on the time series analysis of both T and RH, in order to improve the characterization of zones with a similar microclimate. (3) The proposed methodology in this paper might be implemented in ’real time’ monitoring so that corrective actions might be adopted in the case of inappropriate measurements. (4) Work in progress is currently applying advanced statistical methods to study the relationship between time series of T and the height gradient in a typical church in a Mediterranean climate.

4

Characterization of Temperature Gradients According to Height in a Baroque Church by Means of Wireless Sensors

Ramírez, Sandra.; Zarzo, M.; Perles, A.; Garcia-Diego, F.J. *Characterization of Temperature Gradients According to Height in a Baroque Church by Means of Wireless Sensors*. Sensors 2021, 21 (20): 6921. doi:10.3390/s21206921.

This chapter corresponds to the publication mentioned above, after changing the positions of some graphs and tables.

4.1 | Abstract

The baroque church of Saint Thomas and Saint Philip Neri (Valencia, Spain), which was built between 1727 and 1736, contains valuable paintings by renowned Spanish artists. Due to the considerable height of the central nave, the church can experience vertical temperature gradients. In order to investigate this issue, temperatures were recorded between Aug 2017 and Feb 2018 from a wireless monitoring system comprised by 21 sensor nodes, which were located at different heights in the church from 2 to 13 m from the floor level. For characterizing the temperature at high, medium and low altitude heights, a novel methodology is proposed based on sparse Partial Least Squares regression (sPLS), Linear Discriminant Analysis (LDA), Holt-Winters method, among others which were applied to time series of temperature. This approach is helpful to discriminate temperature profiles according to sensor height. Once characterized the vertical thermal gradients for each month, it was found that temperature reached the maximum correlation with sensor height in the period between August 10th to September 9th. Furthermore, the most important features from the time series that explain this correlation are the mean temperature and the mean of moving range. In the period mentioned, the vertical thermal gradient was estimated to be about $0.043°C/m$, which implies a difference of $0.47°C$ on average between sensor nodes at 2 m from the floor with respect to the upper ones located at 13 m from the floor level. The gradient was estimated as the slope from a linear regression model using height and hourly mean temperature as the predictor and response, respectively. This gradient is consistent with similar reported studies. The fact that such gradient was only found in one month suggests that the mechanisms of dust deposition on walls involved in vertical thermal gradients are not important in this case regarding the preventive conservation of artworks. Also, the methodology proposed here was useful to discriminate the time series at high, medium and low altitude levels. This approach can be useful when a set of sensors is installed for microclimate monitoring in churches, cathedrals, and other historical buildings, at different levels and positions.

Keywords: Autocorrelation; Holt-Winters; LDA; Temperature gradient; sPLS; Wireless sensors.

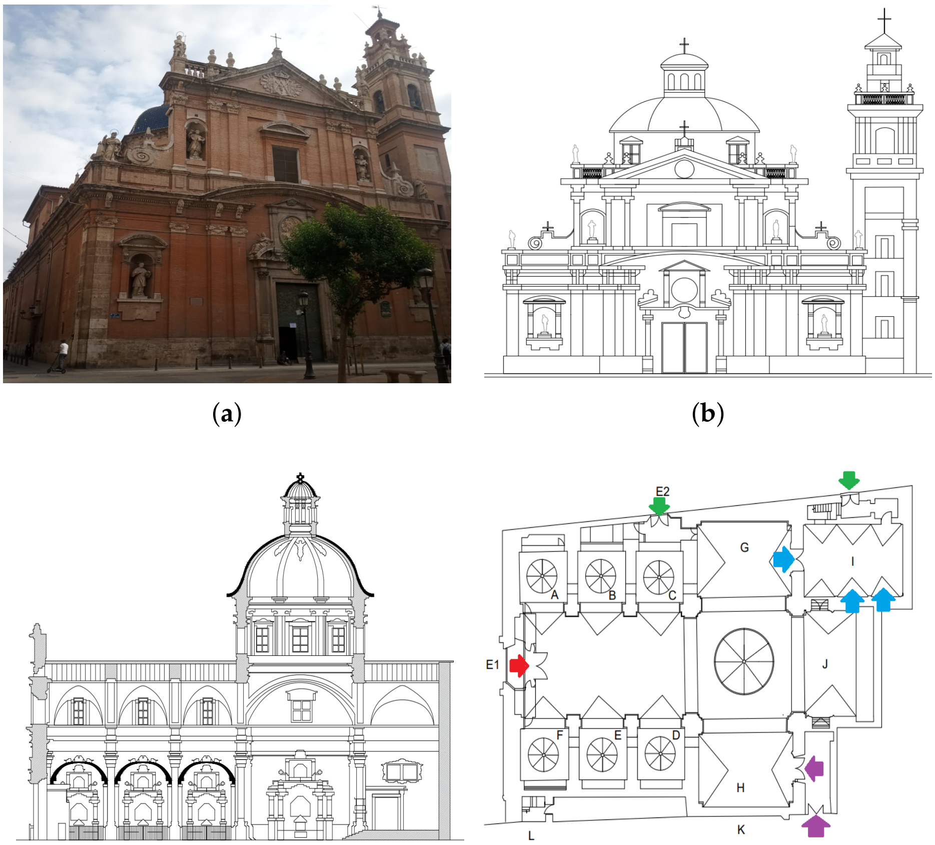
4.2 | Introduction

Cultural heritage is a source of wealth because it promotes tourism, creative art and native culture. Tourists often select places to visit based on the culture and artistic significance of museums, monuments, exhibitions, and historical ruins, among other criteria. The protection and conservation of cultural heritage is a challenge because artworks undergo certain degradation over time. In order to prevent damage, artworks should be maintained in stable and controlled climatic scenarios. However, usually, such conditions are only achieved in museums [3].

Temperature (T) and relative humidity (RH) tend to be more stable inside a building, while outer air conditions present a higher daily and seasonal variability [8]. In fact, the most significant physical factors in the preservation of collections and artefacts are T and RH, which can potentially deteriorate or damage historical or cultural objects [155]. The requirements for an appropriate control of indoor air conditions depend on the type of materials, some of which are very sensitive to sudden variations of T or RH. Thus, particular artworks demand specific microclimatic conditions. Also, certain characteristics of buildings can generate more complex requirements because it is not possible to control the indoor environment [87; 88; 89; 90]. The values of T and RH inside a historical building basically depend on the climatic conditions outside, apart from other factors like construction materials, structure, and dimensions of the building. Variations in T and RH can induce thermal shock [8], air movements, wet-dry cycles [78; 193], and surface or under-surface salt dissolution–crystallisation [194]. Air movements, as well as wet and dry cycles are usually responsible for soiling processes and deterioration [195]. Furthermore, dissolution of alkaline surfaces can be caused by condensation that can be generated by either water vapour coming from open doors, human metabolism, or the use of lit candles. Also, in the presence of high humidity and moderate temperatures, surface condensation and damp can give rise to biological colonization by insects, bacteria or fungi, which can generate biodeterioration in specific areas of the building [196]. In numerous cases, artworks in churches have been affected due to inappropriate microclimatic conditions [197].

In the Mediterranean region, the use of active air conditioning systems has been more common in modern spaces of worship, due to the growing request from the public for comfortable temperatures in the buildings. Such systems must guarantee conditions of wellbeing, safety and energy efficiency [10], but it is a challenge to satisfy at the same time the requirements of human comfort, the preservation of artworks and energy efficiency [10]. In most cases, such requisites cannot be fulfilled optimally. In Spain, as in other Mediterranean countries, the thermal conditions of historical buildings are not considered in current environmental conditioning regulations (e.g., European Standards EN 15757:2010 [97], which is based, on laboratory tests [198] and on case-studies [199; 200]) [201].

In Spain, the majority of ancient churches, cathedrals and other historical buildings do not have air conditioning systems; as a consequence, artworks can experience harmful thermo-hygrometric oscillations, due to the outer climatic conditions. Moreover, these buildings are large, which favors vertical air flows that are related to the deposition of dust and dirt on walls, paintings, frescoes, altarpieces, and other artworks, which can require expensive cleaning and maintenance actions for an appropriate preservation of the cultural heritage. Vertical air movements can be caused by the ventilation, because many of these buildings have windows in the upper part, so the air enters through the main doors and lower inlets and leaves through the upper windows. The presence of vertical thermal gradients is another factor of vertical air movements, because hot air has a lower density and rises up. Therefore, studying the correlation between temperature and sensor height is of interest to assess the vertical air flows, which makes it possible to evaluate whether the gradient of T is acceptable or excessive, regarding the risk for dust deposition in walls and paintings. In case of inappropriate gradients, corrective actions might be proposed. The present research is intended to study vertical temperature gradients in the church of Saint Thomas and Saint Philip Neri in Valencia, Spain (Latitude: 39 30N and Longitude: 000 28W [202]), which has an unheated/natural microclimate indoor (see Figure 4.1). The climate in Valencia is classified as BsK (tropical and subtropical steppe) according to the Köppen classification [202]. The principal source of ventilation is through the main entrance of the church, and there are a few air inlets in the sacristy and the chapel of the Holy Communion though the former is separate from the main nave by a door that usually remains closed. This church contains valuable artworks, among these are paintings by renowned Spanish artists such as Juan de Juanes or Vicente Juan Macip (1507–1579), Jerónimo Jacinto de Espinosa (1600–1667), José Vergara (1726–1799), and Vicente López (1772–1850). These paintings are located in the main chapel, as well as the altarpieces of Saint Joseph and Our Lady of the Unforsaken (see Figures 4.1c and 4.1d).



(**a**) (**b**)

(**c**) (**d**)

Figure 4.1: Church of Saint Thomas the Apostle and Saint Philip Neri in Valencia (Spain). **(a)** Front and side view of the church; **(b)** Front view; **(c)** Longitudinal section; **(d)** Plan of the church, where the different observable structures are indicated: A. Baptismal chapel, B. Chapel of Our Lady of the Forsaken, C. Chapel of the Holy Trinity, D. Chapel of Our Lady of Mount Carmel, E. Chapel of the Calvary, F. Chapel of Saint Anthony of Padua, G. Altarpiece of Saint Joseph, H. Altarpiece of Our Lady of La Salette, I. Chapel of the Holy Communion, J. High altar and main altarpiece, K. Sacristy, L. Bell tower, E1. Main entrance, E2. Side entrance. The small circles indicate the projection of vaults of the internal chapels. The larger circle represents the projection of the main dome of the church. The arrows indicate the air inlet and sources of ventilation in the church.

4.2.1 | Microclimatic Monitoring for the Preservation of Cultural Heritage

In recent years, European governments have funded different initiatives with the goal of preserving artworks in museums and similar buildings. For example, the Collection-

Care Project is working at present on an innovative system of wireless sensors for the preservation of cultural heritage [166]. In this context, experts suggest that it is necessary to implement continuous monitoring systems to identify harmful microclimatic conditions which affect the works of art [99]. Long-term monitoring of indoor air conditions is a key issue, according to the new requirements for preventive conservation [203]. Such systems require maintenance and routine practices [204]. Also, practical solutions need to be proposed for the adaptation of climate change [77]. Furthermore, it is important to define the compatibility between the climate control potentials and the preservation requirements [99].

Many studies about the microclimate monitoring of historical buildings have recorded time series of either T or RH by means of autonomous data loggers [4; 5; 6; 7; 8; 9; 10; 11; 161; 205] or wireless monitoring systems [3; 206]. Some of these research works [8; 9; 10] have been carried out in European churches to investigate the possible consequences derived from traditional heating, in order to improve indoor air conditions for preserving the cultural heritage. Sensors are often located in the historical buildings at a similar distance to the floor level. Regarding the statistical methodology, Principal Component Analysis (PCA) was applied to time series of RH recorded at the Cathedral of Valencia aimed at obtaining clusters of sensors [6; 7]. Using the same data set, a novel methodology was recently proposed for classifying the different time series of RH [161]; it was based on sparse Partial Least Squares Discriminant Analysis (sPLS-DA) [63] using input variables extracted from either Autoregressive Integrated Moving Average models (ARIMA), Holt-Winters method, or functions applied to time series of RH. In a subsequent research, the aforementioned approach was extended using new variables from the Holt-Winters method and Wold decomposition, which was applied to time series of T recorded at the archaeological site of L'Almoina in Valencia [205].

4.2.2 | Microclimatic Studies with Sensors Located at Different Heights

The European Friendly-Heating Project [207] highlights the problems caused by installing heating systems in old worship places [10]. In the Mediterranean region, the heating demand in churches is much lower in winter compared with places in Northern Europe, while the requirements for dehumidifying and cooling are greater in spring and summer because of the outdoor humidity and high temperatures [10].

Merello et al. [11] studied time series of T and RH recorded from dataloggers in specific wall orientations and at different levels (floor vs. upper position) at Ariadne's house in Pompeii (Italy). They applied Analysis of Variance (ANOVA) to either estimates of mean, minimum, or maximum of daily time series of RH and T, in order to study the effect of height and wall orientation where the sensors were located on. Likewise, Aste et al. [193] estimated the vertical gradients of T and RH on the entire volume of the Duomo Cathedral (Milan, Italy). They computed the gradient as the difference of T or RH from the lowest sensors compared with those located at the highest levels. Measurements were recorded at 5, 10, 15, and 20 m, from the floor level. Where the maximum height was 45 m, two further measurements were recorded at 35 and 40 m. They found that the gradients in different points were not relevant except for the areas near the entrance to the North aisle, which undergoes higher changes because the gate is used as a primary entrance by churchgoers. Klein et al. [206] installed a wireless monitoring system at The Cloisters, the medieval branch of the New York Metropolitan Museum of Art, in order to improve long-term microclimate monitoring. Sensors were located at different heights in the galleries (e.g., in the Late Gothic Hall the sensor placement height ranged from 0.5 m up to 11.0 m). They evaluated air moisture levels, the thermal stratification along the height of one gallery, and slight temperature gradients between different galleries. They found higher variations in the Hall at the upper level. Using sensors located in different positions and heights, García-Diego et al. [208] applied ANOVA and contour plots to study the performance of the mean T and RH when the heating system was switched on in order to quantify the effects of the heating system on temperature and RH.

Recently, in contrast to traditional technology, Adán et al. [196] used three-dimensional thermal computer vision-based technologies (3D-TCV) for monitoring climatic conditions. This novel approach records dense thermal information in a 3-D space, resulting a data matrix containing 3-D coordinates with the associated T and time when the values were recorded. This methodology, combined with traditional recordings of T and RH using a wireless monitoring system, was recently applied at the church of Santos Juanes in Valencia [196]. Data were recorded by the wireless sensors at the lower zone of the principal nave of this church and at the upper zone near the domes. Also, the local surface temperature monitoring system obtained data from three different zones. Such information was studied by computing standard deviation of surface T. The datasets from 3D-TCV were analyzed by means of thermal orthoimages at different times and graphs of thermal evolution over time [196].

In total, the present research analyzed 21 hourly time series of T from wireless nodes, for seven months during 2017-2018. Sensors were positioned at different heights, ranging from 2 to 13 m from the floor level in the church of Saint Thomas and Saint Philip Neri in Valencia (Spain). The microclimate monitoring system was developed a few years ago as a test prototype [3].

Different monitoring campaigns for the preventive conservation of cultural heritage have been carried out [209; 210]. Some of them used autonomous data-loggers, e.g., Hobo data-loggers [211] were employed by Visco et al. [212], data-loggers DS1922L [213] by Valero et al. [214], and DS1923 [215] by Merello et al. [158, 216]. There are also studies about microclimates in cultural heritage based on a wired sensor network, composed of different nodes wired to a single microcontroller [6; 158]. A more versatile wired/wireless system [217] can be used to solve the problem when using data-loggers, which requires recordings from them to be downloaded manually. The IoT wireless system employed in this study was developed by Perles et al. [3].

The study of thermal conditions in a building can be approached in different ways; for instance, by analyzing the rate of T changes every two height levels (low to high), or by analyzing changes in the characteristics of time series of T at different height levels in the buildings. The latter approach would correspond to classifying time series of T in different clusters according to the height levels. In the first case, a pronounced rate of T change per height might imply phenomena of dust deposition on the walls and artworks in the building, thus it would be necessary to take corrective actions to reduce risks on works of art. In the second case, classifying time series according to different heights (e.g., high, medium and low levels) could be helpful for monitoring the microclimatic conditions in the building. Possible reasons for classifying a set of sensors incorrectly, might be the malfunctioning of sensors, changes of thermal conditions where the sensors are located, and the classification method performance, which is influenced by the total number of sensors or the number of sensors per clusters. Thus, sensors incorrectly classified should be evaluated to identify possible setbacks for the artworks.

In order to study vertical temperature gradients and to characterize the time series of T per different height level, two methodologies are proposed. The first one is helpful to determine the existence of a vertical gradient, to estimate the gradient, and to establish the period in a year when such gradient is apparent. This methodology is based on Pearson's correlation coefficient [218] and linear regression [219]. The second methodology could be used for characterizing the temperature at high, medium and low altitude heights and to determine the main variables that help establish the changes of temperature by level. This methodology which classifies time series, is based on sparse Partial Least Squares regression (sPLS) and Linear Discriminant Analysis (LDA). They are employed for classifying purposes, using features from time series as input, which are computed with two methods. The first one corresponds to using some traditional time series functions (i.e., Auto Correlation Function ACF, Partial Auto Correlation Function PACF, periodogram, Moving Range MR), and features defined using quantiles [220]. The second, corresponds to using the Holt-Winters method. Finally, with the goal of proposing a plan for long-term monitoring in the church of Saint Tomas and Saint Philip Neri in Valencia, Spain, the thermal condition in this building was analyzed by using both methodologies.

With respect to the first methodology proposed, the temperature gradient has not received much attention yet in the context of art conservation. Some studies have approached the temperature analysis by computing the variation in temperature at different levels of heights [193; 206], by using contour graphs of temperatures or by comparing estimates of parameters such as the maximum and minimum temperature [11; 208].

Regarding the second methodology proposed, which is used here for classifying time series of T according to different levels of height, it is considered as a novel approach in the context of clustering of time series and cultural heritage. The methodology consists of applying both sPLS [53] with LDA, using features extracted from time series as input. The dissimilarity measures calculated for the method were computed according to other approaches employed in the field of clustering of time series (i.e., profiles of time series, dynamic structure of series, assuming specific underlying models, future forecasts, among others) [1; 25; 38]. In this case, the dissimilarity measure (i.e., Mahalanobis or Euclidean distance) was computed using a linear combination of a set of variables. These variables could correspond to different approaches, e.g., assuming specific underlying models and future forecasts, profiles of time series and the dynamic structure of series. In this sense, Elorrieta, Felipe et al. [49] have proposed using several features from the field of astronomy and two features that they designed as input for different classification methods (e.g., logistic regression, CART algorithm, boosting, random forest, support vector machine, artificial neural network, and Lasso regression). Some features were extracted from raw data, while others after fitting a harmonic model [49; 50]. Concerning the classification algorithms and the methods for computing features from series that are proposed in this paper, this is probably the first time that the combination of both algorithms and such methods are used for classifying and clustering of time series. On the other hand, for art conservation, classifying time series has rarely been explored and it has only been analyzed using PCA [6; 7] or sPLS-DA [161; 205].

Finally, this research reports a statistical analysis conducted in the church of Saint Tomas and Saint Philip Neri for the first time, which is of relevant interest since inappropriate conditions of temperature can affect the artworks inside the church. Furthermore, the results found in this study might provide guidelines for establishing a plan for thermal monitoring and preventive conservation in similar churches.

The structure of this paper is as follows. Firstly, section 4.3 describes the monitoring system, the data set, installation of wireless nodes, as well as criteria for determining the stages of the time series of T, methods for computing features from the series, and the regression method for relating temperature values according to sensor height. The most relevant results and discussion of the different analyses are presented in section 4.4. Finally, conclusions can be found in Section 4.5.

4.3 | Materials and Methods

4.3.1 | Description of the Monitoring System

The monitoring system used in this paper is the same as in [3], where the details and descriptions of the general system and their components are provided. Furthermore, the reason why the general system and their different components were used is explained.

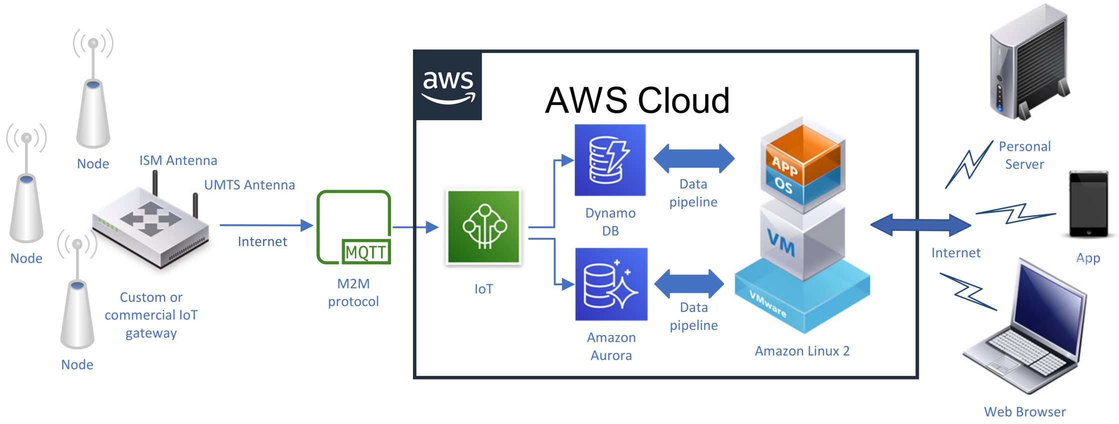


Figure 4.2: Scheme of the wireless microclimate monitoring system. In respect to notation: IoT is Internet of Things, ISM is industrial, scientific, medical band, UTMS is Universal Mobile Telecommunications System, and AWS is Amazon Web Services.

The system (Figure 4.2) was specifically designed for the monitoring needs of cultural heritage buildings and objects. It consists of low-energy wireless sensors, a gateway for collecting the data sampled by the sensors, and a cloud computing infrastructure for data storage, processing and visualisation.

A total amount of 21 wireless sensor nodes were installed at the church Saint Thomas and Saint Philip Neri (Figure 4.3a) for monitoring indoor air conditions. These sensor nodes are built around an ultra-low power C8051F920 microcontroller (Silabs, San José, CA, USA), a CC1101 radio-modem (Texas Instruments, Dallas, USA), a high-density 3.6 V, 1 Ah Lithium-thionyl battery, and a SHT15 chip. The latter is a surface mountable device with a RH sensor and a temperature sensor (Sensirion, Staefa ZH, Switzerland). This device was individually calibrated by the manufacturer (Sensirion). The calibration coefficients are programmed into an inside memory on the chip. To improve the accuracy, these coefficients and the internal voltage regulator are used to calibrate the transmitted signals from the sensors. The accuracy of the SHT15 sensor is ±0.3°C in the range from 10°C to 40°C [221].

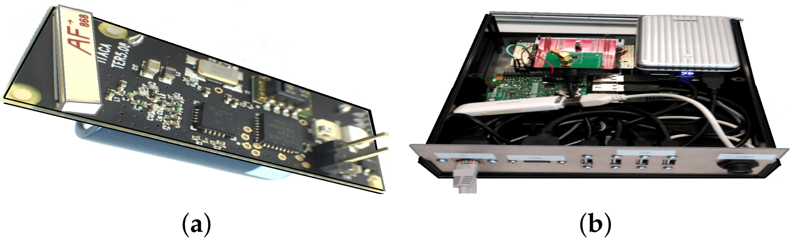


Figure 4.3: **(a)** wireless sensor node (approximate dimensions: $4.1 \times 1.5 \times 1.5$ cm); **(b)** sink gateway.

The sensor nodes were used to sample environmental variables of interest, which were transmitted using GFSK (Gaussian Frequency Shift Keying) modulation in the 868 MHz European unlicensed industrial, scientific, medical (ISM) band. All sensors transmit blindly on the same channel without acknowledgement messages from the gateway. This approach allows to be very energy efficient at the cost of losing some transmissions.

This sensor node is an adaptation of a previous one devoted to the detection of xylophagous [222] and copes adequately with the requirements of life-span and long distances and thick walls of historical buildings [3].

The gateway, shown in Figure 4.3b, was built to be as flexible as possible in order to experiment with different approaches, so it was decided to implement it around a Raspberry Pi 3 board (Broadcom Inc., CA, USA) and the Linux operating systems. To this base system, we added suitable hardware to support the functionality: a CC1101 radio module and an STM32L04 microcontroller (StMicroelectronics n.v, Geneva, Switzerland) to receive the transmissions of the sensor nodes, a 3G USB dongle to provide mobile connectivity to Internet, and, considering that the gateway is connected to the mains power, a rechargeable lithium-ion battery to provide energy to the gateway during power outages. The main task of the gateway is to collect wireless transmissions of the sensor nodes, store it temporarily in a local database and transmit it to Internet when connectivity is available. The data transfer is implemented using the MQTT [223] client server publish/subscribe messaging transport protocol.

For the implementation of the cloud infrastructure, it was decided to choose the offering from Amazon Web Services (AWS). The MQTT messages are processed by the AWS IoT cloud service in order to split the message in sensed magnitudes like temperature, humidity or light level (humidity and light not used in this work), as well as in communication-related parameters (e.g., received signal strength indicator, battery level and message counter). These two types of data flows are stored in a NoSQL (stands for "non SQL" for ones and for "not only SQL" for others) AWS NoSQL DynamoDB database and in an SQL AWS AuroraDB, respectively. In order to allow data access through web browsers, a Linux virtual machine was deployed in the AWS EC2 service which runs a Redash [224] data visualization dashboard. For statistical analysis, all data collected along the monitored period could be downloaded locally using AWS Datapipelines service.

Among the advantages of this monitoring system are (1) it is capable of storing an unlimited volume of data, this cloud helps to increase sample frequency, which means that updating the recorded information can be carried out every time period, as required (i.e., every second, minute, and another time period), (2) the fact that updating the recorded information does not need to be carried out manually.

The cost of these devices is highly dependent on the type of work to be performed. Mass-market devices tend to be cheaper due to the scale of production, which reduces the cost of the bill of materials and dilutes the engineering cost. In the field of cultural heritage specific devices, and in general in the scientific field, this scale does not apply, so engineering costs are easy to estimate in the cost of the devices and other important costs, such as installation costs (e.g. a wired installation is often very expensive) or personnel costs (e.g. a classical data logger will require periodic battery replacement and manual data downloading) which have to be taken into account. In this particular project, wireless sensor nodes were the best fit in terms of simplicity of installation and personnel requirements, but in a different situation, other options might be more suitable [3].

4.3.2 | Experiment for the Calibration of Temperature Sensors

As stated above, the temperature sensor used (SHT15) provides an accuracy of ±0.3°C according to the manufacturer. In order to get the best performance of the deployment, all the sensors were calibrated by comparison before being installed in the church, aimed at estimating their bias and improving the accuracy.

Basically, the set of nodes was located together inside a climate chamber of 23 m^3 that was driven by an air cooler in the ceiling (Küba Comfort DP model DPB034). The temperature was controlled inside the chamber during a period of three hours, increasing from 26°C up to 30°C. Sensors collected temperature at a rate higher than a sample per minute.

By computing the mean temperature recorded in the hot stage of the calibration experiment for each sensor, it was found that node M was the one closer to the overall sample mean. Hence, this sensor was regarded as a reference (i.e., with a null bias). Then, for each sensor, the bias was computed as the difference between the mean T recorded by that node, during the hot stage, and the mean T of this reference node (see Table 4.1).

An independent accurate sensor with a certified calibration would lead to a better estimation of the bias, but, unfortunatelly, such sensor was not available.

This approach is good enough for the purpose of the present study because the main goal is to analyze the relationship between temperature and the height of nodes, and knowing the real bias per node is of little interest to this paper. Bias values range from -0.28 to $+0.28$, which is consistent with the accuracy of ±0.3°C indicated by the sensor manufacturer.

Table 4.1: Temperature bias (°C) per node derived from the calibration experiment.

Node	**B**	**T**	**U**	**S**	**R**	**C**	**D**	**G**	**E**	**O**	**K**
Bias	-0.280	0.097	0.160	-0.003	0.069	-0.088	0.077	0.009	-0.019	-0.089	-0.036
Node	**N**	**L**	**M**	**I**	**J**	**Q**	**A**	**F**	**P**	**H**	
Bias	-0.046	-0.249	**0.000**	0.150	0.189	-0.277	-0.098	0.276	0.335	0.175	

Each value of temperature registered per node during the microclimate monitoring experiment was corrected by subtracting its corresponding bias.

The calibration "in situ" of sensors [212; 225] is an effective technique that consist of putting together all node sensors along with a calibrated sensor inside the building which is being monitored. Thus, it is possible to have a climatic condition reference from the calibrated sensor for comparing the records from all nodes. In this study, calibration "in situ" was not considered because for a massive campaign, the application of this technique requires a greater investment due to the cost of using a calibrated sensor and more time for its implementation. Furthermore, the experiment calibration approach of T used here, was possible given that the goal was to compute the differences of T from sensors, instead of estimating the mean of T.

4.3.3 | Installation of Wireless Nodes

After the calibration experiment, the 21 wireless sensor nodes were located in the church at different heights (h): 2, 2.5, 3.9, 4.0, 4.3, 5.0, 8.7, 12.1, and 13 m from the floor level

(Figure 4.4). The sensors located at each different height are the following:

- $h = 13.0$: nodes I and J were located at the upper part of the retable decorating the presbytery.
- $h = 12.1$: nodes A, F, P, and Q were placed at the upper position, close to the ceiling vaults.
- $h = 8.7$: nodes L and M were also located at the retable.
- $h = 5.0$: nodes G, H and O were placed near to the main altar.
- $h = 4.3$: it corresponds to node U, which was located near the main entrance.
- $h = 4.0$: nodes K and N were also installed at the retable.
- $h = 3.9$: node D was located close to the altarpiece of Saint Joseph.
- $h = 2.5$: nodes B and T were positioned near to the main entrance.
- $h = 2.0$: nodes C, E, R, and S were located as indicated in Figure 4.4 at the lowest level.

Some criteria for establishing the position of nodes were the following: (i) to spread out the sensors in different places of the church, (ii) to locate some nodes close to the main entrance and other openings allowing air exchange from outside, and (iii) to install at least 2 nodes at a comparable height for comparison purposes. Moreover, sensors were not placed too close to the floor level because they might be stolen or manipulated by churchgoers. The ideal scenario would have been to spread out the 21 nodes randomly inside the church. However, restrictions such as the building characteristics, the maximum number of nodes available, and the need to prevent problems caused by the movement of people, among other factors, made it impossible to achieve a random distribution of the nodes.

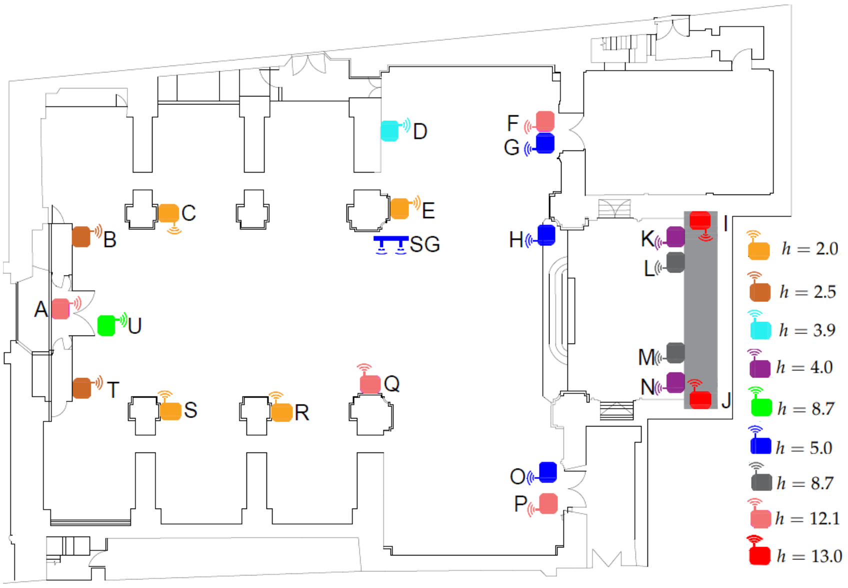


Figure 4.4: Position of the 21 wireless nodes located in the church of Saint Thomas and Saint Philip Neri (Valencia, Spain). Color refers to height (h) of node (in meters, m). SG indicates the position of the sink gateway that receives data wirelessly from the sensor nodes. The light gray rectangle indicates the position of the main altarpiece (retable).

4.3.4 | Data Pretreatment

The experiment of microclimate monitoring was carried out from the 1st of August 2017 until the 28th of February 2018 (7 months, 212 days). When programming the communication of sensor nodes with the sink gateway, the time between two consecutive measurements of T (t_j, t_{j+1}) was established as a random variable following an exponential distribution with a mean of one hour. Due to in fact the church has unheated/natural microclimate indoor, that better explains the selection of the sampling time of 1 hour that in case of a heated microclimate could not be sufficient. The main reason for using this type of distribution was to decrease the probability of data transmission collisions. However, as a drawback, it leads to missing values, which becomes a problem for the methodology of time series analysis applied here.

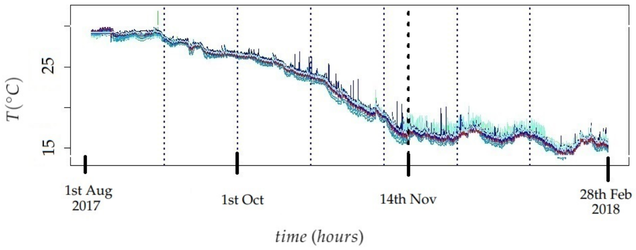
Regarding the missing data resulting from the exponential distribution used for establishing two consecutive measurements of T, it was checked that the percentage of missing values per node was approximately the same for the different nodes. By contrast, when the reason was having problems of wireless communication with the gateway or electrical failures, the percentage of missing values was greater. In particular, such amount was the highest for node R (41.4%). Taking into account that this node was at 12.5 *m* from the gateway, which is not too far away, the problem of wireless communication with the gateway was discarded as a reason for having such big amount of missing values, and the main cause could be a flaw in the electronics. The target was to have a common number of observations per sensor, particularly, one value per hour. For this purpose, all missing values were imputed. Taking into account that the distribution of missing data does not follow any specific pattern (i.e., missing at random [226]), all missing data were imputed using either Stineman interpolation [227] or linear interpolation. The latter was used when the time between two consecutive available measurements of T was less than 2 hours (i.e., a single missing value). For the rest of cases, the Stineman interpolation was used. The interpolation equations were solved for every unknown observation of T between two known values of T. The resulting data were organized as a matrix with 5088 rows (one per hour) by 21 columns (one per node). Finally, each value of temperature registered per node during the microclimate monitoring experiment was corrected by subtracting its corresponding bias. Similar studies have also applied interpolation procedures for the imputation of missing values. Klein et al. [206] estimated temperature and air moisture values using a smooth bivariate interpolant to the scattered sensor data, which is an effective method when the temperature is smoothly varying on short distances. However, this approach most likely loses accuracy near air inlets and outlets in galleries. Trying to overcome this drawback, the authors [206] also applied physics-based models incorporating Computational Fluid Dynamics by prescribing thermal boundary conditions.

4.3.5 | Statistical Methods

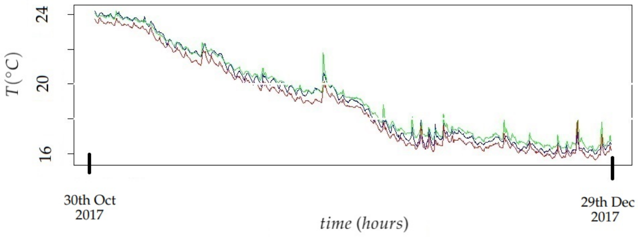
The methodology is comprised of five main steps. First, identification of stages in the time series of T (see Section 4.3.5.1). Second, estimation of the vertical gradient of T (see Section 4.3.5.2). Third, computation of parameters from the time series (e.g., sample mean values of Auto Correlation Function ACF, moving range MR, Partial Auto Correlation Function PACF at the first 4 lags, among others, and the additive seasonal Holt-Winters (SH-W) method (see Section 4.3.5.3). Fourth, analysis of the relationship between T and sensor height, using variables determined in the previous step and sparse Partial Least Squares (sPLS) (see Section 4.3.5.4). Finally, characterization of temperature at high, medium, and low altitude heights using Linear Discriminant Analysis (LDA) and the latent components from sPLS calculated in the previous step [108]. The R software (version 4.3) was used to carry out the statistical analyses. The main packages used were `mixOmics` [68], `klaR` [228], and `spls` [229].

4.3.5.1 | Identification of Stages in the Time Series

Regarding the monitoring experiment, two main stages were visually identified in the different time series of T: firstly, the average temperature slightly decreases until about November 14th and, next, it becomes approximately stationary (see Figure 4.5a). By visually inspecting the evolution over time of the time series of T, all of them are quite parallel (see an example in Figure 4.5b), which can be partly explained by the different position of each node and, moreover, by the bias of each sensor (Figures 4.5a and 4.5b show raw data, prior to the bias correction). The trajectory of M (i.e., the reference node) is depicted in red in Figures 4.5a and 4.5b.



(a)



(b)

Figure 4.5: **(a)** Trajectories of temperature for the 21 nodes, before subtracting their corresponding sensor bias. The thick vertical dotted line (November 14th) indicates a change of trend: T slowly decreases before this date on average while, next, the mean T is rather constant. Each thin vertical dotted line separates two consecutive stages (months) that were considered to split the different time series of T. In total, seven stages were considered. **(b)** Trajectories of temperature (before subtracting their corresponding sensor bias) for the nodes B (in blue), H (in green), and M (in red) corresponding to the period between October 30th 2017 to December 29th 2017.

The observed time series of temperature were denoted as $\boldsymbol{T}$, where $\boldsymbol{T} = (t_1, \ldots, t_j, \ldots, t_n)$.

By using the supF test [115; 120], two potential structural breaks were identified (i.e., changes in the slope of a linear trend), at observation number 763 (September 1st at 6:00 PM, $p - values < 0.02$) and 2797 (November 11th at 12:00 AM, $p - values < 0.01$).

The supF test was applied after calculating the logarithmic transformation and one regular differentiation to the distinct time series. Such logarithmic transformation was employed to stabilize the variance, and the regular differentiation was intended to eliminate the trend of the different time series [107]. The notation employed throughout this article is as follows: $\boldsymbol{r}$ indicates the logarithmic transformation of $\boldsymbol{T}$, and $\boldsymbol{W}$ refers to one regular differentiation of $\boldsymbol{r}$. Thus, each value of $\boldsymbol{W}$ corresponds to $w_j = r_j - r_{j-1}$, where $r_j = \ln(t_j)$. The two structural breaks identified lead to splitting the time series into three stages, but this number seems too low for the target of the present work. In order to extract more features from each time series, which presumably might lead to better results, it was decided to split all time series into seven stages, one per month (see Figure 4.5a). This criterion is consistent with the structural break identified on September 1st, though not with the one found on November 11th, but this issue was considered as a minor drawback.

4.3.5.2 | Estimation of the Vertical Gradient of Temperature for each Month

With the goal of determining if the vertical gradient is apparent, the Pearson correlation test [218] was applied to different periods of the time series of T (i.e., each month as established in Section 4.3.5.1). By using the test, it is possible to determine whether the correlation between temperature and height of sensors is statistically significant. Once a period with statistically significant correlation was identified, the slope of the linear relationship was considered as the gradient estimation. Such slope is the derivative of the function that estimates the mean temperature in the month with respect to height.

Consider that the relationship between temperature, $\boldsymbol{T} = (t_1, \ldots, t_n)$, and height, $\boldsymbol{h} = (h_1, \ldots, h_n)$, is determined by the following linear regression model in Equation 4.1, where, $\boldsymbol{\varepsilon} = \boldsymbol{T} - E(\boldsymbol{T}|h)$, $E(\boldsymbol{T}|h)$ is the conditional expectation of $\boldsymbol{T}$ given $\boldsymbol{h}$, $E(\boldsymbol{\varepsilon})$ is the expectation of the errors $\boldsymbol{\varepsilon}$, and $V(\boldsymbol{\varepsilon})$ is the variance of $\boldsymbol{\varepsilon}$ [219]. Details about computing the estimated values and confidence intervals of β_0 and β_1 can be found in [219].

$$\begin{aligned} t_i &= \beta_0 + \beta_1 h_i + \varepsilon_i, \text{where} \\ i &= 1, \ldots, n \\ E(\boldsymbol{\varepsilon}) &= 0 \\ V(\boldsymbol{\varepsilon}) &= \sigma^2 \end{aligned} \tag{4.1}$$

Considering the linear equation that estimates temperatures as a function of height, the derivative of this function is the slope β_1 of the regression line, which is the gradient estimation. The thermal vertical gradient can be interpreted as the rate of increase of T according to height.

In this study, for each month, the gradient was estimated as the slope of the linear regression model by using height (h) as predictor variable and mean temperature as response. This gradient was expressed as $^{\circ}C/m$.

The existence of a gradient implies that the correlation between T and h is statistically significant, which was checked for each month. If this condition is not fulfilled, there is not enough evidence to affirm that the slope of the regression line is different from zero at the population level. Hence, there is no evidence for a vertical thermal gradient. The proposed method for the calculation of vertical gradients seems reasonable when all sensors are located one above the other, in the same vertical axis, but this is not the case here. However, a preliminary analysis suggested that longitudinal thermal gradients were not relevant in this case, because the ventilation rate of the building is rather limited and because indoor air conditions are not affected by heating or air conditioning systems, which are not installed in this church.

4.3.5.3 | Calculation of Classification Variables

Two methods were used to compute features from the time series, which were applied to the different observed time series ($\boldsymbol{T}$ or $\boldsymbol{W}$) separately per month. As an exception, each complete time series was also used in the second method, in addition to modelling each month independently. Features will be denoted hereafter as classification variables.

1. Method 1: Using Time Series Functions

   This method consists of computing features from the observed time series $\boldsymbol{T}$, in some cases, and from the time series after applying the logarithm transformation and regular differencing to $\boldsymbol{T}$. The goal of using this transformation and differencing was to stabilize the variance and remove the trend of the series in order to extract information about the seasonal component. Features were calculated by means of values of sample Auto Correlation Function (ACF), sample Partial Auto Correlation Function (PACF), periodogram, Moving Range (MR) [18; 104], as well as features defined using quantiles [220]. Each variable was computed for each month and sensor. These correspond to estimates of the following parameters:

a) `mean.ts`: Mean of $\boldsymbol{T}$ recorded in the month. This parameter allows to compare the level of the different time series.

b) `sd.ts`: Standard deviation of $\boldsymbol{T}$, which provides information about the variability of the recorded values.

c) `range.ts`: Range of $\boldsymbol{T}$ (i.e., by subtracting the minimum to the maximum). It reflects the amplitude of the time series of T and gives information about the dispersion.

d) `mean.mr`: Mean of MR values with order 24 of $\boldsymbol{T}$. MR computes the moving range for all sequences of 24 consecutive observations.

e) `median.mr`: Median of MR values with order 24 of $\boldsymbol{T}$. This parameter and the previous one are helpful for capturing the daily variability of the different time series of T.

f) `mean.acf`: Mean of the first 72 lags ($l = 1, \ldots, 72$) of sample ACF applied to $\boldsymbol{W}$ time series. Each value of ACF for $\boldsymbol{W}$ at lag l (acf_l) is the correlation coefficient between the observations that are lagged for a time gap l. It is given by $acf_l = \text{cor}(w_j, w_{j-l})$, i.e., Pearson's correlation coefficient between the time series and the lagged values (i.e., the time gap which is considered). The value 72 was used because sample ACF values computed for $l = 1, 2, \ldots, 72$ were comprehended within the limits of a 95% confidence interval in the correlogram. This parameter provides information about the dynamic structure of the time series.

g) `median.acf`: Median of the first 72 lags of sample ACF applied to $\boldsymbol{W}$. As in the previous case, this parameter can be useful for comparing the dynamic structure of the time series.

h) `sd.acf`: Standard deviation of the first 72 lags of sample ACF of $\boldsymbol{W}$.

i) `pacf`: First 4 lags ($l = 1, \ldots, 4$) of sample PACF applied to $\boldsymbol{T}$. A value of PACF at lag l measures the autocorrelation between the observation t_j and t_{j-l}, which is not accounted for by lags 1 to $l - 1$. The first four values of PACF are usually the most important ones for capturing the most significant autocorrelation information. These four values were computed trying to differentiate the dynamic structure of the different time series.

j) `maximum.I`: Maximum value from the periodogram (I), which is employed for identifying the dominant periods or frequencies of time series of T. This parameter is helpful for recognizing the dominant cyclical behavior in a series.

k) `range.I`: Range of values of the periodogram. This parameter can be useful to compare the impact of the dominant cyclical pattern in the different series.

l) `maximum.slps`: Maximum increase of T in one hour found in the month (i.e., $max(t_{j+1} - t_j)$). This parameter allows the comparison of the maximum changes of $\boldsymbol{T}$ for two consecutive hours, and it is intended to capture the information of abnormal peaks or sudden increases due to occasional events.

m) `median.abs.sd`: Median of absolute values of the deviation between the values of $\boldsymbol{T}$ and the median of $\boldsymbol{T}$. It is given by $median(|\boldsymbol{T} - median(\boldsymbol{T})|)$. This parameter is somewhat related to the variance (i.e., average of the squared deviations with respect to the mean) and, hence, it is another measurement of data dispersion.

n) `t.p.r.m20`: it is computed as $(\boldsymbol{T}_{60} - \boldsymbol{T}_{40})/(\boldsymbol{T}_{95} - \boldsymbol{T}_5)$, being $\boldsymbol{T}_a$ the percentile a of values in the month. Thus, it is the ratio of percentiles (60th–40th) over (95th–5th) of $\boldsymbol{T}$. The numerator is the range of variability corresponding to 20% of the central part of the original time series. The denominator is basically the range of the original time series after removing the lowest 5% and highest 5% . An equivalent interpretation corresponds to the parameters `t.p.r.m35`, `t.p.r.m50`, and `t.p.r.m80` described next.

o) `t.p.r.m35`: it is computed as $(\boldsymbol{T}_{67.5} - \boldsymbol{T}_{32.5})/(\boldsymbol{T}_{95} - \boldsymbol{T}_5)$, which is the ratio of percentiles (67.5th–32.5th) over (95th–5th) of $\boldsymbol{T}$.

p) `t.p.r.m50`: Ratio of percentiles (75th–25th) over (95th–5th) of $\boldsymbol{T}$.

q) `t.p.r.m80`: Ratio of percentiles (90th–10th) over (95th–5th) of $\boldsymbol{T}$.

r) `p.d.f.p`: Ratio of percentiles (95th–5th) over the median of $\boldsymbol{T}$. This parameter divides the amplitude (range) of the time series, after removing the lowest 5% and highest 5% of observations, by the median of $\boldsymbol{T}$.

This list comprises a set of 21 variables that were computed for each one of the seven months, which implies 147 variables in total. They were arranged in a matrix denoted as $\mathbf{X}_1$ comprised of 21 rows (one per node) and 147 columns (one per variable).

2. Method 2: Additive Seasonal Holt-Winters Method (SH-W)

This approach calculates features from time series of T, by using the Holt-Winters method (SH-W) [20], which is an extension of the Holt's method [19]. It captures the level, trend, and seasonality of the different time series and is comprised of the forecast equation and three smoothing equations (i.e., one for the level a_i, one for the trend or slope b_i, and one for the seasonal component s_i) with corresponding smoothing parameters α, β, and γ [21]. According to the additive SH-W, the forecast equation for a time series of T with period length p is given by Equation 4.2 (in this study, p is 24), where k is the integer part of $(l-1)/p$, and $\hat{t}_{i+l|i}$ is the forecast at step $(i+l)$ [21].

$$\begin{aligned}\hat{t}_{i+l|i} &= a_i + lb_i + s_{i+l-p(k+1)}, \text{where} \\ a_i &= \alpha(t_i - s_{i-p}) + (1-\alpha)(a_{i-1}+b_{i-1}) \\ b_i &= \beta(a_i - a_{i-1}) + (1-\beta)b_{i-1} \\ s_i &= \gamma(t_i - a_{i-1} - b_{i-1}) + (1-\gamma)s_{i-p}, \\ \text{where} \quad & 0 \leq \alpha \leq 1, \quad 0 \leq \beta \leq 1, \quad 0 \leq \gamma \leq 1, \text{and} \quad i > s\end{aligned} \tag{4.2}$$

Slope, level and seasonal components at step i are estimated by using the three smoothing equations (i.e., for b_i, a_i, and s_i), respectively. If the algorithm converges, a, b and s_1 to s_p are the estimations for the level, trend or slope and seasonal components. This algorithm was run by using the function `HoltWinters` of the `stats` package [171] of `R` software.

The flow diagram for additive SH-W method is displayed in Figure 4.6. In this diagram, all the steps are repeated with each observation of time series t_i, $i: 1, \ldots, n$. However, in step (1), the initial values of level (a_0), trend, (b_0) and seasonal coefficients (s_0) are only used once to start up the algorithm. The initial conditions are estimated through a simple decomposition in trend and seasonal component by using moving averages. After initialization, steps from (2) to (4) perform the forecast task internally, these values were updated and stored for the next step [171]. In step (2), the estimation of slope requires knowledge of the level at steps i, $(i-1)$, and so on until a_0, as well as slope at steps $i-1$, and so on until b_0. In step (3), as in step (2), the equation is solved recursively. Estimation of level requires knowledge of the level, slope, seasonal components at different steps starting at $i-1$ (for a_i), $i-1$ (for b_i), $i-p$ (for s_i), and finishing when the values are a_0, b_0, and s_0. It also requires values of T at steps i, and so on until t_0, where t_0 is just the oldest data point in the training data set (i.e., a set of observations starting from t_1 until the current observation t_i). Note that the weighting coefficients α, β and γ need to be computed for running steps (2), (3) and (4). Such coefficients are calculated by minimizing the squared one-step prediction error [171]. Now that the level, trend and seasonal component at time step i have been estimated, the forecast $\hat{t}_{(i+l)}$ at step $(i+l)$ with $l = 1, \ldots, 24$ can be estimated by using the three values of components together.

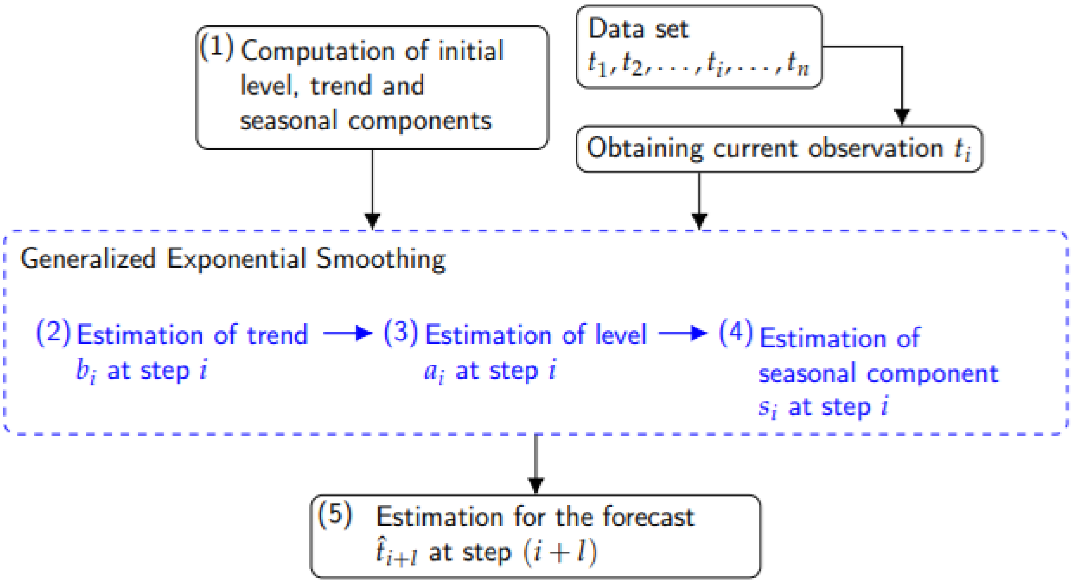

(1) Computation of initial level, trend and seasonal components
Data set $t_1, t_2, \ldots, t_i, \ldots, t_n$
Obtaining current observation t_i
Generalized Exponential Smoothing
(2) Estimation of trend b_i at step i
(3) Estimation of level a_i at step i
(4) Estimation of seasonal component s_i at step i
(5) Estimation for the forecast $\hat{t}_{i+l}$ at step $(i+l)$

Figure 4.6: The flow diagram displays five steps for carrying out the additive SH-W method. Step (1) indicates that the initial conditions for the components are computed. Steps (2), (3), and (4), indicate that the slope, level and seasonal component at step i are estimated. Finally, step (5) indicates that forecasts $(\hat{t}_{i+l})$ at step $(i+l)$ are calculated, where $l : 1, \ldots, 24$.

According to this method, the level, trend, and seasonal components are updated over a historical period. For example, when the method is applied per month, the components are updated every hour over each month. If the algorithm converges, a, b and s_1 to s_{24} are the estimated values for the level, trend and seasonal components at the last instant of time in the month.

The level at a time t corresponds to a weighted average between the seasonally adjusted temperature and the level forecast, based on the level and slope at the previous instance of time $t-1$. This component gives an estimate of the local mean (i.e., mean per hour in this study). Regarding the slope component, it expresses the linear increment of the level, over an hour. Finally, the seasonality component estimates the deviation from the local mean, due to seasonality.

The features calculated per sensor are the following:

a) `a`: Estimated value for the level for each month of the time series.

b) `b`: Estimated value for the trend (slope) for each month.

c) `s1,s2,...,s24`: Estimated values for the seasonal components for each month.

d) `sse`: Sum of squared estimate of errors per month.

e) `maximum.I`: Maximum value of the periodogram computed with the residuals of SH-W for each month.

f) `mean.acf`: Mean of sample ACF of residuals at lags 1 to 72 per month.

g) `median.acf`: Median of sample ACF of residuals at lags 1 to 72 for each month.

h) `range.acf`: Range of sample ACF of residuals at lags 1 to 72 per month.

i) `Dn`: Statistic of the Kolgomorov-Smirnov (*KS*) normality test [127] of the residuals derived from SH-W, per month of the time series. The *KS* normality test was employed to compare the empirical distribution function of the residuals with the cumulative distribution function of the normal model.

j) `Wn`: Statistic of the Shapiro-Wilk test (*SW*) [122] of the residuals per month. This test was used to detect deviations from normality, because of either kurtosis or skewness, or both. The `Dn` and `Wn` statistics were also used as classification variables, because they provide information about deviation from normality for the residuals derived from the SH-W method.

k) `fcast`: 24 forecasts of T (i.e., $\hat{t}_{i+l|i}$, $l = 1, \ldots, 24$) for a unique additive SH-W model that was fitted using the complete time series without splitting it in different months.

Features calculated from (a) to (j) imply a set of 33 variables computed for each month. By including the 24 forecasts as explained in (k), the total number of variables was $33 \times 7 + 24 = 255$, which were organized as a matrix denoted as $\mathbf{X_2}$, comprised of 21 rows (one per sensor) and 255 columns (one per variable).

For both data sets, $\mathbf{X}_1$ and $\mathbf{X}_2$, those variables with a strongly skewed distribution were transformed with the goal of finding a simple transformation leading to normal distribution. For this purpose, standard (simple) Box-Cox transformations [173] were applied to those variables with a Fisher–Pearson standardized moment coefficient of skewness [174], or with a Fisher coefficient of kurtosis [174] outside the intervals of -2.0 to 2.0. For those variables with a negative skewness, absolute values were used instead of their original ones for applying a Box-Cox transformation. The skewness statistic evaluates the asymmetry of the probability distribution. The kurtosis statistic indicates which variables were heavy-tailed or light-tailed, relative to a normal distribution. Furthermore, the estimates of kurtosis were useful measures for identifying outliers in the different variables.

The percentage of outliers in both data sets was 0.73% in $\mathbf{X}_1$ and 0.65% in $\mathbf{X}_2$, which is a small amount. Outliers were discarded, and the resulting missing values were imputed using Non Linear Estimation by Iterative Partial Least Squares (NIPALS, [51; 65])). Given the low percentage of missing values, their estimation is assumed to be appropriate [176]. Next, once the values were imputed, each column of $\mathbf{X}_1$ and $\mathbf{X}_2$ was centered by subtracting its column mean. Also, it was scaled to unitary variance by dividing over its standard deviation.

As both data sets contain more than 100 variables and just 21 rows, a high degree of multicollinearity is expected a priori, which would lead to severely ill-conditioned problems. Furthermore, from a practical point of view, for these high-dimensional data sets, results might be difficult to interpret given the large number of variables. One solution is to extract latent variables that summarize the information using a subset of variables. In this context, many sparse versions [52; 53; 61; 64; 230] have been proposed for feature selection purposes. These versions work properly in regression by introducing penalties in the model such as Lasso [182] and Ridge [231].

4.3.5.4 | sPLS

Since Partial Least Squares (PLS) regression was introduced by Wold [51], it has been employed as an alternative approach to Ordinary Least Squares (OLS) regression in ill-conditioned linear regression models that emerge in many disciplines, such as biology, chemistry and economics [52]. PLS is a dimension reduction technique that relates a regressor matrix $\mathbf{X}$ and a response matrix $\mathbf{Y}$ by computing latent components that correspond to linear combinations of the original variables (predictors). PLS maximizes the covariance between components from two data sets. PLS is computationally fast and the projection of observations on a low-dimensional space allows a graphical representation of observations and variables. Due to these reasons, this method has gained a lot of attention in high-dimensional classification problems [53].

In this study, the data sets $\mathbf{X}_1$ and $\mathbf{X}_2$ were analyzed using sPLS with a regression model in an attempt to identify the main variables correlated with sensor height, which will explain the differences in the time series of T according to the distance to the floor level. The information used by sPLS was the following: the response vector, $\mathbf{Y} \in \mathbb{R}^{n \times 1}$, containing the height of each sensor ($n = 21$), and the regressor matrix, $\mathbf{X} \in \mathbb{R}^{n \times p}$ ($\mathbf{X}_1$ or $\mathbf{X}_2$), which contains the classification variables computed in Section 4.3.5.3.

sPLS modeled $\mathbf{X}$ and $\mathbf{Y}$ as a linear regression, where $\mathbf{X} = \Xi\mathbf{C} + \mathbf{E}_1$ and $\mathbf{Y} = \Xi\mathbf{D} + \mathbf{E}_2 = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}_2$, where $\boldsymbol{\beta} \in \mathbb{R}^{n \times p}$ is the matrix of regression coefficients, $\mathbf{E}_1 \in \mathbb{R}^{n \times p}$ and $\mathbf{E}_2 \in \mathbb{R}^{n \times 1}$ are random errors, $\Xi = (\boldsymbol{\xi_1}, \ldots, \boldsymbol{\xi_H}) \in \mathbb{R}^{n \times H}$ is the matrix of latent component, where $\Xi = \mathbf{XU}$, with $\mathbf{U} \in \mathbb{R}^{p \times H}$ as H direction vectors, with $1 \leq H \leq min\{n, p\}$ and $\mathbf{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_H)$. Furthermore, $(\boldsymbol{u}_h, \boldsymbol{v}_h)$ is the solution of the optimization problem according to Equation 4.3 for $j = 1, \ldots, h-1$, subject to $\|\boldsymbol{u}\|_2 = 1$.

$$\min_{\boldsymbol{u},\boldsymbol{v}} \{ \|\mathbf{M} - \boldsymbol{u}\boldsymbol{v}^\top\|_F^2 + P_{\lambda_1}(\boldsymbol{u}) \} \tag{4.3}$$

The optimization problem minimizes the Frobenius norm $\|\mathbf{M} - \boldsymbol{u}\boldsymbol{v}^\top\|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{p} (m_{ij} - u_i v_j)^2$, where $\mathbf{M} = \mathbf{X}^\top\mathbf{Y}$, $\boldsymbol{u}$ and $\boldsymbol{v}$ are the loading vectors, and $\mathbf{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_H)$. Furthermore, $P_{\lambda_1}(\boldsymbol{u})$ is the Lasso penalty function, where $P_{\lambda_1}(\boldsymbol{u}) = \lambda_1 \|\boldsymbol{u}\|_1$ [63; 64].

This optimization problem is solved based on the PLS algorithm [65] and Singular Value Decomposition (SVD) [66] of a matrix $\tilde{\mathbf{M}}_h$ per dimension h. The SVD decomposition of matrix $\tilde{\mathbf{M}}_h$ is subsequently deflated per iteration h. This matrix is computed as $\mathbf{U}\Delta\mathbf{V}^\top$, where $\mathbf{U}$ and $\mathbf{V}$ are orthonormal matrices, and Δ is a diagonal matrix whose diagonal elements are called the singular values. During the deflation step of PLS, $\mathbf{M}_h \neq \mathbf{X}_h^\top \mathbf{Y}_h$, given that $\mathbf{X}_h$ and $\mathbf{Y}_h$ are computed separately, and the new matrix is called $\tilde{\mathbf{M}}_h$. At each step, a new matrix $\tilde{\mathbf{M}}_h = \mathbf{X}_h^\top \mathbf{Y}_h$ is calculated and decomposed by SVD. Furthermore, in sPLS algorithm, the soft-thresholding function $g(\boldsymbol{u}) = (|\boldsymbol{u}| - \lambda)_+ sign(\boldsymbol{u})$, with $(x)_+ = max(0, x)$, was used in penalizing loading vectors $\boldsymbol{u}$ to perform variable selection in the regressor matrix; thus, $\boldsymbol{u}_{new} = g_\lambda(\tilde{\mathbf{M}}_{h-1}\boldsymbol{v}_{old})$ [61].

The `mixOmics` package [67] offers different functions for carrying out multivariate analysis of data sets, with a specific focus on data exploration, dimension reduction and visualisation [68]. Among the different functions, it proposes some in order to carry out sPLS. Also, it implements Leave-One-Out cross-validation (LOO-CV) to compare the performance of diverse models with different Lasso penalties. Furthermore, in order to perform variable selection, it employs an algorithm that uses the soft-thresholding function $g(\boldsymbol{u})$, according to Equation 4.4. By controlling η instead of the direction vector specific sparsity parameters λ, the method evades combinatorial tuning of the set of sparsity parameters and supplies a bounded range for the sparsity parameter [53].

$$\begin{aligned} g(\boldsymbol{u}) &= (|\boldsymbol{u}| - \eta max_{1 \leq j \leq p}|\ u_j\ |)_{+} sign(\boldsymbol{u}), \text{where} \\ 0 &\leq \eta \leq 1 \\ (x)_{+} &= max(0, x) \end{aligned} \tag{4.4}$$

The algorithm implemented in the `mixOmics` package uses the number of variables denoted as `keepX` for running PLS, instead of the parameter η, while it employs η close to 1. The `keepX` argument in the package functions is employed in order to evaluate different subsets of variables on each latent component and determine the best number of variables that optimizes the objective function of PLS.

The `perf` function was used to determine the optimal number of components. The performance of sPLS was evaluated for 10 components using LOO-CV. The optimal number of components was determined by identifying when the further decrease in Root Mean Square Error of Prediction $RMSEP$ is relatively insignificant [232]. $RMSEP$ is defined in Equation 4.5, where $PRESS_h = \sum_{i=1}^{n} (y_i - \widehat{y}_{h(-i)})^2$, with $\widehat{y}_{h(-i)}$ is the model prediction with 1 to h components across all but the $i - th$ observation.

$$RMSEP_h = \sqrt{\frac{PRESS_h}{n}} \tag{4.5}$$

The main criterion for selecting the optimal number of components was $RMSEP$, while the second one was the goodness-of-fit R^2 ($0 \leq R^2 \leq 1$). The latter is inflationary and rapidly approaches 1 as the number of model parameters increases. Therefore, it is not sufficient to only have a high R^2.

In order to determine the optimal number of variables to select on each component, a grid (`keepX`) of the non-zero elements of the loading vector was assessed on each component, one at a time. The values of three different grids were carefully chosen to achieve a trade-off between resolution and computational time. Firstly, two coarse tuning grids were evaluated before establishing a finer grid. The penalization parameter was chosen by computing the error prediction ($RMSEP$) with LOO-CV, per component. The `tune.spls` function was used to determine the optimal number of variables per component. Once the optimal number of components and variables were determined, the final sPLS method was run.

Variable Importance in Projection VIP_j [187] was used for computing the overall importance of each predictor variable on the response, cumulatively over the total components. This measure was computed using the loading vectors and the sum of squares per component. Variables with $VIP_j > 1$ are the most important ones in the regression model.

Although PLS was not originally designed for classification, it has been employed for that objective, with effective performance [53]. In respect to the ajustment of PLS to classification for high-dimensional data, some approaches have been studied, e.g., SPLS Discriminant Analysis (SPLSDA), Sparse Generalized PLS (SGPLS) [52], and sPLS-Discriminant Analysis (sPLS-DA)[63]. Regarding SPLSDA, different variants have been proposed: SPLSDA-LDA (i.e., with linear discriminant analysis) and SPLSDA-LOG, (i.e., with Logistic Regression). These methods aim to improve the PLS classification a-pproaches by using dimension reduction and variable selection simultaneously. In fact, sPLS-DA has been used in order to classify time series in the context of art conservation [161; 162].

Likewise, this study proposed a statistical methodology based on SPLSDA [52] for classifying different time series of T in the context of preventive conservation of cultural heritage. SPLSDA computes latent components using sparse partial least squares (SPLS) regression [53]. SPLS selects predictors while reducing dimension. Next, a classifier is fitted, either Logistic Regression (LOG) or Linear Discriminant Analysis (LDA) [53]. Chung and Keles [53] suggest using a linear classifier because it might be better from an interpretation point of view. The methodology proposed here consists of using sPLS [61] instead SPLS [53]. Once the latent components are computed, LDA is used subsequently.

When examining time series for art conservation, they are generally very similar in distinct positions or height levels of the same building. In this area of research, it is of interest to develop statistical methodologies that can improve the classification of time series with easy interpretation. Such classification can be useful for characterizing and monitoring microclimatic conditions in different zones and heights in a museum, archaeological site or heritage building, with the goal of avoiding problems such as moisture and dust deposition on walls and artworks.

4.3.5.5 | Linear Discriminant Analysis (LDA)

LDA is a supervised method for the discrimination of qualitative variables in which two or more clusters are known a priori and new observations can be classified into one of them, according to their characteristics [231]. In this study, for separating three clusters ($K = 3$) of sensors according to height, LDA was run by using the matrix $\mathbf{X} \in \mathcal{R}^{n \times d}$, whose elements correspond to values of the d components for n sensors. The components ($d = 2$) were computed from sPLS (either method 1 or method 2). The clusters that were defined according to the heights (h), are the following: 1 ($2.0 \leq h \leq 4.3$), 2 ($4.3 < h \leq 8.7$), and 3 ($8.7 < h \leq 13$). The number of nodes per cluster were 10, 5 and 6, respectively. Clusters 2 and 3 comprise of a vertical difference of 4.4 and 4.3 m respectively, but this value is about half (2.3 m) in cluster 1. This is not an ideal situation, but this criterion was adopted in order to have a similar number of nodes per cluster.

LDA predicts the cluster most appropriate for each of sensor by using Bayes' theorem, which helps computing the posterior probability $P(y = k|x)$, for each cluster k, $k = 1, 2, 3$. Suppose that a predictor $\boldsymbol{x} \in \mathcal{R}^d$ and that the class conditional distribution $P(x|y = k)$ is modeled as a multivariate Gaussian distribution (with mean $\boldsymbol{\mu}_k \in \mathcal{R}^d$ and variance matrix $\Sigma_k \in \mathcal{R}^{d\times d}$), where all clusters have the same covariance matrix Σ. Then, the log posterior ($\delta_k(\boldsymbol{x})$) is given by Equation 4.6, where D is the Mahalanobis distance between the data $\boldsymbol{x}$ and the mean $\boldsymbol{\mu}_k$. LDA classifies a sensor in the cluster k, if the cluster maximizes the log posterior probability $\delta_k(\boldsymbol{x})$ [231]. Thus, this method classifies a sensor, by accounting for the cluster prior probabilities $P(y = k)$, and the cluster whose mean is the closest to the data $\boldsymbol{x}$, according to Mahalanobis distance (D) [231].

$$\begin{aligned} \delta_k(\boldsymbol{x}) &= \log P(y = k|\boldsymbol{x}) \\ \delta_k(\boldsymbol{x}) &= -\frac{1}{2}D + \log P(y = k) + constant, \text{where} \\ D &= (\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k) \end{aligned} \tag{4.6}$$

Equation 4.6 can be written as indicated in Equation 4.7, which implies that this method has a linear decision surface [231].

$$\begin{gathered} \delta_k(\boldsymbol{x}) = \log P(y = k|x) = \omega_{k0} + \boldsymbol{\omega}_k^\top \boldsymbol{x} + constant, \text{where} \\ \boldsymbol{\omega}_k = \Sigma^{-1}\boldsymbol{\mu}_k; \quad \omega_{k0} = -\frac{1}{2}\boldsymbol{\mu}_k^\top \Sigma^{-1}\boldsymbol{\mu}_k + \log P(y = k) \end{gathered} \tag{4.7}$$

Figure 4.7 illustrates the boundary of decision, $D(\boldsymbol{x}) = \delta_{k=0}(\boldsymbol{x}) - \delta_{k=1}(\boldsymbol{x}) = 0$, for classifying one observation (blue point) from two clusters ($K = 2$). The first cluster has $\hat{\mu}_0$ as the estimation of the mean, and the second one has $\hat{\mu}_1$ as the mean. The blue point was classified in cluster 0 because $D(x) > 0$.

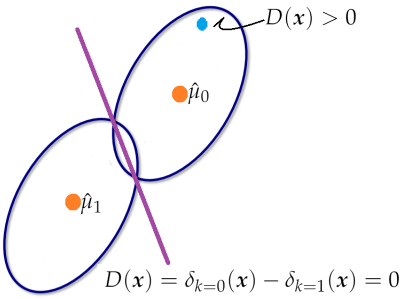


Figure 4.7: The picture displays two clusters (cluster 0 and cluster 1). If $D(x)$ is greater than 0, the blue point is classified as cluster 0 and otherwise as cluster 1. The purple line corresponds to the boundary of decision, $D(\boldsymbol{x}) = \delta_{k=0}(\boldsymbol{x}) - \delta_{k=1}(\boldsymbol{x}) = 0$.

LDA can be carried out by first transforming the data in order to have an identity covariance matrix. Next, LDA assigns $\boldsymbol{x}$ to a cluster k, taking into account prior probabilities of the cluster and the cluster whose mean is the closest to the observation, according to Euclidean distance [231]. Calculating Euclidean distances in d-dimensional space ($\boldsymbol{\mu}_k \in \mathcal{R}^d$) is equivalent to first projecting the data points into an affine subspace of the dimension at a maximum of $K - 1$ [231]. Thus, in this case, LDA determines linear combinations of the components from sPLS for predicting the clusters for the different sensors. This method was run by using the function `train` (with `method="lda"`) of the `caret` package [233], and `partimat` of the `klaR` package [228] of R software.

In this study, the assumption that each component has a normal distribution for each cluster was verified, as well as whether the variance of the components was the same in all clusters. When the normal condition is not fulfilled, LDA loses accuracy but can still reach a relatively good performance [18]. Results from the methodology proposed (sPLS with LDA) were compared with the results from SPLSDA and sPLS-DA. The classification error rates and number of selected variables from each method were compared. SPLSDA was run by using the function `cv.spls` of the `spls` package [229] and sPLS-DA method was run by using the functions `perf` and `tune.splsda` of the `mixOmics` package [68].

4.4 | Results and Discussion

The values of T inside the church are influenced by the climatic conditions outside. Figure 4.8 displays the trajectories of T (days) in the period from August 1st 2017 to February 28th 2018, outside and inside the church of Saint Thomas and Saint Philip Neri. The trajectories of T inside the building correspond to the 21 node sensors employed in this study, while the trajectories of T outside correspond to the minimum and maximum daily temperatures. The trajectories show a similar tendency, as the temperature decreases until November, and T becomes stable after that day. The variability of T, from sensors inside the building is obviously less pronounced than the variability of T outside the church. The values of T inside the church are more influenced by the maximum temperature outside. If the maximum daily temperature is smoothed, it can be observed that the values are quite similar throughout the year to those registered inside the church. This fact is striking, since it would be expected that the temperature inside the temple would be intermediate between the maximum and minimum values of outside air conditions. The main hypothesis is that the maximum outdoor temperature is measured in the shade and under standardized conditions. However, the solar radiation incident on the roof of the church reaches a temperature much higher than that of the surrounding air, which occurs throughout the year because the weather in Valencia is very sunny. This heat is transmitted inside the temple, and would affect the air temperature in the church. A detailed study of heat transmission would be necessary to better study this issue, but it is out of the scope of the present work.

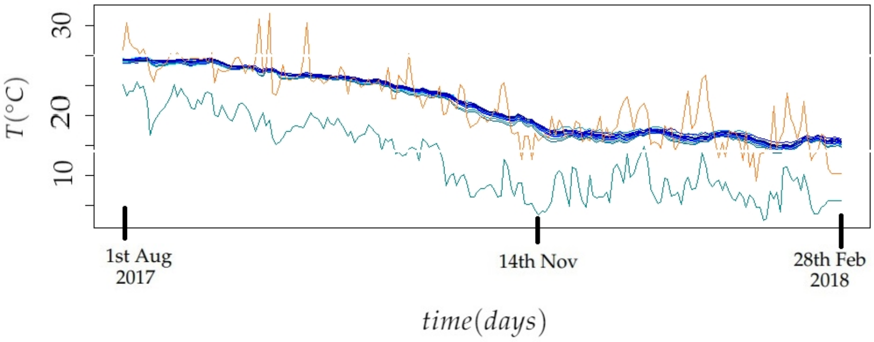


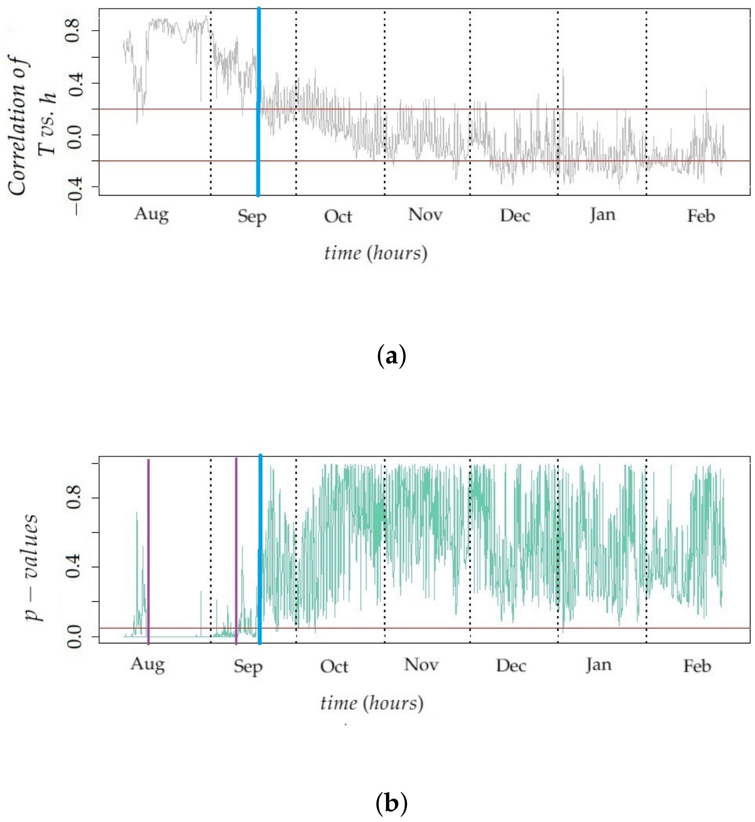
Figure 4.8: Trajectories of daily-mean temperature over time (days) in the period from August 1st 2017 to February 28th 2018. The green and brown trajectories correspond to the minimum and maximum daily temperature, respectively, in the city of Valencia, Spain. The blue trajectories correspond to temperatures recorded by the 21 sensor nodes inside the church of Saint Thomas and Saint Philip Neri.

4.4.1 | Vertical Gradients of Temperature

The vertical gradient was estimated for each month by fitting a linear regression model using height and hourly mean temperature as the predictor and response variables, respectively (see Section 4.3.5.2). The target was to identify in which month the correlation between both variables was statistically significant. It was found that vertical thermal gradients in the church of Saint Thomas and Saint Philip Neri change along the year.

Figure 4.9a shows that the correlation between height and hourly values of T is around $r = 0.8$ in August, but it decreases afterwards, reaching a null value in October, and the correlation becomes slightly negative in winter. Taking into account that July and August are the hottest months of the year in Valencia, this observed correlation suggests that, in summer, the temperature at the upper part of the central nave is higher than at lower levels. The reason could be the hot temperatures reached during the day in summer in the Mediterranean region [10]. By contrast, in winter, the correlation tends to be slightly negative. However, such correlation is not statistically significant, as described below, which implies that there is not enough evidence to affirm that temperatures in the lower positions tend to be higher in winter. Results reveal that vertical gradients of T are not stable throughout the year, and summer is the only period when vertical airflows might be involved in phenomena of dust deposition in walls. For August and September, most values of the correlation coefficient between sensor level and temperature were greater than 0.20 (upper red line in Figure 4.9a). In fact, the maximum value of $r = 0.80$ was found for August. For the period from August 10th at 8:00 AM to September 9th at 11:00 PM, most $p - values$ were less than 0.05 (red line in Figure 4.9b), which implies that the correlation between height and monthly mean T is statistically significant. Thus, it is possible to establish a linear relationship between them. By contrast, since September 17th, most $p - values$ were greater than 0.05 (see Figure 4.9b). As a consequence, August and September are the most relevant months for explaining the relationship between sensor levels and temperature. In particular, the period from August 10th at 8:00 AM to September 9th at 11:00 PM.

For the period from August 10th at 8:00 AM to September 9th at 11:00 PM, the difference between the mean temperature for the maximum sensor level (13 m) and the minimum sensor level (2 m) was 0.39$°C$. Also, estimations of the intercept and slope with their confidence intervals at 95% in the linear regression model using height (predictor variable) and mean of temperatures (response) were 28.31 (28.207, 28.35) and 0.043 (0.030, 0.057) respectively. The coefficient of determination is $R^2 = 70.55\%$. In the period mentioned, at the floor level (height = 0), the estimated mean of T is 28.31$°C$. Also, if the height increases by 1 meter, the mean of T will increase on average approximately 0.043$°C/m$, which implies 0.43$°C$ per 10 m. This result is consistent with the difference previously calculated. In a vertical difference of 11 meters, the model estimates 0.47$°C$ as the thermal difference. This linear increase can be seen in Figure 4.10. Given that the gradient corresponds to the slope of the linear regression fitted to the data (mean temperature per node vs. height), it is possible to compare the results from this study with other works reporting differences of temperatures at different height levels.



(**a**)

(**b**)

Figure 4.9: (**a**) Evolution of the correlation coefficient (r) between sensor height and temperature, over time (hour). Horizontal red lines correspond to values of $-0.20 < r < 0.20$. Dashed vertical lines account for the different months. (**b**) $p - value$ from the correlation test over time (hour). The red horizontal line corresponds to $p - value = 0.05$; for lower values the correlation was regarded as statistically significant. Purple vertical lines correspond to August 10th at 8:00 AM and September 9th at 11:00 PM. The blue line indicates September 17th at 9:00 PM.

The vertical thermal gradient quantified here is consistent with a similar study carried out in the Duomo of Milan [193], where the vertical gradient was estimated as $0.033°C/m$. This Cathedral does not have a heating or air conditioning system inside, which would explain the linear gradient and the small variations of T. If the trajectories of T recorded in the Duomo are compared with those from the church of Saint Thomas and Saint Philip Neri, their characteristics are rather similar (e.g., maximum, minimum, trend, etc.), probably because the indoor microclimate in both churches is unheated (i.e., natural) without any air conditioning system and the climate in Valencia and Milan is rather similar.

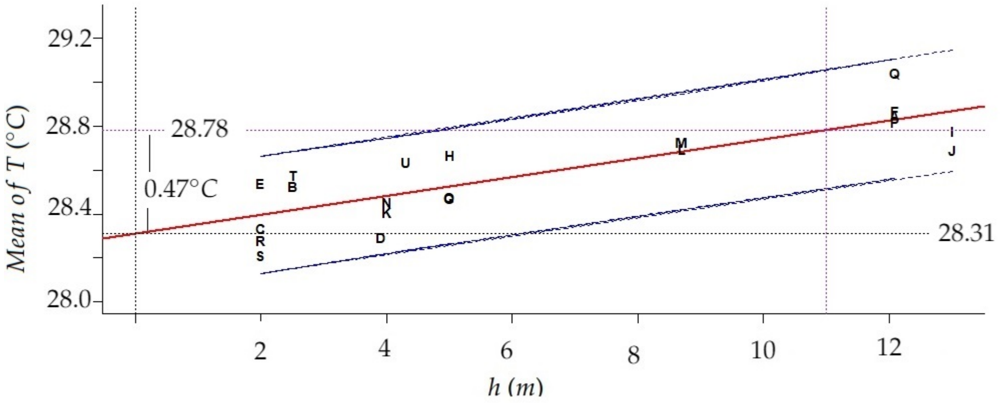


Figure 4.10: Plot of fitted linear model for mean temperature in the period from August 10th at 8:00 AM to September 9th at 11:00 PM, from each node (codes as in Figure 4.4) versus height. Prediction limits (in purple) correspond to 95% confidence level. The vertical gradient estimated from 0 to 11 *m* is about 0.47°*C*.

In the Basilica di Santa Maria Maggiore, in Rome, a study of temperature gradients was carried out at heights of 3, 7, and 11 *m* [234]. A greater vertical gradient was identified in August than in September and December. In August, most time series of temperature underwent an increase by 0.05°*C*/*m* approximately. Regarding the trajectories of T recorded in the church of Santa Maria Maggiore [234], which is relevant for the case of Saint Thomas and Saint Philip Neri, in August and September, higher temperatures were recorded at the maximum height, while the lower ones were found for the minimum height. By contrast, in December, the phenomenon changed, so that lower temperatures were recorded at the maximum height while the opposite occurred near the floor level. The difference between maximum and minimum heights for the sensors were similar in both studies (i.e., 8 and 11 *m*, for the Basilica in Rome and for the church in Valencia, respectively). Furthermore, the gradient found for August at the Basilica of Santa Maria Maggiore was 0.05°*C*/*m*, which is consistent with the confidence interval of 95%, $(0.030, 0.057)$, for the gradient estimated for the period from August 10th to September 9th in church of Saint Thomas and Saint Philip Neri. However, in other periods like May, in Santa Maria Maggiore [234], the increment of temperature per meter was at least 0.25°*C*/*m*. The main reason could be the hot air that came in through the front door of the church [234]. According to reported results and the fact that some studies have displayed the effect of temperature gradient on the dust accumulation process under different temperatures [234; 235], an important conclusion of both works is that the ventilation of churches can be very important for discussing temperature gradient in height. The ventilation rate should be studied and quantified, as it contributes to the deposition of dust on art works. This issue is discussed in Section 4.4.2.

Although many studies have analyzed time series of T in the context of art conservation, their focus has not been on comparing parameters (e.g., mean, maximum and minimum) of temperature at different height levels. For example, Merello et al. [11] compared estimation of parameters such as the minimum and maximum of T in distinct positions in a building instead of different height levels of the sensors. Also, they studied the performance of the mean T by using contour plots, which helped to analyze the change of T at different height levels in the building. However, it is not possible to estimate the vertical gradient from this reported study. The methodology proposed by Merello et al. [11] based on ANOVA could be employed to compare the series of T at different levels in the building. However, this method cannot help to discriminate according to the different characteristics of series of T.

In order to study the temperature gradient and identify the best function that explains the changes of temperature according to the height variable (i.e., linear, quadratic or further polynomial orders), it is necessary to employ temperature measures at distinct height levels. In case of a linear gradient, using a linear regression model seems better than computing the differences between temperatures measured at two levels. The estimation of the model provides a better interpretation of the results. Although linear regression was used in this study, other methods based on smoothing techniques and nonparametric regression [231; 236], which relax the usual assumption in several standard models like the one used here, could be employed. These models are more flexible and they can fit a wide range of structures in the data, e.g., observations from buildings which employed an air conditioning system.

There are several European Standards [92; 93; 94; 95; 96; 97; 98; 237; 238] for providing guidelines for monitoring, elaboration and study of the microclimatic conditions inside heritage buildings, aimed at art conservation [198; 199; 200]. According to European Standards EN 15757, variables such as annual average, seasonal variation, short-term fluctuations, and 7th and 93rd percentiles of short-term fluctuations, can be used as reference for specifying the levels of T or RH in order to avoid physical damage in organic and hygroscopic materials. Seasonal variation are computed by using moving average of 30 days, and short-term fluctuations are calculated by using difference between the instantaneous measures and a moving average [97; 239]. Short-term fluctuations are used instead of seasonal cycles because buildings located in cold climates are expected to be equipped with heating systems, which helps to provide more stable seasonal cycles. As a consequence, the indoor conditions are less dependent on the external conditions. However, there is not a complete study on the application of the EN

15757 in all types of climates, thus, it is necessary to assess its methodology in temperate climates and suggested changes, if required [201]. Silva and Henriques [201] carried out an microclimatic study of the Church of St. Christopher in Lisbon (Portugal) with records from November 2011 to August 2013. They analyzed T and RH from 17 thermocouples or portables sensors located on the church in a vertical profile (5 levels: 0.15, 1.50, 3.90, 7.50, and 10 m), in horizontal profiles (4 profiles at different positions), some surface points of on wall, among others. They studied indoor conditions as indicated by references like the EN 15757:2010 [97], the Italian National Unification UNI 10829 [160] and the American Society of Heating, Refrigerating and Air-Conditioning Engineers ASHRAE specification [239]. This research is of interest because studies about indoor air conditions in historical buildings in temperate climates are scarce [201]. Although Portugal has a Mediterranean climate, due to its proximity to the Atlantic Ocean, it has a particular climate with winters less cold and summers less warm than climates of other countries in southern Europe. Silva and Henriques [201] define an interval for short-term fluctuations of T of 0.8 $^{\circ}C$. This interval or target band was limited by 7th and 93rd percentiles of T. They suggested following a target band for T in the future as a preventive measure. Also, they found, for example, that the maximum temperature from sensor at level 3.90 m was 24.9 $^{\circ}C$ and the temperature minimum was 13.2 $^{\circ}C$. Although the trend of the temperature trajectory found for the Church of St. Christopher may coincide with other temperature trajectories for other buildings in Mediterranean countries, the band, minimum and maximum temperatures can be very different. For example, in this study, both the minimum and maximum temperatures were higher than those determined in the Church of St. Christopher. Regarding to the previous ideas, it is necessary to evaluate the climatic conditions in buildings located in Mediterranean climate in order to have reference values for monitoring indoor conditions. Thus, the methodology proposed in the present work for estimating the temperature gradient, could be useful in order to determine reference measures for historical buildings. Furthermore, for buildings located in Mediterranean countries, the confidence interval (95%) of vertical gradient reported here (0.030 $^{\circ}C/m$, 0.057 $^{\circ}C/m$), could be considered as a reference measure in summer.

Functions to estimate risk damage in cultural heritage are a permanent subject of study and investigation [240; 241]. In [241], a detailed review of such risk damage can be found. However, the quantification of vertical thermal gradients has not received much attention yet regarding the study of risk damage in cultural heritage, though it is well established that T gradients affect dust deposition on walls and works of art [234]. One reason for this can be the difficulty of measuring the speed of the movement of air within the building, which is a consequence of thermal currents dragging particles.

Air speed is not easily quantified since it is necessary to model the speed value at each point [242; 243; 244]. The techniques related to Computational Fluid Dynamics (CFD) analyze physical parameters at each point by using finite elements of volume, mainly T, RH and wind speed. Although these techniques use a computer system for their calculations, they need real measurements to indicate the boundary or input conditions of the problem and, secondly, to verify and validate the results. Therefore, a technique that is capable of quantifying a gradient, such as the one described in the present work, might be useful in a CFD study [242; 243; 244].

4.4.2 | Ventilation of the Church of Saint Thomas and Saint Philip Neri

It has been estimated that the total volume of the church of Saint Thomas and Saint Philip Neri is about 18000 m^3, including the side chapels and the Chapel of the Holy Communion, which is separated from the main nave by a door that always remains open, except in winter. In this chapel there are two tilt-and-turn windows of 1×1.5 m, almost always opened vertically. The Sacrist has ventilation to the outside, but the door that connects the main nave with the sacristy remains closed most of the time. The temple has multiple windows in the upper area, but they do not have openings for ventilation. The main source of ventilation is the large front door, which is rarely fully opened. Ordinarily, the main door gives access to the nartex, which is a wooden structure that serves as a transition between the exterior and interior environment. This nartex has two 2.4×0.9 m doors, which must be pushed to open by the churchgoers. They close automatically by means of springs.

Through the website https://datosclima.es/Aemethistorico/Vientostad.php (accessed on 18 October 2021) it has been found that in Valencia, between August and December 2017, the average wind speed was about 1.4 m/s, which, multiplied by the section of the narthex door (2.2 m^2), is equivalent to an average air flow of about 3.08 m^3/s. Assuming that during work days this door is open for a total of 200 s (taking into account that the temple can be visited for 6.5 h a day), this equates to an average air volume of 616 m^3. Thus, under these conditions, 29 days would be necessary to renew 18,000 m^3 of the total volume.

Assuming that on Sundays the attendance of parishioners is much higher, up to perhaps 10 times, it would take about 3 days to renew the total air volume. In any case, these preliminary calculations show that the ventilation rate of the temple is very low. Air renewal rate is an important aspect to consider in the present study. Actually, the fact that the vertical thermal gradient was basically observed in August, could be related to the low ventilation rate during this month. Perhaps a much higher ventilation rate could have homogenized the vertical profile of temperatures and could have altered the results.

4.4.3 | Application of sPLS to Identify Key Features Correlated with Height

Next, sPLS was employed to identify the main features from the time series that are correlated with sensor height in the church, which is of interest particularly for those periods where the vertical gradient was not statistically significant. When applying sPLS as described in Section 4.3.5.4, according to criteria of $RMSEP$ and R^2, two components seem to be enough, both when using variables from method 1 and method 2. The total number of selected variables for methods 1 and 2 were 7 and 13, respectively. Variables are sorted in Table 4.2 by decreasing value of VIP_j [187], which was computed for determining the overall importance of each predictor variable on the response, cumulatively over the total components. The values of VIP_j for variables in Table 4.2 are greater than 1, which are the most important ones in the model. Regarding method 1, the variables selected by sPLS correspond to the stages: 1 (`mean.ts` and `mean.mr`), 3 (`pacf4`), 4 (`pacf4`), 5 (`pacf3` and `pacf4`), and 7 (`pacf3`). The relevance of `pacf3` and `pacf4` is difficult to interpret, because these variables imply that the time series are autocorrelated with the values observed 3 or 4 hours before. Anyway, the most relevant information is the fact that, in August, both `mean.ts` (mean temperature) and `mean.mr` (mean moving range with order 24) present nearly the same degree of correlation ($r = 0.67$) with sensor height. Thus, not only the mean temperature tends to be higher at the upper position, but also the daily variability. The reason might be the high temperatures reached in Valencia in August during the day, but they become mild at night.

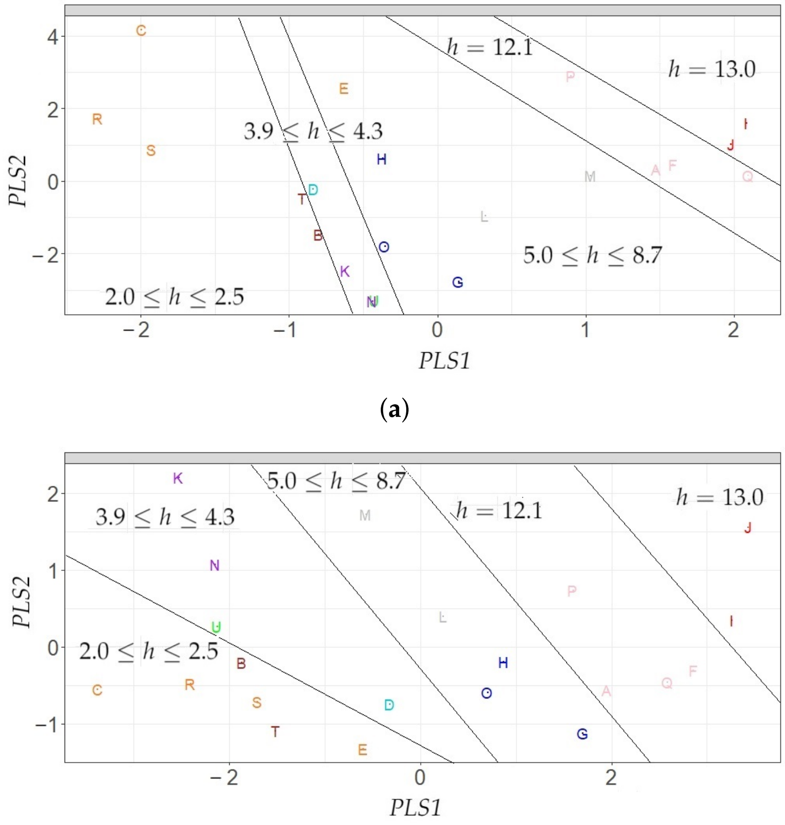
The most relevant stages were 1 and 4, which correspond to August and November. For both methods, August was the most important month. The mean temperature was important for this month, because the overall mean temperature (`mean.ts`) was selected for method 1 and, moreover, the local mean at the last instance of time (`a`, i.e., level) was chosen for method 2. The feature `mean.ts` was the most important, according to the VIP_J for method 1 (see Table 4.2a) and the level was the 10th variable among the selected ones for method 2 (see Table 4.2b). Regarding the selected variables from sPLS, `mean.ts` and `mean.mr` were the only ones with a statistically significant correlation at $\alpha = 1\%$ ($r = 0.86$, $p - value < 0.001$ and $r = -0.65$, $p - value = 0.001$).

Table 4.2: Results from sPLS: variables selected (V), ordered from top to bottom according to VIP_j. The monthly stage is indicated as Stg1 (August) to Stg7 (February). Correlation coefficients (r) of each selected variable vs. sensor height, and the corresponding $p-values$ of the correlation test. Results are presented in accordance with the variables used in sPLS: (**a**) Method 1 and (**b**) Method 2

(a)					**(b)**			
	Stage	V	r	$p-value$	Stage	V	r	$p-value$
1	`Stg1`	`mean.ts`	0.86	0.000	`Stg4`	`a`	-0.33	0.138
2	`Stg5`	`pacf4`	0.28	0.217	`Stg1`	`s7`	0.83	0.000
3	`Stg4`	`pacf4`	0.36	0.108	`Stg1`	`s8`	0.79	0.000
4	`Stg7`	`pacf3`	0.51	0.019	`Stg1`	`s6`	0.80	0.000
5	`Stg3`	`pacf4`	0.43	0.054	`Stg1`	`s19`	-0.77	0.000
6	`Stg1`	`mean.mr`	-0.65	0.001	`Stg5`	`a`	-0.31	0.178
7	`Stg5`	`pacf3`	0.31	0.167	`Stg4`	`s12`	0.74	0.000
8					`Stg1`	`s18`	-0.75	0.000
9					`Stg4`	`s6`	-0.23	0.319
10					`Stg1`	`a`	0.77	0.000
11					`Stg7`	`s20`	-0.41	0.065
12					`Stg7`	`s16`	0.70	0.000
13					`Stg4`	`s23`	-0.69	0.001

In fact, the period from August 10th to September 9th was the most important period for explaining the vertical gradient of temperature. Also, August was the most relevant month for discriminating the temperature according to height. In the same manner, research of the time series of T recorded at the archaeological site of L'Almoina in Valencia found that the most important fluctuations occurred during summer [205], due to the greenhouse effect caused by a skylight that covers part of the ruins. Results reported here are consistent with a similar work that found summer as the most important period for explaining the gradient of T, probably because outdoor temperatures in the Mediterranean region are greater in summer [10].

For method 2, the estimated value for the level (`a`) was found as relevant in stages 1, 4 and 5. However, the correlation between the level and height was only statistically significant for August ($r = 0.77$, $p-value < 0.001$). In fact, August was the unique month with a pronounced correlation ($r = 0.78$, $p-value < 0.001$) between the level and the mean temperature (`mean.ts`), which is strongly correlated with the height ($r = 0.86$, $p-value < 0.001$). By taking a look at the coefficients r in Table 4.2b, the highest values corresponds to `s6`, `s7` and `s8` (stage 1), which implies a seasonality every 7 hours approximately. Also, in this stage, `s18` and `s19` are relevant.



(**a**)

(**b**)

Figure 4.11: Projection of sensors over the two relevant components (PLS1 and PLS2) on the subspace spanned by the regressor data sets from sPLS; (**a**) using variables from Method 1 and (**b**) Method 2. Sensor codes, represented by letters, as in Figure 4.4, which were colored according to their height: 13.0 m in red, 12.1 m in pink, 8.7 m in gray, 5.0 m in blue, 4.3 m in green, 4.0 m in purple, 3.9 m in cyan, 2.5 m in brown, and 2.0 m in orange. Solid tilted lines were inserted to better reflect the distribution of nodes in both plots according to height.

In simple regression, $\boldsymbol{Y} = f(\mathbf{X})$, so that $\boldsymbol{Y}$ depends on the values of $\mathbf{X}$. In the observed correlation between mean temperature and sensor height, the temperature varies according to the sensor height and, hence, temperature should be regarded as the dependent variable ($\boldsymbol{Y}$) and height as the predictor ($\mathbf{X}$). In this linear model, the slope can be interpreted as the gradient, as discussed above. Nonetheless, in order to better understand the differences in the time series recorded at the lower vs. the upper positions, multiple linear regression (MLR) was used to fit sensor height ($\boldsymbol{Y}$) as a function of variables selected from sPLS from method 1. Using these variables in the regression model leads to a high degree of multicollinearity; thus, only two variables were considered in the final model as predictors: mean of T and mean of MR (i.e., moving range of order 24), both from stage 1 (August). The estimation of the height is given by $\widehat{y_i}$ =0.36+3.24·`mean.ts`-1.72·`mean.mr`, $i = 1, \ldots, 21$. Thus, sensor height can be fitted according to the average temperature in August and a measure of daily variability. The R^2 for the model was 87%; $p - values$ (from F-test and t-tests) were less than 0.0001 for determining whether the independent variables in the model are statistically significant. The residual analysis showed that the assumptions of the linear regression model were fulfilled.

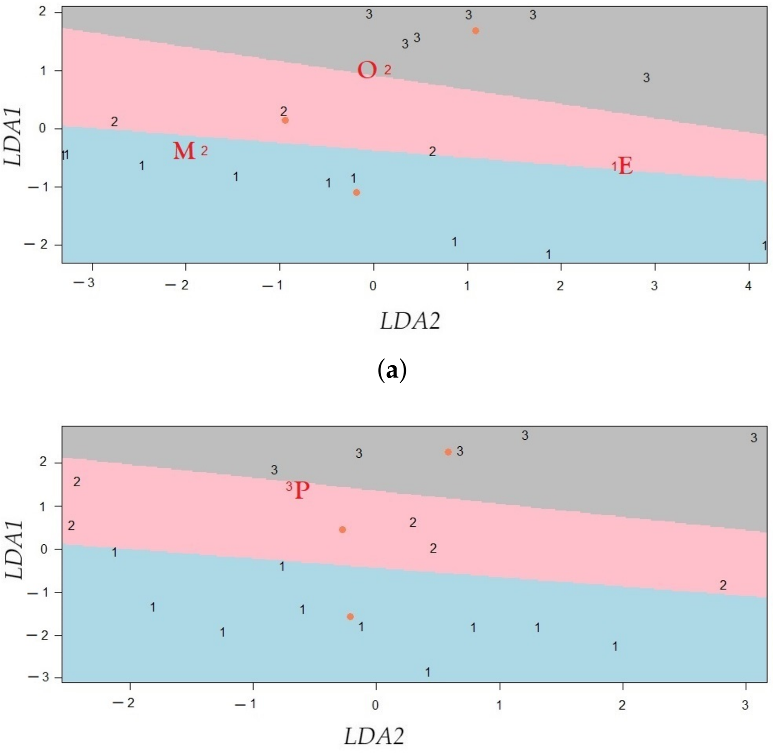
Regarding the sPLS results from both methods, Figure 4.11 shows the projection of sensors over the two relevant components (PLS1 and PLS2) on the subspace spanned by the regressor data sets from sPLS. The projections of sensors were colored according to their height levels (one color per height: 3.0 m in red, 12.1 m in pink, 8.7 m in gray, 5.0 m in blue, 4.3 m in green, 4.0 m in purple, 3.9 m in cyan, 2.5 m in brown, and 2.0 m in orange). According to the tilted solid lines represented in Figure 4.11, it is possible to establish 5 classes of sensor nodes according to both methods. It is noteworthy that the solid lines are markedly tilted, which implies that both the first and second components are necessary to achieve a reasonable discrimination of nodes according to height. These classes are adjacent and appear ordered in the plots. Lines in Figures 4.11a and 4.11b separate the different groups, which were denoted as 1 ($2.0 \leq h \leq 2.5$) in blue, 2 ($3.9 \leq h \leq 4.3$) in pink, 3 ($5.0 \leq h \leq 8.7$) in gray, 4 ($h = 12.1$) in purple, and 5 ($h = 13.0$) in green. The tilted solid lines were drawn by visually checking the positions of points, taking into account similar height levels of the sensors. For method 1, node E was classified incorrectly. Nodes T and B are located in the limit of classes 1 and 2, while J appears in the boundary of groups 4 and 5. By contrast, for method 2, all nodes were classified correctly according to the lines drawn in the plot. However, U was located in the limit between class 1 and 2. This classification can be improved by utilizing LDA, which maximizes the differences between the clusters, being the two first components (LDA1 and LDA2) a linear combinations of PLS1 and PLS2 components, which in turn are linear combinations of predictor variables from methods 1 and 2. The most important variables per method and component were the following: for method 1, PLS1 was mainly determined by `mean.ts` and `mean.mr`, while PLS2 was basically computed by `pacf3` and `pacf4`. For method 2, PLS 1 was determined by the variables `a`, `s6`, and `s7`, while PLS2 was calculated by using `s8`, `s12`, `s16`, `s18`, `s19`, `s20`, and `s23`. These results suggest that for method 1, the first component explained the level and changes of the levels of the time series of T, while the second explained the autocorrelation of time series. Furthermore, for method 2, the first component explained the level of the last observation of the time series of T, while the second component explained the prediction of the last observation of series at 7, 15, and 24 hours past the time of the last observation.

4.4.4 | Discrimination of Sensors in Three Categories by Means of LDA

By considering those variables found as relevant from sPLS with two components, LDA was applied in order to check if is possible to discriminate sensors according to their height, and to better understand the variables most relevant for such discrimination. Three categories were established: low, medium, and high elevation.

Figure 4.12 displays how the sensors are discriminated in three clusters (i.e., blue for cluster 1, red for cluster 2, and gray for cluster 3) by applying LDA. The plot outputs show the projection of sensors over the two relevant components (LDA1 and LDA2). The three lines in the pictures which separate the three clusters, were determined according to the boundary of decisions from LDA (see Section 4.3.5.5). By considering variables from method 1, the nodes E, M and O, were incorrectly classified. Nonetheless, nodes E and O are located close the limit of the correct class. Figure 4.12b shows results from method 2. Only node P was wrongly classified. Although four sensors were classified incorrectly their projection on LDA1 vs LDA2 appear very close to the boundary of decisions. Therefore, the incorrectly classified sensors do not depart too much from the expected performance.

The discriminant approach used here is based on two steps; firstly, sPLS is applied to identify the most relevant variables and, next, LDA is used for the discrimination. Hence, this procedure was referred to as sPLS with LDA. In order to further discuss the results, two additional discriminant methodologies based on a single step were applied: sPLS-DA and SPLSDA. When comparing the results from the three methods, sPLS with LDA led to the minimum error rates, 14.28% and 4.76% for method 1 and 2, respectively (see Table 4.3). Using variables from method 1, sPLS-DA selected 10 variables, while sPLS with LDA used 15 (see Table 4.3a). For method 2, SPLSDA selected 11 variables, less than the other two methods, both of which used 15 (see Table 4.3b). Computational experiments carried out by Chung and Keles [53] suggested that variable selection performance of SPLSDA improves when the sample sizes increase. However, in the context of art conservation, the number of sensors installed is usually rather small due to restrictions in heritage buildings. Then, a combination of sPLS with LDA can be useful to discriminate the time series according to different levels and zones in this type of building.



(a)

(b)

Figure 4.12: Projection of sensors over the two relevant components (LDA1 and LDA2) from LDA; (**a**) using variables from Method 1, (**b**) Method 2. Three classes were considered according to sensor height (h): class 1 ($2.0 \leq h \leq 4.3$), class 2 ($4.3 < h \leq 8.7$), and class 3 ($8.7 < h \leq 13$). Red numbers correspond to sensors wrongly classified.

Table 4.3: Classification error rate and number of selected variables (N) using sPLS-DA, SPLSDA, and sPLS with LDA. Results are presented according to the method used for computing the features from the time series: (**a**) Method 1 and (**b**) Method 2.

	(a)		**(b)**	
Classification method	Error rate (%)	N	Error rate (%)	N
sPLS-DA	35.06	10	18.75	15
SPLSDA	19.04	42	19.04	11
sPLS [61] with LDA	14.28	15	4.76	15

For the three classification methods applied, the classification error rates using variables from method 2 were lower or equal to the rates obtained for method 1. In a similar study carried out in Valencia Cathedral [161] and L'Almoina museum [162], sPLS-DA for method 2 obtained the second best results (lower error rate) when comparing with method 1 and other approaches like ARIMA, ARIMA-GARCH or Wold decomposition. Authors concluded that parameters extracted by applying SH-W has a good performance for a wide range of series [162].

In this study, features defined using quantiles [220] were employed. These variables were not computed in previous studies [161; 162] when sPLS-DA was carried out using input features from original time series. Regarding this type of variables, sPLS-DA selected `f.p.r.m35` (for stage 5), `f.p.r.m80` (for stage 2), and `p.d.f.p` (for stage 1). Furthermore, SPLSDA selected `f.p.r.m20` and `f.p.r.m35` (for stages 1, 2, and 5), `f.p.r.m50` (for stages 1, 5, and 7), `f.p.r.m80` (for stages 1 and 2), as well as `p.d.f.p` (for stages 1, 2 and 6). However, the classification method based on sPLS with LDA did not select any of these features. In summary, having a variety of time series characteristics can help improve results of the classification of sensors and comparisons of results from different classification methods. Furthermore, finding a common subset of variables was the most important outcome in this case for the classification methods, while other variables improved the results. Different studies have shown efficient results using features from SH-W method.

sPLS-DA is one-stage approach that performs, in one step, dimension reduction and selects variables for obtaining the lowest classification error rate. The other methods are two-stage approaches (i.e., SPLSDA, and sPLS with LDA) which only maximize the separation between clusters in the second step. One assumption is that employing two steps instead of one might prevent from obtaining the important variables for classifying the time series. However, the results show that using sPLS with LDA provides the best classification error rates.

Elorrieta, Felipe et al. [49] proposed a new methodology for classifying time series in the field of astronomy by capturing their peaks. The methodology was based on different classification methods (e.g., Lasso regression, random forest, support vector machine, logistic regression, CART algorithm, boosting, and artificial neural network), and on certain features from the field of astronomy. Furthermore, they proposed two new features to be used as input for the different classification methods [49; 50]. The new methodology proposed by Elorrieta, Felipe et al. [49] shares a common aspect with the approach used here. Both employed a classification algorithm using different characteristics from time series as input. Notwithstanding, the features extracted from time series and classification algorithms are different. Time series from the art conservation field hold different characteristics from the ones in time series from astronomy research. Furthermore, the main goal is to classify stars and results do not need to be interpreted. Nonetheless, some features and all algorithms proposed by Elorrieta, Felipe et al. [49] can be employed for analyzing the data from art conservation when the aim is to classify time series.

In the field of art conservation, high-dimensional real-world data sets were analyzed to classify time series by using PCA with raw data as input by Zarzo et al. [7] and García-Diego and Zarzo [6]. The relevant principal components were calculated to identify the patterns encoding the highest variance in time series of T and RH. Although PCA does not maximize the separation between clusters for the different time series, it was found that PC1 and PC2 discriminated several clusters from each other, when applying PCA to different time series. Despite the succes of PCA in this context, LDA can improve the results because this method maximizes the separation among clusters of time series.

The three methodologies (i.e., sPLS-DA, SPLSDA, and sPLS with LDA) were efficient for classifying time series, as they separate the time series clusters. It should be remarked that these methodologies were numerically stable and competitive in terms of computational efficiency. Consistent results were obtained when the processes were repeated, and the time running was alike and little. Besides, using linear combinations of variables extracted from time series can greatly improve their classification.

In order to study the vertical gradient of T and to characterize the temperature at high, medium and low heights, it would have been more convenient to decide the position of sensors according to a statistical design of experiments considering the same number of sensors at the different levels and different positions in the church. A proper statistical design is important to improve the results and conclusions. However, it is not always possible to use the ideal statistical design because there may be some restrictions in the buildings, such as the characteristics of the building itself, the maximum number of nodes available, and the need to prevent problems derived by the movement of people, among other factors. Nonetheless, results reported here will be helpful for encouraging further studies using an adequate statistical design that can be adapted to these restrictions.

According to the result, the different height levels of sensors can explain the vertical thermal pattern in August. In the other months, the factors that might affect the temperature, are the following: some sensors are located close to windows which were exposed to direct sunlight for a continued period of the day, some sensors are positioned close to halogen lamps reaching a high temperature, among other factors. Work in progress is currently being carried out to study these factors in detail.

4.5 | Conclusions

With the goal of proposing a methodology for the multisensor microclimate monitoring in the church of Saint Thomas and Saint Philip Neri (Valencia, Spain), two methodologies were put forward for estimating the vertical gradient of temperature and characterizing the differences between time series at high, medium and low heights.

1. This research reports a microclimatic study in the church of Saint Thomas and Saint Philip Neri in Valencia for the first time, which is of relevant interest because inappropriate conditions of temperature can affect the valuable artworks. The results suggest that temperature gradients in this church were comparable to those estimated at the Duomo in Milan and Santa Maria Maggiore in Rome, Italy. Moreover, it turned out that the identification of such gradients was restricted to a very limited period (August-September) during summertime. Furthermore, the results found in this study might provide guidelines for establishing a plan for thermal monitoring and preventive conservation in similar churches.

2. The first methodology is based on Pearson's correlation coefficient and linear regression. This methodology, which could help to determine reference thermal gradients for art conservation, could be improved using smoothing techniques and nonparametric regression. Furthermore, taking into account that datasets about indoor air conditions in historical buildings in Mediterranean climates are scarce, the confidence interval (95%) of the vertical gradient found in summer $(0.030°C/m, 0.057°C/m)$, could be considered as a reference for further similar studies. Results obtained can be extrapolated to similar scenarios, whether in a heritage building or others, such as an industrial building, warehouse or farm of similar volume and height, with little ventilation, in a similar climate, according to some climate classification criteria (e.g., Köppen [245] and Trewartha [246]).

3. The second methodology proposed here combines sPLS [61] and LDA. Also, it employs variables computed from seasonal H-W method, or functions that are applied to time series. This methodology helped to obtain parsimonious models with a small subset of variables, leading to satisfactory discrimination and easy interpretation of the different clusters of the time series. Also, it was useful for identifying the most important variables for classifying time series. The variables computed from seasonal H-W method yielded better results. In other studies, SH-W has also been shown to provide efficient results. This method was more flexible for fitting the distinct time series and obtaining low values of the classification

error rate. The new methodology proposed allowed an efficient characterization of T at high, medium and low altitude levels. This approach had the best results according to the classification error rate and number of selected variables, when compared to results from SPLSDA [52] and sPLS-DA [63]. When using variables from seasonal H-W as input for either sPLS with LDA, sPLS-DA, or SPLSDA, both the error rate and the number of selected variables were better.

5

General Discussion

The process of separating groups according to the similarities of data that are correlated is a common situation in many scientific fields. From a statistical perspective, designing new methodologies for identifying clusters in correlated data represents an important challenge. There are different statistical approaches for dealing with clustering of time series, according to the characteristic of the problem to solve. Some approaches were presented in the introduction. However, clustering performance is reduced when the clusters are close to each other, which is a problem when examining time series in the context of microclimate monitoring for art conservation.

In this context, highdimensional realworld data sets were ana lyzed using PCA with raw data as input, in order to identify clusters of time series. The principal components were computed in order to represent the patterns encoding the highest variance in a data set. Although this method does not help to maximize the separation between clusters in the data set, it was found that the first principal components separated different subgroups of the samples from each other, when applying PCA to these data sets.

In this dissertation, a methodology for classifying time series is proposed. This methodology consists of applying a classification method using variables from different methods as input. Among the three classification methods, the two first are versions of known sparse methods which have not been applied to time dependent data. The third method is a new proposal that is based on two known algorithms. The variables correspond to parameter estimates from functions, methods, or models commonly found in the area of time series. Also, some variables employed in the field of astronomy (for classifying stars) were proposed.

Regarding the classification methods, the first two methods (SPLSDA and sPLS-DA) were proposed by Chung and Keles [53] and Lê Cao et al. [63] respectively, for adapting sparse partial least squares for classification of high dimensional data. By contrast, the third approach is a new proposal that used both sPLS [61] and linear discriminant analysis. This method is based on SPLSDA that was proposed by Chung and Keles [53]. These methods use variables that are extracted from time series as input, for adapting the classification in the context of clustering of time series.

The methods employed for computing the variables for the three studies presented in this dissertation were the following: The first study computed variables using functions applied to the time series (e.g., ACF, PACF, moving range, quantiles, among others), using different seasonal Holt-Winters methods, and using a common seasonal ARIMA- TGARCH model, per stage (i.e., according to climatic conditions and structural breaks) for each time series. The second study used four methods for determining the variables. The first three methods were very similar to the approach used in the first study and the fourth method computed variables using Wold decomposition and different seasonal ARIMA models. For the second method, in addition to variables from the seasonal Holt-Winters method computed in the first study, 24 forecasts were employed. For the third method, different seasonal ARIMA models per stage and time series were used instead of a common seasonal ARIMA-TGARCH model, per stage. The third study employed two methods, both applied in the previous study. However, for the first method, some variables based on quantiles were employed, which are used in the field of astronomy.

Regarding the performance of sPLS-DA with different methods for achieving the classification of time series, the main results were the following: for the first study, the best results were derived from the ARIMA-TGARCH models and the second from the seasonal Holt-Winters method. For the second study, the best results were obtained from functions applied to the time series and the second best results from the seasonal Holt-Winters methods. For the third study, the best results were found when using the seasonal Holt-Winters method. According to these results, the most relevant method for determining variables for classifying time series was the seasonal Holt-Winters method, because the variables helped to obtain effective results for the different data sets. Expanding the methods in the second study, by using the Wold decomposition and predictions of T from the seasonal Holt-Winters method, did not improve the classification of the time series. The Holt-Winters method performed better in the classification because this approach adapts to a wide range of time series and its flexibility helps to provide more information to capture differences for classifying time series. Due to the need of using the same number of variables for each time series, either a common ARIMA-TGARCH, ARIMA, or ARIMA with Wold decomposition, per stage of time series (i.e., every time series was divided into stages according to seasons or structural breaks) was used. Among these models, ARIMA-TGARCH yielded the best results, the reason might be that the residuals were fitted using TGARCH models and this model captured other differences between the time series that were not determined for the ARIMA models. In fact, when using ARIMA-TGARCH models, the main variables selected for PLS correspond to estimates of parameters of the TGARCH model.

In the second study, the performance of sPLS-DA and random forest methods were compared. Similar results were found when both methods used the variables from functions applied to the time series and the seasonal Holt-Winters methods as input. The best results were obtained when using functions applied to the time series and the second best results when using the seasonal Holt-Winters method. Both the sPLS-DA and random forest methods were useful for classifying the time series. It was possible to obtain parsimonious models with a small subset of variables, leading to satisfactory discrimination. Results from sPLS-DA could be easily interpreted via graphical outputs.

In the third study, the performance of sPLS-DA, SPLSDA with a new approach, sPLS [61] with LDA (which was proposed in this study), were compared. These methods used variables from the functions applied to the time series and the seasonal Holt-Winters method as input. When comparing the results from the three classification methods, sPLS with LDA led to the minimum error rates for both cases analyzed (i.e., functions applied to time series and seasonal Holt-Winters method). For the three classification methods applied, variables from the seasonal Holt-Winters method performed similarly or better than when using the functions applied to the time series.

In respect to the most important variables determined for sPLS-DA, the main results are the following: for the ARIMA-TGARCH models, they were the estimates of ω and α. Also, estimates of mean and median of ACF at lags 1 to 72 of the residuals, variance of the residuals, and maximum of the periodogram of the residuals. For the Holt-Winters method, the most important variables were the estimates of the level, seasonal components 18, 19, 20 and 24. Also, the estimates of SSE and maximum of the periodogram of the residuals. For the ARIMA models, they were the estimates of parameters of MA, AR and SAR. Also, sample PACF at lags 1 to 5, as well as the mean and median of the sample of ACF at the first 72 lags of the residuals. For functions applied to time series, they were the estimates of mean, moving range, PACF at lag 1 to 5. In the fourth study, in addition to the variables mentioned, certain variables based on quantiles were also selected. The majority of variables selected for sPLS-DA were also chosen by SPLSDA. However, sPLS with LDA did not select any variable based on quantiles.

Since both methods, SPLSDA and sPLS with LDA treat dimension reduction and one-stage approach, and selects variables in order to obtain the lowest value of the classification error rate, the other methods are two-stage approaches (SPLSDA and sPLS with LDA) which only maximize the separation between clusters in the second step. It was assumed that using two steps instead of one might fail to select some important variables for classifying the time series. However, the results displayed that using sPLS with LDA, produces the best classification error rates.

For the long term monitoring of microclimate in the context of preventive conservation of artworks, it is necessary to minimize the number of sensors used. In this dissertation, one solution which was proposed was a sampling methodology that captures the relevant information from clusters of time series. The idea is to choose a set of sensors, based on the first two components from sPLS-DA, centroids of the clusters, the distances between each centroid and the position of sensors in the multivariate space (i.e., pair of coordinates using C1 for x-axis and C2 for y-axis for each sensor and the centroid of its cluster). Namely, it proposes characterizing the different zones of a building using the maximum number of sensors possible and the classification methodology proposed, then selecting a subset of sensors with the sampling methodology described. When there are a lot restrictions about the number of sensors, the proposal is to monitor the microclimatic conditions using the subset of sensors and the information from the first analysis for classifying sensors. In cases where it is possible to have at least 30 sensors the proposal is to repeat the classification methodology for characterizing the different zones in the buildings every year.

The methodology proposed in this dissertation was useful for characterizing the differences in climatic parameters (e.g., relative humidity and temperature), measured at different positions and heights (high, medium and low altitude levels). Identifying different levels or zones in a museum, archaeological site or historical building could help to monitor the microclimatic conditions inside. Classification error rates from the classification method might be affected by the malfunctioning of some sensors, problems related to the microclimate where the sensors are located, the performance of the classification method which is influenced by the total number of sensors, the number of sensors per clusters, or prior knowledge of clusters. Thus, the incorrectly classified sensors should be evaluated to identify possible problems in the artworks.

Even though the methodology proposed helps to obtain effective results, the need to obtain well-determined clusters when having a smaller number of time series is a problem that should be improved. Another situation that needs special attention is when there is no information about the possible clusters of series. In the methodologies proposed, the different versions of sparse PLS require to establish the number of clusters before implementing the algorithm. In fact, in the second study, it was necessary to carry out the k-means algorithm with PCA in order to obtain different possible clusters of time series. Once the three clusters were established, they were employed as input for sPLS-DA. Another situation that could be improved is computing other variables, in order to capture more information from time series. In particular, when a unique ARIMA-TGARCH or ARIMA model was employed for all time series in the same stage, it is unlikely that the best values for the classification variables could be obtained. Then, results from any classification method can be affected. Furthermore, in high dimensional classification with PLS, the selection of filtering methods and their tuning are still among open questions. The bibliography on variable filtering is rich with many pros and cons [247].

Different directions can be considered in future studies; the first direction could be to determine classification variables, the second, improving the classification error rate using a sparse version of PLS, and the third, proposing an unsupervised method to help determine possible clusters to be used for a sparse version of PLS. With the goal of obtaining variables that capture more information from the data, flexible models can be applied to time series. Some options for computing variables might be structural time series models [151] and a nonparametric approach of the GARCH as proposed in [152; 153]. In respect to the second direction, another version of sparse PLS DA could be considered, SGPLS [53], employing separate tuning parameters for every class comparison, in order to compare the capability of classifying time series for the different classification methods in different scenarios (i.e., varying the number of time series per cluster and using different or the same number of time series per cluster). According to [53], SGPLS sacrificed some specificity related to selected variables because SGPLS employs a common η to control variable selection for every class comparison. Also, they concluded that if employing separate tuning parameters for every class comparison, the specificity of the method might improve although a significant increase in computational time can be required for tuning. For the third direction, another unsupervised method could also be used to establish the classes before applying sPLS-DA. For example, a novel Bayesian-nonparametric strategy for setting the number of clusters and their labels was proposed by Guha et al. [192].

Finally, the three approaches that have been proposed work very well when n increases. Also, unlike other involved methods, they discriminate the clusters of the time series when the time series are very similar, or if the number of variables from the time series is greater than the number of the time series ($p > n$). Another important point is that these methods are numerically stable and competitive in terms of computational efficiency. Furthermore, using linear combinations of variables extracted from time series as input for these methods can greatly improve their performance.

6

General Conclusion

The methodology proposed was based on versions of sparse PLS (sPLS-DA, SPLSDA or sPLS with LDA) and variables computed from time series (i.e., using methods, models or functions that are applied to time series) and helped to obtain parsimonious models with a small subset of variables, leading to satisfactory discrimination of the different clusters of the time series with easy interpretation.

The different versions of sparse PLS and random forest methods were useful for identifying the most important variables for classifying time series. In particular, sPLS-DA and random forest algorithm had high percentages of common variables among the selected variables. Also, for both methods, the classification error rates were similar when using functions applied to time series and when using seasonal Holt-Winters method.

The variables computed using seasonal Holt-Winters method helped to obtain better results from the different sparse PLS methods. Seasonal Holt-Winters was more flexible for fitting the distinct time series and obtaining low values of the classification error rate.

The methodology proposed can be useful for characterizing and monitoring microclimatic conditions in different zones and heights in a museum, archaeological site or heritage buildings, with the goal of avoiding problems such as moisture and dust concentration on the artworks. Classifying sensors using sPLS-DA could help to select a subset of the most important sensors in the buildings with more restrictions on the maximum number of sensors allowed. Also, clusters of time series that are identified could be used as a reference for identifying possible changes in the climatic conditions in a building, and incorrectly classified sensors should be evaluated to identify possible problems with the artworks.

References

[1] J. Vilar and P. Montero, "TSclust: An R package for time series clustering," *Journal of Statistical Software*, vol. 62, no. 1, pp. 1–43, 11 2014, Accessed on 16 September 2021. [Online]. Available: https://www.jstatsoft.org/v062/i01

[2] T. Mitsa, *Temporal data mining*. Chapman and Hall/CRC, 2010.

[3] A. Perles, E. Pérez-Marín, R. Mercado, J. D. Segrelles, I. Blanquer, M. Zarzo, and F. J. Garcia-Diego, "An energy-efficient internet of things (IoT) architecture for preventive conservation of cultural heritage," *Future Generation Computer Systems*, vol. 81, pp. 566–581, 2018, Accessed on 16 September 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X17313663

[4] E. Angelini, S. Grassini, S. Corbellini, M. Parvis, and M. Piantanida, "A multidisciplinary approach for the conservation of a building of the seventeenth century," *Applied Physics A*, vol. 100, no. 3, pp. 763—-769, 9 2010.

[5] P. B. Lourenço, E. Luso, and M. G. Almeida, "Defects and moisture problems in buildings from historical city centres: a case study in portugal," *Building and Environment*, vol. 41, no. 2, pp. 223–234, 2006, Accessed on 16 September 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360132305000156

[6] F.-J. García-Diego and M. Zarzo, "Microclimate monitoring by multivariate statistical control: the renaissance frescoes of the cathedral of Valencia (Spain)," *Journal of Cultural Heritage*, vol. 11, no. 3, pp. 339–344, 2010, Accessed on 16 September 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1296207410000336

[7] M. Zarzo, A. Fernández-Navajas, and F.-J. García-Diego, "Long-term monitoring of fresco paintings in the cathedral of Valencia (Spain) through humidity and temperature sensors in various locations for preventive conservation," *Sensors*, vol. 11, pp. 8685–8710, 11 2011, Accessed on 16 September 2021. [Online]. Available: https://www.mdpi.com/1424-8220/11/9/8685

[8] M. J. Varas-Muriel, R. Fort, M. I. Martínez-Garrido, A. Zornoza-Indart, and P. López-Arce, "Fluctuations in the indoor environment in Spanish rural churches and their effects on heritage conservation: Hygro-thermal and CO2 conditions monitoring," *Building and*

Environment, vol. 82, pp. 97–109, 2014, Accessed on 16 September 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360132314002625

[9] C. M. Muñoz-González, Á. L. León-Rodríguez, R. C. Suárez Medina, and C. Teeling, “Hygrothermal Performance of Worship Spaces: Preservation, Comfort, and Energy Consumption,” *Sustainability*, vol. 10, no. 11, pp. 1–20, 2018, Accessed on 16 September 2021. [Online]. Available: https://www.mdpi.com/2071-1050/10/11/3838

[10] C. Muñoz-González, A. L. León-Rodríguez, and J. Navarro-Casas, “Air conditioning and passive environmental techniques in historic churches in Mediterranean climate. A proposed method to assess damage risk and thermal comfort pre-intervention, simulation-based,” *Energy and Buildings*, vol. 130, pp. 567–577, 2016, Accessed on 16 September 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378778816307873

[11] P. Merello, F.-J. García-Diego, and M. Zarzo, “Microclimate monitoring of Ariadne’s house (Pompeii, Italy) for preventive conservation of fresco paintings,” *Chemistry Central Journal*, vol. 6, no. 1, pp. 1–16, 12 2012, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1186/1752-153X-6-145

[12] W. Palma, *Time series analysis. Wiley series in probability and statistics*, har/psc ed. Hoboken, NJ, USA: John Wiley and Sons Inc, 2016.

[13] P. J. Brockwell and R. A. Davis, *Introduction to time series and forecasting*, 3rd ed. New York, NY, USA: Springer, 2016.

[14] R. F. Engle, “Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation,” *Econometrica*, vol. 50, no. 4, pp. 987–1007, 1982, Accessed on 16 September 2021. [Online]. Available: https://www.jstor.org/stable/1912773

[15] T. Bollerslev, “Generalized autoregressive conditional heteroskedasticity,” *Journal of Econometrics*, vol. 31, no. 3, pp. 307–327, 1986, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1016/0304-4076(86)90063-1

[16] A. Ghalanos, “Introduction to the rugarch package (Version 1.4-3),” 2020, Accessed on 16 September 2021. [Online]. Available: https://cran.r-project.org/web/packages/rugarch/index.html

[17] K. Johnston and E. Scott, “GARCH models and the stochastic process underlying exchange rate price change,” *Journal of Financial and Strategic Decisions*, vol. 13, pp. 13–24, 2000.

[18] W. N. Venables and B. D. Ripley, *Modern applied statistics with S*, 4th ed. NY, USA: Springer: New York, 2002.

[19] C. C. Holt, “Forecasting seasonals and trends by exponentially weighted moving averages,” *Journal of Economic and Social Measurement*, vol. 20, no. 1, pp. 5–10, 2004, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1016/j.ijforecast.2003.09.015

[20] P. R. Winters, “Forecasting sales by exponentially weighted moving averages,” *Management Science*, vol. 6, no. 3, pp. 324–342, 1960, Accessed on 16 September 2021. [Online]. Available: https://www.jstor.org/stable/2627346

[21] R. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*. OTexts, 2013, online at http://otexts.org/fpp/ (accessed on 15 September 2021).

[22] S. Aghabozorgi, A. Seyed Shirkhorshidi, and T. Ying Wah, “Time-series clustering a decade review,” *Information Systems*, vol. 53, pp. 16–38, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306437915000733

[23] M. Chis, S. Banerjee, and A. Hassanien, *Clustering time series data: an evolutionary approach. In foundations of computational, intelligence*. Springer, 2009, vol. 6.

[24] T. W. Liao, “Clustering of time series data a survey,” *Pattern Recognition*, vol. 38, no. 11, pp. 1857–1874, 2005, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1016/j.patcog.2005.01.025

[25] G. Batista, X. Wang, and E. Keogh, “A complexity-invariant distance measure for time series,” in *Proceedings of the Eleventh SIAM International Conference on Data Mining, SDM11*, 2011, pp. 699–710.

[26] Z. Kovačić, “Classification of time series with applications to the leading indicator selection,” in *Data Science, Classification, and Related Methods – Proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96)*, 1998, pp. 204–207.

[27] Z. R. Struzik and A. Siebes, “The haar wavelet in the time series similarity paradigm,” *Żytkow J.M., Rauch J. (eds) Principles of Data Mining and Knowledge Discovery. PKDD 1999. Lecture Notes in Computer Science*, vol. 1704, pp. 12–22, 1999, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1007/978-3-540-48247-5_2

[28] P. Galeano and D. Peña, “Multivariate analysis in vector time series,” *Statistics and Econometrics Series 15*, pp. 1–19, 2000, Accessed on 16 September 2021. [Online]. Available: http://hdl.handle.net/10016/162

[29] J. Caiado, N. Crato, and D. Peña, “A periodogram-based metric for time series classification,” *Computational Statistics & Data Analysis*, vol. 50, no. 10, pp. 2668–2684, 2006, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1016/j.csda.2005.04.012

[30] C. A. Douzal and P. Nagabhushan, “Adaptive dissimilarity index for measuring time series proximity,” *Journal Advances in Data Analysis and Classification*, vol. 1, no. 1, pp. 5–21, 2007, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1007/s11634-006-0004-6

[31] M. Li, J. H. Badger, X. Chen, S. Kwong, P. Kearney, and H. Zhang, “An information-based sequence distance and its application to whole mitochondrial genome phylogeny,” *Bioinformatics*, vol. 17, no. 2, pp. 149–154, 2001, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1093/bioinformatics/17.2.149

[32] M. Li, X. Chen, X. Li, B. Ma, and P. Vitanyi, “The similarity metric,” *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3250–3264, 2004.

[33] J. Lin, M. Vlachos, E. Keogh, and D. Gunopulos, “Iterative incremental clustering of time series,” in *Bertino E. et al. (eds) Advances in Database Technology - EDBT 2004. EDBT 2004. Lecture Notes in Computer Science*, vol. 2299. Berlin, Heidelberg: Springer, 2004, pp. 106—-122, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1007/978-3-540-24741-8_8

[34] R. C. Cilibrasi and P. M. Vitãnyi, "Clustering by compression," *IEEE Transactions on Information Theory*, vol. 51, pp. 1523–1545, 2005, Accessed on 16 September 2021. [Online]. Available: https://homepages.cwi.nl/~paulv/papers/cluster.pdf

[35] A. M. Brandmaier, "Permutation distribution clustering and structural equation model trees," Ph.D. dissertation, Universitat des Saarlandes, 2011, Accessed on 16 September 2021. [Online]. Available: http://dx.doi.org/10.22028/D291-26289

[36] A. M. Alonso, J. R. Berrendero, A. Hernández, and A. Justel, "Time series clustering based on forecast densities," *Computational Statistics & Data Analysis*, vol. 51, pp. 762–776, 2006, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1016/j.csda.2006.04.035

[37] J. A. Vilar, A. M. Alonso, and J. M. Vilar, "Non-linear time series clustering based on non-parametric forecast densities," *Computational Statistics & Data Analysis*, vol. 54, no. 11, pp. 2850–2865, 2010, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1016/j.csda.2009.02.015

[38] D. Piccolo, "A distance measure for classifying arima models," *Journal of Time Series Analysis*, vol. 11, no. 2, pp. 153–164, 1990, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1111/j.1467-9892.1990.tb00048.x

[39] E. A. Maharaj, "A significance test for classifying ARMA models," *Journal of Statistical Computation*, vol. 54, no. 4, pp. 305–331, 1996, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1080/00949659608811737

[40] Y. Kakizawa, R. H. Shumway, and M. Taniguchi, "Discrimination and clustering for multivariate time series," *Journal of the American Statistical Association*, vol. 93, no. 441, pp. 328–340, 1998, Accessed on 16 September 2021. [Online]. Available: https://www.jstor.org/stable/2669629

[41] E. A. Maharaj, "Cluster of time series," *Journal of Classification*, vol. 17, no. 2, pp. 297–314, 2000, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1007/s003570000023

[42] K. Kalpakis, D. Gada, and V. Puttagunta, "Distance measures for effective clustering of ARIMA time-series," *In data mining, 2001. ICDM 2001. Proceedings IEEE international conference*, pp. 273–280, 2000, Accessed on 16 September 2021. [Online]. Available: https://ieeexplore.ieee.org/document/989529

[43] Y. Xiong and D.-Y. Yeung, "Mixtures of arma models for model-based time series clustering," *2002 IEEE International Conference on Data Mining, 2002. Proceedings*, pp. 717–720, 2002, Accessed on 16 September 2021. [Online]. Available: https://ieeexplore.ieee.org/document/1184037

[44] M. Ramoni, P. Sebastiani, and P. Cohen, "Bayesian clustering by dynamics," *Machine Learning*, vol. 47, no. 1, pp. 91–121, 2002, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1023/A:1013635829250

[45] P. Smyth, "Clustering sequences with hidden Markov models," *In advances in neural information processing systems*, pp. 648–654, 1997, Accessed on 16 September 2021. [Online]. Available: http://papers.nips.cc/paper/1217-clustering-sequences-with-hidden-markov-models.pdf

[46] T. Oates, L. Firoiu, and P. R. Cohen, "Clustering time series with hidden Markov models and dynamic time warping," in *Proceedings of the IJCAI-99 workshop on neural, symbolic and reinforcement learning methods for sequence learning*, 1999, pp. 17–21.

[47] M. Vlachos, D. Gunopulos, and G. Das, *Indexing time-series under conditions of noise*, pp. 67–100, Accessed on 16 September 2021. [Online]. Available: https://www.worldscientific.com/doi/abs/10.1142/9789812565402_0004

[48] A. Khaleghi, D. Ryabko, J. Mary, and P. Preux, "Consistent algorithms for clustering time series," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1–32, 2016, Accessed on 16 September 2021. [Online]. Available: https://dl.acm.org/doi/pdf/10.5555/2946645.2946648

[49] Elorrieta, Felipe, Eyheramendy, Susana, Jordán, Andrés, Dékány, István, Catelan, Márcio, Angeloni, Rodolfo, Alonso-García, Javier, Contreras-Ramos, Rodrigo, Gran, Felipe, Hajdu, Gergely, Espinoza, Néstor, Saito, Roberto K., and Minniti, Dante, "A machine learned classifier for rr lyrae in the vvv survey," *A&A*, vol. 595, p. A82, 2016, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1051/0004-6361/201628700

[50] F. Elorrieta, "Classification and modeling of time series of astronomical data," Ph.D. dissertation, Pontificia Universidad Catolica de Chile, 2018, Accessed on 16 September 2021. [Online]. Available: http://www.mat.uc.cl/archivos/dip/tesis-postgrado/doctorado-eyp/classification-and-modeling-of-time-series-of-astronomical-data.pdf

[51] H. Wold, *In multivariate analysis*. New York: Wiley, 1966.

[52] H. Chun and S. Keleş, "Sparse partial least squares regression for simultaneous dimension reduction and variable selection," *Journal of the Royal Statistical Society: Series B*, vol. 72, pp. 3–25, 2010, Accessed on 16 September 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2810828/

[53] D. Chung and S. Keles, "Sparse partial least squares classification for high dimensional data," *Statistical Applications in Genetics and Molecular Biology*, vol. 9, no. 1, pp. 1–30, 2010, Accessed on 16 September 2021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/20361856/

[54] D. Nguyen and D. Rocke, "Multi-class cancer classification via partial least squares with gene expression profiles," *Bioinformatics*, vol. 18, no. 9, pp. 1216–1226, 2002, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1093/bioinformatics/18.9.1216

[55] D. V. Nguyen and D. M. Rocke, "Tumor classification by partial least squares using microarray gene expression data," *Bioinformatics*, vol. 18, no. 1, pp. 39–50, 1 2002, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1093/bioinformatics/18.1.39

[56] A.-L. Boulesteix, "PLS dimension reduction for classification with microarray data," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, pp. 1–30, 2004, Accessed on 16 September 2021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/16646813/

[57] B. D. Marx, "Iteratively reweighted partial least squares estimation for generalized linear regression," *Technometrics*, vol. 38, no. 4, pp. 374–381, 1996, Accessed on 16 September 2021. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/00401706.1996.10484549

[58] B. Ding and R. Gentleman, "Classification using generalized partial least squares," *Journal of Computational and Graphical Statistics*, vol. 14, no. 2, pp. 280–298, 2005, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1198/106186005X47697

[59] Y. Z. A.I. McLeod and X. Changjiang, “Package ’FitAR’,” https://cran.r-project.org/web/packages/FitAR/index.html, 2013, Accessed on 16 September 2021.

[60] G. Fort and S. Lambert-Lacroix, “Classification using partial least squares with penalized logistic regression,” *Bioinformatics*, vol. 21, no. 7, p. 1104, 2004, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1093/bioinformatics/bti114

[61] K.-A. Lê Cao, D. Rossouw, C. Robert-Granié, and P. Besse, “A sparse PLS for variable selection when integrating omics data,” *Statistical Applications in Genetics and Molecular Biology*, vol. 7, no. 1, pp. 1–29, 2008, Accessed on 16 September 2021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/19049491/

[62] S. Waaijenborg, P. C. Verselewel de Witt Hamer, and A. H. Zwinderman, “Quantifying the Association between Gene Expressions and DNA-Markers by Penalized Canonical Correlation Analysis,” *Statistical Applications in Genetics and Molecular Biology Genetics and Molecular Biology*, vol. 7, no. 1, pp. 1–27, 2008, Accessed on 16 September 2021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/18241193/

[63] K.-A. Lê Cao, S. Boitard, and P. Besse, “Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems,” *Journal of Clinical Bioinformatics*, vol. 12, no. 1, 2011, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1186/1471-2105-12-253

[64] K.-A. Lê Cao, P. G. P. Martin, C. Robert-Granie, and P. Besse, “Sparse canonical methods for biological data integration: application to a cross-platform study,” *Journal of Clinical Bioinformatics*, vol. 10, 2009, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1186/1471-2105-10-34

[65] M. Tenenhaus, *La régression PLS: thórie et pratique*. Paris, France: Editions Technip, 1998.

[66] H. Shen and J. Z. Huang, “Sparse principal component analysis via regularized low rank matrix approximation,” *Journal of Multivariate Analysis*, vol. 99, no. 6, pp. 1015–1034, 2008, Accessed on 16 September 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0047259X07000887

[67] K.-A. Lê Cao and S. Dejean, “mixOmics: Omics Data Integration Project,” http://www.bioconductor.org/packages/release/bioc/html/mixOmics.html, Viena,Austria, 2020, Accessed on 16 September 2021.

[68] F. Rohart, B. Gautier, A. Singh, and K.-A. Lê Cao, “mixOmics: An R package for omics feature selection and multiple data integration,” *PLOS Computational Biology.*, vol. 13, no. 11, pp. 1–19, 2017, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1005752

[69] F. Bach, “Model-consistent sparse estimation through the bootstrap,” 2009, Accessed on 16 September 2021. [Online]. Available: https://arxiv.org/abs/0901.3202

[70] N. Meinshausen and P. Bühlmann, “Stability selection,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 4, pp. 417–473, 9 2010, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1111/j.1467-9868.2010.00740.x

[71] K.-A. Lê Cao, E. Meugnier, and G. J. McLachlan, "Integrative mixture of experts to combine clinical factors and gene markers," *Bioinformatics (Oxford, England)*, vol. 26, no. 9, pp. 1192–1198, 5 2010, Accessed on 16 September 2021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/20223834https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2859127/

[72] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 1, pp. 123–140, 1996, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1007/BF00058655

[73] Y. Freund and R. E. Schapire, "Experiments with a New Boosting Algorithm," in *Machine Learning: Proceedings of the Thirteenth International Conference*, Bari, Italy, 1966, pp. 1–9, Accessed on 16 September 2021. [Online]. Available: https://cseweb.ucsd.edu/~yfreund/papers/boostingexperiments.pdf

[74] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[75] I. T. Utami, B. Sartono, and K. Sadik, "Comparison of single and ensemble classifiers of support vector machine and classification tree," *Journal of Mathematical Sciences and Applications*, vol. 2, no. 2, pp. 17–20, 2014, Accessed on 16 September 2021. [Online]. Available: http://pubs.sciepub.com/jmsa/2/2/1

[76] E. Verticchio, F. Frasca, F.-D. García-Diego, and A. M. Siani, "Investigation on the use of passive microclimate frames in view of the climate change scenario," *Climate*, vol. 7, no. 8, 2019, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.3390/cli7080098

[77] E. Sesana, A. S. Gagnon, C. Bertolin, and J. Hughes, "Adapting cultural heritage to climate change risks: perspectives of cultural heritage experts in Europe," *Geosciences*, vol. 8, no. 8, pp. 1–23, 2018, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.3390/geosciences8080305

[78] D. Camuffo, *Microclimate for cultural heritage*, 1st ed. Amsterdam: Elsevier, 1998.

[79] D. Camuffo, "Indoor dynamic climatology: investigations on the interactions between walls and indoor environment," *Atmospheric Environment (1967)*, vol. 17, no. 9, pp. 1803–1809, 1983, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1016/0004-6981(83)90188-9

[80] D. Camuffo and A. Bernardi, "Study of the microclimate of the hall of the giants in the carrara palace in Padua," *Studies in Conservation*, vol. 40, no. 4, pp. 237–249, 1995, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1179/sic.1995.40.4.237

[81] D. Camuffo, E. Pagan, A. Bernardi, and F. Becherini, "The impact of heating, lighting and people in re-using historical buildings: a case study," *Journal of Cultural Heritage*, vol. 5, no. 4, pp. 409—416, 2004, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1016/j.culher.2004.01.005

[82] A. Bernardi, "Microclimate in the British museum, London," *Museum Management and Curatorship*, vol. 9, no. 2, pp. 169–182, 1990, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1016/0964-7775(90)90055-C

[83] A. Bernardi and D. Camuffo, “Microclimate in the chiericati palace municipal museum Vicenza,” *Museum Management and Curatorship*, vol. 14, no. 1, pp. 5–18, 1995, Accessed on 16 September 2021. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/09647779509515423

[84] D. Camuffo, A. Bernardi, G. Sturaro, and A. Valentino, “The microclimate inside the Pollaiolo and Botticelli rooms in the Uffizi Gallery,” *Journal Culture Heritage*, vol. 3, no. 2, pp. 155–161, 2002, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1016/S1296-2074(02)01171-8

[85] P. Merello, F.-D. García-Diego, P. Beltrán, and C. Scatigno, “High frequency data acquisition system for modelling the impact of visitors on the thermo-hygrometric conditions of archaeological sites: a Casa di Diana (Ostia Antica, Italy) case study,” *Sensors*, vol. 18, no. 2, p. 348, 2018, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.3390/s18020348

[86] F. Frasca, A. M. Siani, G. R. Casale, M. Pedone, L. Bratasz, M. Strojecki, and A. Mleczkowska, “Assessment of indoor climate of Mogiła Abbey in Kraków (Poland) and the application of the analogues method to predict microclimate indoor conditions,” *Environmental Science and Pollution Research*, vol. 24, no. 16, pp. 13 895–13 907, 2017. [Online]. Available: https://doi.org/10.1007/s11356-016-6504-9

[87] Y. Tabunschikov and M. Brodatch, “Indoor air climate requirements for Russian churches and cathedrals,” *Indoor Air Journal*, vol. 14 Suppl 7, pp. 168–174, 11 2004, Accessed on 16 September 2021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/15330784/

[88] D. Camuffo, G. Sturaro, and A. Valentino, “Thermodynamic exchanges between the external boundary layer and the indoor microclimate at the basilica of Santa Maria Maggiore, Rome, Italy: the problem of conservation of ancient works of art,” *Boundary-Layer Meteorology*, vol. 92, pp. 243–262, 11 1999.

[89] E. Vuerich, F. Malaspina, M. Barazutti, T. Georgiadis, and M. Nardino, “Indoor measurements of microclimate variables and ozone in the church of San Vincenzo (Monastery of Bassano Romano Italy): a pilot study,” *Microchemical Journal*, vol. 88, no. 2, pp. 218–223, 2008, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1016/j.microc.2007.11.014

[90] L. G, E. Charpantidou, I. Kioutsioukis, and S. Rapsomanikis, “Indoor microclimate, ozone and nitrogen oxides in two medieval churches in Cyprus,” *Atmospheric Environment*, vol. 40, no. 39, pp. 7457–7466, 2006, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1016/j.atmosenv.2006.07.015

[91] A. Bernardi, V. Todorov, and J. Hiristova, “Microclimatic analysis in St. Stephan’s church, Nessebar, Bulgaria after interventions for the conservation of frescoes,” *Journal of Cultural Heritage*, vol. 1, pp. 281–286, 11 2000.

[92] EN16883, “Conservation of cultural heritage. Guidelines for improving the energy performance of historic buildings,” https://standards.iteh.ai/catalog/standards/cen/189eac8d-14e1-4810-8ebd-1e852b3effa3/en-16883-2017, 2017, Accessed on 16 September 2021.

[93] EN16141, "Conservation of cultural heritage. Guidelines for management of environmental conditions. Open storage facilities: definitions and characteristics of collection centres dedicated to the preservation and management of cultural heritage," https://standards.iteh.ai/catalog/standards/cen/452b6461-3ba0-4f70-8007-f6702a7e804d/en-16141-2012, p. 18, 2012, Accessed on 16 September 2021.

[94] EN16242, "Conservation of cultural heritage. Procedures and instruments for measuring humidity in the air and moisture exchanges between air and cultural property," https://standards.iteh.ai/catalog/standards/cen/12d5d8fb-f047-4d49-a750-9349dff5f7f4/en-16242-2012, p. 34, 2012, Accessed on 16 September 2021.

[95] EN15898, "Conservation of cultural property. Main general terms and definitions," https://standards.iteh.ai/catalog/standards/cen/d8ac0bbb-7dfb-4c68-a51b-9c0b66ae8f63/en-15898-2011, p. 64, 2019, Accessed on 16 September 2021.

[96] EN15758, "Conservation of cultural property. Procedures and instruments for measuring temperatures of the air and the surfaces of objects," https://standards.iteh.ai/catalog/standards/cen/8e796a23-395f-4ed2-8154-33ef479b3231/en-15758-2010, p. 19, 2010, Accessed on 16 September 2021.

[97] EN15757, "Conservation of cultural property. Specifications for temperature and relative humidity to limit climate-induced mechanical damage in organic hygroscopic materials," https://standards.iteh.ai/catalog/standards/cen/ad03d50b-22dc-4c57-b198-2321863f3870/en-15757-2010, p. 18, 2010, Accessed on 16 September 2021.

[98] EN16893, "Conservation of cultural heritage. Specifications for location, construction and modification of buildings or rooms intended for the storage or use of heritage collections," https://standards.iteh.ai/catalog/standards/cen/b53c7f39-b809-4998-be91-31f362a8a2cb/en-16893-2018, p. 58, 2018, Accessed on 16 September 2021.

[99] S. P. Corgnati and M. Filippi, "Assessment of thermo-hygrometric quality in museums: Method and in-field application to the Duccio di Buoninsegna exhibition at Santa Maria della Scala (Siena, Italy)," *Journal of Cultural Heritage*, vol. 11, no. 3, pp. 345–349, 11 2010, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1016/j.culher.2009.05.003

[100] Collectioncare.eu, "https://www.collectioncare.eu/ Accessed on 16 November," 11 2020. [Online]. Available: https://www.collectioncare.eu/

[101] R. Dunia, S. J. Qin, T. F. Edgar, and T. J. McAvoy, "Use of principal component analysis for sensor fault identification," *Computers and Chemical Engineering*, vol. 20, pp. S713–S718, 1996.

[102] D. Zhu, J. Bai, and S. Yang, "A multi-fault diagnosis method for sensor systems based on principle component analysis," *Sensors*, vol. 10, no. 1, pp. 241–253, 2009, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.3390/s100100241

[103] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control (Revised Edition)*. Holden-Day, 1976.

[104] P. J. Brockwell and R. A. Davis, *Time series: theory and methods*, 2nd ed. NY, USA: Springer: New York, 1987.

[105] S. Kovalevsky, “Package ‘QuantTools’,” https://quanttools.bitbucket.io/_site/index.html, 2020, Accessed on 16 September 2021.

[106] E. J. Hannan and J. Rissanen, “Recursive Estimation of Mixed Autoregressive-Moving Average Order,” *Biometrika*, vol. 69, no. 1, 4 1982, Accessed on 16 September 2021. [Online]. Available: https://www.jstor.org/stable/2335856

[107] J. D. Hamilton, *Time series analysis*, 5th ed. NJ, USA: Princeton University Press:Princeton, 1994.

[108] R Core Team, “R: a language and environment for statistical computing,” https://www.r-project.org/about.html, Vienna, Austria, 2014, Accessed on 16 September 2021.

[109] D. Qiu, “Package ‘aTSA’,” https://cran.r-project.org/web/packages/aTSA/index.html, 2015, Accessed on 16 September 2021.

[110] R. J. Hyndman and Y. Khandakar, “Automatic time series forecasting: the forecast package for R,” *Journal of Statistical Software*, vol. 26, no. 3, 2008, Accessed on 16 September 2021. [Online]. Available: https://www.jstatsoft.org/article/view/v027i03

[111] R. Hyndman, G. Athanasopoulos, C. Bergmeir, G. Caceres, L. Chhay, M. OHara-Wild, F. Petropoulos, S. Razbash, E. Wang, and F. Yasmeen, “Package Forecast: forecasting functions for time series and linear models,” https://cran.r-project.org/web/packages/forecast/, 2020, Accessed on 16 September 2021.

[112] A. Ghalanos and T. Kley, “Package ‘rugarch’,” 2020, Accessed on 16 September 2021. [Online]. Available: https://cran.r-project.org/web/packages/rugarch/index.html

[113] A. Zeileis, “Implementing a class of structural change tests: an econometric computing approach,” *Computational Statistics & Data Analysis*, vol. 50, no. 11, pp. 2987–3008, 2006, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1016/j.csda.2005.07.001

[114] A. Trapletti and K. Hornik, “Package tseries: time series analysis and computational finance,” https://cran.r-project.org/web/packages/tseries/index.html, 2019, Accessed on 16 September 2021.

[115] F. Leisch, K. Hornik, and C.-M. Kuan, “Monitoring structural changes with the generalized fluctuation test,” *Journal of Economic Theory*, vol. 16, no. 6, pp. 835–854, 2000, Accessed on 3 October 2021. [Online]. Available: https://doi.org/10.1017/S0266466600166022

[116] G. C. Chow, “Tests of equality between sets of coefficients in two linear regressions,” *Econometrica*, vol. 28, no. 3, pp. 591–605, 1960, Accessed on 16 September 2021. [Online]. Available: https://www.jstor.org/stable/1910133

[117] B. E. Hansen, “Tests for parameter instability in regressions with I(1) processes,” *Journal of Business and Economic Statistics*, vol. 20, no. 1, pp. 45–59, 2002, Accessed on 16 September 2021. [Online]. Available: http://www.jstor.org/stable/1392149

[118] D. W. K. Andrews, “Tests for parameter instability and structural change with unknown change point,” *Econometrica*, vol. 61, no. 4, pp. 821–856, 1993, Accessed on 16 September 2021. [Online]. Available: https://www.jstor.org/stable/2951764

[119] D. W. K. Andrews and W. Ploberger, "Optimal tests when a nuisance parameter is present only under the alternative," *Econometrica*, vol. 62, no. 6, pp. 1383–1414, 1994, Accessed on 16 September 2021. [Online]. Available: https://www.jstor.org/stable/2951753

[120] A. Zeileis, F. Leisch, K. Hornik, and C. Kleiber, "Strucchange: an R package for testing for structural change in linear regression models," *Journal of Statistical Software*, vol. 7, no. 2, pp. 1–38, 2002, Accessed on 16 September 2021. [Online]. Available: https://www.jstatsoft.org/article/view/v007i02

[121] J. D. Cryer and K.-S. Chan, *Time series analysis: with applications in R*, springer s ed., ser. Springer texts in statistics. Springer New York, 2008.

[122] J. P. Royston, "Algorithm AS 181: the W test for normality," *Journal of the Royal Statistical Society Series C (Applied Statistics)*, vol. 31, no. 2, pp. 176–180, 1982, Accessed on 16 September 2021. [Online]. Available: https://www.jstor.org/stable/2347986

[123] D. Gaetano, "Forecast combinations in the presence of structural breaks: evidence from U.S. equity markets," *Mathematics*, vol. 6, no. 3, p. 34, 11 2018, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.3390/math6030034

[124] P. S. P. Cowpertwait and A. V. Metcalfe, *Introductory time series with R*, ser. Springer Series: Use R. NY, USA: Springer: New York, 2009.

[125] R. Hyndman, A. B. Koehler, J. K. Ord, and R. D. Snyder, *Forecasting with exponential smoothing: The State Space Approach*. Germany: Springer: Berlin, 2008.

[126] G. M. Ljung and G. E. P. Box, "On a measure of Lack of fit in time series models," *Biometrika*, vol. 65, no. 2, pp. 297–303, 1978, Accessed on 16 September 2021. [Online]. Available: https://www.jstor.org/stable/2335207

[127] W. J. Conover, *Practical nonparametric statistics*, 3rd ed. Hoboken, NJ, USA: John Wiley and Sons, INC, 1999. [Online]. Available: https://books.google.es/books?id=UBV2VwCxrMcC

[128] J. P. Royston, "An extension of Shapiro and Wilk's W test for normality to large samples," *Journal of the Royal Statistical Society Series C (Applied Statistics)*, vol. 31, no. 2, pp. 115–124, 1982, Accessed on 16 September 2021. [Online]. Available: https://www.jstor.org/stable/2347973

[129] P. Royston, "Remark AS R94: A remark on algorithm AS 181: the W-test for normality," *Journal of the Royal Statistical Society Series C (Applied Statistics)*, vol. 44, no. 4, pp. 547–551, 1995, Accessed on 16 September 2021. [Online]. Available: http://www.jstor.org/stable/2986146

[130] O. D. Anderson, "Time series analysis: forecasting and control (revised edition)," *Journal of the Franklin Institute*, vol. 310, no. 2, 1980.

[131] S. R. Yaziz, N. A. Azizan, R. Zakaria, and M. H. Ahmad, "The performance of hybrid ARIMA-GARCH modeling in forecasting gold price," in *20th International Congress on Modelling and Simulation*, Adelaide, Australia, 2013, pp. 1–6. [Online]. Available: http://malrep.uum.edu.my/rep/Record/my.ump.umpir.4260

[132] J.-J. Tseng and S.-P. Li, "Quantifying volatility clustering in financial time series," *International Review of Financial Analysis*, vol. 23, pp. 11–19, 2012, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1016/j.irfa.2011.06.017

[133] R. Engle, "Risk and volatility: econometric models and financial practice," *American Economic Review*, vol. 94, no. 3, pp. 405–420, 2004, Accessed on 16 September 2021. [Online]. Available: https://www.jstor.org/stable/3592935

[134] R. F. Engle and T. Bollerslev, "Modelling the persistence of conditional variances," *Econometric Reviews*, vol. 5, no. 1, pp. 1–50, 1986, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1080/07474938608800095

[135] F. Yusof and I. L. Kane, "Volatility modeling of rainfall time series," *Theoretical and Applied Climatology*, vol. 113, pp. 247–258, 2013, Accessed on 16 September 2021. [Online]. Available: https://booksc.org/book/21722750/635798

[136] R. S. Yaziz, N. A. Azlinna, M. Ahmad, and R. Zakaria, "Modelling gold price using ARIMA-TGARCH," *Applied Mathematical Sciences*, vol. 10, pp. 1391–1402, 2016, Accessed on 16 September 2021. [Online]. Available: http://www.m-hikari.com/ams/ams-2016/ams-25-28-2016/511716.html

[137] C.-L. Hor, S. J. Watson, and S. Majithias, "Daily load forecasting and maximum demand estimation using ARIMA and GARCH," in *International Conference on Probabilistic Methods Applied to Power Systems*, Stockholm, 2006, pp. 1–6, Accessed on 16 September 2021. [Online]. Available: https://ieeexplore.ieee.org/document/4202249

[138] J. Xing, "The research on stock market volatility in China based on the model of ARIMA-EARCH-M (1, 1) and ARIMA-TARCH-M (1, 1)," *International Conference on Advances in Education and Management. ISAEBD 2011: Education and Management*, vol. 210, pp. 521–527, 2011, Accessed on 16 September 2021. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-23065-3_75

[139] J.-M. Zakoian, "Threshold heteroskedastic models," *Journal of Economic Dynamics and Control*, vol. 18, no. 5, pp. 931–955, 1994, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1016/0165-1889(94)90039-6

[140] T. Bollerslev, "A conditionally heteroskedastic time series model for speculative prices and rates of return," *The Review of Economics and Statistics*, vol. 69, no. 3, pp. 542–547, 1987, Accessed on 16 September 2021. [Online]. Available: https://www.jstor.org/stable/1925546

[141] W. A. Fuller, *Introduction to statistical time series*, 2nd ed. New York, NY, USA: John Wiley and Sons, 1996.

[142] H. Lütkepohl and F. Xu, "The role of the log transformation in forecasting economic variables," *Empirical Economics*, vol. 42, no. 3, pp. 619–638, 2012, Accessed on 16 September 2021. [Online]. Available: https://ssrn.com/abstract=1368131

[143] A. I. McLeod and W. K. Li, "Diagnostic cheking ARMA time series models using squared-residual autocorrelations," *Journal of Time Series Analysis*, vol. 4, no. 4, pp. 269–273, 7 1983, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1111/j.1467-9892.1983.tb00373.x

[144] H. Wold, “Path models with latent variables: the NIPALS approach,” in *Quantitative Sociology*. Elsevier, 1975, pp. 307–357, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1016/B978-0-12-103950-9.50017-4

[145] E. A. Maharaj, “Comparison of non-stationary time series in the frequency domain,” *Computational Statistics & Data Analysis*, vol. 40, no. 1, pp. 131–141, 2002, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1016/S0167-9473(01)00100-1

[146] J. A. Vilar and S. Pértega, “Discriminant and cluster analysis for gaussian stationary processes: local linear fitting approach,” *International Review of Financial Analysis*, vol. 16, no. 3-4, pp. 443–462, 2004, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1080/10485250410001656453

[147] M. Li and P. Vitãnyi, “An introduction to kolmogorov complexity and its applications,” *Text and Monographs in Computer Science*, 2007, Accessed on 16 September 2021. [Online]. Available: https://www.springer.com/gp/book/9781489984456

[148] E. Keogh, S. Lonardi, C. A. Ratanamahatana, L. Wei, S.-H. Lee, and J. Handley, “Compression based data mining of sequential data,” *Data Mining and Knowledge Discovery*, vol. 14, no. 1, pp. 99–129, 2007, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1007/s10618-006-0049-3

[149] E. Otranto, “Clustering heteroskedastic time series by model-based procedures,” *Computational Statistics & Data Analysis*, vol. 52, no. 10, pp. 4685–4698, 2008, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1016/j.csda.2008.03.020

[150] A. Ioannou, K. Fokianos, and V. J. Promponas, “Spectral density ratio based clustering methods for the binary segmentation of protein sequences: A comparative study,” *Biosystems*, vol. 100, no. 2, pp. 132–143, 2010, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1016/j.biosystems.2010.02.008

[151] A. C. Harvey and N. Shephard, “10 Structural time series models,” in *Econometrics*, ser. Handbook of Statistics. Elsevier, 1993, vol. 11, pp. 261–302, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1016/S0169-7161(05)80045-8

[152] P. Bühlmann and A. J. McNeil, “An algorithm for nonparametric GARCH modelling,” *Computational Statistics & Data Analysis*, vol. 40, no. 4, pp. 665–683, 2002, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1016/S0167-9473(02)00080-4

[153] N. Rohan and T. V. Ramanathan, “Nonparametric estimation of a time-varying GARCH model,” *International Review of Financial Analysis*, vol. 25, no. 1, pp. 33–52, 3 2013, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1080/10485252.2012.728600

[154] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *Journal of the Royal Statistical Society: Series B*, vol. 67, no. 2, pp. 301–320, 2005, Accessed on 16 September 2021. [Online]. Available: https://www.jstor.org/stable/3647580

[155] G. Pavlogeorgatos, “Environmental parameters in museums,” *Building and Environment*, vol. 38, no. 12, pp. 1457–1462, 2003, Accessed on 16 September 2021. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360132303001136

[156] Falconaumanni, “Plan of the L’Amoina museum licensed under the Creative Commons Attribution-Share Alike 3.0,” https://commons.wikimedia.org/wiki/File:Plano_Almoina_recorrido.png, 2019, Accessed on 16 September 2021.

[157] A. Fernández-Navajas, P. Merello, P. Beltran, and F.-J. García-Diego, “Multivariate thermo-hygrometric characterisation of the archaeological site of Plaza de L’Almoina (Valencia, Spain) for preventive conservation,” *Sensors*, vol. 13, no. 8, pp. 9729–9746, 7 2013, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.3390/s130809729

[158] P. Merello, A. Fernandez-Navajas, J. Curiel-Esparza, M. Zarzo, and F.-J. García-Diego, “Characterisation of thermo-hygrometric conditions of an archaeological site affected by unlike boundary weather conditions,” *Building and Environment*, vol. 76, pp. 125–133, 6 2014, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1016/j.buildenv.2014.03.009

[159] Ministero per i Beni e le Attività Culturali. DM 10/2001, “Atto di Indirizzo sui Criteri Tecnico-scientifici e Sugli Standard di Funzionamento e Sviluppo dei Musei,” https://www.veneto.beniculturali.it/normativa-e-disposizioni/atto-di-indirizzo-sui-criteri-tecnico%E2%80%93scientifici-e-sugli-standard-di, 10 2001, Accessed on 16 September 2021.

[160] UNI Italian Standard 10829, “Works of Art of Historical Importance-Ambient Conditions for the Conservation-Measurement and Analysis,” Milano, Italy, 1999, Accessed on 18 October 2021. [Online]. Available: http://www.brescianisrl.it/newsite/ita/xprodotti.php?sub_content=1&sub_prodotto=47&hash=db99bf602ca201c16ad187434828d334

[161] Ramírez, Sandra, M. Zarzo, A. Perles, and F.-J. Garcia-Diego, “Methodology for discriminant time series analysis applied to microclimate monitoring of fresco paintings,” *Sensors*, vol. 21, no. 2, pp. 1–28, 2021, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.3390/s21020436

[162] S. Ramírez, M. Zarzo, A. Perles, and F.-J. Garcia-Diego, “sPLS-DA to discriminate time series,” in *JSM Proceedings, Statistics and the Environment*. Alexandria, VA: American Statistical Association.: JSM2020, 2021, pp. 107–135.

[163] Z. Huijbregts, R. Kramer, M. Martens, A. van Schijndel, and H. Schellen, “A proposed method to assess the damage risk of future climate change to museum objects in historic buildings,” *Building and Environment*, vol. 55, pp. 43–56, 9 2012, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1016/j.buildenv.2012.01.008

[164] P. Zítek and T. Vyhlídal, “Model-based moisture sorption stabilization in historical buildings,” *Building and Environment*, vol. 44, no. 6, pp. 1181–1187, 6 2009, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1016/j.buildenv.2008.08.014

[165] J. Stewart, S. Julien, and S. Staniforth, “An integrated monitoring strategy at Chedworth Roman Villa (Gloucestershire),” in *Preserving archaeological remains in situ?: Proceedings of the 2nd conference 12-14 September 2001*. London: Museum of London Archaeology Service, 2004.

[166] Collectioncare.eu, “European Horizon 2020 project Collectioncare: Innovative and affordable service for the Preventive Conservation monitoring of individual Cultural Artefacts during display, storage, handling and transport.” https://www.collectioncare.eu/, 2020, Accessed on 16 September 2021.

[167] testo, “WLAN data loggers from Testo for monitoring temperature and ambient conditions,” https://www.testo.com/en-US/products/wifi-datamonitoring, 2021, Accessed on 16 September 2021.

[168] L. Komsta and F. Novomestky, “Package ’moments’,” https://cran.r-project.org/web/packages/moments/index.html, 2015, Accessed on 16 September 2021.

[169] B. G. Peterson and C. Peter, “Package ’PerformanceAnalytics’: Econometric Tools for Performance and Risk Analysis,” https://cran.r-project.org/web/packages/PerformanceAnalytics/index.html, 2020, Accessed on 16 September 2021.

[170] M. Charrad, N. Ghazzali, V. Boiteau, and A. Niknafs, “Package ’NbClust’: Determining the Best Number of Clusters in a Data Set,” https://cran.r-project.org/web/packages/NbClust/index.html, 2015, Accessed on 16 September 2021.

[171] R Core Team and contributors worldwide, “Package ’stats’,” https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html, 2020, Accessed on 16 September 2021.

[172] A. Taylor B and J. W. Emerson, “Package ’dgof’,” https://cran.r-project.org/web/packages/dgof/index.html, 2013, Accessed on 16 September 2021.

[173] G. E. P. Box and D. R. Cox, “An analysis of transformations,” *Journal of the Royal Statistical Society: Series B*, vol. 26, no. 2, pp. 211–243, 7 1964. [Online]. Available: https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

[174] C. Bacon, *Practical portfolio performance measurement and attribution*, 2nd ed. New Delhi,India: John Wiley and Sons, 2008.

[175] P. J. Ribeiro Jr, P. J. Diggle, O. Christensen, M. Schlather, R. Bivand, and B. Ripley, “Package ’geoR’,” https://cran.r-project.org/web/packages/geoR/index.html, 2020, Accessed on 16 September 2021.

[176] S. Dray, N. Pettorelli, and D. Chessel, “Multivariate analysis of incomplete mapped data,” *Transactions in GIS*, vol. 7, no. 3, pp. 411–422, 6 2003, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1111/1467-9671.00153

[177] M. Charrad, N. Ghazzali, V. Boiteau, and A. Niknafs, “NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set,” *Journal of Statistical Software*, vol. 1, no. 6, 2014, Accessed on 16 September 2021. [Online]. Available: https://www.jstatsoft.org/v061/i06

[178] J. A. Hartigan and M. A. Wong, “Algorithm AS 136: A K-Means Clustering Algorithm,” *Journal of the Royal Statistical Society Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 2 1979, Accessed on 16 September 2021. [Online]. Available: http://www.jstor.org/stable/2346830

[179] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, ser. Fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, Calif.: University of California Press, 1967, pp. 281–297, Accessed on 16 September 2021. [Online]. Available: https://projecteuclid.org/euclid.bsmsp/1200512992

[180] S. Lloyd, “Least squares quantization in PCM,” *IEEE Transactions on Information Theory*, vol. 28, pp. 129–136, 1982.

[181] E. Forgy, “Cluster analysis of multivariate data : efficiency versus interpretability of classifications,” *Biometrics*, vol. 21, pp. 768–769, 1965.

[182] R. Tibshirani, “Regression shrinkage and selection via the Lasso,” *Journal of the Royal Statistical Society: Series B*, vol. 58, pp. 267–288, 1996, Accessed on 16 September 2021. [Online]. Available: http://www.jstor.org/stable/2346178

[183] J. A. Wegelin, “A survey of Partial Least Squares (PLS) methods, with emphasis on the two-block case,” University of Washington, Seattle, Tech. Rep., 2000, Accessed on 16 September 2021. [Online]. Available: www.stat.washington.edu

[184] S. De Jong, “SIMPLS: An alternative approach to partial least squares regression,” *Chemometrics and Intelligent Laboratory Systems*, vol. 18, no. 3, pp. 251–263, 1993, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1016/0169-7439(93)85002-XGetrightsandcontent

[185] A. Lorber, L. E. Wangen, and B. R. Kowalski, “A theoretical foundation for the PLS algorithm,” *Journal of Chemometrics*, vol. 1, no. 1, pp. 19–31, 1 1987, accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1002/cem.1180010105

[186] T. Mehmood, K. H. Liland, L. Snipen, and S. Sæbø, “A review of variable selection methods in Partial Least Squares Regression,” *Chemometrics and Intelligent Laboratory Systems*, vol. 118, pp. 62–69, 2012, Accessed on 16 September 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169743912001542

[187] Sartorius Stedim Data Analytics AB, *SIMCA 15 Multivariate data analysis solution user guide*. Umeå, Sweden: Sartorius Stedim Data Analytics AB, 2017, Accessed on 16 September 2021. [Online]. Available: www.umetrics.com

[188] H. Han, X. Guo, and H. Yu, “Variable selection using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest,” in *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, 2016, pp. 219–224.

[189] K. P. Murphy, *Machine learning a probabilistic perspective*. Cambridge, Massachusetts and London, England: The MIT Press, 2012, Accessed on 16 September 2021. [Online]. Available: https://www.semanticscholar.org/paper/Machine-learning-a-probabilistic-perspective-Murphy/25badc676197a70aaf9911865eb03469e402ba57

[190] B. H. Menze, B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, and F. A. Hamprecht, “A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data,” *BMC Bioinformatics*, vol. 10, no. 213, 2009, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1186/1471-2105-10-213

[191] A. Liaw and M. Wiener, “Package randomForest: Breiman and Cutler’s Random Forests for Classification and Regression,” https://cran.r-project.org/web/packages/randomForest/index.html, 2018, Accessed on 16 September 2021.

[192] A. Guha, N. Ho, and X. Nguyen, “On posterior contraction of parameters and interpretability in Bayesian mixture modeling,” https://arxiv.org/abs/1901.05078, pp. arXiv:1901.05 078–arXiv:1901.05 078, 2019, Accessed on 16 September 2021.

[193] N. Aste, R. Adhikari, M. Buzzetti, S. Della Torre, C. Del Pero, H. Huerto C, and F. Leonforte, "Microclimatic monitoring of the duomo (Milan cathedral): Risks-based analysis for the conservation of its cultural heritage," *Building and Environment*, vol. 148, pp. 240–257, 2019, Accessed on 16 September 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360132318307108

[194] C. Hall, A. Hamilton, W. D. Hoff, H. A. Viles, and J. A. Eklund, "Moisture dynamics in walls: response to micro-environment and climate change," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 467, no. 2125, pp. 194–211, 1 2011, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1098/rspa.2010.0131

[195] E. Lucchi and F. Roberti, "Diagnosi, simulazione e ottimizzazione energetica degli edifici storici energy diagnosis, simulation and optimization of historic buildings," 10 2015, Accessed on 16 September 2021. [Online]. Available: https://www.researchgate.net/publication/311846765_Diagnosi_simulazione_e_ottimizzazione_energetica_degli_edifici_storici_Energy_diagnosis_simulation_and_optimization_of_historic_buildings

[196] A. Adán, V. Pérez, J.-L. Vivancos, C. Aparicio-Fernández, and S. A. Prieto, "Proposing 3D Thermal Technology for Heritage Building Energy Monitoring," 2021, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.3390/rs13081537

[197] D. Camuffo, E. Pagan, S. Rissanen, Ł. Bratasz, R. Kozłowski, M. Camuffo, and A. della Valle, "An advanced church heating system favourable to artworks: A contribution to European standardisation," *Journal of Cultural Heritage*, vol. 11, no. 2, pp. 205–219, 2010, Accessed on 16 September 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S129620740900106X

[198] M. F. Mecklenburg, C. S. Tumosa, and W. D. Erhardt, *Structural response of painted wood surfaces to changes in ambient relative humidity in painted wood: history and conservation, Dorge, V. and Howlett, F. C.* Los Angeles, CA, USA: The Getty Conservation Institute, 1998. [Online]. Available: https://repository.si.edu/handle/10088/35942

[199] L. Bratasz, R. Koztowski, D. Camuffo, and E. Pagan, "Impact of indoor heating on painted wood - monitoring the altarpiece in the church of santa maria maddalena in rocca pietore, italy," *Studies in Conservation*, vol. 52, no. 3, pp. 199–210, 2007. [Online]. Available: https://doi.org/10.1179/sic.2007.52.3.199

[200] L. Bratasz, D. Camuffo, and R. Kozlowski, "Target microclimate for preservation derived from past indoor conditions," in *Contributions to the museum microclimates conference. Copenhagen: The National Museum of Denmark*, 11 2007, pp. 129–134. [Online]. Available: https://www.researchgate.net/publication/239580339_Target_microclimate_for_preservation_derived_from_past_indoor_conditions

[201] H. E. Silva and F. M. Henriques, "Microclimatic analysis of historic buildings: A new methodology for temperate climates," *Building and Environment*, vol. 82, pp. 381–387, 2014, Accessed on 16 September 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360132314002972

[202] weatherbase, "Weatherbase," https://www.weatherbase.com/weather/weather-summary.php3?s=48280&cityname=Valencia,+Spain, 2021, Accessed on 3 October 2021.

[203] F.-J. García-Diego, E. Verticchio, P. Beltrán, and A. M. Siani, "Assessment of the minimum sampling frequency to avoid measurement redundancy in microclimate field surveys in museum buildings," pp. 1–17, 2016, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.3390/s16081291

[204] J. Herráez, M. Enríquez de Salamanca, T. Pastor Arenas, and M. Gil Muñoz, "Manual de seguimiento y análisis de condiciones ambientales. Plan Nacional de Conservación Preventiva," https://www.libreria.culturaydeporte.gob.es/libro/manual-de-seguimiento-y-analisis-de-condiciones-ambientales_2654/, Spain, 2014, Accessed on 16 September 2021.

[205] S. Ramírez, M. Zarzo, and F.-J. García-Diego, "Multivariate time series analysis of temperatures in the archaeological museum of l'almoina (valencia, spain)," *Sensors*, vol. 21, no. 13, 2021, Accessed on 16 September 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/13/4377

[206] L. J. Klein, S. A. Bermudez, A. G. Schrott, M. Tsukada, P. Dionisi-Vici, L. Kargere, F. Marianno, H. F. Hamann, V. López, and M. Leona, "Wireless sensor platform for cultural heritage monitoring and modeling system," *Sensors*, vol. 17, no. 9, 2017, Accessed on 16 September 2021. [Online]. Available: https://www.mdpi.com/1424-8220/17/9/1998

[207] D. Camuffo, "The Friendly Heating Project and the Conservation of the Cultural Heritage Preserved in Churches," in *Developments in Climate Control of Historic Buildings. Proceedings from the International Conference "Climatization of Historic Buildings, State of the Art", Linderhof Palace, Ettal, Germany, 2 December 2010*, Bavaria, Germany, 2011, pp. 699–710, Accessed on 16 September 2021. [Online]. Available: http://www.forschungsallianz-kulturerbe.de/download/3-8167-8637-5-developments-in-climate-control-of-historic-buildings.pdf

[208] F. García-Diego, Á. F. Navajas, P. Beltrán, and P. Merello, "Study of the effect of the strategy of heating on the mudejar church of santa maria in ateca (spain) for preventive conservation of the altarpiece surroundings," *Sensors (Basel, Switzerland)*, vol. 13, pp. 11 407–11 423, 2013, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.3390/s130911407

[209] A. Al-Omari, X. Brunetaud, K. Beck, and A. Muzahim, "Effect of thermal stress, condensation and freezing–thawing action on the degradation of stones on the castle of chambord, france," *Environmental Earth Sciences*, vol. 71, pp. 3977–3989, 2014. [Online]. Available: https://doi.org/10.1007/s12665-013-2782-4

[210] X. Brunetaud, L. De Luca, S. Janvier-Badosa, K. Beck, and M. Al-Mukhtar, "Application of digital techniques in monument preservation," *European Journal of Environmental and Civil Engineering*, vol. 16, no. 5, pp. 543–556, 2012. [Online]. Available: https://doi.org/10.1080/19648189.2012.676365

[211] O. USA, "Hobo data loggers," http://www.onsetcomp.com/, 2021, Accessed on 3 October 2021.

[212] G. Visco, S. Plattner, P. Fortini, S. Di Giovanni, and M. P. Sammartino, "Microclimate monitoring in the carcer tullianum: temporal and spatial correlation and gradients evidenced by multivariate

analysis; first campaign," *Chemistry Central Journal*, vol. 6, no. S11, pp. 3977–3989, 2012. [Online]. Available: https://doi.org/10.1186/1752-153X-6-S2-S11

[213] O. USA, "Temperature logger iButton with 8KB data-log memory," https://datasheets.maximintegrated.com/en/ds/DS1922L-DS1922T.pdf, 2015, Accessed on 3 October 2021.

[214] M. A. Valero, P. Merello, A. F. Navajas, and F.-J. Garcia-Diego, "Statistical tools applied in the characterisation and evaluation of a thermo-hygrometric corrective action carried out at the noheda archaeological site (noheda, spain)," *Sensors*, vol. 14, no. 1, pp. 1665–1679, 2014. [Online]. Available: https://www.mdpi.com/1424-8220/14/1/1665

[215] iButton, "Hygrochron temperature/humidity logger iButton with 8KB data-log memory," https://ibutton.cl/product/datalogger-temperatura-humedad-hygrochron-ds1923/, 2021, Accessed on 3 October 2021.

[216] P. Merello, F.-J. García-Diego, and M. Zarzo, "Diagnosis of abnormal patterns in multivariate microclimate monitoring: A case study of an open-air archaeological site in pompeii (italy)," *Science of The Total Environment*, vol. 488-489, pp. 14–25, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0048969714005786

[217] F.-J. García-Diego, E. Borja, and P. Merello, "Design of a hybrid (wired/wireless) acquisition data system for monitoring of cultural heritage physical parameters in smart cities," *Sensors*, vol. 15, no. 4, pp. 7246–7266, 2015. [Online]. Available: https://www.mdpi.com/1424-8220/15/4/7246

[218] D. J. Best and D. E. Roberts, "Algorithm as 89: The upper tail probabilities of spearman's rho," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 24, no. 3, pp. 377–379, 1975. [Online]. Available: http://www.jstor.org/stable/2347111

[219] T. Hastie, *Statistical Models in S*. Routledge, 1992, ch. 4. [Online]. Available: https://doi.org/10.1201/9780203738535

[220] J. W. Richards, D. L. Starr, N. R. Butler, J. S. Bloom, J. M. Brewer, A. Crellin-Quick, J. Higgins, R. Kennedy, and M. Rischard, "On machine-learned clasification of variable stars with sparse and noisy time-series data," *The Astrophysical Journal*, vol. 733, no. 1, p. 10, Apr 2011, Accessed on 16 September 2021. [Online]. Available: http://dx.doi.org/10.1088/0004-637X/733/1/10

[221] S. the sensor company, "Sensirion," https://www.sensirion.com/fileadmin/user_upload/customers/sensirion/Dokumente/2_Humidity_Sensors/Datasheets/Sensirion_Humidity_Sensors_SHT1x_Datasheet.pdf, 2011, Accessed on 3 October 2021.

[222] A. Perles, R. Mercado, J. V. Capella, and J. J. Serrano, "Ultra-low power optical sensor for xylophagous insect detection in wood," *Sensors*, vol. 16, no. 11, 2016. [Online]. Available: https://www.mdpi.com/1424-8220/16/11/1977

[223] O. Standard, "MQTT: the standard for IoT messaging OASIS," https://docs.oasis-open.org/mqtt/mqtt/v5.0/mqtt-v5.0.html, 2019, Accessed on 3 October 2021.

[224] "Redash," https://redash.io, 2021, Accessed on 3 October 2021.

[225] G. Visco, S. Plattner, P. Fortini, and M. P. Sammartino, “A multivariate approach for a comparison of big data matrices. case study: thermo-hygrometric monitoring inside the carcer tullianum (rome) in the absence and in the presence of visitors,” *Environmental Science and Pollution Research*, vol. 24, no. 13990–14004, pp. 3977–3989, 2017. [Online]. Available: https://doi.org/10.1007/s11356-017-8751-9

[226] L. Roderick J. A. and R. Donald B., *Statistical Analysis with Missing Data*, 2nd ed. Wiley, 2002.

[227] S. Moritz and T. Bartz-Beielstein, “imputeTS: Time Series Missing Value Imputation in R,” *The R Journal*, vol. 9, no. 1, pp. 207–218, 2017, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.32614/RJ-2017-009

[228] C. Roever, N. Raabe, K. Luebke, U. Ligges, G. Szepannek, M. Zentgraf, and D. Meyer, “Package ’klaR: Classification and Visualization’,” https://cran.r-project.org/web/packages/klaR/index.html, 2020, Accessed on 16 September 2021.

[229] D. Chung, H. Chun, and S. Keles, “Package spls: Sparse Partial Least Squares (SPLS) Regression and Classification,” https://cran.r-project.org/web/packages/spls/index.html, 2019, Accessed on 16 September 2021.

[230] I. T. Jolliffe, N. T. Trendafilov, and M. Uddin, “A modified principal component technique based on the LASSO,” *Journal of Computational and Graphical Statistics*, vol. 12, no. 3, pp. 531–547, 9 2003, Accessed on 16 September 2021. [Online]. Available: https://doi.org/10.1198/1061860032148

[231] T. Hastie, R. Tibshirani, J. Friedman, and J. Freedman, *The elements of statistical learning: data mining, inference and prediction*, 2nd ed. Standford CA: Springer, 2009.

[232] B.-H. Mevik and R. Wehrens, “The pls package: Principal component and partial least squares regression in r,” *Journal of Statistical Software, Articles*, vol. 18, no. 2, pp. 1–23, 2007, Accessed on 16 September 2021. [Online]. Available: https://www.jstatsoft.org/v018/i02

[233] M. Kuhn, J. Wing, S. Weston, and A. Williams, “Package ’caret’,” https://cran.r-project.org/web/packages/caret/caret.pdf, 2021, Accessed on 3 October 2021.

[234] D. Camuffo, *Microclimate for Cultural Heritage Conservation, Restoration, and Maintenance of Indoor and Outdoor Monuments*, 2nd ed. Elsevier Science, 2014.

[235] Y. Jiang and L. Lu, “A study of dust accumulating process on solar photovoltaic modules with different surface temperatures,” *Energy Procedia*, vol. 75, pp. 337–’342, 2015, clean, Efficient and Affordable Energy for a Sustainable Future: The 7th International Conference on Applied Energy (ICAE2015). Accessed on 16 September 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1876610215011467

[236] T. Hastie and R. Tibshirani, *Generalized Additive Models*, ser. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Boca Raton, London, New York, Washington, D.C.: Taylor & Francis, 1990. [Online]. Available: https://books.google.es/books?id=qa29r1Ze1coC

[237] UNE-EN-15759-1, “Conservación del patrimonio cultural. Clima interior. Parte 1: Recomendaciones para la calefacción de iglesias, capillas y otros lugares de culto,” https://www.une.org/encuentra-tu-norma/busca-tu-norma/norma?c=N0050451, p. 24, 2012, Accessed on 16 September 2021.

[238] UNE-EN 15757, “Conservación del patrimonio cultural. Especificaciones de temperatura y humedad relativa para limitar los daños mecánicos causados por el clima a los materiales orgánicos higroscópicos,” https://tienda.aenor.com/norma-une-en-15757-2011-n0046628, p. 15, 2011, Accessed on 16 September 2021.

[239] A. ASHRAE Standard 55-2007, “Thermal environment conditions for human occupancy,” https://www.ashrae.org/technical-resources/bookstore/standard-55-thermal-environmental-conditions-for-human-occupancy, 2020, Accessed on 16 September 2021.

[240] M. Andretta, F. Coppola, A. Modelli, N. Santopuoli, and L. Seccia, “Proposal for a new environmental risk assessment methodology in cultural heritage protection,” *Journal of Cultural Heritage*, vol. 23, no. C, pp. 22–32, 2017.

[241] K. Fabbri and A. Bonora, “Two new indices for preventive conservation of the cultural heritage: Predicted risk of damage and heritage microclimate risk,” *Journal of Cultural Heritage*, vol. 47, pp. 208–217, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1296207420304453

[242] J.-F. García-Diego, C. Scatigno, P. Merello, and E. Bustamante, “Preliminary data of CFD modeling to assess the ventilation in an archaelogical building,” in *Proceedings of the of the 8th International Congress on Archaeology, Computer Graphics, Cultural Heritage and Innovation ‘ARQUEOLÓGICA 2.0’ in Valencia (Spain)*, 2016, pp. 504–507.

[243] S. Albero, C. Giavarini, M. L. Santarelli, and A. Vodret, “Cfd modeling for the conservation of the gilded vault hall in the domus aurea,” *Journal of Cultural Heritage*, vol. 5, no. 2, pp. 197–203, 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1296207404000275

[244] T. Oetelaar, “Cfd, thermal environments, and cultural heritage: Two case studies of roman baths,” in *2016 IEEE 16th International Conference on Environment and Electrical Engineering (EEEIC)*, 2016, pp. 1–6.

[245] W. Köppen, *(eds) Handbuch der Klimatologie. 1. C. Gebr, Borntraeger*. Verlag Von Gebrü Borntraeger, 1936, ch. Das geographisca system der Klimate 1–44. [Online]. Available: https://upload.wikimedia.org/wikipedia/commons/3/3c/Das_geographische_System_der_Klimate_(1936).pdf

[246] G. Trewartha and L. Horn, *An introduction to climate*, 5th ed. McGraw-Hill, 1980.

[247] A.-L. Boulesteix, C. Strobl, T. Augustin, and M. Daumer, “Evaluating microarray-based classifiers: an overview,” *Cancer informatics*, vol. 6, pp. 77–97, 2008, Accessed on 16 September 2021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/16646813/