A COMPARISON OF PRE-PROCESSING APPROACHES FOR REMOTELY SENSED TIME SERIES CLASSIFICATION BASED ON FUNCTIONAL ANALYSIS

Mattia Balestra[a,*], Roberto Pierdicca[b], Lorenzo Cesaretti[d], Giacomo Quattrini[a],
Adriano Mancini[c], Andrea Galli[a], Eva Savina Malinverni[b], Simona Casavecchia[a], Simone Pesaresi[a]

[a]Department of Agricultural, Food and Environmental Sciences (D3A), Via Brecce Bianche, Ancona, 60131, Italy, IT
[b]Department of Civil, Building Engineering and Architecture (DICEA), Via Brecce Bianche, Ancona, 60131, Italy, IT;
[c]Department of Information Engineering (D3A), Via Brecce Bianche, Ancona, 60131, Italy, IT;
[d]CREA Centro di ricerca Foreste e Legno, Viale Santa Margherita, Arezzo, 52100, Italy, IT

KEY WORDS: Time-series, Data Pre-processing, Outliers Detection, Generalized Additive Model (GAM), Functional Principal Component Analysis (FPCA).

ABSTRACT:

Satellite remote sensing has gained a key role for vegetation mapping distribution. Given the availability of multi-temporal satellite data, seasonal variations in vegetation dynamics can be used trough time series analysis for vegetation distribution mapping. These types of data have a very high variability within them and are subjected by artifacts. Therefore, a pre-processing phase must be performed to properly detect outliers, for data smoothing process and to correctly interpolate the data. In this work, we compare four pre-processing approaches for functional analysis on 4-years of remotely sensed images, resulting in four time series datasets. The methodologies presented are the results of the combination of two outlier detection methods, namely`tsclean` and `boxplot` functions in R and two discrete data smoothing approaches (Generalized Additive Model "GAM" on daily and aggregated data). The approaches proposed are: `tsclean`-GAM on aggregated data (M01), `boxplot`-GAM on aggregated data (M02), `tsclean`-GAM on daily data (M03), `boxplot`-GAM on daily data (M04). Our results prove that the approach which involves `tsclean` function and GAM applied to daily data (M03) is ameliorative to the logic of the procedure and leads to better model performance in terms of Overall Accuracy (OA) which is always among the highest when compared with the others obtained from the other three different approaches.

1. INTRODUCTION

In the last four decades, satellite remote sensing has gained a key role for vegetation and habitat distribution mapping (Zlinszky et al., 2015). The habitat distribution can be properly represented by vegetation map, which, when repeated over time, are useful to assess their preservation (Dash and Ogutu, 2016, Viciani et al., 2016). Given the availability of multi-temporal satellite data, seasonal and inter-annual variations in vegetation dynamics (phenology) can be quantified trough time series analysis and used for vegetation distribution mapping (Caparros-Santiago et al., 2021). These types of data have a very high variability within them, due to the fact that they are derived from satellite images acquired at different time periods, using different sensors, capturing constantly changing dynamics, with ever-changing weather conditions and with solar radiance inclination which creates shadows and reflections caught by the sensors (Meraner et al., 2020). Furthermore, if we add to this the technological issues which may occur when dealing with artificial instrumentation (e.g., of sensor malfunctions), we can immediately realize that the data must be fixed before being used for different purposes. The most challenging spectral reflectance abnormal values are often caused by adverse weather conditions, undetected sub-pixel cloud cover, atmospheric dust and gaseous absorbers but also seasons lighting variations, soil-induced disturbances, shading, or sensor glitches (Alvera-Azcárate et al., 2012). All of them are generally referred to as artifacts (Du et al., 2003) and must be removed from the time-series. These artifacts can alter the temporal pattern of reflectance values, causing a reduction in the accuracy of the phenological estimations (Clark et al., 2002). Therefore, in order to process the data correctly, it is necessary to implement a data pre-processing phase. Given the very large variability of the data and, consequently, the outliers occurrence, it is essential to operate through diversified methods for the detection of those anomalies to obtain a representative output of the vegetation to classify (Jackson and Chen, 2004). A knowledge about what is intended to classify is fundamental because outliers detection also comes through a deep awareness concerning the expected value of a given class. Without such background, which only the user's experience can provide, information might be lost that, instead, is essential to the classifier. Hence, smoothing of the discrete data becomes a key phase to achieve accurate classification models which describe detailed vegetation dynamics and reduce as good as possible the outlier (Zeng et al., 2020). Considering these issues, it is necessary to use appropriate approaches (i.e data smoothing and outliers detection methods) that will mitigate the noise effects of time-series from satellite imagery (Santos et al., 2021). The greater the care and attention in this step, the higher the quality of the data that will be processed and, hence, the output that will be obtained. In statistics, smoothing consists in the application of a function filter aimed to highlight relevant patterns by mitigating noises generated by environmental, computational, or physiological artifacts (Atkinson et al., 2012). The presence of time gaps, given by the missing data due to both the satellite temporal resolution and the artifacts, makes it necessary to deal also with a data interpolation phase that will allow to obtain a continuous function throughout the year. The interpolation process consists to average the data in a series with contiguous values to describe a pattern, but considering these values in a cyclic way given by

* m.balestra@pm.univpm.it

their annual nature. Interpolation process can either precede or even be followed by a data aggregation phase. The data, in order to provide valuable insights for the resulting classification, can be processed through functional data analysis (FDA) (Hurley et al., 2014). The latter must represent discrete data as functions, as the FDA needs to analyze the data as a single function rather than a set of point values spread over time (Pesaresi et al., 2022). Therefore, to work by functional analysis, a proper outlier detection step, data smoothing phase and data interpolation can not be ignored.

With this papers, we want to figure out a suitable strategy to adopt when processing time-series, through the comparison among the classification accuracies obtained. Thus, we compared the outputs of four data pre-processing approaches. The approaches presented are the results of the combination of two outlier detection options (`tsclean` function in `forecast` package (Hyndman and Khandakar, 2008) and `boxplot` function in graphics package (Murrell, 2005)) and two discrete data smoothing approaches (Generalized Additive Model "GAM" (Wood, 2006) on daily and aggregated data). The approaches proposed are: `tsclean`-GAM on aggregated data (M01), `boxplot`-GAM on aggregated data (M02), `tsclean`-GAM on daily data (M03), `boxplot`-GAM on daily data (M04).

The main contribution of this work is to perform four different satellite image time series pre-processing approaches, combining two outliers detection methodology (`tsclean` and `boxplot`) and a data smoothing technique (GAM) applied to differently aggregated datasets, testing them within 2 study areas with a combination of predictors and vegetational indicies.

2. MATERIALS AND METHODS

2.1 Study Areas

2.1.1 Frasassi Gorge The study area overlaps with the Special Area of Conservation "Gola di Frasassi IT5320003", covering an area of 728 ha within the Rossa and Frasassi Gorge Regional Natural Park between the municipalities of Genga and Fabriano, in the province of Ancona. The Frasassi gorge is placed inside the anticline of Mount Valmontagnana - Mount Frasassi, in the pre-Apennine mountain belt. The peak elevation of the site is 931.2 m.a.s.l. at "Monte di Valmontagnana", while the minimum elevation measured is 200 m.a.s.l. at the edge of the Esino river's left bank.

2.1.2 Conero Mount This study area is part of the territory in between the Special Areas of Conservation "Monte Conero IT5320007" and "Portonovo e falesia calcarea a mare IT5320007". It covers an area of 650 hectares within the Conero Regional Natural Park in the province of Ancona, between the municipalities of Sirolo and Ancona. Conero mount is a limestone promontory of 582 m.a.s.l., being the only stretch of limestone coastline from Trieste to Gargano, it interrupts the continuous low and sandy shoreline typical of the Adriatic coast.

2.2 Dataset Collection and Processing

The dataset used are the ones validated on two previous works (Pesaresi et al., 2020, Pesaresi et al., 2022). The ecological variability of this data will then allow to define the best applicable method in different scenarios. In the Conero mount study area, on the basis of the vectors used in (Pesaresi et al., 2022) paper (Figure 1), the vegetation categories identified are 4 and are distributed over 175 points: thermophilic woods with a prevalence of *Quercus ilex (Cyclamino hederifolii - Quercetum ilicis)*; mesophilic woods with a prevalence of *Quercus ilex (Cephalanthero longifoliae - Quercetum ilicis* sub. *ruscetosum hypoglossi); Ostrya carpinifolia*-dominated woodlands (*Asparago acutifolii - Ostryetum carpinifoliae*); coniferous reforestation(Biondi, 1986, Biondi, 1982).

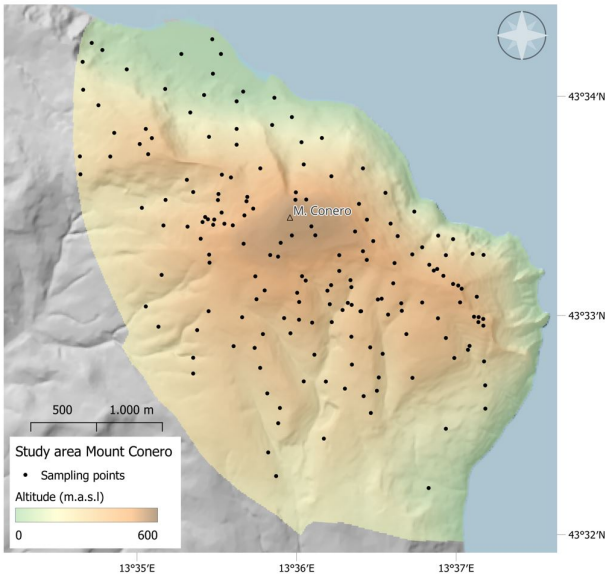

Figure 1. Conero mount study area.

In the Frasassi gorge study area, according to the vectors used in (Pesaresi et al., 2022) research (Figure 2), the described vegetation categories are 8 and distributed in 242 points: broom bushes (*Spartio juncei - Cytisetum sessilifolii* var. *a Spartium junceum)*; junipers bushes (*Spartio juncei - Cytisetum sessilifolii* var. with *Juniperus oxycedrus* and *Juniperus communis*); forests dominated by *Quercus ilex (Cephalanthero longifoliae - Quercetum ilicis)*; mosaic of garrigue and scrub; *Ostrya carpinifolia*-dominated woodlands (*Scutellario columnae - Ostryetum carpinifoliae)*, grasslands (*Asperulo purpureae - Brometum erecti); Quercus pubescens*-dominated woodlands (*Cytiso sessilifolii - Quercetum pubescentis*); *Pinus* sp. reforestation(Biondi and Casavecchia, 2002, Allegrezza et al., 2020).

The images have been acquired by the Sentinel-2A and Sentinel-2B satellite platforms, both managed by the European Space Agency (ESA) as part of the European Copernicus plan (Pesaresi et al., 2022). For each study area, 93 L2A images referable to the period between April 2017 and March 2020 have been collected using the `sen2r` package. The satellite imagery acquisition frequency permits to aggregate in accordance with defined temporal intervals. They can be aggregated by year, month (semester, trimester, bimester), week (weeks, biweeks) or days of the year (DoY). Sentinel-2 acquisition frequency allows for a weekly aggregation time, allowing for time-series composed of 52 values. The images have been co-registered, cropped with the shapefiles matching the limits of the study areas and masked by the cloud cover. Seven vegetation indices have been computed for each of these images and each one has been used as prediction variable. Additionally we used the FPCA to group the information related to the curves' temporal variability into a set of main components. The coef-

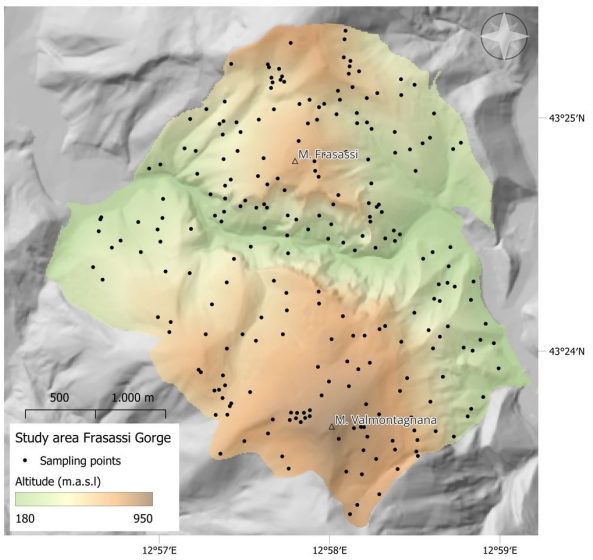

Figure 2. Frasassi gorge study area.

ficients quantifying the weight of each component and maintaining the chronological order of the functional variations are referred to as "scores" (Pesaresi et al., 2020). We used them, in total and in a fraction thereof, as second and third prediction variables in this classification. Moreover, we removed the last component resulting from the FPCA and we reconstructed the time-series and we used them as the fourth prediction variable. The seven vegetation indices are: Normalized Difference Vegetation Index (NDVI), Modified Chlorophyll Absorption in Reflectance Index (MCARI), Green Normalized Difference Vegetation Index (GNDVI), Normalized Difference Red/Green Redness Index (RI), Normalized Difference Red-Edge (NDRE), Normalized DIfference Moisture Index (NDMI), Modified Normalized Difference Water Index (MNDWI). The proposed classification algorithm is Random Forest. In order to ease the readability of the manuscript, the general workflow is reported in Figure 3.

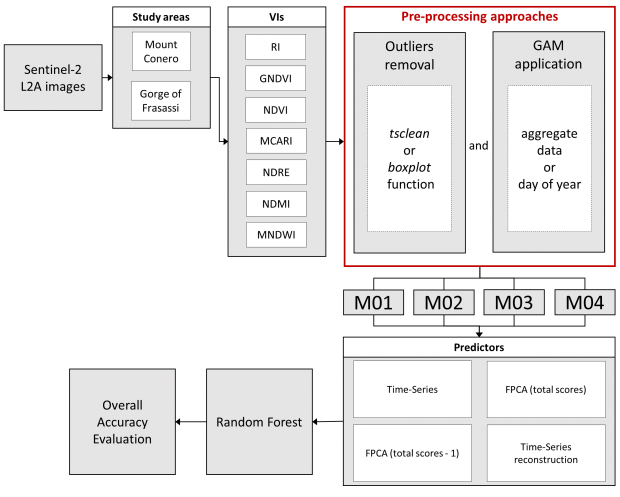

Figure 3. Workflow for the overall accuracy evaluation, combining the dataset to be subjected to one of the 4 pre-processing approaches.

2.2.1 Outliers Removal Temporal information is obtained by extracting the capture date and then converting it into the corresponding day of the year (DoY). Using this information, data are chronologically sorted and the dataset is thus obtained, which will be subjected to the outlier detection process. The detection methods for the anomalous point values are the most cited in the literature (Willsky et al., 1980, Hu et al., 2021, Venkatasubramanian et al., 2003). Among them, the most widely used are the so-called *model-based* techniques (Mehrang et al., 2015, Basu and Meckesheimer, 2007), followed by the *density-based* (Tang and He, 2017, Tian et al., 2016, Angiulli and Fassetti, 2007) and the *histogramming* ones (Blázquez-García et al., 2021, Muthukrishnan et al., 2004). In this paper two different methods of point outlier detection are compared: the `tsclean` function of the `forecast` package (Hyndman et al., 2020) and the `boxplot` function of the graphics package (Murrell and Murrell, 2020). Both are part of model-based techniques, meaning that a point x at time t can be declared an outlier if the distance from its expected value x is greater than a predefined threshold τ (Formula 1).

$$|x_t - \hat{x}_t| > \tau \tag{1}$$

The `tsclean` function is used to process univariate time series and the detection of outliers is different for seasonal and non-seasonal time series (Kandanaarachchi et al., 2020). In this study the interest is in the former type, where a significant seasonal component is identified in the variation of the phenomenon. Specifically, the function uses a time series decomposition method: *Seasonal and Trend decomposition using LOESS (STL)* (Cleveland et al., 1990). The STL method employs location-adapted regression models to deconstruct a time series into trend, seasonal, and residual components. The STL algorithm smooths the time series using the *locally estimated scatterplot smoothing (LOESS)* method in two cycles: an inner and an outer cycle. During the inner one, the seasonal and trend components are calculated. The residual is then found by subtracting these from the time series (Cleveland, 1990). For each time series, outliers are identified and replaced by interpolation.

In the context of outlier detection methods, the `boxplot` belongs to the techniques using basic statistics. Pixels considered outliers are those having values placed over X times the interquartile range from the first and third quartiles, and represented as isolated points in the plot. In this study, the coefficient X is equal to 1.5, considered as the default value in the `boxplot` graphics package function of R. Although groups of pixels can be processed simultaneously, the function is configured by considering the single pixel as a univariate time series (Bernard et al., 2012), making the outputs comparable to those obtained with the `tsclean` function. As opposed to the latter, the `boxplot` function permits to analyze the data by selecting a chosen time span. Therefore, it is necessary to set a vector containing the time span that best considers, in an independent way, the seasonal variability of the data (a value detected anomalous on winter is not necessarily anomalous on summer). In this study, monthly time span has been chosen, based on the density of observations in the DoYs and their seasonal distribution.

2.2.2 Data Smoothing The proposed smoothing algorithm is a Generalized Additive Model (GAM) based on Cyclic Cubic Spline Regression. GAMs are extensions of linear models in which the predictor is the sum of regular functions plus a conventional parametric component (Wood and Wood, 2015).

GAMs allow the configuration of both complex non-linear relationships and inferential statistics by understanding and explaining the inherent structure of the discrete data dispersion model (Azzalini and Scarpa, 2012). The Cyclic Cubic Spline Regression summarises the multi year variability of vegetation surfaces from satellite imagery into one artificial/ideal year. The Cyclic Cubic Spline Regression is a function composed of several connected polynomials designed to interpolate a set of points into defined intervals, called knots. The number of knots in which the dataset is divided is set through a process of cross validation. Separating the dataset into subsets, localized smoothing is achieved and overfitting (given by the global influence of each point on the fit) is avoided (Faraway, 1992). The Cyclic Cubic Spline Regression provides the connection between the first and last knot to consider the temporal nature of the analyzed curves which describe a continuous trend between December and January.

Interpolation process is performed by the `gam` function, of the `mgcv` package (Wood and Wood, 2015). The "GAM on Aggr" approach involves the GAM application on data previously aggregated by weeks, while the "GAM on DoY" one applies the GAM directly on cleaned daily data, not yet aggregated. In the first case, an early data grouping is carried out, in order to compensate the double collection of images in the same days but in different years, and subsequently aggregated by weeks. In the second case, instead, the compensation of double collected values is carried out directly by GAM, to obtain time-series fitted with 365 values. The latter are then aggregated by weeks, making the two approaches comparable.

2.3 Predictors

Once that the outliers removal and the data smoothing processes have been performed with the different approaches, each pixel value (recorded throughout the 4 years) has been fit in a yearly time series. The latter presents slight differences according to the pre-processing approach used, as can be noticed in the figure 4.

The aggregated time-series resulting from approaches M01, M02, M03, and M04 are the first predictors used in this study for classification. These are subsequently subjected to FPCA, a statistical method for variance analysis of functional data (Ullah and Finch, 2013). It is well adapted to time-series, where individual observations are not independent, but rather constrained by their chronological order. It provides an estimation of dataset complexity by determining the minimum number of components needed to represent the content of the dataset without loss of information. The "scores" are the parameters quantifying the similarities among the time-series: they provide information on the position, shape and variation of each curve observed in the space (Shang, 2014). The main FPCA outputs, besides the scores, are the eigenvalues and the eigenfunctions. The eigenvalues represent the variation explained by each component for each time-series. Their sum is equal to the overall data variance. The components explaining up to 99% of the total variance are considered to extract a reduced number of scores. Eigenfunctions, on the other hand, estimate the scores' value by describing the largest functional variances for each component (Hurley et al., 2014). Through these outputs, the original data (X_{iK}) can be rebuilt as a summation of the product of each score (A_{ik}) by its eigenfunction (ϕ_k), plus the mean value (μ) (Wang et al., 2015) (Formula 2):

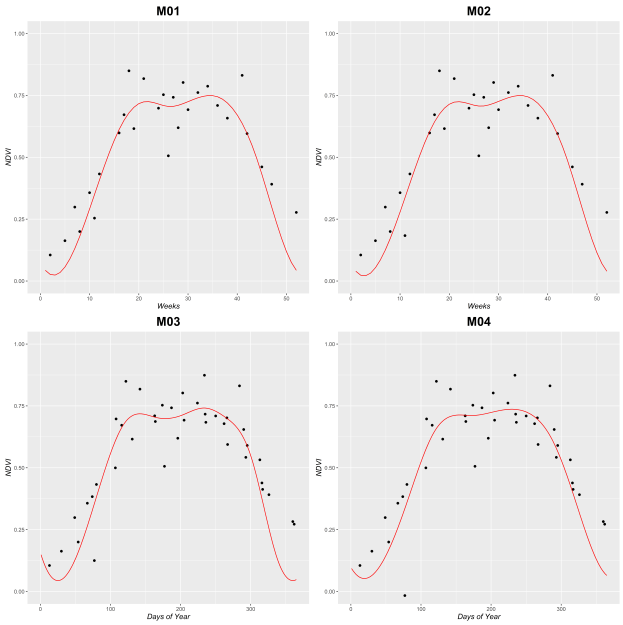

Figure 4. NDVI values for a randomly chosen pixel which fit differently in the function in the different approaches, the black dots are the single pixel values recorded in the 4 years satellite imagery used in this paper.

$$X_{iK}(t) = \mu(t) + \sum_{k=1}^{K} A_{ik}\phi_k(t) \tag{2}$$

Through this process the four predictors used in this study were obtained: time-series (TS), scores (ST), reduced scores (SR) and rebuilt time-series (TSR). The classification is performed by Random Forest. The overall accuracy (OA) is defined by the proportion of correctly classified pixels by the algorithm and the total number of pixels used to train the model. The number of decision trees set in the algorithm is 1500. The Random Forest training is performed through repeated cross validation. The latter has been set by the authors and the number of folders [k-fold] in which the dataset is divided is 10, while the number of repetitions is 5. All functions used in this section belong to the `caret` (Classification and Regression Training) package, which is specific to the construction of classification models (Kuhn, 2015).

3. RESULTS

In this chapter, the tested combinations for defining our best pre-processing approach for remote sensed time-series are shown through boxplot graphs. These various methods, obtained from the 4 pre-processing approaches, using 4 predictors from 7 different vegetation indices in 2 different study areas resulted in a total of 224 models. The charts summarize the variability in accuracies generated by the models for the 4 approaches with respect to the different variables. Therefore, for each plot, each method is compared individually to either a predictor, an index, or a study area. All of the accuracies obtained from the models of each method for each predictor/index/study area being analyzed are then grouped together. In particular, each approach/predictor comparison chart groups 56 (224/4) accuracies. Each approach/index comparison chart groups 32

(224/7) accuracies. Each approach/area comparison chart contains 112 (224/2) accuracies.

3.1 Approaches and Predictors

The results obtained from the predictors' classification reveal a substantial difference between the accuracies achieved. Models obtained through TS, ST, and SR predictors achieved higher accuracies than those based on TSR (Figure 5). For each method and index, in both study areas, the TSR thus demonstrate an insufficient models' accuracy and it cannot be compared to the other predictor models (Table 1). Therefore, the results obtained from the TSR are not included in the subsequent charts showing the accuracies achieved by the different approaches (M01, M02, M03, M04) when compared to the vegetation indices and the study areas. The combination `tsclean` and GAM on DoY (M03) proves to perform better than the other approaches for TS, ST and SR.

	M01	M02	M03	M04
TS	72.55	71.35	73.60	72.25
ST	73.30	71.90	76.70	75.20
SR	73.45	72.90	77.50	74.50
TSR	40.10	40.60	40.30	40.35

Table 1. Median accuracy values from models comparing approaches and predictors.

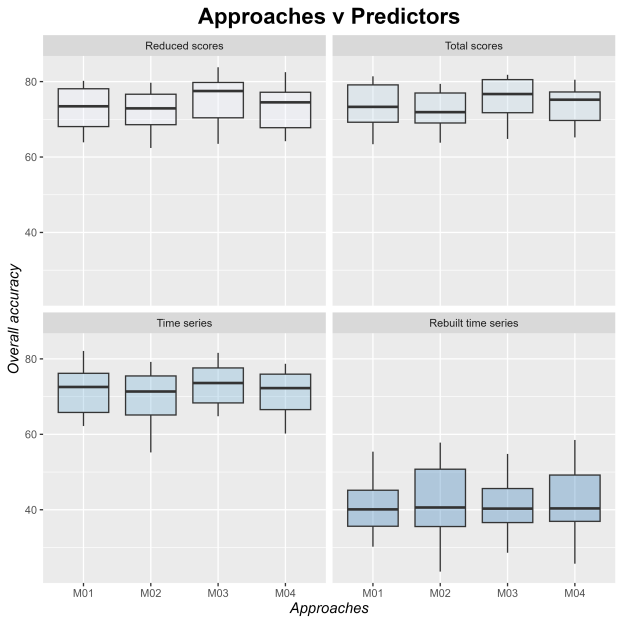

Figure 5. Approaches versus Predictors.

3.2 Approaches and Vegetation Indices

Analyzing the 32 results obtained for each vegetation index (Table 2), it can be noted that M03 is the best performing method for 5 of the 7 indices used (MCARI, NDVI, NDMI, NDRE, RI) (Figure 6). The M04 method is the best performing for GNDVI and MNDWI.

3.3 Approaches and Study Areas

As a result of the 112 results obtained for each study area, the chart (Figure 7) reveals that the M03 method continues to be the most accurate (Table 3). Each of the 224 combinations of approaches has generated a map of the analyzed area, along with its corresponding confusion matrix and OA. By selecting the models that achieved the highest OA for the two study areas, in Figure 8 we can observe the classification obtained in Frasassi gorge using the M03 approach, with SCR and the vegetation index RI. In Figure 9 we can observe the classification of Conero mount obtained using the M03 approach, SCR and the vegetation index NDRE. The maps were produced from the best results obtained through the combination of different approaches. However, a few classes were misclassified for labels that were not initially represented in our classes.

	M01	M02	M03	M04
GNDVI	68.95	69.55	71.05	71.20
MCARI	65.75	64.15	67.85	64.70
MNDWI	63.65	63.75	65.25	66.90
NDMI	65.60	65.80	70.55	68.15
NDRE	73.35	73.20	77.10	75.00
NDVI	71.35	70.25	71.40	69.90
RI	73.40	72.80	77.30	74.70

Table 2. Median accuracy values obtained from models comparing Approaches and vegetation indices.

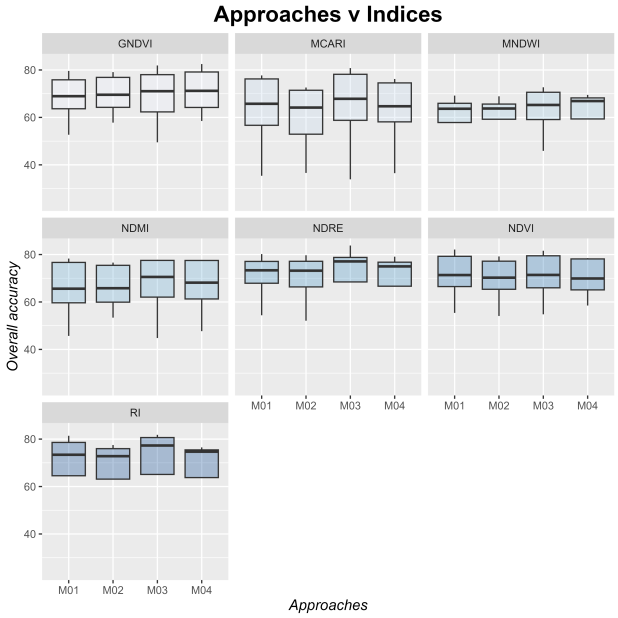

Figure 6. Approaches versus Vegetation Indices.

	M01	M02	M03	M04
CO	76.25	75.35	77.70	76.20
VM	65.65	65.80	67.85	67.25

Table 3. Median accuracy values obtained from models comparing approaches and study areas.

4. DISCUSSION

According to the results obtained in this work, the method resulting in higher model performance is M03, which involves the `tsclean` function and the application of GAM on the daily data. The `tsclean` function allows proper dataset cleaning from outliers with limited computation time and proves to be an important tool when processing time-series from vegetation indices, as earlier proved by (Pesaresi et al., 2020, Pesaresi et al., 2022). Nevertheless, the `boxplot` function has the merit of allowing for outliers removal only (Kerandel et al., 2020) and, moreover, although computation time is slightly stretched, in terms of

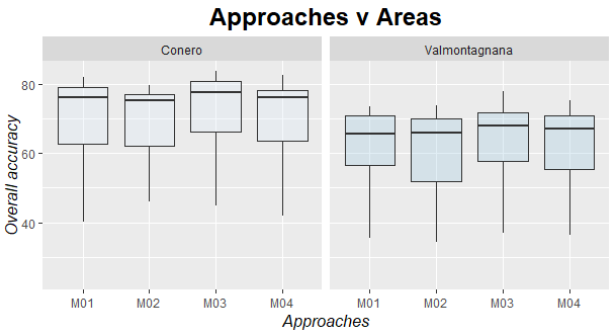

Figure 7. Approaches versus study areas.

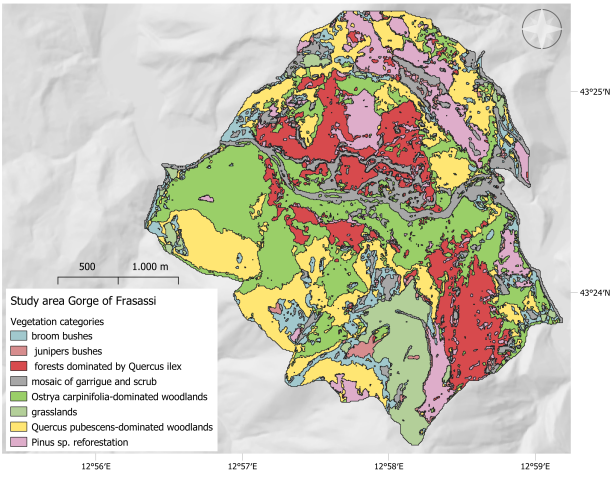

Figure 8. Best model classification obtained in Frasassi gorge, using M03 approach with SCR and RI as vegetation index.

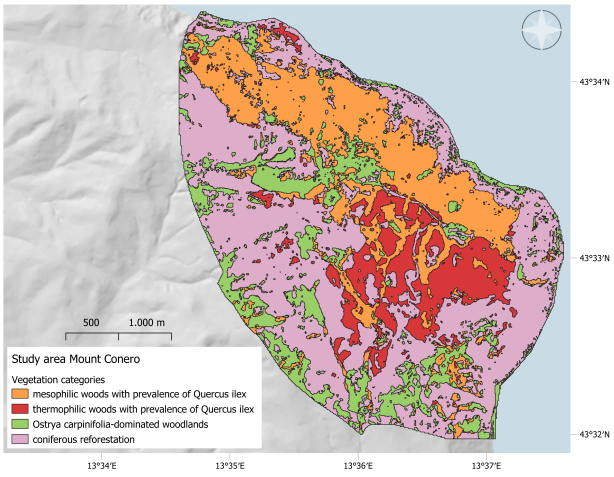

Figure 9. Best model classification obtained in Conero mount, using M03 approach with SCR and NDRE as vegetation index.

coding it is easily handled and outliers can be detected within the time series over different periods of the year (i.e. month, bimester, trimester). Within more heterogeneous environments such as the Frasassi gorge, the `boxplot` function performs better for certain indices when compared to the `tsclean` function. In literature, different techniques have been used to carry out the smoothing and correction phase of time-series satellite data, such as the curve fitting (Pickers and Manning, 2015), the Fourier decomposition (Mingwei et al., 2008), the asymmetric Gaussian function (Jonsson and Eklundh, 2002), the double logistic functions (Atkinson et al., 2012, Eklundh and Jönsson, 2015), the Whittaker smoother (Shao et al., 2016, Kandasamy et al., 2013), the Savitzky–Golay filter (Huang et al., 2021), the high order spline with roughness damping (Hermance et al., 2007), the spatio-temporal tensor completion method (Chu et al., 2021) and other spatio-temporal combination methods such as the adaptive spatio-temporal weighted method (Li et al., 2017) and hybrid Generalised Additive Model (GAM)-geostatistical space-time model (Poggio et al., 2012) wich are even useful to fill temporal gaps. GAM utilization for regression model fitting is widely demonstrated in literature (Hua et al., 2021). Applying GAM on the daily data permits to perform data aggregation during the subsequent steps of the work. Being able to manipulate the dataset starting with the DoY is an advantage in the procedure's logic and actually an efficient way to proceed. Computational times are stretched since the range of values within which the data can be interpolated (from 1 to 365) is greater. However, a single subsequent aggregation step is necessary to compensate positively the times, thus allowing for easy variation in aggregation time (i.e. weekly or biweekly). As a result, the process proves to be both elastic and adaptable as required. Within the two study areas, differences in the produced model accuracy values are evident and substantial, making them coherent with those reported by (Pesaresi et al., 2020, Pesaresi et al., 2022). These results are related to the different complexity levels of vegetation phytocoenoses and land cover in general. In the Conero mount study area there are 4 classes identified, while in the Frasassi gorge study area 8 classes are defined. Those are situated in a much more complex geomorphological and topographical context. This emphasizes the need to test the described methodologies in different and diverse contexts, so as to further assess their reliability. From this standpoint, this work has succeeded in providing a suitable comparative analysis among 4 approaches for time-series preprocessing. The processes involved can be replicated in other areas in order to enhance and validate the mapping accuracy.

5. CONCLUSION

This method is indeed proven to be fast and efficient. In this paper, 4 time series preprocessing approaches were compared, combining 2 outlier detection methods (`tsclean` function `forecast` package and `boxplot` function graphics package) and 2 interpolation algorithm application methods (GAM on aggregate data and GAM on daily data). Therefore, this research intended to stress the preprocessing part of the data that will be subjected to FPCA to identify which of the proposed methodologies performed best in terms of outputs and computational time. From the results obtained, the approach which involves `tsclean` function and GAM applied to daily data (M03) is ameliorative to the logic of the procedure and leads to better model performance in terms of Overall Accuracy. Although the algorithms implemented with the GAM have demonstrated the ability to adequately interpolate aggregate and daily data, the application of other techniques is also desirable to improve the construction of the time series. Other solutions, in the outlier detection phase, will be subject to further analysis since there are several methodologies which can be applied to clean the time series. As a result of the results obtained and the identification of this methodological approach for mapping, it will therefore be possible to periodically repeat these tests to produce maps up-to-date and, thus, to comply with EU regulations.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-1-W1-2023-33-2023 | © Author(s) 2023. CC BY 4.0 License.

REFERENCES

Allegrezza, M., Pesaresi, S., Ballelli, S., Tesei, G., Ottaviani, C., 2020. Influences of mature Pinus nigra plantations on the floristic-vegetational composition along an altitudinal gradient in the central Apennines, Italy. *iForest-Biogeosciences and Forestry*, 13(4), 279.

Alvera-Azcárate, A., Sirjacobs, D., Barth, A., Beckers, J.-M., 2012. Outlier detection in satellite data using spatial coherence. *Remote Sensing of Environment*, 119, 84-91.

Angiulli, F., Fassetti, F., 2007. Detecting distance-based outliers in streams of data. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 811–820.

Atkinson, P. M., Jeganathan, C., Dash, J., Atzberger, C., 2012. Inter-comparison of four models for smoothing satellite sensor time-series data to estimate vegetation phenology. *Remote sensing of environment*, 123, 400–417.

Azzalini, A., Scarpa, B., 2012. *Data analysis and data mining: An introduction*. OUP USA.

Basu, S., Meckesheimer, M., 2007. Automatic outlier detection for time series: an application to sensor data. *Knowledge and Information Systems*, 11(2), 137–154.

Bernard, J., Wilhelm, N., Scherer, M., May, T., Schreck, T., 2012. Projection-based explorative analysis of multivariate time series data. *Time series paths*.

Biondi, E., 1982. L'Ostrya carpinifolia Scop. sul litorale delle Marche (Italia centrale). *Fitosociologia*.

Biondi, E., 1986. *La vegetazione del Monte Conero:(con carta della vegetazione alla scala 1: 10.000)*. Regione Marche-Assessorato All'Ambiente.

Biondi, E., Casavecchia, S., 2002. Inquadramento fitosociologico della vegetazione arbustiva di un settore dell'Appennino settentrionale. *Fitosociologia*, 39(1), 65–73.

Blázquez-García, A., Conde, A., Mori, U., Lozano, J. A., 2021. A review on outlier/anomaly detection in time series data. *ACM Computing Surveys (CSUR)*, 54(3), 1–33.

Caparros-Santiago, J. A., Rodriguez-Galiano, V., Dash, J., 2021. Land surface phenology as indicator of global terrestrial ecosystem dynamics: A systematic review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 171, 330–347.

Chu, D., Shen, H., Guan, X., Chen, J. M., Li, X., Li, J., Zhang, L., 2021. Long time-series NDVI reconstruction in cloud-prone regions via spatio-temporal tensor completion. *Remote Sensing of Environment*, 264, 112632.

Clark, R. N., Swayze, G. A., Livo, K. E., Kokaly, R. F., King, T. V., Dalton, J. B., Vance, J. S., Rockwell, B. W., Hoefen, T., McDougal, R. R., 2002. Surface reflectance calibration of terrestrial imaging spectroscopy data: a tutorial using aviris. *Proceedings of the 10th Airborne Earth Science Workshop*, 2, Jet Propulsion Laboratory Pasadena, CA, USA.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., Terpenning, I., 1990. STL: A seasonal-trend decomposition. *J. Off. Stat*, 6(1), 3–73.

Cleveland, S., 1990. A seasonal-trend decomposition procedure based on Loess (with discussion). *J. Off. Stat*, 6(6), 3.

Dash, J., Ogutu, B. O., 2016. Recent advances in space-borne optical remote sensing systems for monitoring global terrestrial ecosystems. *Progress in Physical Geography*, 40(2), 322–351.

Du, Y., Vachon, P. W., Van der Sanden, J. J., 2003. Satellite image fusion with multiscale wavelet analysis for marine applications: preserving spatial information and minimizing artifacts (PSIMA). *Canadian Journal of Remote Sensing*, 29(1), 14–23.

Eklundh, L., Jönsson, P., 2015. Timesat: A software package for time-series processing and assessment of vegetation dynamics. *Remote sensing time series*, Springer, 141–158.

Faraway, J. J., 1992. On the cost of data analysis. *Journal of Computational and Graphical Statistics*, 1(3), 213–229.

Hermance, J. F., Jacob, R. W., Bradley, B. A., Mustard, J. F., 2007. Extracting phenological signals from multiyear AVHRR NDVI time series: Framework for applying high-order annual splines with roughness damping. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10), 3264–3276.

Hu, Z., Chen, B., Chen, W., Tan, D., Shen, D., 2021. Review of model-based and data-driven approaches for leak detection and location in water distribution systems. *Water Supply*, 21(7), 3282–3306.

Hua, J., Zhang, Y., de Foy, B., Shang, J., Schauer, J. J., Mei, X., Sulaymon, I. D., Han, T., 2021. Quantitative estimation of meteorological impacts and the COVID-19 lockdown reductions on NO2 and PM2. 5 over the Beijing area using Generalized Additive Models (GAM). *Journal of environmental management*, 291, 112676.

Huang, A., Shen, R., Di, W., Han, H., 2021. A methodology to reconstruct LAI time series data based on generative adversarial network and improved Savitzky-Golay filter. *International Journal of Applied Earth Observation and Geoinformation*, 105, 102633.

Hurley, M. A., Hebblewhite, M., Gaillard, J.-M., Dray, S., Taylor, K. A., Smith, W., Zager, P., Bonenfant, C., 2014. Functional analysis of normalized difference vegetation index curves reveals overwinter mule deer survival is driven by both spring and autumn phenology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1643), 20130196.

Hyndman, R. J., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., OHara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., Yasmeen, F., 2020. Package forecast: Forecasting functions for time series and linear models. *R Core Team*.

Hyndman, R. J., Khandakar, Y., 2008. Automatic time series forecasting: the forecast package for R. *Journal of statistical software*, 27, 1–22.

Jackson, D. A., Chen, Y., 2004. Robust principal component analysis and outlier detection with ecological data. *Environmetrics: The official journal of the International Environmetrics Society*, 15(2), 129–139.

Jonsson, P., Eklundh, L., 2002. Seasonality extraction by function fitting to time-series of satellite sensor data. *IEEE transactions on Geoscience and Remote Sensing*, 40(8), 1824–1832.

Kandanaarachchi, S., Muñoz, M. A., Hyndman, R. J., Smith-Miles, K., 2020. On normalization and algorithm selection for unsupervised outlier detection. *Data Mining and Knowledge Discovery*, 34(2), 309–354.

Kandasamy, S., Baret, F., Verger, A., Neveux, P., Weiss, M., 2013. A comparison of methods for smoothing and gap filling time series of remote sensing observations–application to MODIS LAI products. *Biogeosciences*, 10(6), 4055–4071.

Kerandel, N. A., MENSAH, E. P., Jérome, D. D. et al., 2020. Method for automatically processing outliers of a quantitative variable. *International Journal of Advanced Computer Science and Applications*, 11(7).

Kuhn, M., 2015. Caret: classification and regression training. *Astrophysics Source Code Library*, ascl–1505.

Li, X., Fu, W., Shen, H., Huang, C., Zhang, L., 2017. Monitoring snow cover variability (2000–2014) in the Hengduan Mountains based on cloud-removed MODIS products with an adaptive spatio-temporal weighted method. *Journal of hydrology*, 551, 314–327.

Mehrang, S., Helander, E., Pavel, M., Chieh, A., Korhonen, I., 2015. Outlier detection in weight time series of connected scales. *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 1489–1496.

Meraner, A., Ebel, P., Zhu, X. X., Schmitt, M., 2020. Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166, 333–346.

Mingwei, Z., Qingbo, Z., Zhongxin, C., Jia, L., Yong, Z., Chongfa, C., 2008. Crop discrimination in Northern China with double cropping systems using Fourier analysis of time-series MODIS data. *International Journal of Applied Earth Observation and Geoinformation*, 10(4), 476–485.

Murrell, P., 2005. *R graphics*. Chapman and Hall/CRC.

Murrell, P., Murrell, M. P., 2020. Package 'rgraphics'.

Muthukrishnan, S., Shah, R., Vitter, J. S., 2004. Mining deviants in time series data streams. *Proceedings. 16th International Conference on Scientific and Statistical Database Management, 2004.*, IEEE, 41–50.

Pesaresi, S., Mancini, A., Quattrini, G., Casavecchia, S., 2020. Mapping mediterranean forest plant associations and habitats with functional principal component analysis using Landsat 8 NDVI time series. *Remote Sensing*, 12(7), 1132.

Pesaresi, S., Mancini, A., Quattrini, G., Casavecchia, S., 2022. Functional Analysis for Habitat Mapping in a Special Area of Conservation Using Sentinel-2 Time-Series Data. *Remote Sensing*, 14(5), 1179.

Pickers, P., Manning, A., 2015. Investigating bias in the application of curve fitting programs to atmospheric time series. *Atmospheric Measurement Techniques*, 8(3), 1469–1489.

Poggio, L., Gimona, A., Brown, I., 2012. Spatio-temporal MODIS EVI gap filling under cloud cover: An example in Scotland. *ISPRS journal of photogrammetry and remote sensing*, 72, 56–72.

Santos, L. A., Ferreira, K. R., Camara, G., Picoli, M. C., Simoes, R. E., 2021. Quality control and class noise reduction of satellite image time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 177, 75–88.

Shang, H. L., 2014. A survey of functional principal component analysis. *AStA Advances in Statistical Analysis*, 98(2), 121–142.

Shao, Y., Lunetta, R. S., Wheeler, B., Iiames, J. S., Campbell, J. B., 2016. An evaluation of time-series smoothing algorithms for land-cover classifications using MODIS-NDVI multi-temporal data. *Remote Sensing of Environment*, 174, 258–265.

Tang, B., He, H., 2017. A local density-based approach for outlier detection. *Neurocomputing*, 241, 171–180.

Tian, H.-x., Liu, X.-j., Han, M., 2016. An outliers detection method of time series data for soft sensor modeling. *2016 Chinese Control and Decision Conference (CCDC)*, IEEE, 3918–3922.

Ullah, S., Finch, C. F., 2013. Applications of functional data analysis: A systematic review. *BMC medical research methodology*, 13(1), 1–12.

Venkatasubramanian, V., Rengaswamy, R., Yin, K., Kavuri, S. N., 2003. A review of process fault detection and diagnosis: Part I: Quantitative model-based methods. *Computers & chemical engineering*, 27(3), 293–311.

Viciani, D., Dell'Olmo, L., Ferretti, G., Lazzaro, L., Lastrucci, L., Foggi, B., 2016. Detailed Natura 2000 and CORINE Biotopes habitat maps of the island of Elba (Tuscan Archipelago, Italy). *Journal of Maps*, 12(3), 492–502.

Wang, J.-L., Chiou, J.-M., Müller, H.-G., 2015. Review of functional data analysis. *arXiv preprint arXiv:1507.05135*.

Willsky, A., Chow, E., Gershwin, S., Greene, C., Houpt, P., Kurkjian, A., 1980. Dynamic model-based techniques for the detection of incidents on freeways. *IEEE Transactions on automatic control*, 25(3), 347–360.

Wood, S. N., 2006. *Generalized additive models: an introduction with R*. chapman and hall/CRC.

Wood, S., Wood, M. S., 2015. Package 'mgcv'. *R package version*, 1(29), 729.

Zeng, L., Wardlow, B. D., Xiang, D., Hu, S., Li, D., 2020. A review of vegetation phenological metrics extraction using time-series, multispectral satellite data. *Remote Sensing of Environment*, 237, 111511.

Zlinszky, A., Deák, B., Kania, A., Schroiff, A., Pfeifer, N., 2015. Mapping Natura 2000 habitat conservation status in a pannonic salt steppe with airborne laser scanning. *Remote Sensing*, 7(3), 2991–3019.