engineering proceedings

MDPI

Proceeding Paper

Time Series Modelling and Predictive Analytics for Sustainable Environmental Management. A Case Study in El Mar Menor (Spain) †

Rosa Martínez Álvarez-Castellanos *, Ivan Felis Enguix, Mercedes Navarro Martínez and Juan Carlos Sanz González

Centro Tecnológico Naval y del Mar, 30320 Fuente Álamo, Murcia, Spain; email1@email.com (I.F.E.); email2@email.com (M.N.M.); email3@email.com (J.C.S.G.)

* Correspondence: rosamartinez@ctnaval.com; Tel.: +34-968-19-75-21

† Presented at the 10th International Electronic Conference on Sensors and Applications (ECSA-10), 15–30 November 2023; Available online: https://ecsa-10.sciforum.net/.

Abstract: In this study of data science and machine learning, time series analysis plays a key role in predicting evolving data patterns. The Mar Menor, located in the Region of Murcia, represents an urgent case due to its unique ecosystem and the challenges it faces. This paper highlights the need to study the environmental parameters of the Mar Menor and to develop accurate predictive models and a standardised methodology for time series analysis. These parameters, which include water quality, temperature, salinity, nutrients, chlorophyll, and others, show complex temporal variations influenced by different activities. Advanced time series models are used to gain insight into their behaviour and project future trends, facilitating effective conservation and sustainable development strategies. The results show that some models are valid for handling the environmental behaviour of the Mar Menor, with SARIMA and LSTM standing out as the best models for most datasets.

Keywords: time series; statistical models; machine learning; ARIMA; seasonality; LSTM

Citation: Álvarez-Castellanos, R.M.; Enguix, I.F.; Martínez, M.N.; González, J.C.S. Time Series Modelling and Predictive Analytics for Sustainable Environmental Management. A Case Study in El Mar Menor (Spain). *Eng. Proc.* **2023**, *56*, x. https://doi.org/10.3390/xxxxx

Academic Editor(s): Name

Published: 15 November 2023

Copyright: © 2023 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The Mar Menor is a coastal lagoon in the Region of Murcia (Spain) that faces a series of major environmental and ecological problems, which has generated the need to analyse and understand its evolution, as well as its indicators trend over time. Time series analysis in the context of the Mar Menor provides valuable information on the changes and dynamics of this lagoon. These data provide key information for data collection in the management and conservation of this ecosystem, as well as for the implementation of protection and restoration measures. However, time series analysis presents unique challenges. Time series can be complex and influenced by factors as diverse as seasonal cycles, weather events and human activities. In addition, there may be irregularities, missing data and noise that make time series difficult to interpret and model. In this context, the application of the above methodology seeks to address these problems and provide a time series dynamic and present pattern deep understanding.

In the field of data science and machine learning, time series analysis plays a crucial role in studying and predicting data that evolves over time. The main objective of time series analysis is to understand its performance and predict its evolution, but there are a variety of approaches and algorithms available, different models have different assumptions, characteristics and capabilities, and their performance can vary significantly.

Therefore, there is a need to identify a standard process for selecting predictive models appropriate to a time series characteristic, allowing the best approach to be identified for each situation, maximising model performance and minimising prediction error. For

this reason, time series characteristics that may have an influence have been analysed, these characteristics include trend, seasonality, or time dependence.

Several approaches can be used for this purpose, ranging from classical statistical models such as autoregressive (AR), moving average (MA), autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA) or autoregressive integrated moving average with seasonality component (SARIMA) models, to another type of models such as the Facebook Prophet model, or machine learning models such as recurrent neural networks (RNN), where Long Short-Term Memory (LSTM) models.

Through the Mar Menor Data Web, historical data of all the parameters provided have been downloaded. The time interval of these series is about five years, from 2017, and the data come from different monitoring stations along the Mar Menor.

2. Materials and Methods

2.1. Mar Menor Dataset

Data from Mar Menor parameters provide key information to management and conservation decision making of this ecosystem as well as to implement adequate protection and restauration measurements. From the Servidor de Datos Científicos del Mar Menor data of the provided parameters have been downloaded with pre-treatment as interpolation which makes this data easier to process. The next parameters have been selected to the application of the methodology: Chlorophyll [mg/L], salinity [PSU], Oxygen [mg/L], Phycoerythrin [ppm], Water temperature [°C] and Transparency [m].

The time interval of these historical series is about 5 years, form 2017 until 2022. Data comes from different monitoring stations scattered throughout the Mar Menor and have subsequently been standardised by the supplier to a common grid, as shown in **Error! Reference source not found.** where is represented, in blue, OISMA (Oficina de Impulso Socioeconómico del Medio Ambiente) stations and, in red, the Servicio de Pesca stations.

Figure 1. OISMA and Servicio de Pesca stations where measurements have been made.

2.2. Time Series and Machine Learning Models

Time series are sequential observations recorded at regular intervals and analysed for patterns or components such as trend or seasonality, in this context the development of accurate and effective predictive models is essential to obtain reliable results. As mentioned in the introduction, two approaches of time series analysis have been evaluated to study the behaviour of these environmental parameters of the Mar Menor, statistical and machine learning models.

- Statistical models

Autoregressive, moving average models and its combination are used. AR(p) models calculate future values by linear combination of past values (p) determined by the partial autocorrelation function. Then, p is the order of the process that refers to the number of previous time steps in the time series that are used to predict the future value of the series. In MA(q) models the current value of a time series depends only on a small number of past values. This model calculates their current value takes a mean of past errors and adding once are multiplied by their respective coefficients. Them, model order (q) indicates the errors used to obtain the current value. A combination of the properties of the AR and MA processes, where the stationarity of the time series is assumed. The resulting process is stochastic and stationary, called by ARMA(p, q) [1]. In addition, there are an "integrated" version of a stationary series, they are called ARIMA(p,d,q), and are considered stationary after differentiation. They are the most general classical models in time series forecasting, where the parameter d represents the differencing order, which is the number of times the series is differenced to achieve stationarity. Also, SARIMA models consider seasonal patterns and improve forecasting accuracy and it is necessary to realize a deseasonalisation or seasonal difference (denoted by SARIMA(p,d,q)(P,D,Q), where P, D and Q represents the seasonal autoregressive, differencing and moving average order, respectively) [2]. Ultimately, Facebook prophet model, based on the fitting curve technique of the Bayesian model, is appropriate when there is a large seasonality, is robust to missing data or trend variations. This model is an autoregressive additive model with no lineal and observations have been made each hour, day, and month for one year or more [3].

- Machine-learning models

A Recurrent Neural Network (RNN) is a type of artificial neural network that specialises in pre-processing sequential data or time series. These networks have been trained to learn and are characterised by their "memory" of past inputs and use this information in decision making, both in the input stages and in the generation of results. In fact, RNN results depend on the sequence of past elements, allowing time dependencies in the data to be captured. In this paper, a Long Short-Term Memory (LSTM) algorithm, a type of RNN with an input layer, an intermediate layer, and an output layer, was used to introduce different time series as input and to train the network with these data [4].

2.3. Methodology

The standardised process is based on a systematic and objective approach. Clear criteria and relevant evaluation metrics have been used. A methodology has been followed to guide users through the various steps, from initial exploration and data pre-processing to the selection and tuning of appropriate predictive models.

This methodology has been divided in 5 phases, from data cleaning and visualisation, through pattern identification, data transformation for tunning models, model selection and pattern explanation, to model implementation. In the first phase, a detailed examination of the time series was carried out with the aim of identifying trends, possible missing data, thus, correction techniques such as interpolation or rolling averaging have been used where necessary and do not affect the subsequent analysis; data have been deleted where errors have been made; or original data have been retained where they provide information. Continuedly, during the *pattern identification* phase, Dickey-Fuller test were conducted to assess stationarity (null hypothesis is that the series is not stationary, and the alternative hypothesis is that it is stationary); and to assess seasonality, Partial Autocorrelation were implemented. Moreover, transformation techniques such as differentiation and deseasonalisation of the time series have been carried out according to the characteristics previously identified. Once the series have been studied, information criteria methods such as Akaike (AIC) and Schwarz Bayesian (BIC) were implemented to select the most appropriate model for each time series, which is set by the lower value of AIC and BIC. Finally, once the best model is selected, predictive models are applied to the series. For this, the dataset for the time series has been split into training and testing sets

in an 85/15 ratio. Predictions for statistical models were made using a predictive horizon of 7 and the training set is updated at each time step. Meanwhile, both the Facebook Prophet and LSTM models made predictions with a predictive horizon of 7 days on the entire 15% test set. Lastly, in order to assess how well the models fitted, several error metrics were implemented: *root mean square error* (RMSE), *mean absolute error* (MAE) and *mean absolute percentage error* (MAPE).

3. Application

Hereunder, the methodology is depicted on the Mar Menor dataset. It is worth mentioned that for the sake of simplicity and relevancy, only phycoerythrin (PE) and water temperature (T) parameters are shown as visual examples of the implementation. For the data cleaning and visualisation step, from the 1462 total data, only the 4.7% is removed as presented in Table 1 (Outliers). This is due to the processing, as mentioned above, these data are interpolated, and it has been found that empty or atypical data are not significant.

Table 1. Size, range, and outliers of Mar Menor datasets.

Dataset	Total Data	Range	Outliers
Chlorophyll	975	06/2017–01/2020	47
Salinity	2041	04/2017–10/2022	0
Oxygen	1737	04/2017–12/2021	0
PE	487	07/2021–10/2022	22
Temperature	974	04/2017–11/2019	0
Transparency	2252	09/2016–10/2022	0

Following data cleaning, pattern identification phase is conduced, in which, on the one hand, Dickey-Fuller test is applied assuming as null hypothesis then non-stationarity and as alternative hypothesis the stationarity of the series. This has been made with a reliability threshold of 95%, then have been considered that if p-value of series exceeds 0.05 the null hypothesis is cancelled, and the series is non-stationary. On the other hand, to seasonality analysis, partial autocorrelation has been applied, to which a value greater than 0.5 has been considered significant. Table 2 shows the results of these analysis.

Table 2. Seasonality and stationarity for the different dataset.

Dataset	*p*-Value	Correlation Value
Chlorophyll	0.065	0.558
Salinity	0.067	0.824
Oxygen	0.055	0.692
PE	0.232	0.371
Temperature	0.072	0.818
Transparency	0.061	0.807

It can then be concluded that, under the establishes criteria, no dataset has been stationary, on the other hand, except for the PE data, the data does have seasonality present. Afterwards, a differentiation or deseasonalisation has been applied to try to improve the data fit to the models and be able to estimate the suitable model, increasing the quality and decreasing the error rate of the forecasts.

To determine the best model, i.e., the model with the lowest value of AIC and BIC, we looked at different model fits by making combinations of the hyperparameters p and q, varying them from 0 to 4. Table 3 presents the models obtained for each parameter and the lower AIC and BIC valued reached. Thereby, predictions have been applied using these statistical models, in addition to Prophet and LSTM models, which are applied to the dataset as well.

Table 3. Most appropriate statistical model for the Mar Menor datasets based on the AIC and BIC.

Dataset	Model	AIC	BIC
Chlorophyll	SARIMA(2,1,1) (0,1,1)	393.866	417.483
Salinity	SARIMA(3,1,2) (0,1,1)	−6482.535	−6443.294
Oxygen	SARIMA(2,1,3) (0,1,1)	−3023.798	−2985.705
PE	ARIMA(2,1,1)	−2418.08	−2401.08
Temperature	SARIMA(2,1,0) (0,1,1)	−1686.45	−1667.05
Transparency	SARIMA(1,1,2) (0,1,1)	−4900.312	−4871.730

4. Results

Finally, the results of the prediction models are depicted in the following table. As mentioned above, predictions were undertaken with a horizon of 7 days for the statistical models and for Prophet and LSTM model, to which it is applied to the test data (15%). These results are shown in terms of the error metrics: *RMSE*, *MAE* and *MAPE* and can be seen in Table 4 and in graphics term in Figure 2, where the predictions made with the best model for each dataset are shown.

Table 4. Error metrics for different models and datasets.

Dataset	Evaluation	Statistical Model	Prophet Model		LSTM Model	
		Horizon 7	*Horizon 7*	*15%*	*Horizon 7*	*15%*
Chlorophyll	RMSE	**2.420**	4.377	5.621	9.63	6.431
	MAE	1.465	3.519	4.146	1.608	**1.311**
	MAPE	0.243	0.707	0.504	0.670	**0.102**
Salinity	RMSE	0.180	0.587	1.009	0.475	**0.152**
	MAE	**0.134**	0.488	0.840	0.550	0.359
	MAPE	**0.003**	0.012	0.021	0.028	0.008
Oxygen	RMSE	0.316	0.544	1.157	**0.025**	0.133
	MAE	0.214	0.435	1.014	**0.116**	0.315
	MAPE	**0.036**	0.078	0.184	0.198	0.058
PE	RMSE	0.170	0.334	0.333	0.009	**0.002**
	MAE	0.130	0.305	0.293	0.067	**0.041**
	MAPE	0.317	0.921	0.822	0.904	**0.114**
Temperature	RMSE	0.996	0.919	0.932	**0.130**	0.313
	MAE	0.697	0.743	0.758	**0.277**	0.499
	MAPE	0.032	0.036	0.035	0.256	**0.019**
Transparency	RMSE	0.219	1.153	3.060	0.037	**0.018**
	MAE	0.124	0.983	2.860	0.153	**0.121**
	MAPE	0.045	0.307	0.772	0.286	**0.035**

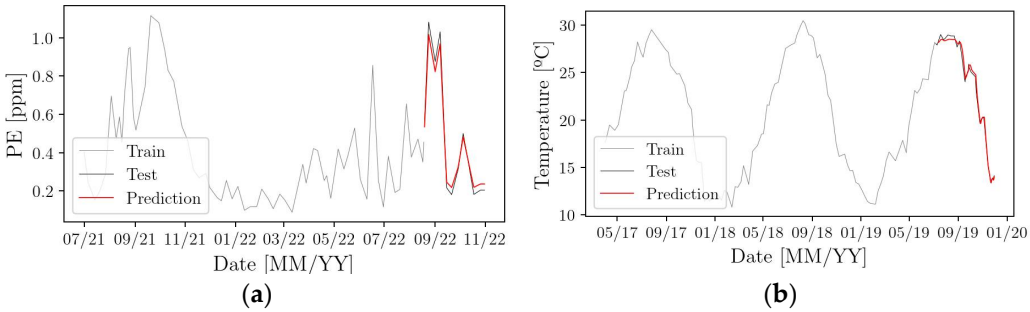

Figure 2. (**a**) Prediction of PE with LSTM (15%); (**b**) Prediction of WT[a] with LSTM (Horizon 7).

5. Conclusions

First, the importance of understanding the time series and defining the objectives of the analysis has been highlighted. The two most important characteristics in time series analysis have been identified and it is concluded that this part of the paper is fundamental to understand the data, reduce errors, obtain more accurate **predictions,** and prepare the series for further analysis. In other hand, it is important to know and understand both the key and error metrics, as well as having clean data, in order to make the correct model selection.

Finally, in case of the predictions has been seen that, in general, the models show good results but the best models for these data are statistical and LSTM models. For example, to PE data, the smallest error is obtained for a LSTM model with a predictive horizon of 7 days with a RMSE of 0.002.

6. Discussion

The general results and predictions of this paper are good, indicating that the methodology used is correct. However, there were limitations in terms of forecasting horizons, as it was not possible to make predictions beyond the forecasting horizon selected with the statistical and machine learning models. It was also not possible to make a longer-term forecast for the statistical models although it has been possible with machine learning models. On the other hand, it can be observed that the preprocessing task has not been as tough as it usually is as the data comes with a previous preprocessing and standardisation by the server (L4 level).

Finally, the chlorophyll errors obtained are high, due to the fact that the training data are very different from the data to be predicted, with small values at the beginning and increasing significantly at the end of the series.

Author Contributions: Conceptualization and methodology, R.M.Á.-C. and I.F.E.; data curation, M.N.M.; formal analysis, M.N.M. and J.C.S.G.; validation, R.M.Á.-C. and I.F.E., writing—original draft preparation, M.N.M. and J.C.S.G.; writing—review and editing, R.M.Á.-C. and I.F.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the *Instituto de Fomento de la Región de Murcia* (INFO) under the Program of grants aimed at Technological Centres of the Region of Murcia for the realization of non-economic R&D activities. Modality 1: Independent R&D Projects, with File No.: 2022.08.CT01.000040.

Institutional Review Board Statement:

Informed Consent Statement:

Data Availability Statement: The data presented in this study are openly available in https://marmenor.upct.es/thredds/catalog/L4/catalog.html (accessed on).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Box, G.; Jenkins, G.; Reinsel, G.; Ljung, G. *Time Series Analysis: Forecasting and Control*, 5th ed.; Wiley: Hoboken, NJ, USA, 2016.
2. Peña, D. *Análisis De Series Temporales*, 2nd ed.; Alianza Edityorial, SA.: Madrid, Spain, 2010.
3. Jha, B.K.; Pande, S. Time Series Forecasting Model for Supermarket Sales using FB-Prophet. In Proceedings of the 5th International Conference on Computing Methodologies and Communication, ICCMC 2021, Institute of Electrical and Electronics Engineers Inc., Erode, India, 8–10 April 2021; pp. 547–554. https://doi.org/10.1109/ICCMC51019.2021.9418033.
4. Bagnato, J.I. Pronóstico de Series Temporales con Redes Neuronales en Python | Aprende Machine Learning. Available online: (accessed on 21 February 2023).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.