Scientific Report on

Statistical activity in the assessment of water quality

Apoio Especial Verão Com Ciência 2022 (3982)

Rita Rodrigues (University of Lisbon)

September 2022

Contents

1 Introduction

This report describes the work performed during the one-month research fellowship for R&D initiation (BII) on the project "Statistical activity in the assessment of water quality" within the project Verão Com Ciência 2022 (3982) with Prof. Clara Cordeiro (FCT–UAlg and CEAUL), Dra. Sónia Cristina (UAlg–CIMA) and Prof. Soraia Pereira (FCUL and CEAUL) through the Center of Statistics and its Applications of the University of Lisbon (CEAUL), funded by Fundação para a Ciência e a Tecnologia (FCT).

The work is divided into four steps, in which the initial steps include the importation of the data to the R software, following up an exploratory analysis through statistical summary, identification of extreme values (outliers) and time series to observe the variation of chlorophyll-a concentration (Chl-a), sea surface temperature (SST), nutrients (ammonium (NH_4^+), phosphate (PO_4^{-3}), nitrate (NO_3^-), iron(Fe) and silicate(SiO_4^{-4})), North Atlantic oscillations (NAO) and upwelling index (UI; São Vicente and Cadiz) over the 18 years of data in the South coast of Portugal (off Sagres and Guadiana). The exploratory analysis also includes the identification of seasonality and the use of functions to remove it, such as *stl.fit* and *decompose*. All these steps are studied for both Guadiana and Sagres. The last step during this research fellowship includes the evaluation and interpretation of generalized linear models (GLMs) and generalized additive models (GAMs). This step will only be studied for Sagres site.

The goal of this work is to verify statistically which variables best model chlorophyll-a.

2 Roadmap

The work performed during this one-month summer research fellowship was divided into four steps. The first step was mainly the importation of the data to the R software, following up an initial data exploration of the datasets obtained for both stations in the south coast of Portugal (off Sagres and Guadiana) in order to verify the existence of extreme values (outliers) and understand the variation of the data over 18 years of data. The second step consisted in an initial data analysis through time series plots, statistical summaries and an autocorrelation function (ACF). The exploratory analysis of the data continued in the third step, in which stl.fit and decompose function were explored to remove the seasonality of the variables and create a new data frame for both stations. This data frame has variables that had its seasonality removed and the others that never had seasonality (confirmed by the ACF function), then scatterplots were plotted to observe the relation between the independent (SST, nutrients, UI and NAO) and dependent (Chl-a) variable after the remotion using stl.fit. This step also included the initial multiple regression as a pre-preparation for the GLMs and GAMs. The last step involved the evaluation and interpretation of the models obtained.

3 Stage 1

3.1 Data extraction

The first activities carried out involved the extraction of the satellite remote sensing data and initial analysis of the dataset obtained for both stations in Algarve - off Sagres (37.010522°, -8.878317°) and off Guadiana (37.148036°, -7.369129°). Monthly time series of Chl-a, SST and NO_3^-, NH_4^+, PO_4^{-3}, SiO_4^{-4} and Fe were downloaded from the EU Copernicus Marine Environmental Monitoring Service https://resources.marine.copernicus.eu/products on 6th September 2022. Chl-a Level 4 (L4) database is from 2002 to 2019, has a resolution of 4km and is referenced as "OCEANCOLOUR_GLO_BGC_L4_MY_009_104"; the SST L4 database is from 2002 to 2019, has a resolution of 0.05 km and is referenced as"SST_ATL_SST_L4_REP_OBS_FULL_TIME_SERIE". The SST database initially had monthly data, referred to as "MULTIOBS_GLO_PHY_TSUV_3D_MYNRT_0 and was downloaded on the same day as the other variables, but did not have the SST values for Sagres and Guadiana. For this reason, the SST database was updated on 22nd September 2022 for a daily database which was converted to monthly values after. The nutrients were obtained from "IBM_MULTIYEAR_BGC_005_003" and the database is from 2002 to 2019. Both Chl-a and SST databases are satellite remote sensing data, while the nutrients data are retrived by nummerical models. Additionally, Upwelling Index (UI) data were obtained from the Spanish Institute of Oceanography http://www.indicedeafloramiento.ieo.es/afloramiento_en.html, in which monthly and daily data, from 2002 to 2019, were obtained for the stations of Cape São Vicente and Cadiz chosen due to its proximity to Sagres and Guadiana, respectively. The NAO were downloaded from National Weather Service Climate Prediction Center of the National Oceanic and Atmospheric Administration (NOAA) `https://www.cpc.ncep.noaa.gov/products/precip/CWlink/pna/nao.shtml`, the monthly NAO index was obtained for the period of 2002 to 2019.

Table 1: Designations for all variables used in this work with their respectively abbreviations and units.

Designation	Abbreviations	Unit
Chlorophyll-a	Chl-a	$mg\ m^{-3}$
Sea Surface Temperature	SST	C°
Nitrate	NO_3^-	µM
Ammonium	NH_4^+	µM
Phosphate	PO_4^{-3}	µM
Silicate	SiO_4^{-4}	µM
Iron	Fe	µM
Upwelling Index	UI	$m^3\ s^{-1}\ km^{-1}$
North Atlantic Oscillations	NAO	–

3.2 Importation and definition of the databases

The statistical analysis were performed using R version 4.2.1 and the packages Rcmdr, weathermetrics, ggplot2 and forecast were used for exploratory data analysis.

The data was imported to RStudio in order to create dataframes for Sagres and Guadiana, in which the varibles were Chl-a, SST, NAO, nutrients (NO_3^- , NH_4^+, PO_4^{-3}, SiO_4^{-4} and Fe) and UI (São Vicente for Sagres and Cadiz for Guadiana). A table was created for a better comprehension of the variables abbreviations and units (Table 1). The R script search for the available data for each parameter at the closest geographical coordinates set for each site location based on the resolution of the datasets. The minimum distance obtained for Chl-a was 1.93 km and 1.3 km, SST 1.63 km and 2.5 km, and nutrients 15.03 km and 3.7 km off Sagres and Guadiana, respectively. Since the database for SST had daily data, the mean was calculated for each month. In the end, the dataframe contained all these variables for the period of January of 2002 until December of 2019. All variables were defined as an object time series (ts) and none of the datasets had missing values (NAs).

3.3 Initial analysis of the datasets

The initial analysis of the datasets were performed to observe if there is outliers and to understand the variation of each variable over the years. For this, time series plot were performed to observe the variables.

In this monthly time series from 2002 to 2019, it's possible to initially observe that between 2006 and 2016 there's a general uptrend in PO_4^{-3} and

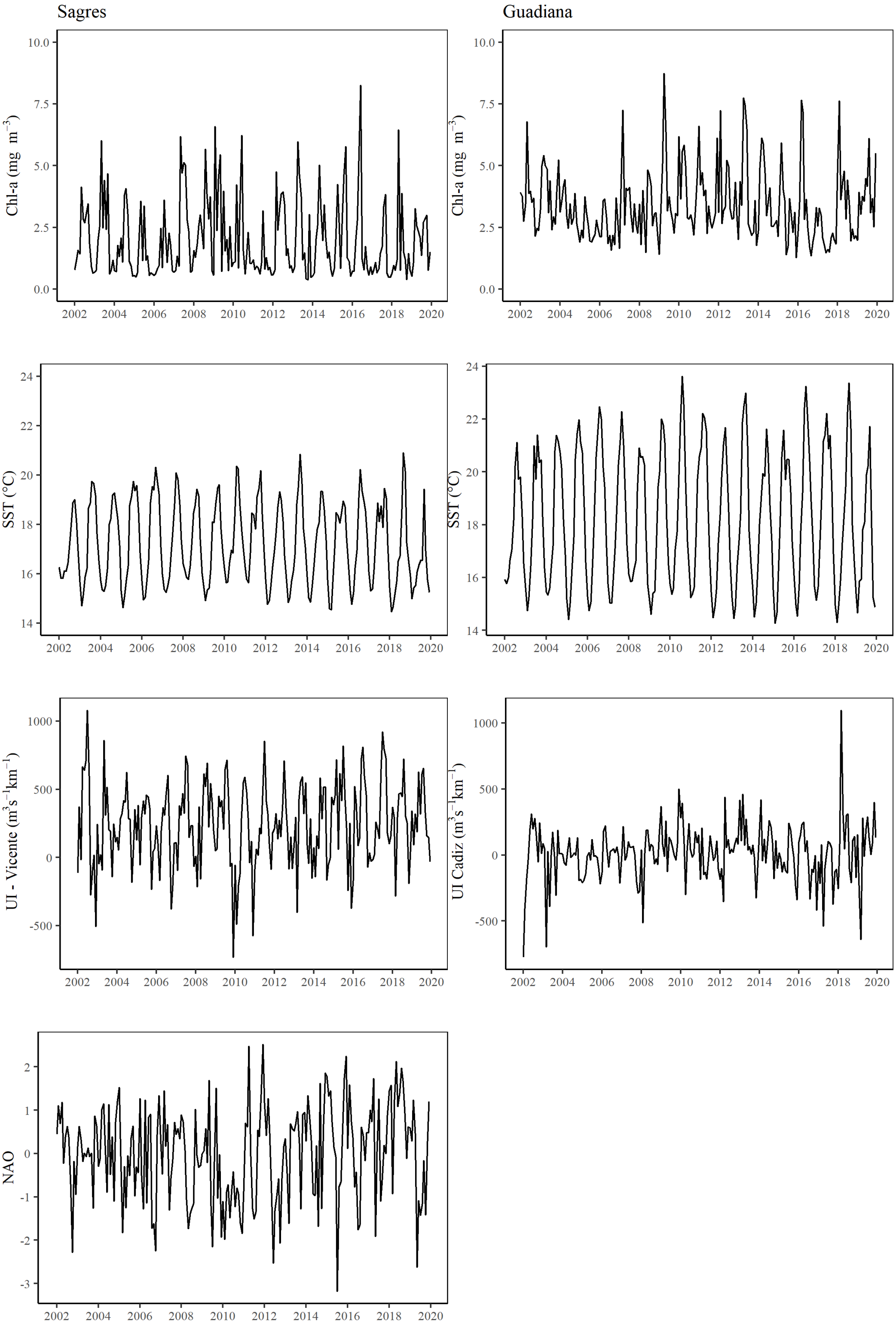

Figure 1: Time series of Chl-a, SST, UI and NAO off Sagres (left) and Guadiana (right).

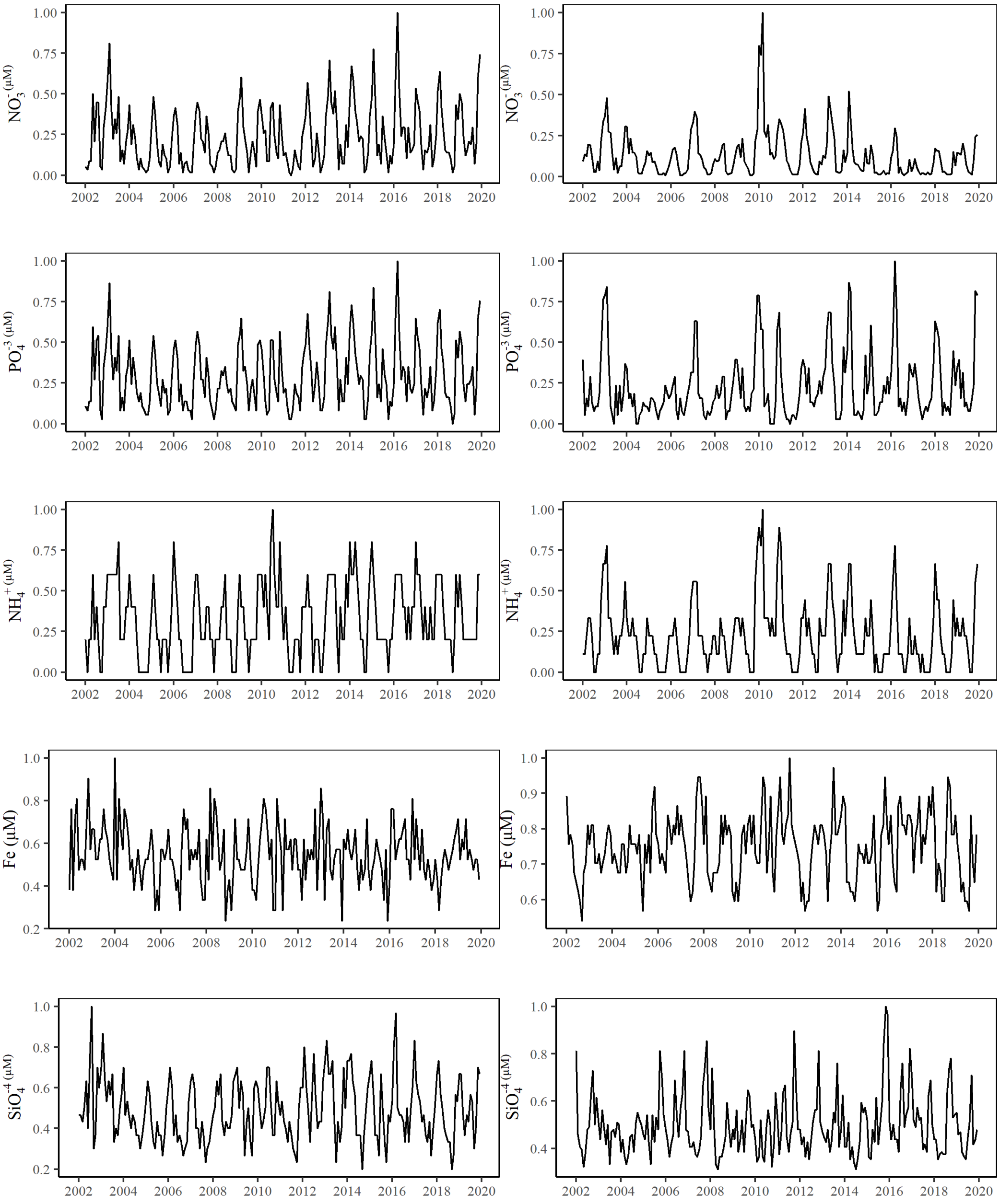

Figure 2: Time series of the nutrients off Sagres (left) and Guadiana (right).

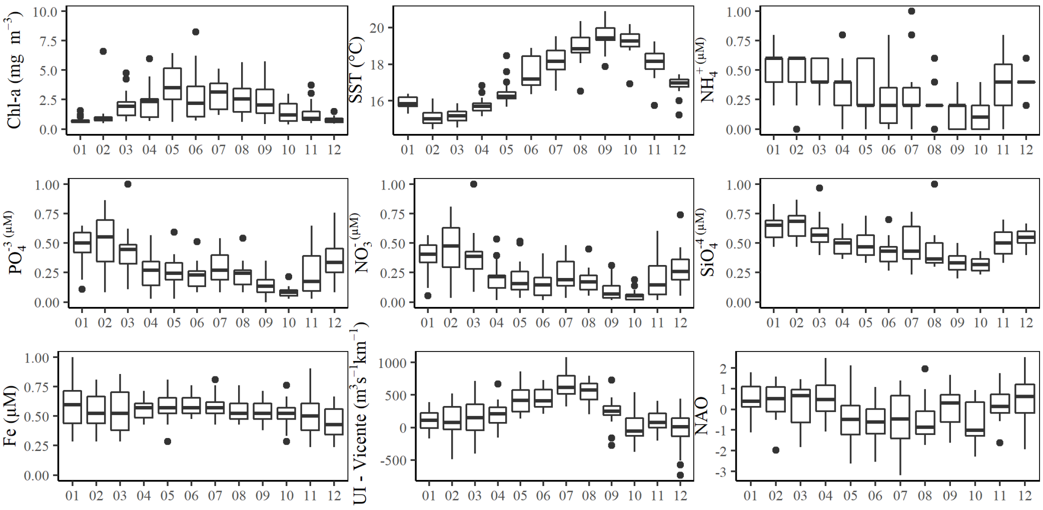

Figure 3: Boxplots off Sagres.

SiO_4^{-4} for both locations, and NO_3^- for Sagres (Fig. 2). Then two patterns were identified. On the one hand, there is an uptrend in the first half of the year and a downtrend in the second half, as seen in Chl-a and SST for both stations and also UI-Vicente and all nutrients (NO_3^-, NH_4^+, PO_4^{-3}, SiO_4^{-4} and Fe) for Sagres. On the other hand, there's a downtrend in the first half of the year and a uptrend in the second half, as seen in NAO for both stations and Fe and all nutrients (NO_3^- , NH_4^+, PO_4^{-3}, SiO_4^{-4} and Fe) for Guadiana (Fig. 2). There are also some outliers, for example, NO_3^- in 2010 and UI-Cadiz in 2018 for Guadiana, UI-Vicente in 2010 and NAO in 2015 for Sagres. Additionally, boxplots were made to observe the extreme values in the datasets. It is important to refer that to make all nutrients have the same scale the values of each variable was divided by its maximum value obtaining a scale from 0 to 1.

There is an increase in Chl-a concentration between spring and the end of summer, coinciding with the higher SST off Sagres (Fig. **??**, as well as the UI of São Vicente also has an increase in summer. Meanwhile, the increase of the nutrients concentrations occurs mostly in winter months (December to Febuary). Phosphate, nitrate and silicate has a noticeable pattern of increasing concentrations in winter, while iron and ammonium hardly do. The NAO vary throughout the year, not having a visible distribution over the seasons.

In the boxplots of Guadiana (Fig. 4) the median concentrations of Si and Fe are higher than those of other nutrients (NO_3^- , NH_4^+ and PO_4^{-3}). It's also possible to observe that during the months of summer the concentrations

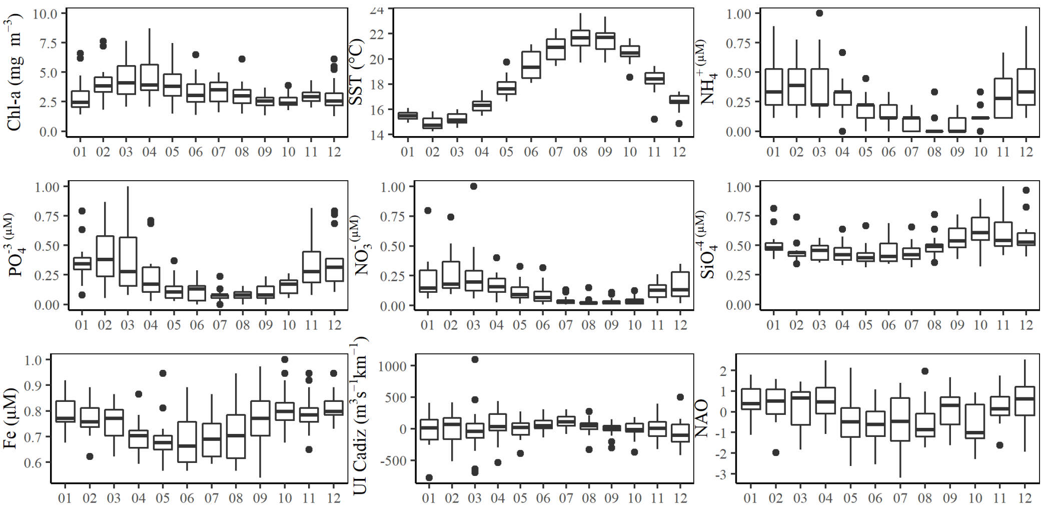

Figure 4: Boxplots off Guadiana.

of NO_3^-, NH_4^+ and PO_4^{-3} are smaller than the rest of the year, while the temperature is higher than the rest of the year. The nutrients NO_3^-, PO_4^{-3}, SiO_4^{-4} and SST and UI have less range than the rest of the variables, meaning the less spread the values are. Another remarkable information about the data sets is that all variables for both stations had outliers.

Moreover, the nummerical summary was performed to obtain the minimum, 1st quartile, median, 3rd quartile, and maximum values for both Sagres and Guadiana.

The summary obtained for both datasets (Table 2 Table 3) did not have extreme variations between locations. SST had similar values for both stations, whereupon the Chl-a had a higher concentration off Guadiana. The nutrients had elevated concentrations in Guadiana compared to Sagres, especially NO_3^- and SiO_4^{-4}, in which the maximum values had a difference of 14.100 and 6.600 µM respectively. Fe, PO_4^{-3} and NH_4^+ had the lowest concentrations for both locations presenting maximum values under 1 µM. NAO has the same values for Sagres and Guadiana.

4 Stage 2

4.1 Detecting patterns

In order to see a relationship between the variables, an ACF was performed using *ggAcf()* from the package *forecast* since it measures the linear relationship between lagged values of a time series. The ACF plots were important

Table 2: Descriptive statistics summary of Sagres

	Min	**1st quartile**	**Median**	**3rd quartile**	**Max**
Chl-**a**	0.378	0.778	1.368	2.831	8.246
SST	14.463	15.780	16.958	18.748	20.892
NO_3^-	-0.000	0.500	1.100	2.100	5.800
NH_4^+	0.000	0.100	0.100	0.225	0.500
SiO_4^{-4}	0.600	1.100	1.400	1.800	3.000
Fe	0.001	0.001	0.001	0.001	0.002
PO_4^{-3}	0.000	0.050	0.090	0.163	0.370
NAO	-3.179	-0.894	0.138	0.731	2.521
UI-**Vicente**	-731.619	43.746	243.796	445.137	1079.957

Table 3: Descriptive statistics summary of Guadiana

	Min	**1st quartile**	**Median**	**3rd quartile**	**Max**
Chl-**a**	1.283	2.375	3.080	4.030	8.737
SST	14.27	15.85	18.06	20.51	23.62
NO_3^-	0.200	0.600	1.900	3.500	20.200
NH_4^+	0.000	0.100	0.200	0.300	0.900
SiO_4^{-4}	3.000	3.900	4.600	5.225	9.600
Fe	0.002	0.002	0.002	0.003	0.003
PO_4^{-3}	0.000	0.030	0.060	0.120	0.380
NAO	-3.179	-0.894	0.137	0.731	2.521
UI-**Cadiz**	-773.75	-91.57	21.44	117.04	1096.02

Table 4: Comparasion of the value of the error measure using the mean absolute error (MAE) in *stl.fit* and *decompose* functions for Sagres and Guadiana.

	Sagres		**Guadiana**	
	stl.fit	**decompose**	**stl.fit**	**decompose**
Chl-**a**	0.687	0.886	0.675	0.841
SST	0.322	0.390	0.381	0.481
NO_3^-	0.399	0.495	0.896	1.170
NH_4^+	0.045	0.058	0.072	0.086
PO_4^{-3}	0.027	0.036	0.030	0.038
SiO_4^{-4}	0.171	0.226	0.617	0.676
Fe	*	*	< 0.001	< 0.001
UI	119.321	152.860	*	*

to understand which variables presented a seasonality by having a sinusoidal form.

The autocorrelation off Sagres illustrates which variables has a seasonality (Fig. 5), whereas SST has the most evident sinusoidal form compared to the other variables. The pattern of seasonality is visible, in a lower scale, for the other variables, except Fe and the NAO. These variables doesn't show seasonal pattern and it is evident that there is no sinusoidal form.

In the case of Guadiana (Fig. 6), the sinusoidal form is clear in SST, NO_3^- , NH_4^+ and PO_4^{-3}. In a smaller scale it's also possible to see a sinusoidal form in Chl-a, SiO_4^{-4} and Fe. UI and NAO do not have any pattern.

In conclusion, based in the figures above, all variables, except Fe and NAO, have seasonality in Sagres, while in Guadiana all variables, except UI-Cadiz and NAO, have seasonality.

4.2 Estimating seasonality

The initial estimation of seasonality was focused on removing the seasonality from those variables that had it, as seen in the autocorrelation plot for Sagres and Guadiana. For this reason, the *stl.fit* (based on Loess) was downloaded from https://github.com/ClaraCordeiro/stl.fit for being a useful tool to remove the seasonality of time series data. Additionally, the *decompose* function from the package *stats* was also performed as a comparison with the value of the error measure obtained from *stl.fit* function, in order to use the function that has the lowest error measures for all variables.

According to the Table 4, the value of the mean absolute error (MAE)

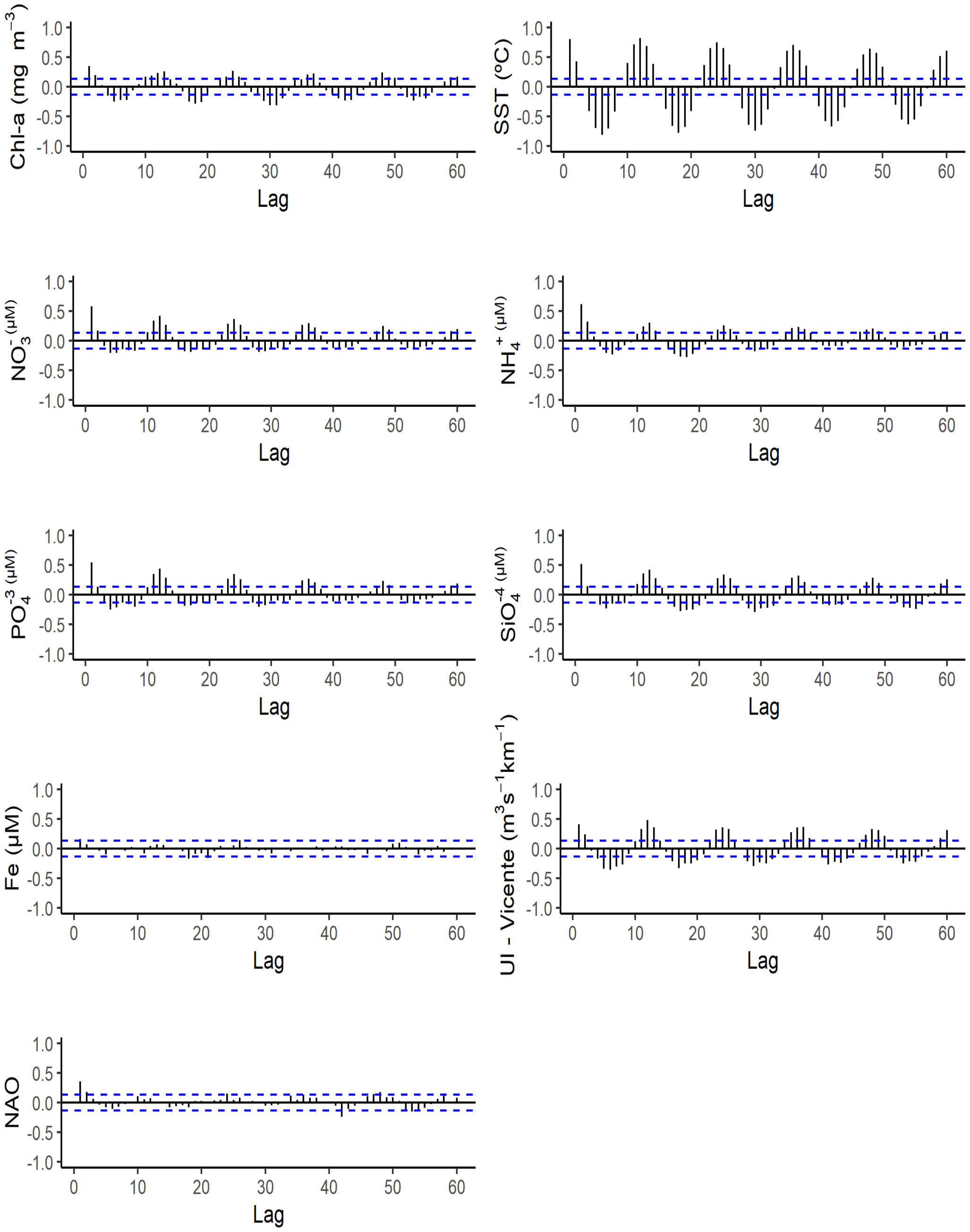

Figure 5: Autocorrelation function of Sagres.

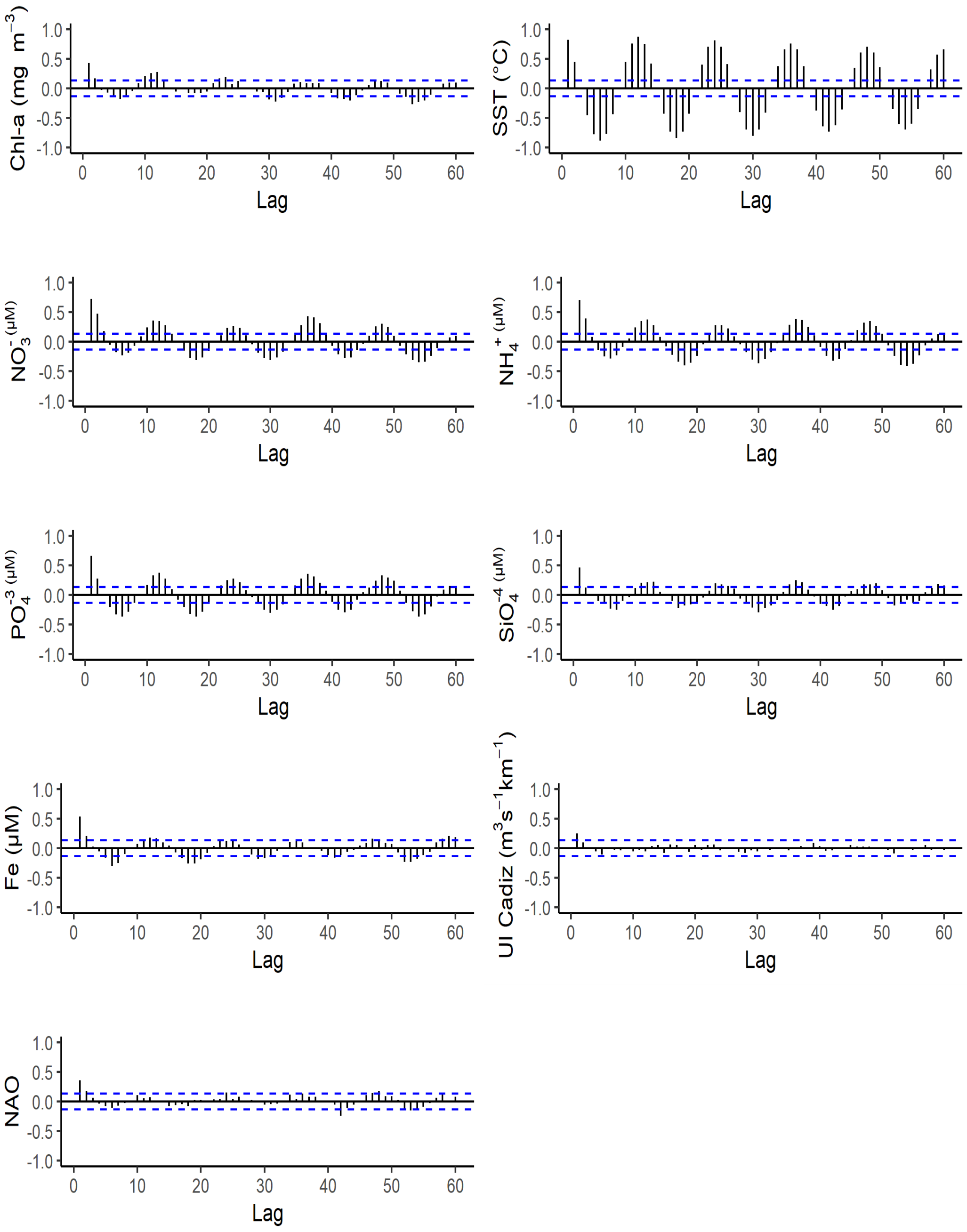

Figure 6: Autocorrelation function of Guadiana.

between the variables for Sagres and Guadiana did not have much difference, most variables had a similar error considering both functions (*stl.fit* and *decompose*). Meanwhile, it is evident that *stl.fit* has a lower error measure than decompose. The comparison with the value of the error measures between the stl.fit and *decompose* demonstrated that *stl.fit* was more accurate, by this means it was preferred to use *stl.fit* in our analysis. In order to confirm that the seasonality was removed, the autocorrelation function was performed again.

Based on the ACF plots, it is evident that the seasonality pattern was removed from all variables off Sagres (Fig. 7), for instance SST had a clear sinusoidal form which is possible to see that it does not exist anymore. Fe and NAO did not have a seasonality before, and for this reason it was not needed to plot these variables again.

It's not possible to see the sinusoidal form that was evident before off Guadiana (Fig. 8). This implies that the seasonality was in fact removed from the variables that had it. There was no need to plot UI neither NAO because those variables didn't have seasonality.

5 Stage 3

5.1 First steps in modelling

From now on all the analysis will just focus on Sagres site. A new data frame was created with the variables that was possible to remove the seasonality and the others that did not have seasonality. In which, only Fe and NAO did not have seasonality. Scatterplots were performed in order to see the relation between independent variables (SST, nutrients, NAO and UI) and the dependent variable (Chl-a).

The nutrients shows different correlations with the Chl-a in Sagres (Fig. 9). An uphill relation is evident between Chl-a and SiO_4^{-4}, showing a low positive correlation. As well, NO_3^- has a slighly positive correlation. Chl-a and NH_4^+ seems to have a weak negative correlation, but as there is many outliers it is not much evident. On other hand, NAO does not increase or decrease as it moves horizontally, not showing a clear correlation. SST and UI of São Vicente does not seem to have a correlation with Chl-a since the data is more spread out. As there is no signs of trends, there is no correlation between these variables and the Chl-a.

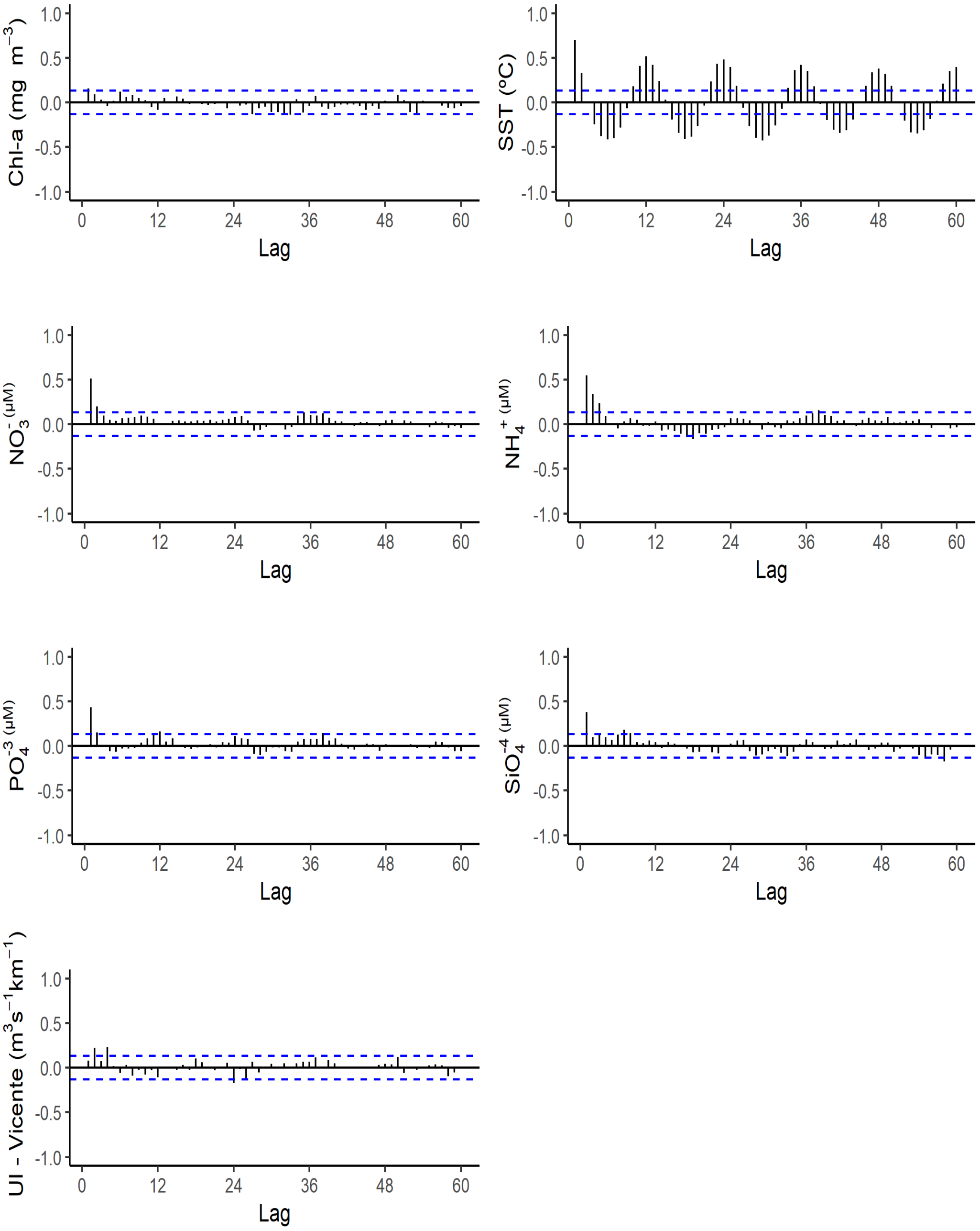

Figure 7: Autocorrelation function of Sagres without seasonality for all variables, except Fe and NAO

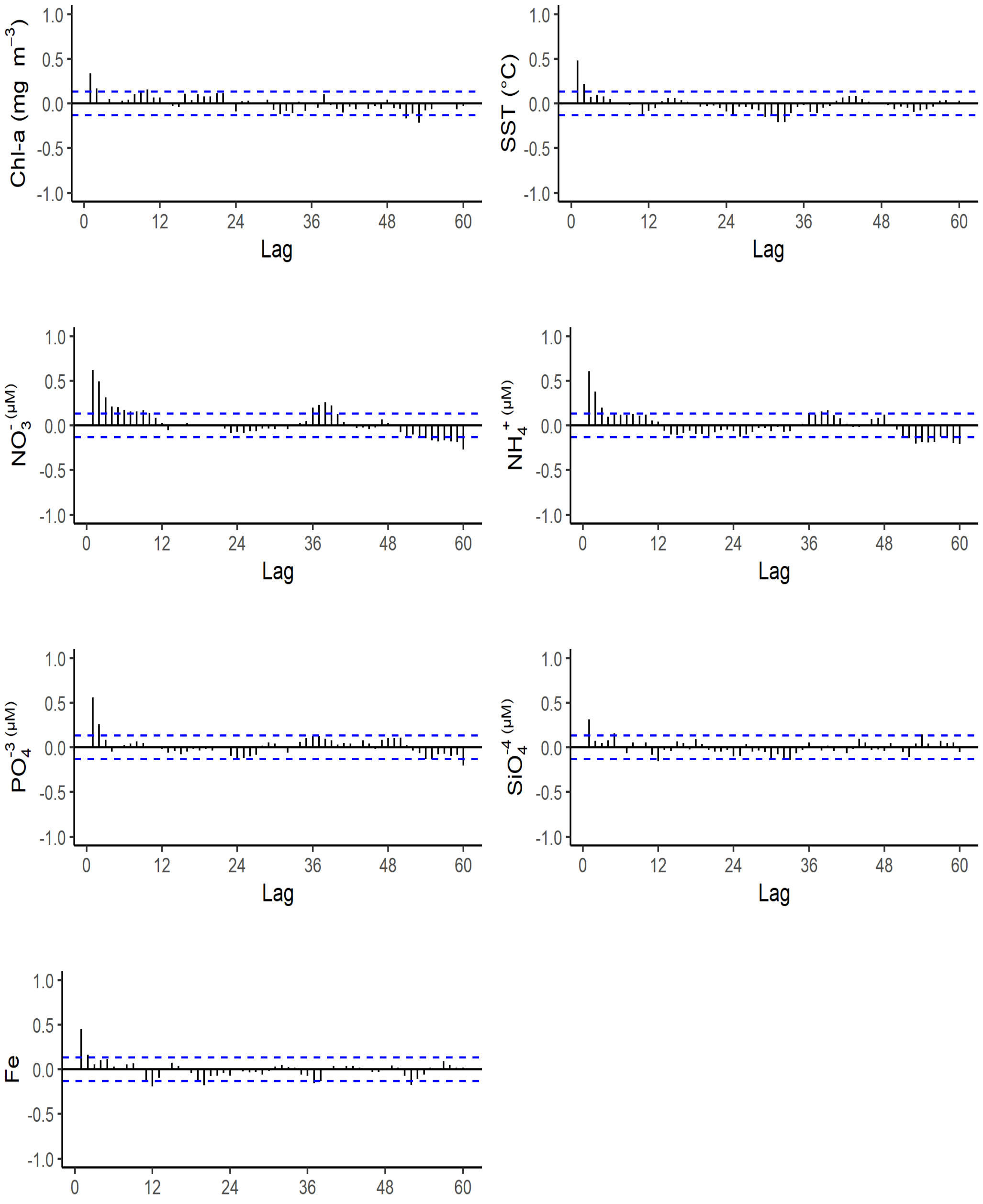

Figure 8: Autocorrelation function of Guadiana without seasonality for all variables, except UI and NAO

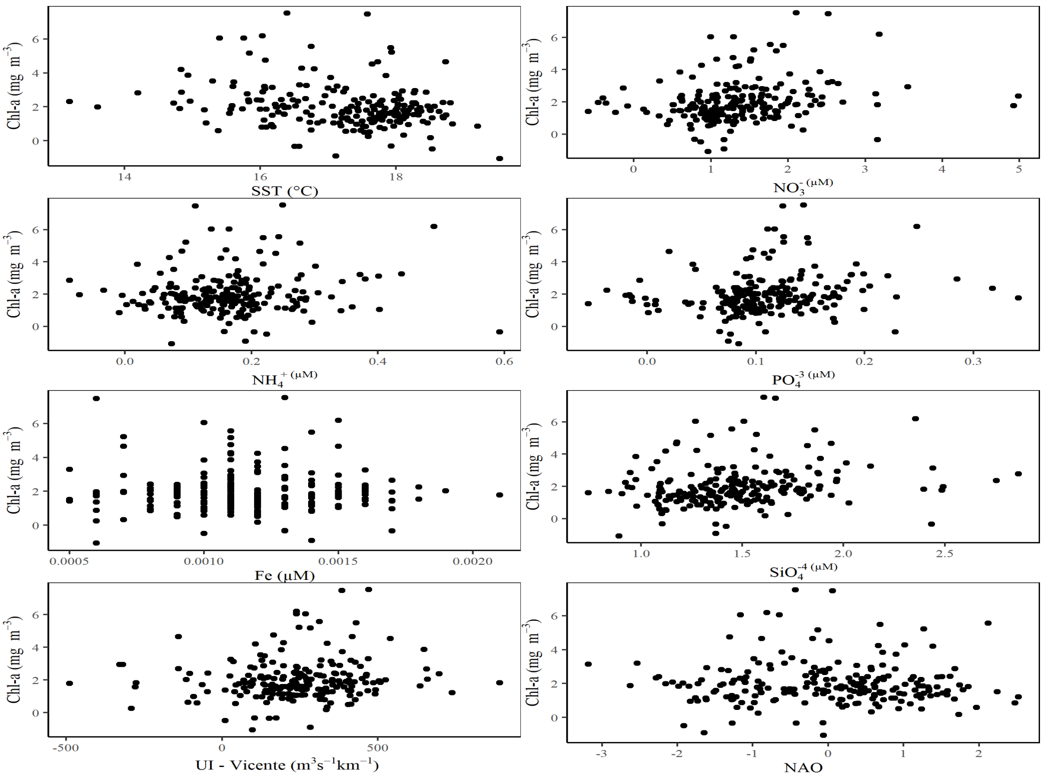

Figure 9: Scatterplots of Sagres.

Table 5: Correlation matrix of Sagres.

	SST	NO_3^-	PO_4^{-3}	SiO_4^{-4}	NH_4^+	Fe	NAO	UI
Chl-**a**	-0.050	0.168	0.209	0.192	0.096	-0.018	-0.030	0.112
SST	1.000	-0.300	-0.309	-0.234	-0.194	-0.024	0.262	-0.164
NO_3^-	-0.300	1.000	0.935	0.757	0.673	0.279	-0.044	0.182
PO_4^{-3}	-0.309	0.935	1.000	0.746	0.652	0.319	-0.077	0.198
SiO_4^{-4}	-0.234	0.757	0.746	1.000	0.549	0.400	-0.118	0.167
NH_4^+	-0.194	0.673	0.652	0.549	1.000	0.310	-0.152	-0.005
Fe	-0.024	0.279	0.319	0.400	0.310	1.000	0.001	0.132
NAO	0.262	-0.044	-0.077	-0.118	-0.152	0.001	1.000	0.314
UI	-0.164	0.182	0.198	0.167	-0.005	0.132	0.314	1.000

Table 6: Correlation matrix of Sagres without nitrate

	SST	PO_4^{-3}	SiO_4^{-4}	NH_4^+	Fe	NAO	UI
Chl-**a**	-0.050	0.209	0.192	0.096	-0.018	-0.030	0.112
SST	1.000	-0.309	-0.234	-0.194	-0.024	0.262	-0.164
PO_4^{-3}	-0.309	1.000	0.746	0.652	0.319	-0.077	0.198
SiO_4^{-4}	-0.234	0.746	1.000	0.549	0.400	-0.118	0.167
NH_4^+	-0.194	0.652	0.549	1.000	0.310	-0.152	-0.005
Fe	-0.024	0.319	0.400	0.310	1.000	0.001	0.132
NAO	0.262	-0.077	-0.118	-0.152	0.001	1.000	0.314
UI	-0.164	0.198	0.167	-0.005	0.132	0.314	1.000

5.1.1 Multiple Linear Regression

A correlation matrix was performed for a better understanding of the relation between the indepent variables (SST, nutrients, NAO and UI). The objective is to look for a linear relation, and for this purpose the Pearson coefficient was used.

If two variables have a correlation greater than 0.8 or less than -0.8 then exists multicollinearity. Off Sagres (Table 5), PO_4^{-3} and NO_3^- have a correlation of 0.935 (> 0.80). For this reason, NO_3^- was removed because is has the lowest correlation with Chl-a ($0.168 < 0.209$). The matrix will be remade without NO_3^-. Now all variables have a correlation between [-0.8, 0.8] (Table 6). The variables that are not statistical useful in helping to describe Chl-a will be removed. For this purpose, a process called Backward Selection will be used. The P-to-leave is set to 0.15, in which the initial analysis is to perform the full model, Chl-a $\sim NH_4^+ + PO_4^{-3} + SiO_4^{-4}$ + SST + UI + Fe + NAO. In this model NAO had the greatest p-value, 0.720, which is greater than 0.15, so this variable was removed. Then the model Chl-a $\sim NH_4^+ + PO_4^{-3} + SiO_4^{-4}$ + SST + UI + Fe was performed. This time PO_4^{-3} had the greatest p-value ($0.722 > 0.15$). It was removed and the model Chl-a $\sim NH_4^+ + SiO_4^{-4}$ + SST + UI + Fe was built. In this case NH_4^+ had the greatest p-value ($0.713 > 0.15$). Once again, it was removed and the model was rebuilt Chl-a $\sim SiO_4^{-4}$ + SST + UI + Fe. The greatest p-value this time was 0.599 by UI. It was removed and the model was performed as Chl-a $\sim SiO_4^{-4}$ + SST + Fe. Fe still had a p-value greater than 0.15, 0.628. Then, the model built was Chl-a $\sim SiO_4^{-4}$ + SST. Si was removed considering the level of significance, 0.05, with a p-value of 0.14. The final model was Chl-a $\sim$ SST in which SST had a p-value of $1.60e - 10(< 0.15)$. In order to check the normality of the model, a quantile-comparison plot (Fig. 10) was plotted with the residuals of the last model (Chl-a $\sim$ SST).

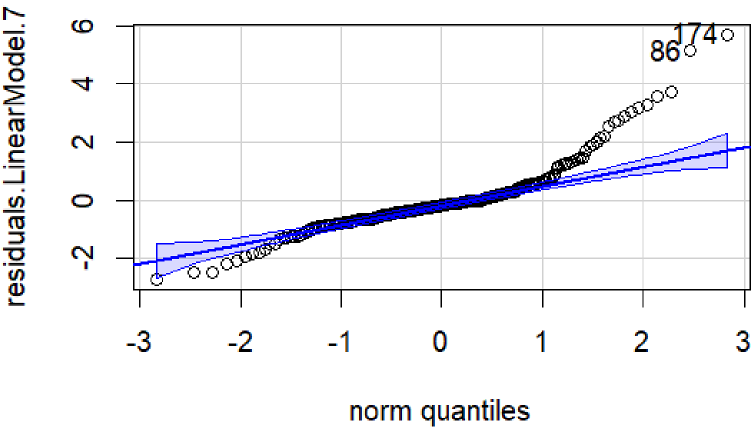

Figure 10: Q–Q plot (quantile-quantile plot) of the final linear model between Chl-a $\sim$ SST at Sagres station

As the normality fails, it is needed to transform the Chl-a column to log(Chl-a). In this new dataset, the multicollinearity will be verified again.

Just like in the previous case, NO_3^- will be removed (Table 7).

Now there is no multicollinearity (Table 8). The same process is followed as before. The order of elimination of variables was PO_4^{-3}, Fe, UI, NAO, NH_4^+, SiO_4^{-4} and the final model was log(Chl-a) $\sim$ SST, in which SST had a p-value of $3.84e - 13$, which is less than 0.15. The residuals were analysed

Table 7: Correlation matrix of Sagres between the dependent variable (log(Chl-a) with the independent variables (SST, nutrients, NAO and UI).

	SST	NO_3^-	PO_4^{-3}	SiO_4^{-4}	NH_4^+	Fe	NAO	UI
Chl-**a**	-0.083	0.103	0.127	0.172	0.098	0.003	-0.008	0.111
SST	1.000	-0.213	-0.160	-0.029	-0.169	0.151	-0.146	-0.256
NO_3^-	-0.213	1.000	0.928	0.775	0.706	0.247	-0.047	0.062
PO_4^{-3}	-0.160	0.928	1.000	0.748	0.666	0.287	-0.069	0.098
SiO_4^{-4}	-0.029	0.775	0.748	1.000	0.609	0.363	-0.121	0.004
NH_4^+	-0.169	0.706	0.666	0.609	1.000	0.260	-0.106	-0.023
Fe	0.151	0.247	0.287	0.363	0.260	1.000	-0.126	-0.037
NAO	-0.146	-0.047	-0.069	-0.121	-0.106	-0.126	1.000	0.314
UI	-0.256	0.062	0.098	0.004	-0.023	-0.037	0.314	1.000

Table 8: Correlation matrix of Sagres of log(Chl-a) without the nitrate

	SST	PO_4^{-3}	SiO_4^{-4}	NH_4^+	Fe	NAO	UI
Chl-**a**	-0.083	0.127	0.172	0.098	0.003	-0.008	0.111
SST	1.000	-0.160	-0.029	-0.169	0.151	-0.146	-0.256
PO_4^{-3}	-0.160	1.000	0.748	0.666	0.287	-0.069	0.098
SiO_4^{-4}	-0.029	0.748	1.000	0.609	0.363	-0.121	0.004
NH_4^+	-0.169	0.666	0.609	1.000	0.260	-0.106	-0.023
Fe	0.151	0.287	0.363	0.260	1.000	-0.126	-0.037
NAO	-0.146	-0.069	-0.121	-0.106	-0.126	1.000	0.314
UI	-0.256	0.098	0.004	-0.023	-0.037	0.314	1.000

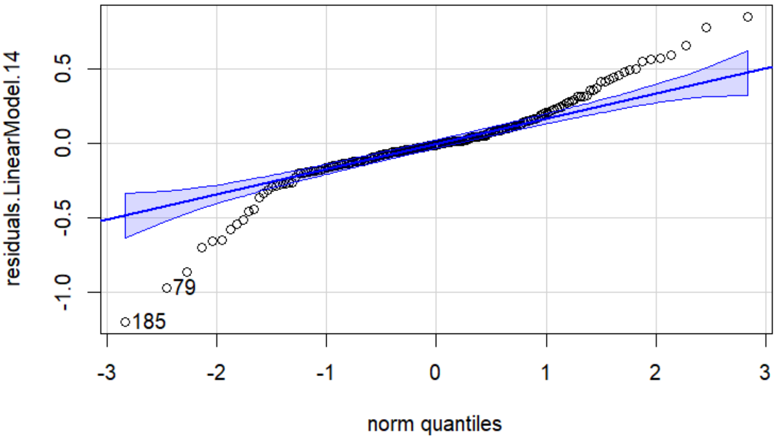

Figure 11: Q–Q plot (quantile-quantile plot) of the final linear model between log(Chl-a) ∼ SST at Sagres station

Table 9: Results of Shapiro-Wilk normality test on the residuals obtained from the LM of Chl-a and log(Chl-a) off Sagres

Residuals of LM	*p-value*
Chl-a	$<< 0.001$
Log(Chl-a)	$<< 0.001$

(Fig. 11).

Once again, the normality fails for Sagres (Fig. 11; Table 9).

5.1.2 Generalized Linear Models

The first step in this stage was plotting an histogram so that the distribution of Chl-a is identified. In this stage the transformation $Chl - a - min(Chl - a) + 1$ was applied to Chl-a so that all values are positive.

Chlorophyll-a seems to follow a gamma distribution (Fig. 12). This distribution will be used in the following models. But before that, a correlation matrix will be performed to check the multicollinearity between the independent variables and Chlorophyll-a.

Since PO_4^{-3} and NO_3^- have a correlation of 0.950, NO_3^- will be removed (Table 10) because it has the lowest correlation with Chl-a ($0.209 < 0.245$). The matrix will be calculated again but this time without NO_3^-.

Now none of the values are greater than 0.8 or less than -0.8 (Table 11). The GLMs can now be performed with all variables except NO_3^-. As

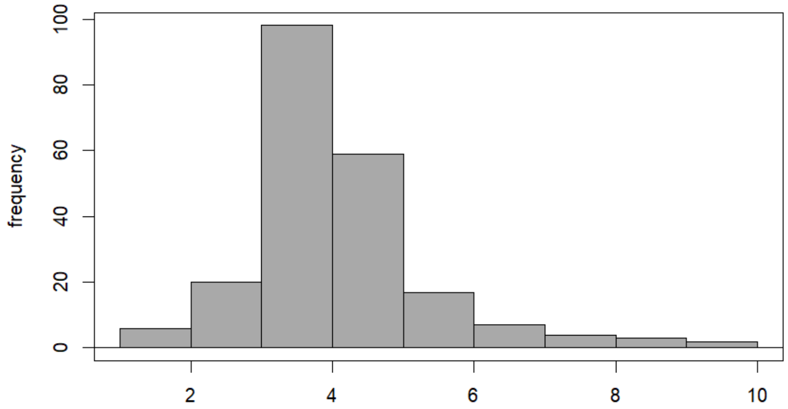

Figure 12: Histogram of chlorophyll-a at Sagres

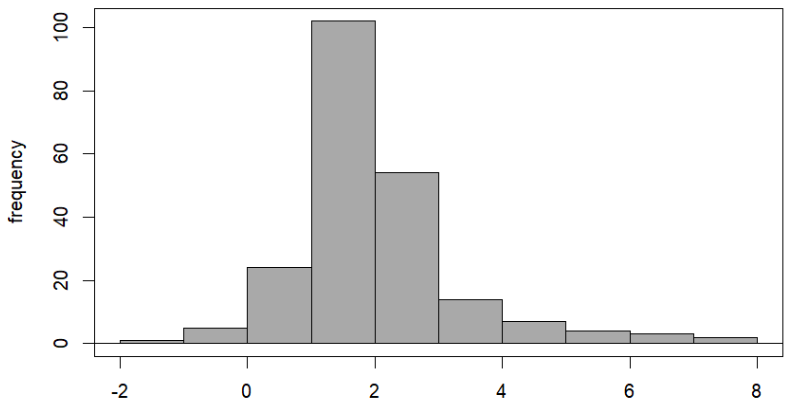

Figure 13: Histogram of log(chlorophyll-a) at Sagres

Table 10: Correlation matrix of Sagres.

	SST	NO_3^-	PO_4^{-3}	SiO_4^{-4}	NH_4^+	Fe	NAO	UI
Chl-**a**	-0.050	0.209	0.245	0.225	0.096	-0.001	-0.018	0.118
SST	1.000	-0.213	-0.160	-0.029	-0.169	0.151	-0.146	-0.256
NO_3^-	-0.213	1.000	0.950	0.775	0.706	0.247	-0.047	0.062
PO_4^{-3}	-0.160	0.950	1.000	0.748	0.666	0.287	-0.069	0.098
SiO_4^{-4}	-0.029	0.775	0.748	1.000	0.609	0.363	-0.121	0.004
NH_4^+	-0.169	0.706	0.666	0.609	1.000	0.260	-0.106	-0.023
Fe	0.151	0.247	0.287	0.363	0.260	1.000	-0.126	-0.037
NAO	-0.146	-0.047	-0.069	-0.121	-0.106	-0.126	1.000	0.314
UI	-0.256	0.062	0.098	0.004	-0.023	-0.037	0.314	1.000

Table 11: Correlation matrix of Sagres without NO_3^-

	SST	PO_4^{-3}	SiO_4^{-4}	NH_4^+	Fe	NAO	UI
Chl-**a**	-0.050	0.245	0.225	0.096	-0.001	-0.018	0.118
SST	1.000	-0.160	-0.029	-0.169	0.151	-0.146	-0.256
PO_4^{-3}	-0.160	1.000	0.748	0.666	0.287	-0.069	0.098
SiO_4^{-4}	-0.029	0.748	1.000	0.609	0.363	-0.121	0.004
NH_4^+	-0.169	0.666	0.609	1.000	0.260	-0.106	-0.023
Fe	0.151	0.287	0.363	0.260	1.000	-0.126	-0.037
NAO	-0.146	-0.069	-0.121	-0.106	-0.126	1.000	0.314
UI	-0.256	0.098	0.004	-0.023	-0.037	0.314	1.000

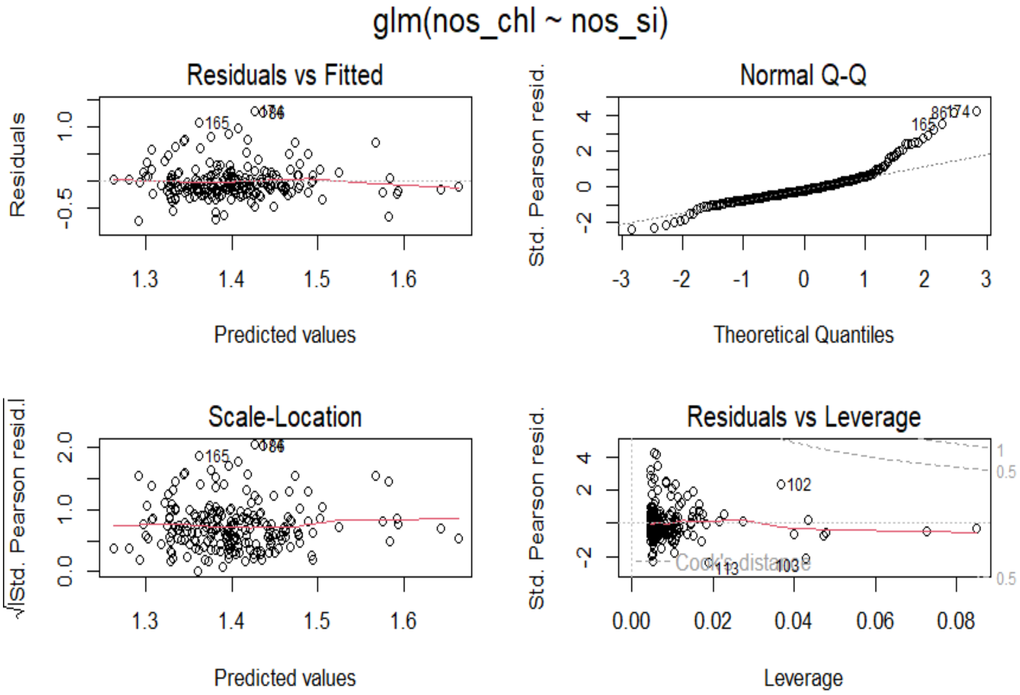

Figure 14: Plot for the GLM using Gamma(log) for Chl-a ∼ SiO_4^{-4}

mentioned before the family of the models will be Gamma which has three link functions (log, inverse and identity). All of them will be tested in order to compare and decide the best model. These models were built in Rcmdr and p-value is set to 0.05.

Log function

In the initial model (Chl-a ∼ NH_4^+ + PO_4^{-3} + SiO_4^{-4} + SST + UI + Fe + NAO) all variables had a p-value greater than 0.05. Because of that, the variable that had the greatest p-value will be removed. This process will continue until all variables have a p-value less than or equal to 0.05. In this case, the order with the correspondent p-values was SST (0.8918), NAO (0.6700), NH_4^+ (0.668), PO_4^{-3} (0.6651), Fe (0.22031), UI (0.13943). The final model was Chl-a ∼ SiO_4^{-4} in which Si had a p-value of 0.00563.

In this case QQ plot fails normality (Fig. 14).

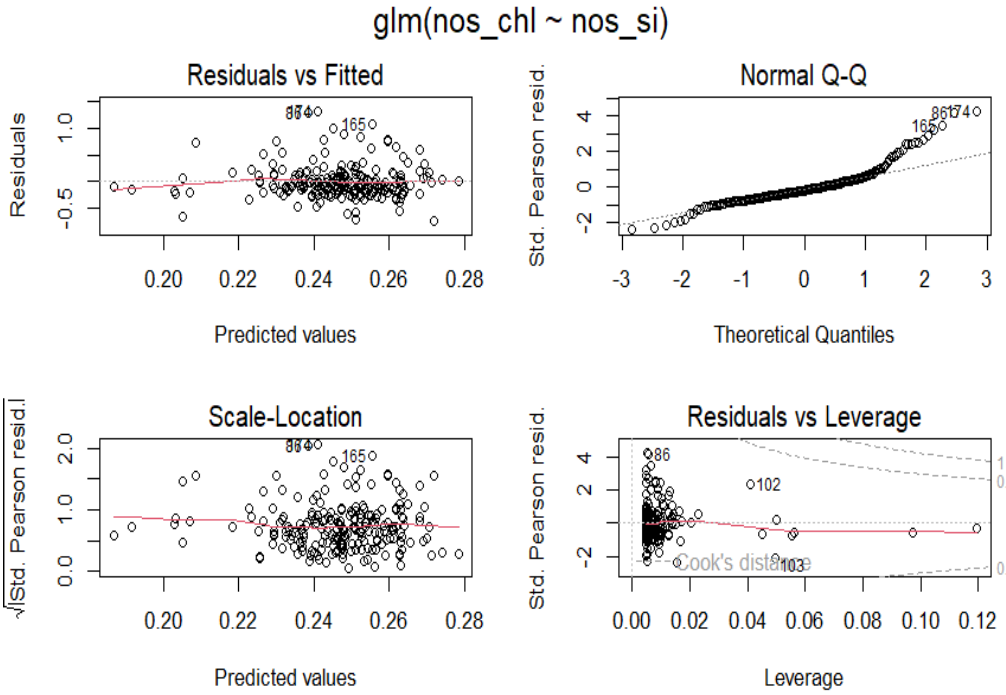

Figure 15: Plot for the GLM using Gamma(inverse) for Chl-a ~ SiO_4^{-4}

Inverse function

The p-values of the initial model with inverse link function were all again higher than 0.05. The order of elimination was SST (0.9057), NH_4^+ (0.643), PO_4^{-3} (0.648), NAO (0.57649), Fe (0.29063), UI (0.12820). The final model was Chl-a ~ SiO_4^{-4}. SiO_4^{-4} had a p-value of 0.00455.

Once again, QQ plot fails normality (Fig. 15).

Identity function

The p-values for all variables were also higher than 0.05. The variables were eliminated by the following order: SST (0.8801), NAO (0.7681), NH_4^+ (0.7255), Fe (0.1598), SiO_4^{-4} (0.00222), UI (0.15350). The final model was Chl-a ~ SiO_4^{-4}, with a p-value of 0.00631.

The normality still fails (Fig. 16).

The final model was the same in all the cases. There's no clear conclusion so the GAM will be analysed.

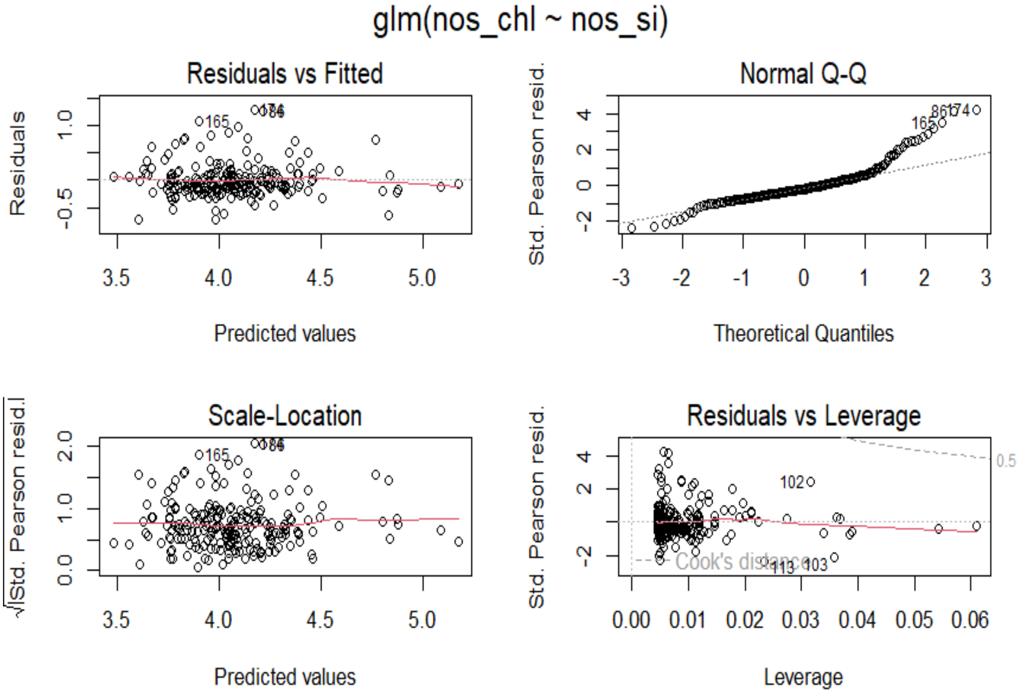

Figure 16: Plot for the GLM using Gamma(identity) for Chl-a ~ SiO_4^{-4}

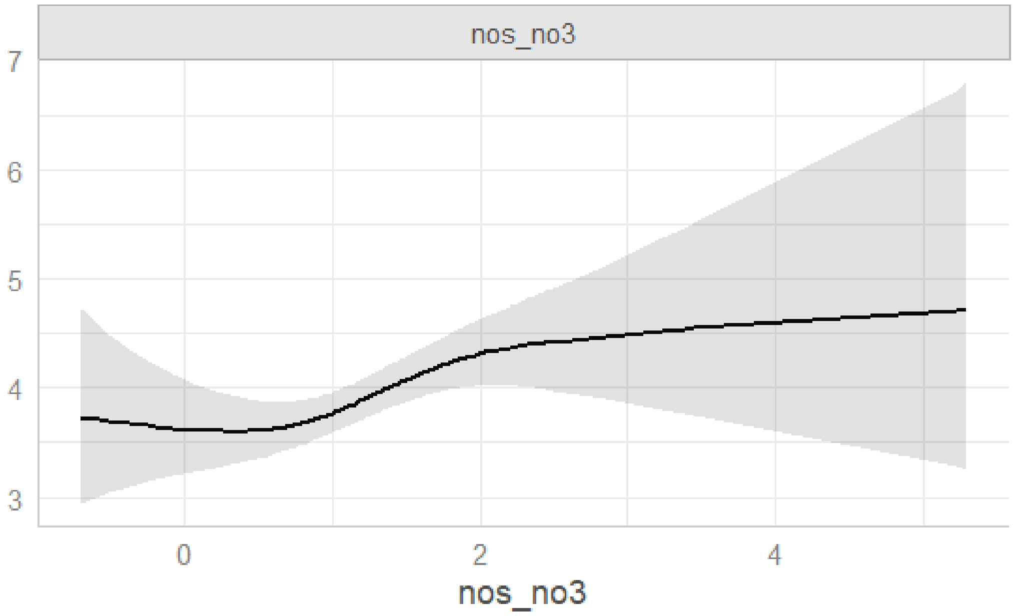

Figure 17: Plot for the GAM using gamma(log) for Chl-a $\sim NO_3^-$

5.1.3 Generalized Additive Models

Just like in the GLM, GAM will be analysed with the three link functions, log, inverse and identity. In all the models, independent variables have a smoothing spline fit in the GAM definition. P-value is set to 0.05.

Log function

The first model was Chl-a $\sim NH_4^+ + NO_3^- + PO_4^{-3} + SiO_4^{-4}$ + SST + UI + Fe + NAO. The method used to eliminate the variables was the same as before. The first variable removed was SST (0.712), followed by NAO (0.456), PO_4^{-3} (0.489), SiO_4^{-4} (0.286), Fe (0.354), NH_4^+ (0.266), UI (0.206). The final model was Chl-a $\sim NO_3^-$, in which NO_3^- had a p-value of 0.002.

NO_3^- doesn't seem to have a linear relation (Fig. 17).

Inverse function

The steps were the same as in the previous case and the final model is also the same: Chl-a $\sim NO_3^-$. This time, NO_3^- had a p-value of $< 2e - 16$.

NO_3^- also doesn't seem to have a linear relation in this case (Fig. 18).

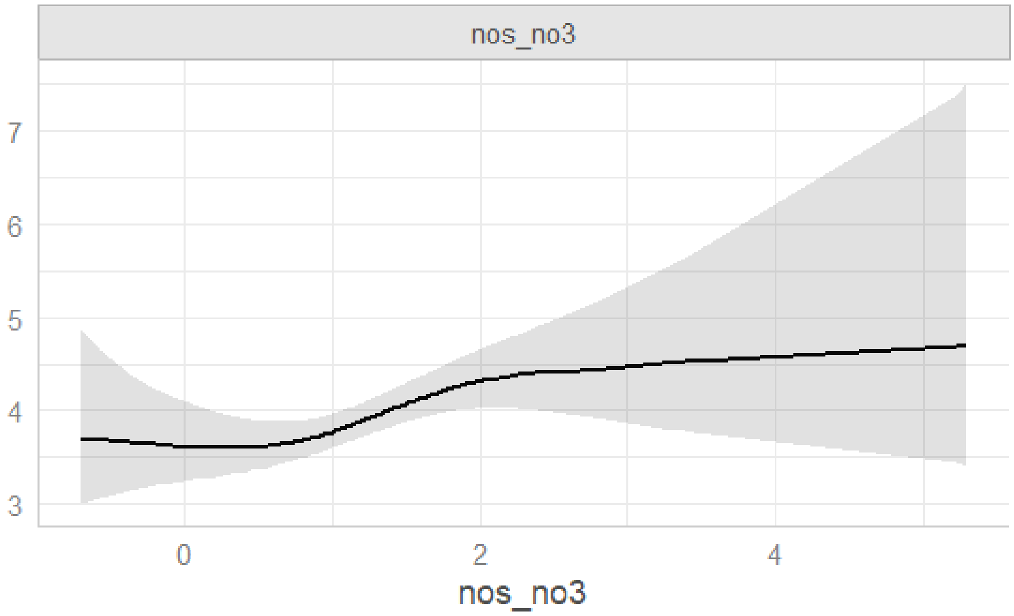

Figure 18: Plot for the GAM using gamma(inverse) for Chl-a $\sim NO_3^-$

Identity function

The steps and results are the same as the ones in log and inverse function (Fig. 19).

6 Stage 4

After analysing different models (e.g. linear models (LM), generalized linear models (GLM) and generalized additive models (GAM)) through different link functions (log, inverse and identity), this section aims to choose the model that is best to analyze the relationship between Chl-a and the independent variables: SST, nutrients (NO_3^-, NH_4^+, PO_4^{-3}, SiO_4^{-4} and Fe), UI and NAO. In order to choose the best model, the Akaike Information Criterion (AIC) will be calculated. This criterion estimates the information lost by the model, thus the less information it loses (corresponding to a smaller number), the best the model is. It is important to refer that the LM with the log transformation of Chl-a will not be considered for the AIC since the this transformation can best model selection. For this reason, the AIC selection will be between the LM (without the log transformation), GLM's and GAM's performed in during the third stage of work.

NO_3^-, SiO_4^{-4} and SST seem to be the best variables to model Chl-a.

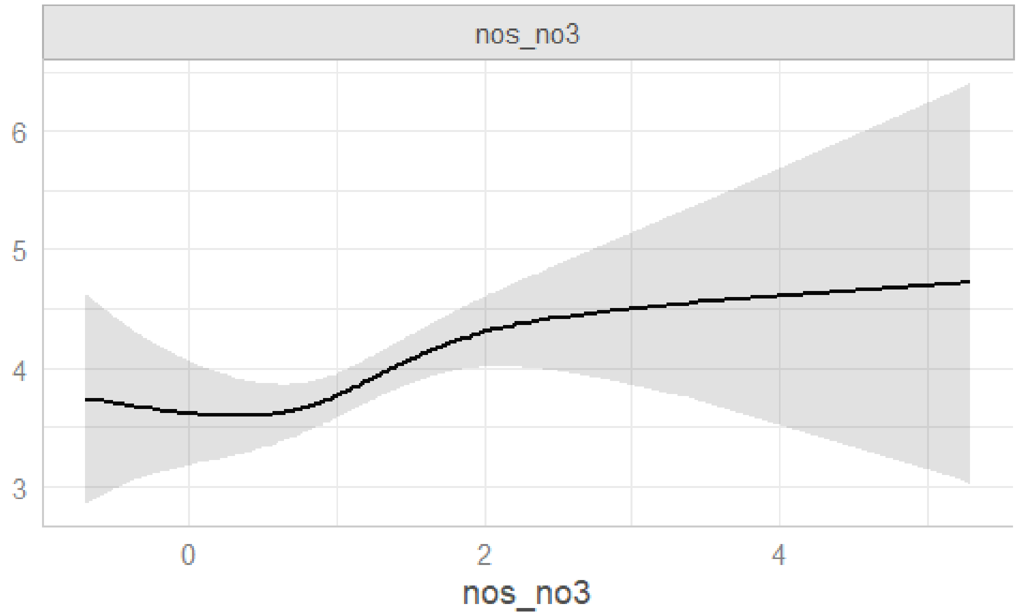

Figure 19: Plot for the GAM using gamma(identity) for Chl-a ∼ NO_3^-

GAM doesn't seem to be the best because none of the variables, in any link function, have a linear relation. Looking at all QQ-plots, it seems that the ones from GLM's have more outliers than the ones from LM's.

The values from AIC correspond to the Table 12. GLM's and GAM's have, in fact, higher numbers. The best model might be the Multiple Linear Regression.

7 Concluding remarks

In conclusion, to find which variables best model chlorophyll-a, the data sets were downloaded and an exploratory analysis was performed. The analysis confirms that between May and July chlorophyll-a is higher and during the same time the temperature starts to rise and NO_3^- and SiO_4^{-4} are lower. Then the seasonality was estimated and removed from every variable except Fe and NAO because these did not have it. Finally, the models were tested. The final linear model was between chlorophyll-a and SST but normality failed, so chlorophyll-a was transformed to log(chlorophyll-a). Because of this transformation, this model was not used to access the best model. GLM were also tested with the gamma distribution with three link functions (log, inverse, and identity). The variable that was most related to chlorophyll-a

Table 12: AIC for all models of Sagres

	AIC
LM	677.95033
GLM Gamma(log)	681.5892
GLM Gamma(identity)	681.3928
GLM Gamma(inverse)	681.9014
GAM Gamma(log)	678.1453
GAM Gamma(identity)	678.0485
GAM Gamma(inverse)	678.2538

was SiO_4^{-4}. The final models that were tested were GAM, also with Gamma and its link functions and the variable that was most related was NO_3^-. Considering all the previous analysis, the conclusion is that SST, SiO_4^{-4} and NO_3^- are the variables that best model chlorophyll-a.

Acknowledgements

This report would not have been possible without the support of CEAUL neither the teaching given by Prof. Clara Cordeiro (FCT–UAlg and CEAUL), Dra. Sónia Cristina (UAlg–CIMA) and Prof. Soraia Pereira (FCUL and CEAUL).

References

Anderson GB, Bell ML, Peng RD. “Methods to calculate the heat index as an exposure metric in environmental health research.” Environemtental Health Perspectives, 121(10), 1111-1119 (2013).`http://ehp.niehs.nih.gov/1206273/`

Cordeiro, C.: Function stl.fit (2022). `https://github.com/ClaraCordeiro/stl.fit`

Fox, J., Bouchet-Valat, M.: Rcmdr: R Commander. R package version 2.8-0 (2022).

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O’Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., Yasmeen, F.: Forecast: Forecasting functions for time series and linear models. R package version 8.17.0 (2022). `https://pkg.robjhyndman.com/forecast/`

R Core Team. R: A language and environment for statistical computing. R Found. for Statistical Computing, Vienna, Austria (2022). `https://www.R-project.org/`

Wickham, H.: ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York (2016).