sPLS-DA to discriminate time series

Sandra Ramírez * Manuel Zarzo† Fernando-Juan García-Diego ‡
Angel Perles §

Abstract

The main goal of this research is to propose a methodology for classifying time series. Two approaches were used in this methodology: (A1) methods based on the parameters from models, and (A2) methods based on the features of time series. Approach A2 was used in method 1 (M1) and both approaches A1 and A2 were used in methods 2 and 3 (M2 and M3). (M1) Features based on functions such as *spectral density*, *sample Auto Correlation Function (Sample ACF)*, *sample Partial Auto Correlation Function (Sample PACF)* and *rolling ranges*, (M2) Estimates of parameters and features based on a *Seasonal Autoregressive Integrated Moving Average (Seasonal ARIMA)* model with a *Threshold Generalized Autoregressive Conditional Heteroskedastic (TGARCH)* model and a *Student distribution for residuals* (*Seasonal ARIMA-TGARCH-Student*) and (M3) Estimates of parameters and features based on a *Additive Seasonal Holt-Winters prediction function* (*Additive SH-W*). For M2 and M3: Firstly, estimates of the parameters of models were calculated. Secondly, features of residuals from the models, such as the maximum of the *spectral density* and mean of the values of *Sample PACF* were computed. The *Sparse Partial Least Squares Discriminant Analysis* (*sPLS-DA*) was used to identify groups of time series using the *classification variables* (features or parameter estimates) from the three methods. The *centroid distance* and the *Balanced classification Error Rate* (*BER*) were used to apply the *sPLS-DA*. The methodology is described using time series data from a study carried out in the Metropolitan Cathedral of Valencia in 2008 and 2010. The time series data corresponds to the time series of relative humidity from sensors positioned at different points of the apse (positions: cornice or ribs $\mathcal{RC}$, walls $\mathcal{W}$ and frescoes $\mathcal{F}$) in the Cathedral in 2008 and 2010. The sensors were monitored with the goal of assisting conservation of the renaissance frescoes in the Cathedral. The *classification variables* in our study were calculated separately for various seasons of the year (winter, spring and summer) for both 2008 and 2010. For methods 1,2 and 3 in 2008 and M1 and M3 in 2010, the first component from *sPLS-DA* showed that the time series that are situated in the $\mathcal{RC}$ and $\mathcal{W}$ positions were classified according to their location. Also, for M1 (2010) the time series in $\mathcal{RC}$, $\mathcal{F}$ and $\mathcal{W}$ were classified according to their positions. The methodology proposed in this research would be appropriate when there are no major differences between the time series of different groups, and when, according to the characteristics and context of the problem, it is possible to indicate the class of the time series.

Key Words: ARIMA, Art conservation, Auto correlation function, Diagnosis sensor, Holt Winters, Microclimate Spectral density, TGARCH, Student

1. Introduction

The famed Renaissance frescoes in Valencia's cathedral had been kept at a relatively consistent temperature until the year 2004 when restoration began [García-Diego and Zarzo, 2010]. In order to maintain the preservation of the frescoes, a unique monitoring system

*Department of Applied Statistics, Operations Research and Quality, Universitat Politècnica de València, Camino de Vera s/n 46022 Valencia, Spain. Department of Natural Sciences and Mathematics, Pontificia Universidad Javeriana seccional Cali, Colombia

†Department of Applied Statistics, Operations Research and Quality, Universitat Politècnica de València, Camino de Vera s/n 46022 Valencia, Spain

‡Department of Applied Physics, (U.D.Agriculture Engieneering), Universitat Politècnica de València, Camino de Vera s/n 46022 Valencia, Spain

§Department of Computer and Systems Informatics, Universitat Politècnica de València, Camino de Vera s/n 46022 Valencia, Spain

was introduced to regulate both temperature and humidity. Sensors were located at different points at the vault of the apse and inserted into the painting's surface itself [Zarzo et al., 2011, García-Diego and Zarzo, 2010]. The system was designed to identify both water entering from the roof and general humidity in the vault itself [García-Diego and Zarzo, 2010]. The statistical analysis directed by Zarzo et al. [2011], García-Diego and Zarzo [2010] of data on relative humidity (*RH*) displayed the importance of humidity sensors in maximizing protection and preventing deterioration of the frescoes.

In the preliminary study, *RH* data was analyzed [Zarzo et al., 2011, García-Diego and Zarzo, 2010] from the sensors (positions: cornice $\mathcal{C}$, ribs $\mathcal{R}$, walls $\mathcal{W}$ and frescoes $\mathcal{F}$) located in the cathedral of Valencia. These are shown in Figure 1. The researchers used a *Principal Components Analysis* (*PCA*). This method was applied to the *RH* data (RH_h data or RH_d data or RH_m data), where RH_h corresponds to an average of 60 values of *RH* per hour, RH_d to the average of the values of *RH* per day and RH_m to the average of the values of *RH* per month. The researchers concluded that the study of *RH*, using *PCA* as well as the interpretation of the first two components of the *PCA*, can be a very powerful method for the preventive conservation of frescoes. Also, they came to the conclusion that *PCA* can be used to identify abnormal conditions of the paintings and an abnormal performance in sensors [Zarzo et al., 2011, García-Diego and Zarzo, 2010].

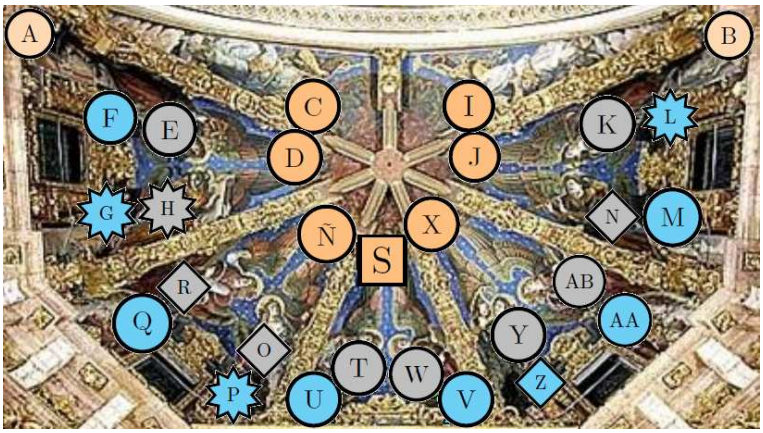


Figure 1: Position of the 29 sensors for monitoring the *RH* of the air inside the Cathedral. Seven of the sensors are located on the ribs (orange), two probes at the cornice (light orange), ten probes on the walls below the severies (blue) and ten probes on the frescoes (gray). In this research the sensors that were used were: 18 common sensors in both 2008 and 2010 (circular), 4 sensors only used in 2008 (diamond), one sensor only used in 2010 (square) and 4 sensors which were not used (star).

This study will propose a methodology for discriminating time series from sensors located in the Cathedral of Valencia. Research that may be related to this study correspond to *time series clustering*. According to Vilar and Montero [2014], *clustering* is an *unsupervised learning task* that separates a set of unlabeled data objects into homogeneous clusters. Separation is carried out in such a way that objects in the same cluster are more alike each other than objects in different clusters, according to some defined criterion. There are five approaches to *time series clustering* methods based on: (A1) parameters from models, (A2) serial features extracted from the original time series, (A3) complexity of a time series (A4) the properties of the predictions and (A5) the comparison of raw data [Liao, 2005].

A crucial aspect in cluster analysis is establishing a suitable *similarity or dissimilarity measure* between two objects. Different approaches for defining a *dissimilarity* between time series have been proposed. Liao [2005] presented an investigation about a *similarity*

or dissimilarity measure and Fu [2011] provided an overview on time series data mining directions, such as *measures of dissimilarity*, clustering procedures and visualization tools. Vilar and Montero [2014] also produced a description on time series clustering and created a `R` package `TSclust` where they implemented a large set of well-established, peer-reviewed, time series dissimilarity measures. They also described the main features of `TSclust` and presented examples of how it is used. Some of these are shown below.

Several measures have been proposed for approach 1 (A1) [Piccolo, 1990, Maharaj, 1996, 2002, Kakizawa et al., 1998, Vilar and Pértega, 2004]. The most commonly contemplated criterion has been to assume that the time series are generated by*ARIMA* processes. Piccolo [1990] introduced the *Euclidean distance* between their corresponding *Autoregressive (AR) expansion* as the metric and used a complete linkage clustering algorithm to construct the *dendrogram*. One problem of this metric is related to the effective numerical computations of the *AR* coefficients. For *Autoregressive Moving Average models ARMA* processes, Maharaj [2000] developed an *agglomerative hierarchical clustering* procedure that is based on the *p-value* of a *test of hypothesis* applied to every pair of given *stationarity* time series. Kalpakis et al. [2000] studied the clustering of *ARIMA* time series, by using the *Euclidean distance* between the *Linear Predictive Coding LPC spectrum* of two time series as their *dissimilarity measure*. Xiong et al. [2002] clustered univariate *ARIMA* time series and assumed that the time series are generated by *ARMA* models, although they used mixtures of *ARMA* models. They derived an *Expectation Maximization (EM)* algorithm for learning the mixing coefficients as well as the parameters of the component models. One problem with the method is that if the underlying clusters are very close to each other, the clustering performance might diminish significantly. Researchers in machine learning and speech recognition have proposed models such as *Markov chains (MC)* or *hidden Markov (HM)* [Ramoni et al., 2002, Smyth, 1997, Oates et al., 1999, Bagnall et al., 2003]. In relation to approach 2 (A2), some *dissimilarity measures* extracted from the original time series features are *ACF*, *cross-correlations*, *spectral features* [Kovačić, 1996, Struzik and Siebes, 1999, Peña and Galeano, 2001, Caiado et al., 2006, Douzal and Nagabhushan, 2007]. According to approach 3 (A3), most measures based on complexity of a time series are supported on the notion of *Kolmogorov complexity* or *algorithmic entropy* [Li and Vitãnyi, 2007]. There are two useful approaches for evaluating complexity differences between two time series, the first, uses algorithms based on data compression [Li et al., 2001, 2004, Cilibrasi and Vitãnyi, 2005, Keogh et al., 2007] and the second, considers differences between permutation distributions [Brandmaier, 2011]. In relation to approach 4 (A4), studies by Alonso et al. [2006], Vilar et al. [2010] focused on the notion of dissimilarity governed by the performance of future forecasts. Two time series are similar if their forecasts for a specific future time are close. Measures such as *Minkowski*, *Fréchet*, *Euclidean* or *Manhattan* are used in approach 5 (A5). Other measures of this type are described by Batista et al. [2011].

In the opinion of Liao [2005], beyond all these criteria for defining *dissimilarity measures* between time series they are often adapted to the problem at hand. According to the problem in a specific context, certain properties of the time series are highlighted in a *dissimilarity measure*.

In this study a methodology for discriminating time series is proposed, where approaches A1 and A2 are used. The *measure* that is used to predict the class of each time series is called *prediction distance*. It is applied to *latent variables* (components) that are calculated using a linear combination of estimates of the parameters of one of the models aforementioned and features of the time series. One problem is the amount of information that can be used in the *latent variables* because this might make the results difficult to interpret. As a consequence, a methodology is proposed to identify the clusters of time series

with the optimal information of the time series. In this research, the term *classification variables* is utilized to refer to the estimates of the parameters of the models and to refer to the features extracted from the time series data.

As the classification data (related to *classification variables*) in this study, is characterized by more variables than the number of the time series (sensors), they often imply a high degree of multicollinearity, and this might lead to severely ill-conditioned problems. One solution is to perform feature selection, or introduce artificial variables that summarize most of the information.

Support Vector Machines (*SVM*) [Vapnik, 1999] and *Classification and Regression Trees* (*CART*) [Deconinck et al., 2005], do not necessarily require variable selection for predictive purposes. However, the results are often difficult to interpret when there are a large number of variables. In order to solve this problem, some methods have been suggested: *Nearest Shrunken Centroids* (*NSC*) [Tibshirani et al., 2002], *Optimal Feature Weighting* (*OFW*) [Lê et al., 2009, 2007], *Random Forests* (*RF*) [Breiman, 2001], *Recursive Feature Elimination* (*RFE*) [Guyon and Elisseefi, 2003], *Linear Discriminant Analysis* (*LDA*), *Principal Component Analysis* (*PCA*) [Bair et al., 2006, Jombart et al., 2010], *Partial Least Squares Regression* (*PLS*) [Wold, 1966], and *PLS* with discrimination purposes [Antoniadis et al., 2003, Boulesteix, 2004, Dai et al., 2006].

LDA has often been shown to produce the best classification results. However, for large data sets with a large number of correlated predictors, *LDA* uses too many parameters that are estimated with a high variance. While *Sparse LDA* produces a parsimonious model. Another limitation of the approaches cited above is the deficiency of interpretability when dealing with a large number of variables.

In order to select the relevant variables in *PLS* [Lê Cao et al., 2008, 2009, Chun and Keleş, 2010], the penalties ℓ_1 (*Lasso regression* [Hoerl and Kennard, 1970]) or ℓ_2 (*Ridge regression* [Hoerl and Kennard, 1970]) are applied to the variable weight vectors. Chung and Keles [2010] extended the *Sparse PLS* from Chun and Keleş [2010] for multiclass classification problems and demonstrated that both *Sparse PLS Discriminant Analysis* (*sPLS-DA*) and *Sparse PLS* (*sPLS*) with an incorporated generalized framework, improved classification accuracy compared to classical *PLS* [Fort and Lambert-Lacroix, 2005].

sPLS-DA has very satisfying predictive performances and is able to select informative variables easily. While the approach proposed by Chung and Keles [2010] uses a two-staged procedure, *sPLS-DA* proposed by Lê Cao et al. [2011] performs variable selection and classification in a one-step procedure.

In this study, in order to identify a small subset of components and *classification variables* and to recognize groups of the time series, the *sPLS-DA* [Lê Cao et al., 2011] was employed. The *prediction distance* proposed was *centroid distance* and *Balanced classification error rate (BER)* was used to evaluate the results.

Three different methods were applied to the time series to find *classification variables*: (M1) Features based on functions such as *spectral density*, *sample Auto Correlation Function (sample ACF)*, *sample Partial Auto Correlation Function (sample PACF)* and *rolling ranges* [Palma, 2016, Venables and Ripley, 2002, Brockwell and Davis, 1991a, Box and Jenkins, 1976, Bloomfield, 1976, Brockwell and Davis, 1991b, Kovalevsky, 2018], (M2) Estimates of parameters and features based on a *Seasonal Autoregressive Integrated Moving Average (Seasonal ARIMA)* model with a *Threshold Generalized Autoregressive Conditional Heteroskedastic (TGARCH)* model and a *Student distribution for residuals* (*Seasonal ARIMA-TGARCH-Student*) [Kovalevsky, 2018] and (M3) Estimates of parameters and features based on *Additive Seasonal Holt-Winters prediction function* (*Additive SH-W*) [Holt, 2004, Winters, 1960].

Estimates of parameters of models applied to time series of *RH* were calculated, for M2

and M3. Secondly, features of residuals from models such as the maximum of the *spectral density* and mean of the values of *partial autocorrelation function* were computed for the same methods.

The databases used in this research correspond to subsets of the data sets of *RH* used in the study conducted by [Zarzo et al., 2011]. This study focuses on the data from 23 sensors for 2008 and 20 for 2010. Furthermore, the values of *RH* are not *missing values*. The *RH* data from each sensor corresponds to a time series of *RH*. The *sPLS-DA* was carried out separately for various seasons of the year (winter, spring and summer) for both 2008 and 2010. In this supervised classification framework, it is assumed that the time series are partitioned into *K=3* groups according to the position of sensors, $\mathcal{RC}$, $\mathcal{W}$ and $\mathcal{F}$ in the Cathedral. The groups $\mathcal{R}$ and $\mathcal{C}$ were joined as a new group $\mathcal{RC}$ to ensure a balanced number of sensors per group. The groups were selected according to physical interest in the Cathedral's microclimate.

The `R` software [R Core Team, 2014] was used to carry out the analysis (versions `3.6.2` and `4.0`). The most important packages of `R` employed in this piece of work were `struc change` [Zeileis, 2006a], `rugarch` [Kovalevsky, 2018], `mixOmics` [Rohart et al., 2017b], `QuantTools` [Kovalevsky, 2018], `aTSA` [Qiu, 2015], `forecast` [Hyndman et al., 2020], `stats` [R Core Team, 2014] and `tseries` [Trapletti and Hornik, 2019].

This research aims to bring forward a supervised methodology for discriminating time series according to approaches A1 and A2. This methodology has two stages: (1) obtain the *classification variables* using three methods (M1, M2, and M3) and (2) classify the time series using *sPLS-DA* as a discriminant technique. This technique is applied to *classification variables* estimated in stage (1).

This article is structured as follows: In Sect. 2, characteristics of the data sets and the sensors are displayed. Statistical tests to identify *structural breaks* of the time series, methods for calculating *classification variables*, and the *sPLS-DA* are introduced in Sect. 3. The results of the methods applied to time series and estimates from the *sPLS-DA* are presented in Sect. 4. The most notable results from *sPLS-DA* are presented in Sect. 5.

2. Data and Materials

2.1 Data

The databases correspond to subsets of the time series of *RH* (for each sensor) used in the study conducted by [Zarzo et al., 2011]. The time periods of the subsets were selected in such a way as to avoid the *missing values*. The *missing values* were not used in order to have a simple methodology.

In respect to the notation $RH_h = \{RH_{h_t}, t \in \mathbb{Z}\}$ or $RH_d = \{RH_{d_t}, t \in \mathbb{Z}\}$ represent two real-valued processes, where RH_{h_t} corresponds to an average of 60 measures of *RH* per hour at time t and RH_{d_t} corresponds to the average of measures per day at time t. Also, $\mathbf{RH_h} = (RH_{h_1}, \ldots, RH_{h_{n_h}})^\top$ and $\mathbf{RH_d} = (RH_{d_1}, \ldots, RH_{d_{n_d}})^\top$ represent the partial realizations (observed time series) of the RH_h and RH_d processes. For each partial realization the lengths are n_h and n_d respectively. The notation $\mathbf{RH}$ refers to both: $\mathbf{RH_h}$ and $\mathbf{RH_d}$, and RH refers to both: RH_h and RH_d.

The observed time series $\mathbf{RH_h}$ in 2008 consists of 3,851 (n_h) observations and it consists of 3,414 (n_h) in 2010. In 2008: the realization of time series $\mathbf{RH_h}$ for winter, spring and summer consists of 1,430 and 2,099 and 322 observations respectively. In 2010: the same time series for winter, spring and summer consist of 636 and 2,178 and 600 observations respectively.

This research analyzed 23 sensors of *RH* in the year 2008 and 20 sensors in 2010. Among these, 18 sensors were common in both 2008 and 2010 (circular), 4 sensors were only used in 2008 (diamond), one sensor was only used in 2010 (square), 4 sensors were not used (star). On the other hand, there are an unbalanced number of sensors per position $\mathcal{RC}$ (orange) ($\mathcal{R}$ or $\mathcal{C}$), $\mathcal{W}$ (blue) and $\mathcal{F}$ (gray). In 2008 the number of sensors was: 7, 9 and 7 and in 2010: 8, 6 and 6 (see Figure 1).

The sensors `G`, `H`, `P` and `L` were discarded because in this work the analysis was separated by season and it was imposed as a condition of having time series with at least 300 observations.

2.2 Materials

For each probe there is a sensor in the apse of the Cathedral. Each probe contains an integrated circuit model DS2438 (Maxim Integrated Products, Inc.) that incorporates an analogue-to-digital voltage converter. Characteristics of the probes and sensors, details of the curves of calibration, as well as installation description of sensors and probes, are described in [Zarzo et al., 2011, García-Diego and Zarzo, 2010].

3. Methodology

3.1 Structural breaks of time series

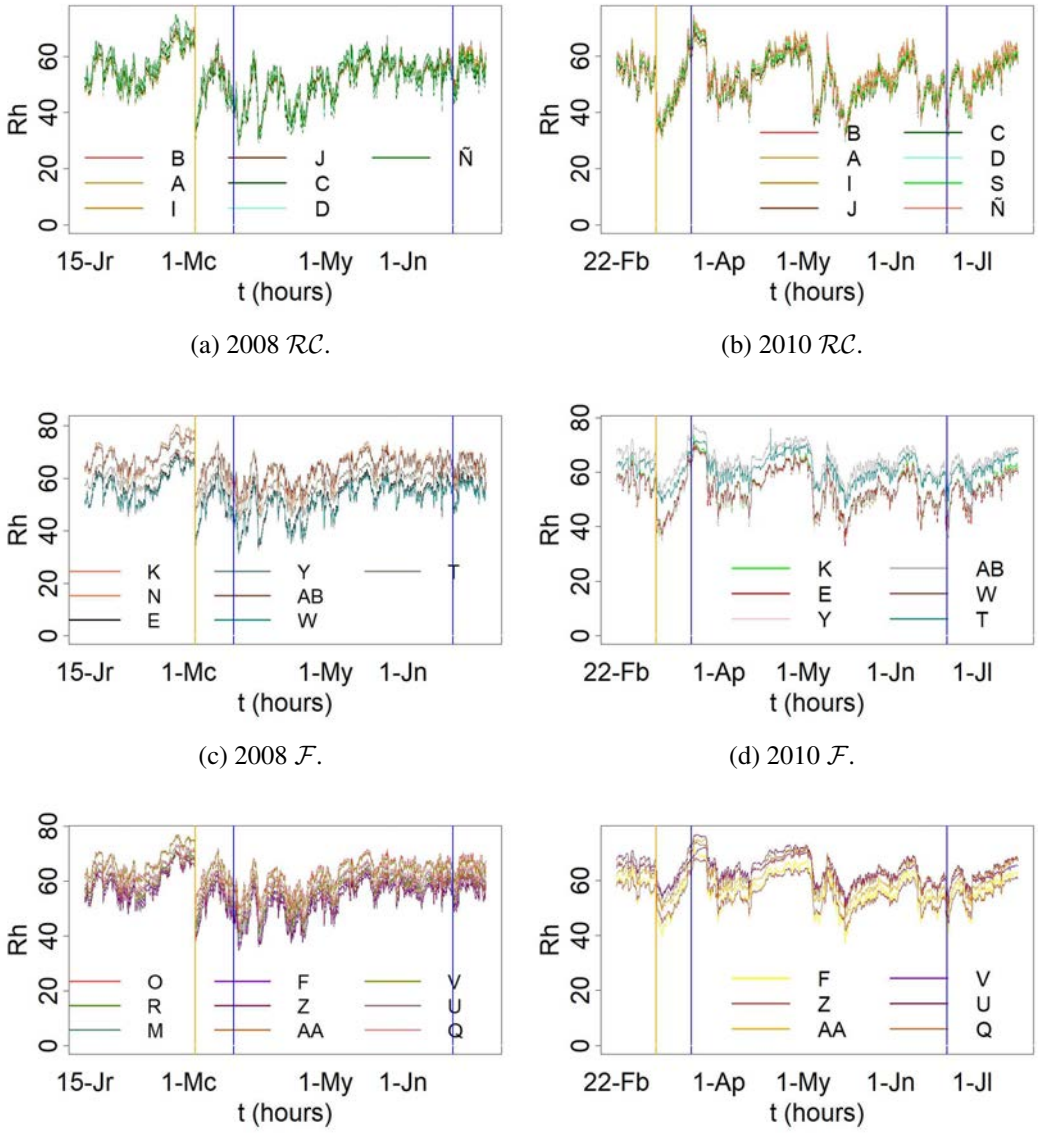
Some models such as the *ARMA* [Box and Jenkins, 1976], *ARCH* and *GARCH* [Palma, 2016] assume that the mean of an observed time series over a period of time is constant throughout [Palma, 2016]. However, an *observed time series* that corresponds to real situations can often present breaks in the mean, due to changes in external factors. If there is a change in the slope of the linear trend without a discontinuity of the trend then there is a *structural break* in the time series [Palma, 2016].

The most important classes of test on *structural breaks* are: (1) tests from the generalized fluctuation test framework (e.g., the *CUSUM* and *MOSUM* tests, among others) [Leisch et al., 2000] and (2) tests based on *F statistics* (e.g., *Chow* [Chow, 1960] and the *supF* [Zeileis et al., 2002] tests, etc)[Hansen, 2002, Andrews, 1993, Andrews and Ploberger, 1994] The tests from class 1 test *empirical fluctuation processes* and the tests from class 2 compute and test sequences of *F statistics* [Zeileis et al., 2002].

Figures from the observed time series $\mathbf{RH_h}$ for both years, 2008 and 2010, suggest potential *structural breaks* in at least two points (see Figure 2). The *supF* and *CUSUM* tests were used to to assess the significance of these potential *structural breaks*. The null hypothesis (H_0) and the alternative hypothesis (H_1) are as follows: "no structural change" and "the coefficient vector varies over time" [Zeileis et al., 2002].

In order to apply both tests, the observed time series $\mathbf{RH_h}$ was used after applying the logarithmic *transformation*, $r_t = \log(RH_{h_t})$, which has been used by other works to stabilize the variance of the time series [Cryer and Chan, 2008]. Furthermore, the time series were also *differentiated*, $W_t = r_t - r_{t-1}$ in order to remove the trend of the time series [Cryer and Chan, 2008]. The *supF* and *CUSUM* tests were applied to six groups: winter 2008 (group 1), spring 2008 (group 2), summer 2008 (group 3), winter 2010 (group 4), spring 2010 (group 5) and summer 2010 (group 6). Each group j ($j = 1, \ldots, 6$) has two variables $\mathbf{x_j}$ and $\mathbf{y_j}$, where $\mathbf{x_j} = [w_{j_2}, w_{j_3}, \ldots, w_{j_{n_j}}]$ and $\mathbf{y_j} = [w_{j_1}, w_{j_3}, \ldots, w_{j_{n_j-1}}]$. The elements of $\mathbf{x_j}$ and $\mathbf{y_j}$ are elements of a vector $\mathbf{W_{t_j}}$ given by $\mathbf{W_{t_j}}$= $\mathbf{W_j} = [w_{j_1}, w_{j_2}, \ldots, w_{j_{n_j}}]$. Where n_j is the number of observations of group j. Furthermore, $n_1 = 1,430$, $n_2 = 2,099$, $n_3 = 322$, $n_4 = 636$, $n_5 = 2,178$ and $n_6 = 600$.

On the other hand, the higher the probability that the H_0 will be rejected the more the sample size increases [Thiese et al., 2016, Wasserstein and Lazar, 2016]. According to this argument, the *significance level* was determined by the sample size. Then, for summer (2008) and winter and summer (2010), the *significance level* was 0.05 and for other groups, this was 0.02. Descriptions of the tests are presented below.



(a) 2008 $\mathcal{RC}$.

(b) 2010 $\mathcal{RC}$.

(c) 2008 $\mathcal{F}$.

(d) 2010 $\mathcal{F}$.

(e) 2008 $\mathcal{W}$.

(f) 2010 $\mathcal{W}$.

Figure 2: Column 1 corresponds to 2008 (from January the 15th to the 4th of July) and column 2 to 2010 (22nd of Febraury to the 18th of July). Graphics correspond to $\mathbf{RH_h}$ separated by their positions in the apse in the Cathedral: cornice and ribs ($\mathcal{RC}$), walls ($\mathcal{W}$) and frescoes ($\mathcal{F}$). Also, by the seasons (winter, spring and summer). Separation by season is indicated with the vertical blue line. Winter is divided into 2 parts (winter 1 and winter 2) by a structural break that was identified in the series according to the *SupF* and *CUSUM*. The *structural break* is shown by the yellow line. The results by positions are presented as follows: (a) y (b) $\mathcal{RC}$, (c) y (d) $\mathcal{F}$ and (e) y (f) $\mathcal{W}$.

3.1.1 The SupF test

The statistic is the *maximum F-statistic* ($supF$, across the grid of all potential *change points*) from the *Chow test* [Chow, 1960] in an interval given according to the data [Zeileis et al., 2002].

To explain the test simply, $\mathbf{x}$, $\mathbf{y}$ and n were used instead of $\mathbf{x_j}$, $\mathbf{y_j}$ and n_j. To carry out the test the *standard linear regression model (SLRM)* must be defined. For this *SLRM* the response variable is $\mathbf{y}$ and predictor variable is $\mathbf{x}$. The *SLRM* is given by $y_i = \mathbf{x_i}^\top \beta_i + u_i$, where $i = 1, \ldots, n$, $\mathbf{x_i} = (1, x_i)$ and u_i is an *independent and identically distributed (i.i.d.)* $(0, \sigma^2)$.

An *Ordinary Least Squares* (*OLS*) model was fitted, where the regression coefficients are estimated twice: once for the observations before the *change point* (i_0) and once for those after the i_0. Then, β_i is the 2×1 vector of regression coefficients and it is defined as follows: $\beta_i = \beta_A$ if $1 \leq i \leq i_0$ and $\beta_i = \beta_B$ if $i_0 < i \leq n$. Where i_0 is a *break point* in the interval $(1, n-1)$.

The test statistic $supF$ is the maximum of the values of *F statistics* for any potential *break point* (i_0) in an interval $1 < \underline{i} \leq i_0 \leq \bar{i} < n-1$. The $supF$ is defined as $supF = Sup_{\underline{i} \leq i \leq \bar{i}} F_{i_0}$ where $F_{i_0} = \frac{\widehat{u}^\top \widehat{u} - \widehat{e}^\top \widehat{e}}{\widehat{e}^\top \widehat{e}/(n-2)}$. The *OLS* residuals from the regression model are $\widehat{e} = (\widehat{u}_A, \widehat{u}_B)^\top$ and $\widehat{u}$ is the residuals from the model where the parameters were fitted once for all observations. The null hypothesis is rejected when the $supF$ gets too large.

3.1.2 The CUSUM test

Ploberger and Kramer [1992] suggested basing a *structural change test* on cumulative sums of the common *OLS* residuals instead of *recursive residuals* [Zeileis et al., 2002]. The *OLS-CUSUM* type empirical fluctuation process is defined by $W_n^0(t) = \frac{1}{\widehat{\sigma}\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} \widehat{u}_i$, $0 < t < 1$. The limiting process for $W_n^0(t)$ is $W^0(t) = W(t) - tW(1)$, where $W(t)$ is a *Wiener Process* [Durrett, 2000]. It starts on 0 at $t = 0$ and it also returns to 0 for $t = 1$. Under a single structural shift alternative, the path should have a peak around t_0 [Zeileis et al., 2002]. In the `R` software, the results can be computed with functions `Fstats`, `efp` from the `strucchange` package.

According to the $supF$ test, in winter, the H_0 should be rejected, this means that there is a *structural break* after the 1,058th (2008) and 338th (2010) observations. The cumulative sum of the residuals *CUSUM* should fluctuate around zero, however, significant deviation occurs from the 1,058th (2008) and 338th (2010) observations. The dates (and time) of the *structural breaks* are: the 27th of March 2008 (7:00 AM) and the 8th of March 2010 (1:00 PM).

A fixed parameter model cannot be expected to forecast well if the true parameters of the model change over time. Ignoring *structural breaks* can lead to negative implications such as inconsistency of the parameter estimates and forecast failures [Gaetano, 2018]. The interest of this study is to determine the estimates of the parameters of the models of time series such as *ARIMA* and *GARCH* and as a consequence, the time series data was separated according to the *structural break* detected. Also, it might be better to separate the analysis per season and year, in congruence with the physical characteristics of the data. According to both considerations, the analysis was separated into four periods: winter 1, winter 2, spring and summer. Winter 1 corresponds to the period (in winter) before the *structural break* and winter 2, after the *structural break*. In 2008, winter 1 corresponds to observations from 1 to 1,058, and winter 2 to observations from 1,059 to 1,430. In 2010, winter 1 corresponds to observations from 1 to 338 and winter 2 to observations from 339 to 636 (see Figure 2).

3.2 Methods to determine classification variables

In this study, three different methods were applied to find *classification variables*: (M1) Features based on functions such as *spectral density*, *ACF*, *PACF* and *rolling ranges* [Palma, 2016, Venables and Ripley, 2002, Brockwell and Davis, 1991a, Box and Jenkins, 1976, Bloomfield, 1976, Brockwell and Davis, 1991b, Kovalevsky, 2018], (M2) Estimates of parameters and features based on a *Seasonal ARIMA-TGARCH-Student* [Kovalevsky, 2018] and (M3) Estimates of parameters and features based on a *Additive SH-W* [Holt, 2004, Winters, 1960]. The parameters were estimated for M2 and M3 and secondly features of residuals from models such as the mean of the values of *partial autocorrelation function* and the maximum of the *spectral density* were calculated for the same methods.

The analyses were carried out separately for various groups (winter 1, winter 2, spring and summer) for both 2008 and 2010. A description of the methods are presented below:

3.2.1 M1: Features based on functions

In this method the functions, *spectral density*, *ACF*, *PACF* and *rolling ranges* were used to study the mean, variance, correlation structure, and seasonal components of the time series of *RH*. Before computing the *classification variables* using the first two functions, logarithm transformation and regular differencing were employed, to stabilize the variances and remove the trend of the observed time series $\mathbf{RH}$, i.e, $r_t = \log(RH_{h_t})$ and $w_t = r_t - r_{t-1}$. Thus, the fourth and fifth functions were applied to $\mathbf{RH_h}$ and $\mathbf{RH_d}$ and the first two functions were applied to $\mathbf{W}$, where $\mathbf{W} = (w_1, \ldots, w_{n_W})^\top$ and $\mathbf{r} = (r_1, \ldots, r_{n_r})^\top$. A brief explanation of them is presented below.

1. *Rolling range*: the *rolling range over n past values* (moving ranges with order n) is the difference between the *maximum* and *minimum* over n past values [Kovalevsky, 2018]. In this study, $n = 2$ and the $\mathbf{RH}$ was used to calculate *rolling range*. In the `R` software, the estimate of *rolling range* can be computed with the function `rollrange` from the `QuantTools` package.

2. *Sample Auto Correlation Function (Sample ACF)*: a value of the *Sample ACF* is the correlation between a value of the time series with a value of the same time series at previous points (called lags and denotes with l) [Metcalfe and Cowpertwait, 2009]. The covariance is estimated for $l > 0$ from $n_W - l > 0$ for observed pairs (w_{1+l}, w_1), $\ldots, (w_{n_W}, w_{n_W-l})$. Under assumption of second-order stationarity the subseries $(w_{1+l}, \ldots, w_{n_W})$, $(w_1, \ldots, w_{n_W-l})$ have the same mean and variance and the estimator of *Sample ACVF* at lag l, $acvf_l$, is presented in equation 1. Where n_W is the number of observations of the observed time series $\mathbf{W}$ and $\overline{W}$ is the sample mean of the observations of the $\mathbf{W}$. The sample size n_W is used even though there are $|n_W - l|$ terms.

$$acvf_l = \frac{1}{n_W} \sum_{s=max(1,-l)}^{min(n_W-l,n_W)} \left(w_{s+l} - \overline{W}\right)\left(w_s - \overline{W}\right) \tag{1}$$

   The *Sample ACVF* of observed time series $\mathbf{W}$ at lag 0, $acvf_0$, equals the sample variance of $\mathbf{W}$ calculated with a denominator of n_W. *ACF* is the correlation of a variable with itself at different time lags and it is given by $acf_l = acvf_l/acvf_0$ [Venables and Ripley, 2002]. In the `R` software, the *Sample ACF* can be computed with the function `acf` from the `stats` package.

3. *Sample Partial Autocorrelation Function (Sample PACF)*: the value of *PACF* at lag l correspond to the linear correlation of a time series RH_{h_s} and a lagged version of itself $RH_{h_{s+l}}$ with the linear dependence of $RH_{h_l}^{l-1}$ removed, where $RH_{h_l}^{l-1}$ denotes $\{RH_{h_{s-1}}, RH_{h_{s-2}}, \ldots, RH_{h_{s-(l-1)}}\}$. If $l = 1$, the *Sample PACF* is the correlation between RH_{h_1} and RH_{h_0}. If $l \geq 2$, *Sample PACF* is the correlation between $RH_{h_l} - RH_{h_l}^{l-1}$ and $RH_{h_0} - RH_{h_0}^{l-1}$ [Box and Jenkins, 1976, Brockwell and Davis, 1991b].

   The *Sample ACF* and *Sample PACF* plots are important diagnostic tools for helping to select the proper order of p and q in *ARMA* (p, q) models [Box and Jenkins, 1976]. In the `R` software, the *Sample PACF* can be computed with the function `pacf` from the `tseries` package.

4. *Spectral density function*: *spectral density* and the *autocovariance* function expressed the same information in different ways. The *spectral density* function is estimated using the *periodogram*. The latter displays information about the seasonal components of a time series and the strengths of the various frequencies for explaining the seasonal components. To study the seasonal component of the time series, the maximum of the estimates of *spectral density* and corresponding frequencies are identified [Venables and Ripley, 2002]. A brief description of the *spectral density* that is explained in [Venables and Ripley, 2002] is presented below.

   A covariance-stationary process W_t with mean $\mu := E[W_t]$ and jth autocovariance $\gamma_j := [E(W_t - \mu)(W_{t-j} - \mu)]$. The *population spectrum* of W at frequency $\omega \in \mathbb{R}$ is given by

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

   This function is well defined, provided that the sequence $\{\gamma_k : k \in \mathbb{Z}\}$ is absolutely summable. From the properties of the complex exponential function, it becomes clear that the *population spectrum* is symmetric around 0 and periodic with period π. In addition, it can be shown that: (1)$\int_{-\pi}^{\pi} s_W(\omega) e^{iwk} d\omega = \gamma_k$ and (2) $s_W(\omega) \geq 0, \quad \omega \in [-\pi, \pi]$.

   Hence, the *autocovariance function* and the *population spectrum* function contain the same information about W. In particular, $\gamma_0 = V[w_t]$ can be computed as follows:

$$\gamma_0 = \int_{-\pi}^{\pi} s_W(\omega) d\omega = 2 \int_0^{\pi} s_W(\omega) d\omega.$$

   In fact, this is nothing but a particular case of a far deeper result. Recall the *spectral representation theorem*: any covariance-stationary process $\{W_t\}_{-\infty}^{\infty}$ with absolutely summable autocovariances can be represented as

$$W_t = \mu + \int_0^{\pi} \{\alpha(\omega) \cos(\omega t) + \delta(\omega) \sin(\omega t)\} d\omega,$$

   where $\alpha(.)$ and $\delta(.)$ have zero means. Heuristically, this theorem says that W_t can be decomposed in terms of frequencies. It can be proved that, for any given $\pi_0 \in [0, \pi]$, the portion of $V[W_t]$ associated with frequencies lower than π_0 is precisely $2 \int_0^{\pi_0} s_W(\omega) d\omega$.

   In respect to estimating the *population spectrum*, the most basic estimator of $s_W(.)$ is the so called periodogram:

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

where T is the sample size and $\hat{\gamma}_j$ is the jth sample autocovariance. This is an unbiased but an unacceptably noisy estimator of $s_W(\omega)$. However, if it is assumed that s_W is smooth, the values of this naive estimator can be averaged over frequencies near ω to get a much more precise estimator of $s_W(\omega)$, namely:

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

where $\omega_j = 2\pi j/T$, and l takes the role of a bandwidth indicating how many different frequencies can be considered close to ω_j, and $\mathcal{W}_T(.)$ is a weighting function that must have the following properties:

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

In the `R` software, the estimate of $s_W(\omega_j)$ can be computed with the function `spectrum` from the `stats` package. By default, this function assumes that $\mathcal{W}_T(\omega_j) \propto 2 - I(j \in \{-l, l\})$.

The *classification variables* according to type of data are the following: (1) For $\mathbf{RH_h}$ and $\mathbf{RH_d}$: the means of *rolling ranges* [Kovalevsky, 2018] of order 2 for the $\mathbf{RH_d}$ and $\mathbf{RH_h}$ (`rMd` and `rMh`). These variables correspond to the parameters (`HMV` and `DMV`) considered in the preliminary investigations of this project [Zarzo et al., 2011, García-Diego and Zarzo, 2010]. Other variables are variance of *rolling ranges* of order 2 for the $\mathbf{RH_d}$ and $\mathbf{RH_h}$ (`rVd` and `rVh`). Also, the estimates of the *sample PACF* of the $\mathbf{RH_h}$ for the first four lags (`pacf1`, `pacf2`, `pacf3` and `pacf4`) and (2) for $\mathbf{W}$: maximum of *spectral density* (`spec.mx`), frequency corresponding to maximum of *spectral density* (`freq`), estimates of the mean (`acf.m`), range (`acf.r`) and variance (`v.acf`) of the *sample ACF* for the first 72 lags.

3.2.2 M2: Seasonal ARIMA-TGARCH-Student model

The *ARIMA* model aims to describe the autocorrelations in time series [Palma, 2016]. A time series follows an *ARIMA(p,d,q)* process where p is the number of autoregressive (*AR*) terms, d is the number of difference taken and q is the number of moving average (*MA*) terms. Although *ARIMA* is flexible and powerful in forecasting, it is not able to manage the continuous changing of variance and nonlinearity that some time series can have in their behaviour [Roslindar et al., 2016].

If a time series follows an *ARIMA* process, the conditional variance must be constant. When it is not constant, the process is known as a *conditional variance process* [Cryer and Chan, 2008]. As a consequence the data is affected by nonlinear characteristics of the variance, often referred to as *volatility* or *variance clustering* [Laux et al., 2011]. Time series with periods of high *volatility* and periods of low *volatility* are said to exhibit *volatility clustering* and this implies unconditional standard deviations which are not constant [Laux et al., 2011].

The most important models for studying *volatility* are the *Autoregressive Conditional Heteroskedasticity* (*ARCH*) and *Generalized ARCH (GARCH)* models Engle [1982], Bollerslev [1986], Engle and Bollerslev [1986]. Some types of *GARCH* are studied in Ghalanos [2020]. Among these models are the family *GARCH* model and Threshold *GARCH* (*TGARCH*) model of Zakoian [1994b] which belongs to this family [Ghalanos, 2020].

Thus, instead of using an *ARIMA* model to study the conditional mean of future values, it is necessary to use a hybrid of the *ARIMA* and *GARCH* models which can simultaneously analyze both the conditional mean and the conditional heteroscedasticity of the process [Roslindar et al., 2016]. In recent years, hybrid models have been proposed to analyse forecasting of rainfall [Yusof and Kane, 2013], of daily load patterns of energy (voltage) [Hor et al., 2006], stock market value [Xing, 2011] and of the price of gold [Roslindar et al., 2016].

The *ARIMA-GARCH* model is applied in two steps. In the first step, the most successful *ARIMA* model is used to analyze the linear data of the time series. In the second step, the most successful *GARCH* model is used to fit the nonlinear patterns of the residuals. In this model, the error term of the *ARIMA* model is said to follow a *GARCH* process of orders r and s [Weiss, 1984]. To check if the *ARIMA-GARCH* model applied, fits well for time series data, the residuals of the *ARIMA* model and the residuals of the *GARCH* model need to be analyzed [Weiss, 1984].

The Box-Jenkins methodology [Box and Jenkins, 1976] consists of three iterative steps: the estimation of the parameters, the inspection of diagnostics and forecasting. A description of the two first steps is presented below.

Step 1 consists of three stages: Firstly, checking the condition of *stationarity*: A time series is *stationary* if its statistical characteristics are preserved across the time periods. If the mean and variance of the time series are constant and regardless of the moment at which it is evaluated, the relative dependence of an observation remains the same in respect to past values [Palma, 2016]. *Stationarity* is a crucial assumption in time series analysis [Palma, 2016]. Secondly, there are some techniques developed for transforming nonstationary data into *stationary* data. Variance stabilization, trend estimation through linear regression and differentiation of the series are often employed [Palma, 2016]. Variance stabilization is usually obtained by a *Box-Cox transformation* of the data, *Linear models* are tools for removing a deterministic trend from the data and *differentiation* is used to remove a trend in the data when the underlying trend is assumed to be stochastic [Palma, 2016]. Thirdly, determining the order of the *ARIMA* model: the values of *Sample ACF* and *Sample PACF* of the observed time series are employed to determine the order (p, q) of the *ARIMA* model.

Step 2 consists of four stages: Firstly, verifying whether the residuals are *white noise*: a time series follows a *white noise* process if the variance of the process is constant, its mean is constant and its observations are not correlated [Palma, 2016]. Whiteness testing procedures do not usually involve checking for independence unless a time series is assumed to have *Normal* distribution [Palma, 2016]. The *Box-Ljung* test can be employed to verify whether the errors are *white noise* or not [Palma, 2016]. Secondly, checking the condition of the residuals *independently*: if there is a significant autocorrelation between lags of the square root (or square) of the residuals, there is evidence against the hypothesis of them being *independently* and *identically* distributed [Cryer and Chan, 2008]. The slow decay of the *ACF* of the square root (or square) of the residuals in the *ACF* plot suggests that the distribution of the residual time series is not independent [Tsay, 2005]. The *Box-Ljung* test [Zeileis, 2006b, Ljung and Box, 1978, Harvey, 1993] can be employed to check the assumption that the errors (or squared errors) are not autocorrelated. Thirdly, determining if *Arch* effects exist in the residuals: *Volatility clustering* can be analyzed using *ACF* plots of the square root (or square) of the time series. If these plots slowly decay as a function of time lag, then the time series is said to show *volatility clustering*. If the values of the *ACF* decay are relatively slow, the effect of *volatility clustering* is high [Jie-Jun and Sai-Ping, 2012]. The *Lagrange Multiplier* test [Engle, 1982] can be employed to examine the conditional variance of the error and research whether *Arch* effects are present. Fourthly, verifying distribution of the residuals: *Q-Q normal scores* plot, tests of normality (*Shapiro-*

Wilk [Royston, 1982a,b] and *Kolgomorov-Smirnov* [Birnbaum and Tingey, 1951, Conover, 1994]) and values of *kurtosis* [Hein and Spudeck, 1988] were used to check whether the distribution of the errors are *Normal* or not.

Once it is identified that the *ARCH* effect exists in the residuals of the *ARIMA* model, a *GARCH* model is fitted to the residuals. The *GARCH* model selected is adequate if its parameters are statistically significant and its residuals satisfy step 2 [Tsay, 2005].

In this research the *ARIMA-TGARCH* models were fitted. In order to stabilize the variance of the time series before fitting the *ARIMA* models, data was transformed using the logarithm transformation on $\mathbf{RH_h}$, $r_t = \log(RH_{h_t})$. Because the variability of the data $\mathbf{RH_h}$ becomes more homogeneous using logarithm transformation, the observed time series $\mathbf{r}$ was used Lütkepohl and Xu [2012]. Furthermore, *regular differencing* was applied to remove the trend, i.e., $\mathbf{W}$ was used. After selecting the most successful *ARIMA-TGARCH* model for each sensor, the *Q-Q normal scores* plots of the residuals displayed a heavy-tailed distribution, the values of *kurtosis* are greater than three and the normality tests rejected the hypothesis of the normality. Then, a *Student* distribution was used to fit the residuals of the *TGARCH* model. Then, it is assumed that the time series r_t follows the *Seasonal ARIMA* (p, q) *-TGARCH*(s, m)*-Student* processes. It is also assumed that the time series r_t can be modelled as an *ARMA*(p, q) process whose error term, in turn, follows a *TGARCH*(s, m)*-Student* model. The *TGARCH* model considered residuals with a *Student distribution* with $\mathcal{V}$ as its parameter of shape. For each group, a common model was applied to the data of each sensor. A brief introduction of the *Seasonal ARIMA* (p, q) *-TGARCH*(s, m)*-Student* process is presented below.

1. The *ARMA model*: the time series r_t follows an *Autoregressive Moving Average process* with parameters p and q (denoted by *ARMA*(p, q) [Palma, 2016]) if it can be written as equation 2, where B is the *backshift operator*.

$$\begin{aligned} \boldsymbol{\phi_p}(B) r_t &= \boldsymbol{\theta_q}(B)\varepsilon_t, \\ \boldsymbol{\phi_p}(B) &= 1 - \phi_1 B - \cdots - \phi_p B^p, \\ \boldsymbol{\theta_q}(B) &= 1 + \theta_1 B + \cdots + \theta_q B^q, \end{aligned} \tag{2}$$

   and $\{\varepsilon_t\}$ is a *White Noise* (*WN*) process with mean 0 and variance σ^2. It can be written as $r_t = \phi_1 r_{t-1} + \ldots + \phi_p r_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$, where $\varepsilon_t \sim WN(0, \sigma^2)$.

   If the solution of the equation $0 = 1 - \phi_1 x - \ldots - \phi_p x^p$ is outside the unit circle (unit root tests) then r_t is stationary [Hamilton, 1994].

2. The *Seasonal ARIMA model*: a *Seasonal ARIMA* model $(p, d, q)(P, D, Q)_S$ [Box and Jenkins, 1976] has a non-seasonal component (p, d, q) and a seasonal component $(P, D, Q)_S$, where S is the number of observations per day ($S = 24$). Thus, a multiplicative *Seasonal ARIMA* model is obtained which has the form of equation 3, where B is the *backshift operator*,

$$\boldsymbol{\Phi_P}(B^S)\boldsymbol{\phi_p}(B)\nabla_S^D \nabla^d r_t = \boldsymbol{\Theta_Q}(B^S)\boldsymbol{\theta_q}(B)\varepsilon_t, \tag{3}$$

   where $\phi_p(B)$ is the regular *AR* operator of order p, $\theta_q(B)$ is the regular moving average operator of order q, $\Phi_P(B^S)$ is the seasonal *AR* operator of order P, $\Theta_Q(B^S)$ is the seasonal moving average operator of order Q and $\{\varepsilon_t, t \in \mathbb{Z}\}$ is a *WN* process. Furthermore, $\nabla_S^D = (1 - B^S)^D$ represents the seasonal differences and $\nabla^d = (1 - B)^d$ represents the regular differences [Palma, 2016].

The non-seasonal components (*AR* and *MA*) and the seasonal components (*SAR* and *SMA*) are presented in equation 4, where B is the *backshift operator*.

$$\begin{aligned} \boldsymbol{\phi_p}(B) &= 1 - \phi_1 B - \cdots - \phi_p B^p, \\ \boldsymbol{\theta_q}(B) &= 1 + \theta_1 B + \cdots + \theta_q B^q, \\ \boldsymbol{\Phi_P}(B^S) &= 1 - \Phi_1 B^S - \Phi_P B^{PS}, \\ \boldsymbol{\Theta_Q}(B^S) &= 1 + \Theta_1 B^S + \cdots + \Theta_Q B^{QS}. \end{aligned} \tag{4}$$

Furthermore, if $w_t = \nabla_S^D \nabla^d r_t$, is obtained by differentiating the series regularly d times and D times seasonally, then w_t follows a multiplicative *Seasonal ARIMA process* given by equation 5 [Hyndman et al., 2020].

$$\boldsymbol{\Phi_P}(B^S)\boldsymbol{\phi_p}(B) w_t = \boldsymbol{\Theta_Q}(B^S)\boldsymbol{\theta_q}(B)\varepsilon_t. \tag{5}$$

In order to choose a *Seasonal ARIMA model* for each sensor (concerning selecting an appropriate model *order*, that is the values p, q, P, Q, D and d) and the *estimations* of the parameters of model, the `arima` function from `stats` package was used.

With the objective of estimating the parameters (for given values of p, d, q, P, D and Q) of the *Seasonal ARIMA* model, *the Maximum Likelihood Estimation* (*MLE*) method was applied. Furthermore, the *corrected Akaike's Information Criterion* ($AICc$) was useful for determining the order of a *Seasonal ARIMA* model. In this study, the values of D and d that were used were those that accommodate $D + d < 2$ and $d \leq 1$. The *sampleACF* plot and the *sample PACF* plot of **r** were used to determine appropriate values for p, q, P and Q. The most successful model for each time series was chosen according to the lowest $AICc$ value and the results from *the analysis of the residuals* from the chosen models. The $AICc$ values were compared for models which have the same orders of differencing with the same values of d and D.

3. The *TGARCH-Student* model: statistical studies and papers have proposed several specifications for σ_t for the *TGARCH* process [Palma, 2016, Zakoian, 1994a, Kovalevsky, 2018]. It is beyond the scope of this report to explain them all. Instead, this study will explain the particular conditionally heteroskedastic processes used. The innovations $\{\varepsilon_t, t \in \mathbb{Z}\}$ follow a conditionally heteroskedastic process, if it can be written as $\varepsilon_t = \sigma_t \epsilon_t$. Where errors $\{\epsilon_t, t \in \mathbb{Z}\}$ are an *i.i.d.* process with mean 0 and variance 1. Furthermore, the conditional mean (μ_t) and the conditional variance of process $\{\varepsilon_t, t \in \mathbb{Z}\}$ can be written as $\mu_t = E(\epsilon_t | \epsilon_{t-1}, \epsilon_{t-2}, \ldots)$ and $\sigma_t^2 = E(\epsilon_t^2 | \epsilon_{t-1}, \epsilon_{t-2}, \ldots)$.

   The innovations $\{\varepsilon_t, t \in \mathbb{Z}\}$ follow a process of the family *GARCH* model (*fGARCH*) [Kovalevsky, 2018] if it can be written as equation 6, where the conditional mean and variance are used to scale the residuals $z_t = \frac{\epsilon_t - \mu_t}{\sigma_t}$.

$$\begin{aligned} \varepsilon_t &= \sigma_t \epsilon_t, \\ \sigma_t^\lambda &= (\omega + \Sigma_{j=1}^N \varsigma_j V_{jt}) \\ &\quad + \Sigma_{j=1}^q \alpha_j \sigma_{t-j}^\lambda (|z_{t-j} - \eta_{2j}| - \eta_{1j}(z_{t-j} - \eta_{2j}))^\delta \\ &\quad + \Sigma_{j=1}^p \beta_j \sigma_{t-j}^\lambda. \end{aligned} \tag{6}$$

Equation 6 is a *Box-Cox transformation* for the conditional standard deviation whose *shape* is determined by λ, and the parameter δ transforms the absolute value function, which subjects it to rotations and shifts through the η_{1j} and η_{2j} parameters respectively.

N represents the number of *external regressors* V_j (which are passed pre-lagged) [Kovalevsky, 2018]. If $\lambda = \delta = 1$, $\eta_{2j} = 0$, $|\eta_{1j}| \leq 1$ the innovations $\{\varepsilon_t, t \in \mathbb{Z}\}$ can be written as equation 7 and $\{\varepsilon_t, t \in \mathbb{Z}\}$ follows a *Threshold Generalized Autoregressive Conditional Heteroskedastic* process with parameters s and m (denoted *TGARCH*(s, m)). In this work, *external regressors* were not considered [Kovalevsky, 2018].

$$\begin{aligned} &\varepsilon_t = \sigma_t \epsilon_t, \\ &\sigma_t = \omega + \Sigma_{j=1}^{m} \alpha_j \sigma_{t-j}(|z_{t-j}| - \eta_{1j} z_{t-j}) + \Sigma_{j=1}^{s} \beta_j \sigma_{t-j}. \end{aligned} \tag{7}$$

In this study, $z_t = \frac{\epsilon_t}{\sigma_t}$ and *external regressors* were not considered, thus σ_t can be written as equation 8. Furthermore, it was considered that ϵ_t follows a *Student* distribution with parameter v. The *TGARCH-Student model* was used as an alternative to *Normal* distribution for fitting the standarized *errors* (ϵ_t) [Kovalevsky, 2018].

$$\sigma_t = \begin{cases} \omega + \Sigma_{j=1}^{m} \alpha_j \epsilon_{t-j}(1 - \eta_{1j}) + \Sigma_{j=1}^{s} \beta_j \sigma_{t-j} & if \quad \epsilon_{t-j} \geq 0 \\ \\ \omega - \Sigma_{j=1}^{m} \alpha_j \epsilon_{t-j}(1 + \eta_{1j}) + \Sigma_{j=1}^{s} \beta_j \sigma_{t-j} & if \quad \epsilon_{t-j} < 0 \end{cases} \tag{8}$$

The analyses were carried out separately for various groups: winter 1, winter 2, and spring and summer, for both 2008 and 2010. The final model was applied to each group. The models are detailed according to equations 5 and 8 below.

□ For winter 1 (2008): the time series r_t follows a *Seasonal ARIMA* $(1,1,0)$ $(2,0,0)_{24}$ *-TGARCH-Student* $(1,1)$ *process*.

The *Seasonal ARIMA*$(1,1,0)(2,0,0)_{24}$ *process* is given by:

$$\begin{aligned} \boldsymbol{\Phi_2}(B^{24})\boldsymbol{\phi_1}(B) w_t &= \varepsilon_t, \\ AR : \boldsymbol{\phi_1}(B) &= 1 - \phi_1 B, \\ SAR : \boldsymbol{\Phi_2}(B^{24}) &= 1 - \Phi_1 B^{24} - \Phi_2 B^{2(24)}. \end{aligned}$$

Thus, w_t follows the process $w_t = -\Phi_2\phi_1 w_{t-49} + \Phi_2 w_{t-48} - \Phi_1\phi_1 w_{t-25} + \Phi_1 w_{t-24} + \phi_1 w_{t-1} + \varepsilon_t$ and ε_t follows a *TGARCH-Student* $(1,1)$ *process* given by

$$\begin{aligned} &\varepsilon_t = \sigma_t \epsilon_t, \\ &\sigma_t = \begin{cases} \omega + \alpha_1 \epsilon_{t-1}(1 - \eta_{11}) + \beta_1 \sigma_{t-1} & if \quad \epsilon_{t-1} \geq 0 \\ \\ \omega - \alpha_1 \epsilon_{t-1}(1 + \eta_{11}) + \beta_1 \sigma_{t-1} & if \quad \epsilon_{t-1} < 0 \end{cases} \\ &\epsilon_t \sim Student(v), \end{aligned}$$

□ For winter 1 (2010): the time series r_t follows an *ARIMA* $(1,1,2)-$*TGARCH-Student*$(1,1)$ *process*.

The *ARIMA*$(1,1,2)$ process is given by:

$$\begin{aligned} \boldsymbol{\phi_1}(B) w_t &= \boldsymbol{\theta_2}(B)\varepsilon_t, \\ AR : \boldsymbol{\phi_1}(B) &= 1 - \phi_1 B, \\ MA : \boldsymbol{\theta_2}(B) &= 1 + \theta_1 B + \theta_2 B^2. \end{aligned}$$

Thus, W_t follows the process $w_t = \phi_1 w_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_1 \varepsilon_{t-1} + \varepsilon_t$ and ε_t follow a *TGARCH-Student* $(1,1)$ *process*.

☐ For winter 2 (2008 and 2010): the time series r_t follows a *Seasonal ARIMA* $(1,1,1) - (2,0,0)_{24}$-*TGARCH-Student*$(1,1)$ *process*.

The *ARIMA* $(1,1,1) - (2,0,0)_{24}$ *process* is given by

$$\begin{aligned} \boldsymbol{\Phi_2}(B^{24})\boldsymbol{\phi_1}(B)w_t &= \boldsymbol{\theta_1}(B)\varepsilon_t, \\ AR : \boldsymbol{\phi_1}(B) &= 1 - \phi_1 B, \\ MA : \boldsymbol{\theta_1}(B) &= 1 + \theta_1 B, \\ SAR : \boldsymbol{\Phi_2}(B^{24}) &= 1 - \Phi_1 B^{24} - \Phi_2 B^{2(24)}. \end{aligned}$$

Thus, W_t follows the process $w_t = -\Phi_2\phi_1 w_{t-49} + \Phi_2 w_{t-48} - \Phi_1\phi_1 w_{t-25} + \Phi_1 w_{t-24} + \phi_1 w_{t-1} + \theta_1\epsilon_{t-1} + \epsilon_t$ and ε_t follows a *TGARCH-Student* $(1,1)$ *process*.

☐ For spring (2008 and 2010): the time series r_t follows a *Seasonal ARIMA* $(1,1,2) - (0,0,2)_{24}$-*TGARCH-Student*$(1,1)$ *process*.

The *ARIMA*$(1,1,2) - (0,0,2)_{24}$ *process* is given by

$$\begin{aligned} \boldsymbol{\phi_1}(B)w_t &= \boldsymbol{\Theta_2}(B^{24})\boldsymbol{\theta_2}(B)\varepsilon_t, \\ AR : \boldsymbol{\phi_1}(B) &= 1 - \phi_1 B, \\ MA : \boldsymbol{\theta_2}(B) &= 1 + \theta_1 B + \theta_2 B^2, \\ SMA : \boldsymbol{\Theta_2}(B^{24}) &= 1 + \Theta_1 B^{24} + \Theta_2 B^{2(24)}. \end{aligned}$$

Thus, W_t follows the process $w_t = \phi_1 w_{t-1} + \theta_2\Theta_2\varepsilon_{t-50} + \theta_1\Theta_2\varepsilon_{t-49} + \Theta_2\varepsilon_{t-48} + \theta_2\Theta_1\varepsilon_{t-26} + \theta_1\Theta_1\varepsilon_{t-25} + \Theta_1\varepsilon_{t-24} + \theta_2\varepsilon_{t-2} + \theta_1\varepsilon_{t-1} + \varepsilon_t$ and ε_t follows a *TGARCH-Student* $(1,1)$ *process*.

☐ For summer (2008 and 2010): the time series r_t follows a *Seasonal ARIMA* $(1,1,1) - (1,0,0)_{24}$-*TGARCH-Student*$(1,1)$.

The *ARIMA*$(1,1,1) - (1,0,0)_{24}$ *process* is given by

$$\begin{aligned} \boldsymbol{\phi_1}(B)\boldsymbol{\Phi_1}(B^{24})w_t &= \boldsymbol{\theta_1}(B)\varepsilon_t, \\ AR : \boldsymbol{\phi_1}(B) &= 1 - \phi_1 B, \\ MA : \boldsymbol{\theta_1}(B) &= 1 + \theta_1 B, \\ SAR : \boldsymbol{\Phi_1}(B^{24}) &= 1 - \Phi_1 B^{24}. \end{aligned}$$

Thus, W_t follows the process $w_t = \Phi_1 w_{t-24} - \phi_1\Phi_1 w_{t-25} + \phi_1 w_{t-1} + \theta_1\varepsilon_{t-1} + \varepsilon_t$ and ε_t follows a *TGARCH-Student* $(1,1)$ *process*.

In this study, *the Dickey-Fuller stationarity* test [Fuller, 1996] was applied to residuals from the model. In this test, the H_0 is: "the errors from the models are not stationary". The test was computed using the `adf.test` function from the `aTSA` package. Also, the *Ljung Box serial autocorrelation* test [Box and Pierce, 1970, Ljung and Box, 1978], with H_0: "the errors from the models do not have serial autocorrelation" applied to the residuals. The test was computed using the `Box.test` function from the `stats` package. The *Shapiro Wilk normality* test [Royston, 1982a,b] with H_0 : "the errors from the model are distributed normally" was applied to the residuals. The test was computed using the `shapiro.test`

function from the `stats` package. Furthermore, the *Lagrange Multiplier* test (*LM*) was applied to the square root of the residuals from the chosen model to determine *Autoregressive Conditional Heteroscedasticity (Arch effects)* [Tsay, 2010]. The H_0 is: "the errors from model do not have an *Arch effect*". The test was computed using the `arch.test` function from the `aTSA` package. The `ugarchfit` function from the `rugarch` package was used to fit the *TGARCH-Student* model for the residual of the *Seasonal ARIMA* model.

When analyzing the residuals of the *ARIMA-TGARCH-Student* models for 2008 and 2010, there were some models that did not satisfy the conditions of step 2. To extract the information that the models did not capture, some features of the residuals were computed. The *classification variables* for the this method corresponds to the estimates of parameters and features from residuals: (1) Estimates of parameters of *ARIMA* model: the first parameter of the autoregressive component (`ar1`), the first parameter of the moving average component (`ma1`), the first parameter of the seasonal autoregressive component (`sar1`), the second parameter of the seasonal autoregressive component (`sar2`), (2) Estimates of parameters of *TGARCH* model: the *ARCH* parameter (`alpha1`), the rotation parameter (`eta11`), the *GARCH* parameter (`beta1`), the variance intercept parameter (`omega`) and the Student parameter of the residuals (`shape`), and (3) Features calculated from residuals: maximum of *spectral density* (`spec.mx`), frequency corresponding to maximum of *spectral density* (`freq`), variance of the *sample ACF* (`acf.v`), mean of the *sample ACF* (`acf.m`), median of the *sample ACF* (`acf.md`), range of the *sample ACF* (`acf.r`) and variance of the residuals (`res.v`).

3.2.3 M3: Additive Seasonal Holt-Winters (SH-W) prediction function

The *Holt-Winters (H-W)* method [Holt, 2004] is an extended Holt's method [Winters, 1960]. This method is an algorithm for producing point forecasts only [Hyndman et al., 2008]. The *Seasonal H-W (SH-W)* method is based on three smoothing equations: for the level component (a_t) at time t, the trend component (b_t) at time t, and for seasonality components (S_t) at time t. Each component corresponds smoothing parameters α, β and γ. Furthermore, s denotes the frequency of the seasonality [Hyndman et al., 2008]. On the other hand, there are two different *SH-W* methods, depending on whether seasonality is modeled in a multiplicative or additive way [Hyndman et al., 2008]. The estimates of parameters a, b, S were determined by minimizing the *squared prediction error* [Holt, 2004, Winters, 1960].

In this study an *Additive SH-W prediction function* for each observed time series $\mathbf{r}$ was studied for winter (1 and 2), spring and summer (2008 and 2010). The *Additive SH-W* method was fitted to both the observed time series $\mathbf{r}$ and $\mathbf{RH_h}$. The best results were obtained with $\mathbf{r}$. The frequency of the seasonality considered was 24 hours.

The *Additive SH-W prediction* function (for observed time series $\mathbf{r}$ with period length s) is given by

$$\begin{aligned} \hat{r}_{t+h|t} &= a_t + hb_t + S_{t\oplus h}, \\ t \oplus h &= t - s + 1 + (h-1) \bmod s, \end{aligned}$$

where a_t, b_t and S_t are given by

$$\begin{aligned} a_t &= \alpha(r_t - S_{t-s}) + (1-\alpha)(a_{t-1} + b_{t-1}), \\ b_t &= \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1}, \\ S_t &= \gamma(r_t - a_t) + (1-\gamma)S_{t-s}, \end{aligned}$$

where $0 \leq \alpha \leq 1$, $0 \leq \beta \leq 1$ and $0 \leq \gamma \leq 1$. The previous equations were computed where t is greater than s. When presenting outcomes from the algorithm, a corresponds

to the last level of a_t and b is the last slope of b_t and S_1 to S_{24} are the last seasonal of S_t, $t = 1, \ldots, 24$.

The `HoltWinters` function from the `stats` package was used to fit the *Additive SH-W model*. The `shapiro.test` function from the stats package and `ks.test` from the `dgof` package were used to apply the *normality tests*.

According to the analyses of the residuals for all groups, for at least 80% of the cases the hypothesis that the errors from the method do not have serial autocorrelation (lag: from 2 to 49) was rejected. Also, the normality tests rejected the normality of the errors for at least 80% of the cases. In order to extract more information, some features of the residuals were calculated.

The *classification variables* for this method correspond to the estimates of the parameters and features from residuals: (1) Estimates of parameters of the *SH-W* method: *level component* (`a`), *trend component* (`b`), and *seasonal component* (`s1, s2, ..., s24`). (2) Features obtained from residuals: mean of the *sample ACF* (`acf.m`), median of the *sample ACF* (`acf.md`), range of the *sample ACF* (`acf.r`) and variance of the *sample ACF* (`acf.v`), *sum of squared estimate of errors* (`sse`) [Holt, 2004, Winters, 1960], maximum of *spectral density* (`spec.mx`), frequency corresponding to maximum of *spectral density* (`freq`), statistic of the *Shapiro-Wilk* normality test (`shap.w`) and statistic of the *Kolgomorov-Smirnov* normality test (`kolg.d`).

In this research the *classification data* consists of the values of the *classification variables* determined by the three aforementioned methods (per year). In 2008 the order of the data sets (number of sensors $\times$ number of variables) by methods 1, 2 and 3 are as follows 23×60, 23×141, 23×49. In 2010 the order of the data sets are 20×60, 20×141, 20×49 respectively. The following classification method uses the *classification data sets*.

3.3 Method of classification Sparse PLS-DA

The *sPLS-DA* is a special variation of the *sPLS* and this in turn is a variation of the *PLS*. The *sPLS* performs simultaneous variable selection in the design and response matrix using a *penalization* in the *PLS*. In this study, the response matrix $\mathbf{Y}$ only has one column. When applying the *sPLS-DA*, the response $\mathbf{Y}$ (that is qualitative) is modified as a dummy block matrix called $\mathbf{Z}$. The *sPLS* regression is then run as if $\mathbf{Z}$ is a continuous matrix. A description of the *sPLS-DA* is presented below using *PLS* and *sPLS* as references as well as a description of the *prediction distances*, *algorithm of imputation*, *M-fold cross validation* and *classification error rate*.

3.3.1 Model

1. The *Partial Least Squares* (*PLS*) *regression*: regression [Wold, 1966] is a popular alternative to *OLS* when handling multicollinearity. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be a design matrix whose elements correspond to the values of the p variables and n sensors from the *classification data*, and $\mathbf{Y} \in \mathbb{R}^{n \times q}$ is a response matrix. *PLS* sequentially finds 2 lists of H orthonormal vectors, $(u_1, \ldots, u_H)$ and $(v_1, \ldots, v_H)$, such that the pair (u_h, v_h) solves

$$\max_{\mathbf{u},\mathbf{v}} cov(\mathbf{X}_{h-1}\mathbf{u}, \mathbf{Y}_{h-1}\mathbf{v}), \quad s.t. \quad \|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1, \tag{9}$$

where $\mathbf{X}_{h-1}$ is the orthogonal projection of $\mathbf{X}$ on $span\{\xi_1, \ldots, \xi_{h-1}\}^{\perp}$ and $\xi_h = \mathbf{X}_{h-1}\mathbf{u}_h$. It can be shown that $\mathbf{u}_h$ and $\mathbf{v}_h$ are equal to the 1st left and right singular vectors of $\mathbf{M}_{h-1} = cov(\mathbf{X}_{h-1}, \mathbf{Y}_{h-1})$. Once $\Xi = [\xi_1, \ldots, \xi_H]$ is computed, the matrices $\mathbf{Y}$ and $\mathbf{X}$ are modelled as $\mathbf{X} = \Xi\mathbf{C} + \mathbf{E}_1$ and $\mathbf{Y} = \Xi\mathbf{D} + \mathbf{E}_2$, where $\mathbf{C}$ and $\mathbf{D}$ are regression coefficients, while $\mathbf{E}_1$ and $\mathbf{E}_2$ are random errors. From this perspective, *PLS* can

be understood as another dimensionality reduction technique. However, unlike similar methods like *PCA*, *PLS* takes the covariance between covariates and the responses into account.

2. The *Sparse PLS* (*sPLS*): although *PLS* achieves a dimensionality reduction, its output is difficult to interpret, because each component of $\boldsymbol{\Xi}$ is a combination of all the original variables. This problem can be solved by adding *sparsity-promoting penalties* (e.g. *lasso penalties*) to the objective function of problem (9). However, this *naive approach* produces a difficult problem to solve. Recall, however, that the 1st left and right singular vectors of a matrix are also related to the their rank-1 approximation. Indeed, consider the problem

$$\min_{\mathbf{u},\mathbf{v}} \|\mathbf{M}_{h-1} - \mathbf{u}\mathbf{v}^\top\|_F^2, \quad s.t. \quad \|\mathbf{u}\|_2 = 1.$$

It can be shown that one solution to this problem is $\hat{\mathbf{u}} = \mathbf{u}_h$ and $\hat{\mathbf{v}} \propto \mathbf{v}_h$ [Lê Cao et al., 2011].

Exploiting this, Lê Cao et al. [2011] suggested replacing the pair $(\mathbf{u}_h, \mathbf{v}_h)$ with the solution of the following problem:

$$\min_{\mathbf{u},\mathbf{v}} \|\mathbf{M}_{h-1} - \mathbf{u}\mathbf{v}^\top\|_F^2 + P_{\lambda_1}(\mathbf{u}) + P_{\lambda_2}(\mathbf{v}), \quad s.t. \quad \|\mathbf{u}\|_2 = 1,$$

where P_{λ_1} and P_{λ_2} are given by $P_{\lambda_j}(.) = sign(.)(|.| - \lambda_j)_+, \quad j = 1, 2$. Lê Cao et al. [2011] show that this problem can be solved in a fairly efficient way. From now on, here it will be assumed that $\lambda_2 = 0$.

3. *Sparse PLS-DA* (*sPLS-DA*): An indicator matrix $\mathbf{Z} \in \{0,1\}^{n \times K}$ was created, where $z_{ik} = I(Y_i = k)$, with $k = 1, 2, \ldots, K$. *PLS-DA* correspond to compute the *PLS* as if $\mathbf{Z}$ were the response matrix and $\boldsymbol{\Xi}$ as the design matrix [Lê Cao et al., 2011].

There are two version of *Sparse SPLS* [Lê Cao et al., 2011, Rohart et al., 2017b,a]. Lê Cao et al. [2011], Rohart et al. [2017a] proposed a version that use one-step procedure and Rohart et al. [2017a] developed a version of *sPL-DA* with the penalty ℓ_1 (*Lasso*) on the loading vector $\mathbf{u}$ to shrink some coefficients to zero. The latter is applied in this research.

In this study, the number of classes (K) is 3 and their corresponding values are: 1 is $\mathcal{F}$, 2 is $\mathcal{RC}$ and 3 is $\mathcal{W}$. The elements used to apply *sPLS-DA* were: (1) a matrix with dimension $n \times p$ called $\boldsymbol{X}$ where p is the number of *classification variables* and n is the number of the time series and $\boldsymbol{X}$ is the *classification data*, (2) a factor vector of length n called $\boldsymbol{Y}$, this vector indicates the class of each time series (or sensor), (3) a dummy matrix called $\boldsymbol{Z}$ with dimension $n \times K$, where n the number of the time series and K the number of classes. The matrix $\boldsymbol{Z}$ is defined using the vector $\boldsymbol{Y}$. The values for each column of $\boldsymbol{Z}$ are either 0 or 1. For the first column, if a sensor is in position $\mathcal{RC}$ the value is 1 otherwise it is 0. For the second column, if a sensor is in position $\mathcal{F}$ the value is 1 otherwise it is 0 and finally, in the third column, if a sensor is in position $\mathcal{W}$ the value is 1 otherwise it is 0. Furthermore, the main outputs from the analysis were: (1) a set of components associated with $\boldsymbol{X}$ and $\boldsymbol{Z}$, (2) a set of loading vectors. Each of their components are assigned to a variable to define each component. Loading vectors are obtained to maximize the covariance between a linear combination of the variables from $\boldsymbol{X}$ (the $\boldsymbol{X} - component$) and from $\boldsymbol{Z}$ (the $\boldsymbol{Z} - component$), (3) a list of selected variables from $\boldsymbol{X}$ that are associated with each component, (4) the values of the *BER*

for each component and (5) the prediction class for each time series. In the `R` software, the *sPL-DA* can be computed with the functions `tune.splsda` and `splsda` from the `mixOmics` package.

3.3.2 Prediction distances

To predict the position of the sensors the following *prediction distances* can be used: the *maximum* distance, *centroid* distance and *Mahalanobis* distance. In this study, *centroid* distance was selected. A brief description of this distance is presented below.

The matrix $\mathbf{Z}$ is a dummy matrix of size $n \times K$. Furthermore, $\mathbf{Z}_{new}$ and $\mathbf{X}_{new}$ are derived from the $\mathbf{X}$ and $\mathbf{Z}$ training data sets with new observations (n_{new}). The prediction $\widehat{\mathbf{Z}}_{new}$ from a model with H components can be estimated as:

$$\widehat{\mathbf{Z}}_{new} = \mathbf{X}_{new}\boldsymbol{\Xi}^*(\mathbf{C}^{*\top}\boldsymbol{\Xi}^*)^{-1}\mathbf{D}^*,$$

where $\boldsymbol{\Xi}^*$, $\mathbf{C}^*$ and $\mathbf{D}^*$ are derived from the $\mathbf{X}$ and $\mathbf{Z}$ training data sets. $\boldsymbol{\Xi}^*$ is a $P \times H$ matrix containing the loading vectors associated to $\mathbf{X}$, $\mathbf{C}^*$ is a $P \times H$ matrix containing the regression coefficients of $\mathbf{X}$ on its H latent components and $\mathbf{D}^*$ is a $H \times K$ matrix containing the regression coefficients of $\mathbf{Z}$ on the H latent components associated with $\mathbf{X}$ [Rohart et al., 2017a, Lê Cao et al., 2011].

The predicted *components* $\mathbf{T}_{\mathbf{pred}}$ of size $n_{new} \times H$ are given by:

$$\mathbf{T}_{\mathbf{pred}} = \mathbf{X}_{new}\boldsymbol{\Xi}^*(\mathbf{C}^{*\top}\boldsymbol{\Xi}^*)^{-1},$$

and the prediction distance *centroid* is given by:

$$dist(\boldsymbol{T_{pred}}, \boldsymbol{G_k}) = \sqrt{\sum_{h=1}^{H}\left((\boldsymbol{T_{pred}})_h - (\boldsymbol{G_k})_h\right)^2},$$

where $\boldsymbol{G_k}$ is a *centroid* of all the learning set samples belonging to the class $k \leq K$ based on the H latent components associated with $\mathbf{X}$. The predicted position of a new time series is the result from the following equation:

$$\text{argmin}_{1\leq k\leq K} dist(\mathbf{T}_{\mathbf{pred}}, \boldsymbol{G_k}).$$

Details of *Mahalanobis* and *maximum* distances can be viewed in [Rohart et al., 2017a].

3.3.3 Algorithm of imputation

In this research, each *anomalous* value of the each *classification variable* was considered as a *missing* value when applying the *sPLS-DA*. The *outlier* and *anomalous* values were identified using box plots for each variable.

In 2008: 1.04% (M1), 1.06% (M2) and 1.24% (M3) correspond to the percentage of the *missing* values of the *classification data sets*. In 2010 the corresponding percentages were 1.39%, 0.49% and 0.5%. Before carrying out the *sPLS-DA*, *classification data sets* were normalized and *missing* values were imputed.

The *missing* values were substituted using the *Non linear estimation by iterative partial least squares* (*NIPALS*, [Wold, 1966]). The *NIPALS* is applied internally with the functions `splsda` and `tune.splsda` in `R` software. Details of the algorithm *NIPALS* can be found in [Tenenhaus, 1998, Rohart et al., 2017b].

3.3.4 M-fold cross validation and classification error rate

In order to evaluate the results of the *sPLS-DA* method, repeated *M-fold cross-validation* [Rohart et al., 2017a] was applied for the maximum number of ten components. It was performed with stratified subsampling where all positions ($\mathcal{RC}$, $\mathcal{F}$ and $\mathcal{W}$) are represented in each fold, where $M = 3$. Thus, *M-fold cross-validation* was repeated 1,000 times for each fold.

With the objective of assessing the optimal number of components, a plot with results from the *sPLS-DA* was used. The plot outputs are the mean of the *Overall classification error rate* and *Balanced classification error rate* (*BER*) [Rohart et al., 2017b] and the *standard deviation* according to three *prediction* distances (*maximum*, *centroid* and *Mahalanobis*) [Rohart et al., 2017b]. *BER* calculates the average proportion of wrongly classified samples in each class, weighted by the number of samples in each class. *BER* is less biased towards majority classes during the performance assessment. Each result was carried out with *M-fold cross -validation* repeated 1,000 times for each component.

The optimal number of components was achieved by determining the best performance (the lowest error), based on *BER* and *centroid* distance. Also, the optimal number of variables for each component was obtained using a grid of the number of variables to keep values that will be assessed on each component in $X - loadings$, (one component at a time. Similar to above, *M-fold cross-validation*, repeated 1,000 times) with a *centroid* distance prediction. Based on the results of the optimal number of components and the optimal number of variables, the final *sPLS-DA* model was applied.

4. Results

4.1 Methods to determine classification variables

In respect to the residual analysis: (1) For the *Arima-TGARCH-Student* models for 2010 and 2008 (in brackets) at least 70% (53%) of time series satisfied all tests in the residual analysis. (2) For the *Additive SH-W* methods the hypothesis of the stationarity of the errors (*Dickey Fuller test*) was accepted for all groups. However, the hypothesis of serial non-correlation of errors (*Ljung Box test*) was rejected for all groups.

The *significance level* values used for each statistical test were: 0.02 for winter 1 and spring (2008), 0.05 for winter 2 and summer (2008). As well as 0.05 for winters 1 & 2 and summer (2010) and 0.02 for spring (2010).

4.2 sPLS-DA

The final model for M2 and M3 (2008) includes 1 component, and 5 and 10 selected variables respectively. The final model for M1 (2008) includes 2 components and 15 selected variables for both components. The final model for M1, M2 and M3 (2010) includes 1 component, and 15 selected variables for all components (see Table 1).

Table 1 shows the values of the *BER* from the *sPLS-DA* for both years 2008 and 2010. When applying *sPLS-DA* for each method (M1, M2 and M3), all the variables (for each season and year) were used. In this table, the variables which were determined for each of the components were ordered from highest to lowest, according to the *absolute value* of their loading weights. Variables with loadings of negative values are shown in blue. The three most important variables for the first component for each method are: (2008) For M1 they were `spec.mx`, `res.v` and `omega`, for M2 they were `sse`, `spec.mx` and `kolg.d` and for M3 they were `spec.mx`, `rMh` and `rVh`. (2010) for M1 they were `res.v`,

`spec.mx` and `omega`, for M2 they were `spec.mx` , `sse` and `kolg.d`, for M3 they were `spec.mx` , `rMd` and `rMh`.

The *classification variables* selected for M2 and M3 were: (from the residuals) `res.v`, `sse`, `kolg.d`, `shape`, `spec.mx`, `acf.m` and `acf.md` and (from the models) `omega`, `alpha`, `b`, `s18`, `s19`, `s20` and `s24` (see Table 1). From the residuals, the first two features are aimed at explaining the variance that wasn't explained by the models, the second and third features study the distribution of residuals, and the three last features are intended to represent the dynamic structure of each time series. When looking at the models, for M1, the first feature aimed to explain the changes in the mean of the *volatility* and the second feature aimed to quantify the impact of the rotation on the *volatility*. For M2, the third feature is related to the *trend component* of the time series and the fourth to the seventh features are related to the *seasonal components* of the time series.

The *classification variables* selected for M1 were: `rMh`, `rMd`, `rVh`, `rVd`, `spec.mx`, `freq`, `pacf1`, `pacf2`, `pacf3`, `pacf4`, `acf.v` and `acf.r` (see Table 1). The two first features are aimed at explaining the changes in the mean of the time series and the third and fourth features are intended to explain the changes in the variance of the time series. Finally, the fifth to the twelfth features represent the dynamic structure of each time series.

In 2008, the values of the *BER* from M1, M2 and M3 are as follows: 30.02%, 22.60% and 24.05% and in 2010, the values of the *BER* are: 24.08%, 12.81% and 21.17% (see Table 1a).

Table 1: Results from *sPLS-DA* (2008 and 2010): variables selected per component (C) and per Method (M). Optimal number of components and of variables, when using all variables ($Wr1$ stands for winter 1, $Wr2$ for winter 2, Sp for spring and Sm for summer). The variables which determine each of the components are ordered from highest to lowest, according to the *absolute value* of their loading weights. Variables with weights of negative values are shown in blue. For each component it shows the values of the *Balanced classification Error Rate* (BER).

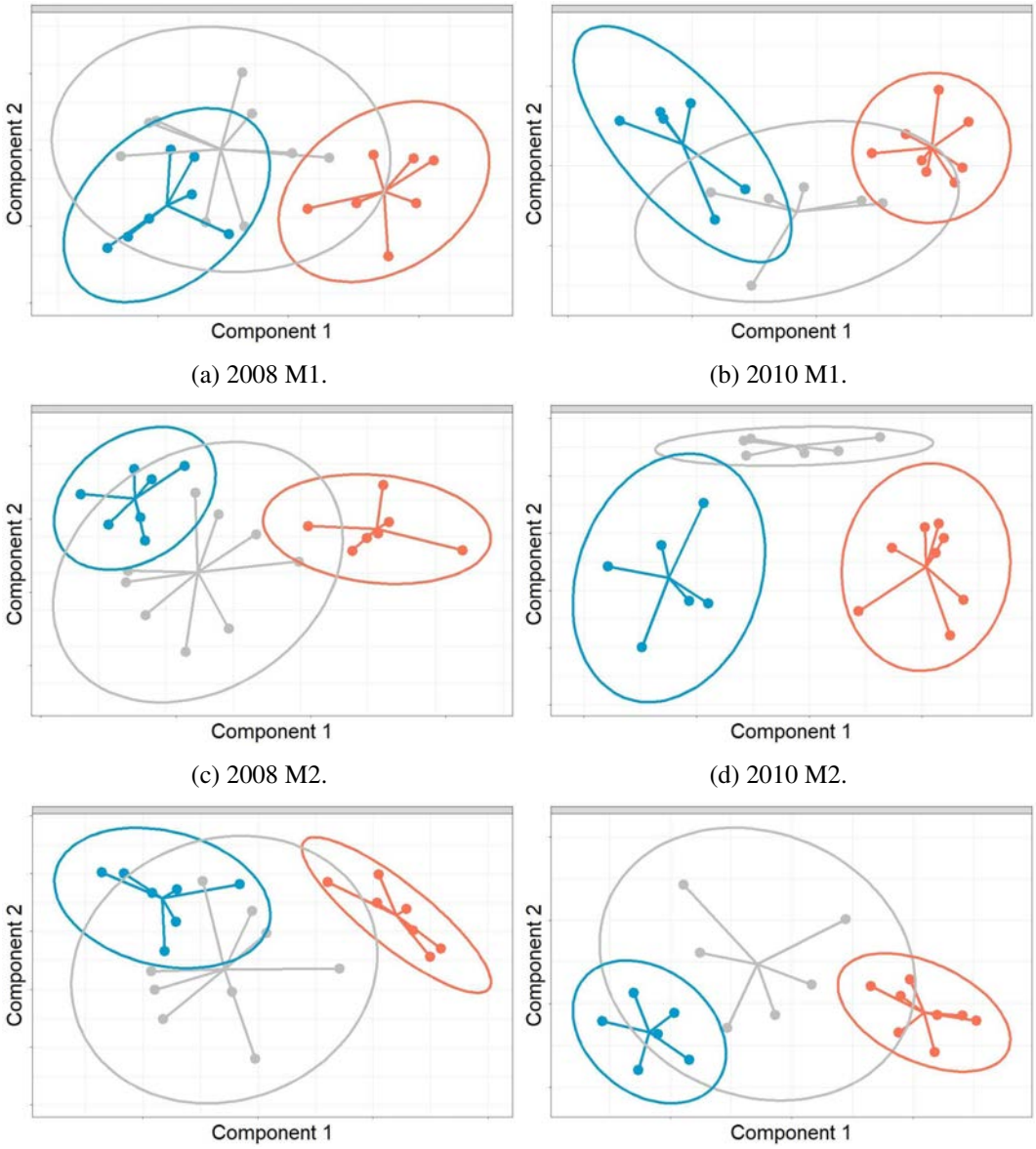
M	C	Variables	BER
1	1st	$Wr1_{spec.mx}$, $Wr1_{rMh}$, $Wr2_{rMh}$, Sp_{rMh}, Sm_{rMh}, $Wr1_{rVh}$,$Wr2_{rVh}$ Sp_{rVh}, Sm_{rVh}, $Wr1_{rMd}$, $Wr2_{rMd}$, Sp_{rMd}, Sm_{rMd}, $Wr1_{rVd}$, $Wr2_{rVd}$	30.02%
1	2nd	$Wr2_{pacf3}$, Sp_{pacf1}, $Sm_{acf.r}$, $Sm_{acf.v}$, Sm_{pacf1}, Sm_{freq},$Sm_{acf.r}$, $Sm_{acf.v}$ $Wr2_{pacf2}$,Sm_{pacf4}, $Sm_{spec.mx}$,Sm_{pacf2}, $Wr1_{pacf2}$, $Sp_{spec.mx}$, $Wr2_{pacf4}$	
2	1st	$Wr1_{spec.mx}$,$Wr1_{res.v}$, $Wr2_{spec.mx}$, $Wr2_{res.v}$, $Wr1_{omega}$	22.60%
3	1st	$Wr1_{sse}$, $Wr1_{spec.mx}$,$Wr2_{sse}$,Sp_{sse}, $Wr1_{kolg.d}$, $Sp_{spec.mx}$, $Wr2_{kolg.d}$ $Sm_{kolg.d}$, Sm_{sse}, $Sm_{spec.mx}$	24.05%

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

M	C	Variables	BER
1	1st	$Wr1_{spec.mx}$, $Wr1_{rMd}$, $Wr2_{rMd}$, Sp_{rMd} ,Sm_{rMd}, $Wr1_{rMh}$, $Wr2_{rMh}$ Sp_{rMh}, Sm_{rMh}, $Wr1_{rVh}$, $Wr2_{rVh}$, Sp_{rVh} ,Sm_{rVh}, $Wr2_{spec.mx}$, $Wr2_{pacf2}$	24.08%
2	1st	$Sp_{res.v}$, $Sm_{res.v}$, $Sm_{spec.mx}$, $Wr1_{res.v}$, $Wr2_{res.v}$, $Sp_{spec.mx}$, Sp_{omega} Sm_{omega}, $Wr1_{spec.mx}$, $Wr2_{omega}$, $Wr2_{spec.mx}$, $Wr1_{alpha}$, $Wr1_{shape}$ $Sp_{acf.md}$,$Sp_{acf.m}$	12.81%
3	1st	$Wr2_{spec.mx}$, Sm_{sse}, $Sp_{kolg.d}$,Sp_{sse}, $Wr1_{kolg.d}$, $Wr2_{s1}$, $Sm_{kolg.d}$, $Wr2_{sse}$ $Wr2_{s24}$, $Sm_{spec.mx}$, Sm_{s19}, Sm_{s18} ,$Wr1_{b}$, Sm_{s20}, Sp_{s24}	21.17%

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

Components from the *sPLS-DA* are linear combinations of variables that might correspond to different groups either winter 1 or winter 2, or spring or summer. For the three methods (M1, M2 and M3) the results from *sPLS-DA* only had one component, except for M1 (2008). In the following paragraphs the variables that determine component 1 are explained.



(a) 2008 M1. (b) 2010 M1.

(c) 2008 M2. (d) 2010 M2.

(e) 2008 M3. (f) 2010 M3.

Figure 3: Column 1 corresponds to 2008 and column 2 to 2010. The positions are Frescoes ($\mathcal{F}$), Cornice and Ribs ($\mathcal{RC}$) and Wall ($\mathcal{W}$). $\mathcal{F}$ is in gray, the $\mathcal{RC}$ is in orange, $\mathcal{W}$ is in blue. Graphics correspond to estimates of comparisons of the sample representation using the first 2 latent variables from the *sPLS-DA* when using all variables for one year for each method. A star plot displays arrows from each group centroid, towards each sensor. The results are presented by method, as follows: (a) y (b) M1, (c) y (d) M2 and (e) y (f) M3.

M1 (2008): component 1 is a linear combination of variables from the four groups. In these groups, the common variable are `rMh`, `rVh` and `rMd`. The variable `rVd` is a common variable in winter 1 and winter 2. The variable `spec.mx` is only seen in winter 1. In 2010, `rMh`, `rMd` and `rVh` are the common variables in the four groups, `spec.mx` is the common variable in winter 1 and winter 2. `pacf2` is only found in winter 2 (see Table 1a).

M2 (2008): the variables that determined the component, correspond to the groups winter 1 and winter 2. In both, winter 1 and winter 2, the common variables are `spec.mx` and `res.v`. While `omega` is only a variable for winter 1 (see Table 1a). While in 2010

this component for the same method is a linear combination of variables from all groups: winter 1, winter 2, spring and summer. For these four groups the common variables are `spec.mx` and `res.v`. For winter 2, spring and summer the common variable is `omega`. The variables `alpha` and `shape` only correspond to winter 1. The variables `acf.md` and `acf` can only be found in spring (see Table 1b).

M3 (2008): the variables that determined the component correspond to the groups winter 1, winter 2, spring and summer. In the these groups, the common variable is `sse`. The variable `kolg.d` is a common variable in winter 1, winter 2 and summer. A common variable for winter 1, spring and summer is `spec.mx`. In 2010, `kolg.d` is a common variable in winter 1, spring and summer, `s24` is a common variable in winter 2, spring and summer, the variable `spec.mx` is a common variable in winter 2 and summer. While, the variable `b` only corresponds to winter 1, the variable `s1` is only found in winter 2, and the variables `s18`, `s19` and `s20` can only be found in summer (see Table 1a).

In the results for M1 (2008) from *sPLS-DA* a second component was necessary. Component 2 is a linear combination of variables from the four groups. The variable `pacf2` is the common variable in winter 1, winter 2 and summer. The variable `pacf4` is the common variable in winter 2 and summer. The variables `spec.mx`, `pacf1`, `acf.r` and `acf.v` are the common variables in spring and summer. The variable `pacf3` is only found in winter 2 and `freq` is only seen in summer (see Table 1a).

The results shown in the Figure 3 correspond to the sample plots on the first two components from the *sPLS-DA* applied to *classification data sets*. *Confidence ellipses* for each class are plotted to highlight the strength of the discrimination (confidence level is 95%). In 2008, the first component for each method showed that two groups of sensors were discriminated, $\mathcal{RC}$ from $\mathcal{W}$. In 2010, the first component for M2 displayed a clearer discrimination between sensors located on the three positions $\mathcal{RC}$, $\mathcal{F}$ and $\mathcal{W}$. While, for M3 and M1, two groups of sensors were discriminated, $\mathcal{RC}$ from $\mathcal{W}$ (see Figure 3).

5. Discussion

In this article, a methodology for classifying time series has been proposed. The methodology consists of the following two steps. For step 1, three methods are used to obtain the *classification variables*. Method 1 (M1) utilizes functions such as *sample ACF*, *sample PACF* and *spectral density* to calculate features of the time series. Method 2 (M2) employs the *ARIMA-TGARCH-Student* model and its parameters are estimated and features from the residuals are calculated (e.g., mean, variance or range of the values of the *ACF*, among others). Method 3 (M3) computes the *Additive S-HW*, its parameters are estimated, features from the residuals are obtained (i.e., functions such as *sample ACF*, *sample PACF*, *spectral density* and statistics from *Kolgomorov's* test are employed to calculate the features). For step 2, the *classification variables* determined in step (1) are used to apply the *sPLS-DA*.

According to M2 and M3, a low percentage of the parameters of models were considered as essential in the classification of the time series, while most of the essential variables were features from the residuals of models. This is most likely due to the similarity of the time series studied. In consequence, the characteristics that weren't explained by the models were decisive for capturing the differences between the time series. In relation to the *BER* from M2 and M3, M2 showed better performance than M3, possibly because the analysis of the residuals of M2 was better than the same analysis for M3. The features calculated from the residuals of the models were insufficient to capture the information the model didn't recognise. The variables from these methods explain the variance, the mean, the distribution of residuals and the dynamic structure of the time series.

In respect to M1, where a model-free approach is used, the *classification variables*

explain the changes in the mean and the variance of the time series. This method had the lowest performance. This might be due to the method capturing less information than methods M2 and M3.

One limitation of M2 is using a unique *Seasonal-ARIMA-TGARCH-Student* model to fit all of the time series for the same group. Selecting a unique model may not achieve the best estimates of the parameters. One limitation of method M1 is that if the time series require dividing by periods for analysis, the number of parameters grows considerably. This might make the results difficult to interpret.

In respect to the advantages of the methods: for M2, *Seasonal ARIMA-TGARCH-Student* captures more information about different characteristics of the time series than other methods. For M3, *Additive SH-W* is easier to fit than in method M1. For M1, the calculation of the characteristics from the common functions that are applied to the time series allows quick and easy calculations. The three methods use known functions and models in the time series that allow their easy computational implementation.

In the preliminary study Zarzo et al. [2011] about *RH* in Valencia's Cathedral, the time series were clustered according to the mean of *RH* and `rVd` (they called this DMV). In this study, the classification of the time series was improved for method M1, which also used the variable `rVd`. This might be a consequence of using: *sPLS-DA* and features related to *sample ACF*, *sample PACF*, *spectral density* and *frequency* corresponding to *spectral density*.

In respect to classification technique, one disadvantage of using the *sPLS-DA* is that it is only applicable to cases where the context of the problem of group classification is possible for each time series. However, this method makes it possible to classify time series that have very similar characteristics.

The proposed methodology might be improved by adding *classification variables* related to model predictions and other features of the time series such as *spectral features*. Additionally, *penalties* different from the ones proposed for the *sPLS-DA* could be explored, then the selection of essential *classification variables* for clustering time series could be evaluated. Additionally, it would be advisable to carry out a study in controlled scenarios, where time series can be simulated and different characteristics of the time series are controlled.

The methodology proposed here might be a good option to consider when there are no major differences between the time series of different groups. As, according to the characteristics and context of the problem it is possible to indicate the categories of the time series. This methodology can lead to the discovery of interesting patterns in time series datasets. Furthermore, it might help researchers understand the structure of data, clusters, anomalies, and other regularities in datasets, to develop prediction models, among others.

Funding: This study was given funding from the *European Union's Horizon 2020 research* and the *innovation programme* under grant agreement No 814624. The research was also given support from the *Instituto Colombiano de Crédito Educativo y Estudios Técnicos en el Exterior, ICETEX* through the *Programa crédito Pasaporte a la Ciencia*. As well as, *Pontificia Universidad Javeriana seccional Cali* (Nit 860013720-1) via the Convenio de Capacitación para Docentes O. J. 086/17.

Acknowledgments: The authors of the paper would like to acknowledge A. Muñoz-Sánchez from the Environmental department of the Universitat Politècnica de València for providing meteorological data.

References

AM Alonso, JR Berrendero, A Hernandez, and A Justel. Time series clustering based on forecast densities. *Computational statistics and data analysis*, 51:762–776, 2006.

Donald W. K Andrews. Tests for parameter instability and structural change With unknown change point. *Journal of the econometric society*, 61(4):821–856, 1993.

Donald W. K Andrews and Werner Ploberger. Optimal tests when a nuisance parameter is present only under the alternative. *Journal of the econometric society*, 62(6):1383–1414, 1994.

A Antoniadis, S Lambert-Lacroix, and F Leblanc. Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics*, 19(5):563–570, 2003.

AJ Bagnall, G Janacek, B de la Iglesia, and M Zhang. Clustering time series from mixture polynomial models with discretised data. In *Proceedings of the 2nd Australasian data mining workshop*, page 105–120, University of technology Sydney, 2003.

Eric Bair, Trevor Hastie, Debashis Paul, and Robert Tibshirani. Prediction by supervised principal components. *Journal of the american statistical association*, 101(473):119–137, 2006.

GEAPA Batista, X Wang, and EJ Keogh. A complexity-invariant distance measure for time series. In *Proceedings of the eleventh SIAM international conference on data mining, SDM11*, pages 699–710, 2011.

Z. W Birnbaum and Fred H Tingey. One-sided confidence contours for probability distribution functions. *Annals of mathematical statistics*, 22(4):592–596, 12 1951.

Peter Bloomfield. *Fourier analysis of time series: an introduction*. Wiley, San Francisco, second edition, 1976.

Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3): 307–27, 1986.

Anne-Laure Boulesteix. PLS dimension reduction for classification with microarray data. *Statistical applications in genetics and molecular biology*, 3, 2004.

E. P. G. Box and A. David Pierce. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the american statistical association*, 1970.

George E. P Box and Gwilym M Jenkins. *Time series analysis: forecasting and control (revised Edition)*. Holden-Day, revised edition, 1976.

AM Brandmaier. *Permutation distribution clustering and structural equation model trees*. PhD thesis, Universitat des Saarlandes, 2011.

L Breiman. Random forests. *Machine learning*, 45, 2001.

P. J Brockwell and R. A Davis. *Time series: theory and methods*. Springer Verlag, New York, second edition, 1991a.

P. J Brockwell and R. A Davis. *Time series: theory and methods*. Springer Verlag, 2 edition, 1991b.

J Caiado, N Crato, and D Peña. A periodogram-based metric for time series classification. *Computational Statistics and Data Analysis*, 50(10):2668–2684, 2006.

Gregory C Chow. Tests of equality between sets of coefficients in two linear regressions. *Journal of the econometric society*, 28(3):591–605, 1960.

Hyonho Chun and Sunduz Keleş. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the royal statistical society Series B (Statistical Methodology)*, 72:3–25, 2010.

D Chung and S Keles. Sparse partial least squares classification for high dimensional data. *Statistical applications in genetics and molecular biology*, 9(17), 2010.

R Cilibrasi and PMB Vitãnyi. Clustering by compression. *IEEE transactions on information theory*, 51: 1523–1545, 2005.

William J Conover. *Practical nonparametric statistics*. John Wiley and Sons, 1 edition, 1994.

J.D. Cryer and K.S. Chan. *Time series analysis: with applications in R*. Springer Texts in Statistics. Springer New York, 2008. URL `https://books.google.ca/books?id=bHke2k-QYP4C`.

Jian J Dai, Linh Lieu, and David Rocke. Dimension reduction for classification with gene expression microarray data. *Statistical applications in genetics and molecular biology*, 5, 2006.

E Deconinck, T Hancock, D Coomans, D.L Massart, and Y. Vander Heyden. Classification of drugs in absorption classes using the classification and regression trees (CART) methodology. *Journal of pharmaceutical and biomedical analysis*, 39(1):91–103, 2005.

Chouakria A Douzal and PN Nagabhushan. Adaptive dissimilarity index for measuring time series proximity. *Advances in data analysis and classification*, 1(1):5–21, 2007.

Rick Durrett. *Probability: theory and examples*. Cambridge University Press, fourth edition, 2000.

Robert F Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Journal of the econometric society*, 50(4):987–1007, 1982.

Robert F. Engle and Tim Bollerslev. Modelling the persistence of conditional variances. *Econometric reviews*, 5(1):1–50, 1986.

Gersende Fort and Sophie Lambert-Lacroix. Classification using partial least squares with penalized logistic regression. *Bioinformatics*, 21(7):1104, 2005.

TC Fu. A review on time series data mining. *Engineering applications of artificial intelligence*, 24(1):164–181, 2011.

W. A. Fuller. *Introduction to statistical time series*. John Wiley and Sons, New York, second edition, 1996.

Davide Gaetano. Forecast combinations in the presence of structural breaks: evidence from U.S. equity markets. *Mathematics*, 6:34, 03 2018.

Fernando-Juan García-Diego and Manuel Zarzo. Microclimate monitoring by multivariate statistical control: the renaissance frescoes of the cathedral of Valencia (Spain). *Journal of cultural heritage*, 11(3):339 – 344, 2010.

Alexios Ghalanos. *Introduction to the rugarch package (Version 1.4-3)*, 2020. R package version 1.4-3.

Isabelle Guyon and André Elisseefi. An introduction to variable and feature selection. *Journal of machine learning research*, 3(7-8), 2003.

James D Hamilton. *Time series analysis*. Princeton university press, 1 edition, 1994.

Bruce E. Hansen. Tests for parameter instability in regressions with i(1) processes. *Journal of business and economic statistics*, 20(1):45–59, 2002.

A. C. Harvey. *Time series models*. Harvester Wheatsheaf, New York, second edition, 1993.

Scott E. Hein and Raymond E. Spudeck. Forecasting the daily federal funds rate. *International journal of forecasting*, 4(4):581 – 591, 1988.

Arthur E Hoerl and Robert W Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Charles C Holt. Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, 20(1):5–10, 2004.

C. Hor, S. J. Watson, and Shanti MajithiaS. Daily load forecasting and maximum demand estimation using arima and garch. In *2006 International conference on probabilistic methods applied to power systems*, pages 1–6, 2006.

Rob Hyndman, George Athanasopoulos, Christoph Bergmeir, Gabriel Caceres, Leanne Chhay, Mitchell O´Hara-Wild, Fotios Petropoulos, Slava Razbash, Earo Wang, and Farah Yasmeen. *forecast: forecasting functions for time series and linear models*, 2020. R package version 8.12.

Rob J Hyndman, Anne B Koehler, J Keith Ord, and Ralph D Snyder. *Forecasting with exponential smoothing: the state space approach*. Springer series in statistics. Springer, 2008.

Tseng Jie-Jun and Li Sai-Ping. Quantifying volatility clustering in financial time series. *International review of financial analysis*, 23:11 – 19, 2012.

Thibaut Jombart, Sébastien Devillard, and Francois Balloux. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BiomMed central genomics*, 11(94), 2010.

Y Kakizawa, RH Shumway, and M Taniguchi. Discrimination and clustering for multivariate time series. *Journal of the American statistical association*, 93(441):328–340, 1998.

K Kalpakis, K Gada, and V Puttagunta. Distance measures for effective clustering of arima time-series. *In data mining, 2001. ICDM 2001. Proceedings IEEE international conference*, page 273–280, 2000.

E Keogh, S Lonardi, CA Ratanamahatana, L Wei, SH Lee, and J Handley. Compression based data mining of sequential data. *Data mining and knowledge discovery*, 14(1):99–129, 2007.

ZJ Kovačić. Classification of time series with applications to the leading indicator selection. In *In data science, classification, and related methods – proceedings of the fifth conference of the international federation of classification societies (IFCS-96)*, page 204–207, 1996.

Stanislav Kovalevsky. *QuantTools: enhanced quantitative trading modelling*, 2018. R package version 0.5.7.

P. Laux, S. Vogl, W. Qiu, H. R. Knoche, and H. Kunstmann. Copula-based statistical refinement of precipitation in RCM simulations over complex terrain. *Hydrology and earth system sciences*, 15(7):2401–2419, 2011.

Cao KA Lê, Olivier Goncalves, Philippe Besse, and Sébastien Gadat. Selection of biologically relevant genes with a wrapper stochast algorithm. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

Cao KA Lê, A Bonnet, and S Gadat. Multiclass classification and gene selection with a stochastic algorithm. *Computational statistics and data analysis*, 53(10):3601 – 3615, 2009.

Kim-Anh Lê Cao, Debra Rossouw, Christèle Robert-Granié, and Philippe Besse. A sparse PLS for variable selection when integrating omics data. *Statistical applications in genetics and molecular biology*, 7(1), 2008.

Kim-Anh Lê Cao, Pascal GP Martin, Christele Robert-Granie, and Philippe Besse. Sparse canonical methods for biological data integration: application to a cross-platform study. *BiomMed central bioinformatics*, 10 (34), 2009.

Kim-Anh Lê Cao, Simon Boitard, and Philippe Besse. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BiomMed central bioinformatics*, 12(1): 253, 2011.

Friedrich Leisch, Kurt Hornik, and Chung-Ming Kuan. Monitoring structural changes with the generalized fluctuation test. *Econometric theory*, 16(6):835–854, 2000.

M Li and P Vitãnyi. An introduction to kolmogorov complexity and its applications. *Text and Monographs in*

Computer Science, 2007.

M Li, JH Badger, X Chen, S Kwong, P Kearney, and H Zhang. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, 17(2):149–154, 2001.

M Li, X Chen, X Li, and PMB Ma, B Vitãnyi. The similarity metric. *IEEE transactions on information theory*, 50(12):3250–3264, 2004.

Warren Liao. Clustering of time series data a survey. *Pattern Recognition*, 38(11):1857 – 1874, 2005.

G. M Ljung and G. E. P Box. On a measure of Lack of fit in time series models. *Biometrika*, 65(2):297–303, 1978.

Helmut Lütkepohl and Fang Xu. The role of the log transformation in forecasting economic variables. *Empirical Economics*, 42(3):619–638, 2012.

EA Maharaj. A significance test for classifying ARMA models. *Journal of statistical computation and simulation*, 54(4):305–331, 1996.

E.A Maharaj. Cluster of time series. *Journal of classification*, 17(2):297–314, 2000.

EA Maharaj. Comparison of non-stationary time series in the frequency domain. *Computational statistics and Data Analysis*, 40(1):131–141, 2002.

Andrew V. Metcalfe and Paul S.P. Cowpertwait. *Introductory time series with R*. Springer series: Use R. Springer-Verlag, 2009.

T Oates, L Firoiu, and PR 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*, pages 17–21, 1999.

Wilfredo Palma. *Time series analysis. Wiley series in probability and statistics*. John Wiley and Sons Inc, har/psc edition, 2016.

Daniel Peña and Pedro Galeano. Multivariate analysis in vector time series. DES - Working Papers. Statistics and Econometrics. WS ws012415, Universidad Carlos III de Madrid. Departamento de Estadística, 2001. URL `https://ideas.repec.org/p/cte/wsrepe/ws012415.html`.

D Piccolo. A distance measure for classifying arima models. *Journal of time series analysis*, 11(2):153–164, 1990.

Werner Ploberger and Walter Kramer. The cusum test with OLS residuals. *Journal of the econometric society*, 60(2):271–285, 1992.

Debin Qiu. *aTSA: alternative time series analysis*, 2015. R package version 3.1.2.

R Core Team. *R: a language and environment for statistical computing*. R foundation for statistical computing, Vienna, Austria, 2014.

M Ramoni, P Sebastiani, and P Cohen. Bayesian clustering by dynamics. *Machine learning*, 47(1):91–121, 2002.

Florian Rohart, Benoît Gautier, Amrit Singh, and Kim-Anh Lê Cao. S1 Supplementary Information mixOmics: an R package for 'omics feature selection and multiple data integration. *PLoS computational biology*, 13 (11):e1005752, 2017a.

Florian Rohart, Benoît Gautier, Amrit Singh, and Kim-Anh Lê Cao. mixOmics: An R package for omics feature selection and multiple data integration. *PLoS computational biology*, 13(11):e1005752, 2017b.

Yaziz Siti Roslindar, Azizan Noor Azlinna, Ahmad Maizah Hura, and Zakaria Roslinazairimah. Modelling gold price using ARIMA-TGARCH. *Applied mathematical sciences*, 10:1391–1402, 2016.

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)*, 31(2):115–124, 1982a.

J. P Royston. Algorithm AS 181: The W Test for Normality. *Journal of the royal statistical society. Series C (applied Statistics)*, 31(2):176–180, 1982b.

P Smyth. Clustering sequences with hidden Markov models. *In advances in neural information processing systems*, page 648–654, 1997.

ZR Struzik and A Siebes. The haar wavelet in the time series similarity paradigm. In *In principles of data mining and knowledge discovery – proceedings of the third European conference, PKDD-99*, page 12–22, 1999.

M Tenenhaus. *La régression PLS: thórie et pratique*. Technip, 1998.

MS Thiese, B Ronna, and U Ott. P value interpretations and considerations. *Journal of thoracic disease*, 8(9): E928–E931, 2016.

Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. In *Proceedings of the national academy of sciences*, pages 6567–6572, 2002.

Adrian Trapletti and Kurt Hornik. *tseries: time series analysis and computational finance*, 2019. R package version 0.10-47.

Ruey S Tsay. *Analysis of financial time series*. John Wiley and Sons, 2005.

Ruey S Tsay. *Analysis of financial time series*. Wiley, Cambridge, Mass, 3rd edition, 2010.

VN Vapnik. *The nature of statistical learning theory (information science and statistics)*. Springer, 1999.

W. N. Venables and B. D. Ripley. *Modern applied statistics with S*. Springer-Verlag, fourth edition, 2002.

JA Vilar and S Pértega. Discriminant and cluster analysis for gaussian stationary processes: local linear fitting approach. *Journal of nonparametric statistics*, 16(3-4):443–462, 2004.

JA Vilar, AM Alonso, and JM Vilar. Non-linear time series clustering based on nonparametric forecast densities. *Computational statistics and data analysis*, 54(11):2850–2865, 2010.

José Vilar and Pablo Montero. TSclust: An R package for time series clustering. *Journal of statistical software*, 62(1), 12 2014.

Ronald L Wasserstein and Nicole A Lazar. The asa statement on p-values: context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016. doi: 10.1080/00031305.2016.1154108.

Andrew A. Weiss. ARMA models with ARCH errors. *Journal of time series analysis*, 5(2):129–143, 1984.

Peter R Winters. Forecasting sales by exponentially weighted moving averages. *Management science*, 6(3): 324–342, 1960.

H Wold. *In multivariate analysis*. Wiley, New York, 1966.

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). *Zhou M. (eds) Education and management. ISAEBD 2011. Communications in computer and information science*, 210, 2011.

Y Xiong, D-Y Yeung, and V Puttagunta. Mixtures of arma models for model-based time series clustering. *In data mining, 2002. ICDM 2003. Proceedings IEEE international conference*, page 717–720, 2002.

F. Yusof and I.L. Kane. Volatility modeling of rainfall time series. *Theoretical and applied climatology*, 113: 247–258, 2013.

Jean-Michel Zakoian. Threshold heteroskedastic models. *Journal of economic dynamics and control*, 18(5): 931 – 955, 1994a.

J.M Zakoian. Threshold heteroskedastic models. *Journal of economic dynamics and control*, 18(5):931–955, 1994b.

Manuel Zarzo, Angel Fernández-Navajas, and Fernando-Juan 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 (Basel, Switzerland)*, 11:8685–710, 12 2011.

Achim Zeileis. Implementing a class of structural change tests: an econometric computing approach. *Computational statistics and data analysis*, 50:2987–3008, 2006a.

Achim Zeileis. Implementing a class of structural change tests: an econometric computing approach. *Computational statistics and data analysis*, 50:2987–3008, 2006b.

Achim Zeileis, Friedrich Leisch, Kurt Hornik, and Christian Kleiber. Strucchange: an R package for testing for structural change in linear regression models. *Journal of statistical software*, 7(2):1–38, 2002.