# Metadata of the chapter that will be visualized in SpringerLink

| | | |
|---|---|---|
| Book Title | Computational Science and Its Applications – ICCSA 2024 Workshops | |
| Series Title | | |
| Chapter Title | Spatial and Multivariate Statistics in Assessing Water Quality in the North Sea | |
| Copyright Year | 2024 | |
| Copyright HolderName | The Author(s), under exclusive license to Springer Nature Switzerland AG | |
| Author | Family Name | **Ody** |
| | Particle | |
| | Given Name | **Christopher** |
| | Prefix | |
| | Suffix | |
| | Role | |
| | Division | |
| | Organization | Universidade Aberta |
| | Address | Lisbon, Portugal |
| | Email | onleft@gmail.com |
| Author | Family Name | **Ramos** |
| | Particle | |
| | Given Name | **M. Rosário** |
| | Prefix | |
| | Suffix | |
| | Role | |
| | Division | |
| | Organization | Universidade Aberta |
| | Address | Lisbon, Portugal |
| | Division | CEAUL - Centro de Estatística e Aplicações, Faculdade de Ciências |
| | Organization | Universidade de Lisboa |
| | Address | Lisbon, Portugal |
| | Email | mariar.ramos@uab.pt |
| | ORCID | http://orcid.org/0000-0001-9114-0807 |
| Corresponding Author | Family Name | **Carolino** |
| | Particle | |
| | Given Name | **E.** |
| | Prefix | |
| | Suffix | |
| | Role | |
| | Division | H &TRC- Health and Technology Research Center, ESTeSL- Escola Superior de Tecnologia da Saúde |
| | Organization | Instituto Politécnico de Lisboa |
| | Address | Lisbon, Portugal |
| | Email | etcarolino@estesl.ipl.pt |

ORCID http://orcid.org/0000-0003-4165-7052

Abstract

The Southern North Sea region plays a vital role for both the economy and society of the surrounding countries. Analyzing the quality of your water is a critical process that involves an assessment of physical, chemical, and biological parameters, essential to guarantee environmental sustainability and the health of local communities and marine ecosystems. Using Multivariate and Spatial Statistics methods, this study seeks to identify spatial patterns and autocorrelations to assess water quality in that region. The data set used was taken on a scientific cruise carried out in December 2020 aboard the RV Meteor vessel, led by a team of German researchers. The raw data went through pretreatment guided by the Data Quality Control protocol of SeaDataNet, an international oceanography project aimed at making European maritime data available. Spike and gradient tests were performed, in addition to data standardization and imputation through inverse distance weighting interpolation. For a better understanding of the scientific area, the data were aggregated by zones for certain analyses, and were sometimes considered globally. An exploratory spatial data analysis (ESDA) was carried out in order to summarize its main characteristics. A reduction in the dimensionality of the original data was carried out through principal component analysis as an auxiliary tool for spatial analysis. The Spatial autocorrelation is analyzed by calculating global and local Moran's *I* Statistics. The outcomes indicate a significant spatial autocorrelation for all variables considered in the freshwater areas and a notable range flattening of the variables in the open sea areas, which possibly caused the lack of significant spatial autocorrelation in those areas.

Keywords (separated by '-')

Exploratory spatial data analysis - North Sea - Principal components - Spatial correlation - Water quality

# Spatial and Multivariate Statistics in Assessing Water Quality in the North Sea

Christopher Ody[1], M. Rosário Ramos[1,3], and E. Carolino[2(✉)]

[1] Universidade Aberta, Lisbon, Portugal
mariar.ramos@uab.pt
[2] H&TRC- Health and Technology Research Center, ESTeSL- Escola Superior de Tecnologia da Saúde, Instituto Politécnico de Lisboa, Lisbon, Portugal
etcarolino@estesl.ipl.pt
[3] CEAUL - Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal

**Abstract.** The Southern North Sea region plays a vital role for both the economy and society of the surrounding countries. Analyzing the quality of your water is a critical process that involves an assessment of physical, chemical, and biological parameters, essential to guarantee environmental sustainability and the health of local communities and marine ecosystems. Using Multivariate and Spatial Statistics methods, this study seeks to identify spatial patterns and autocorrelations to assess water quality in that region. The data set used was taken on a scientific cruise carried out in December 2020 aboard the RV Meteor vessel, led by a team of German researchers. The raw data went through pretreatment guided by the Data Quality Control protocol of SeaDataNet, an international oceanography project aimed at making European maritime data available. Spike and gradient tests were performed, in addition to data standardization and imputation through inverse distance weighting interpolation. For a better understanding of the scientific area, the data were aggregated by zones for certain analyses, and were sometimes considered globally. An exploratory spatial data analysis (ESDA) was carried out in order to summarize its main characteristics. A reduction in the dimensionality of the original data was carried out through principal component analysis as an auxiliary tool for spatial analysis. The Spatial autocorrelation is analyzed by calculating global and local Moran's $I$ Statistics. The outcomes indicate a significant spatial autocorrelation for all variables considered in the freshwater areas and a notable range flattening of the variables in the open sea areas, which possibly caused the lack of significant spatial autocorrelation in those areas.

AQ1

**Keywords:** Exploratory spatial data analysis · North Sea · Principal components · Spatial correlation · Water quality

## 1 Background and Motivation

The North Sea is a vital region for many countries, including Germany, the Netherlands, Denmark and the United Kingdom. It is a crucial center for trade,


O. Gervasi et al. (Eds.): ICCSA 2024 Workshops, LNCS 14816, pp. 1–17, 2024.
https://doi.org/10.1007/978-3-031-65223-3_12

fishing and energy production, significantly involved in the German economy and the economies of other neighboring North Sea countries. The region's importance has led to extensive scientific exploration, with researchers studying the area's unique ecosystem and the impact of human activities on the environment.

Water analysis is essential as it provides information about the quality of North Sea water, nutrient levels and the possible presence of pollutants. Recent studies show that North Sea water quality has improved over the past few decades, thanks to efforts to reduce pollution and improve wastewater treatment. [4,8] showed that changes in carbon dioxide (CO2) and pH are influenced by circulation patterns, Atlantic inflow, local climate conditions and water mass properties components. They also demonstrated nutrient changes influenced by circulation patterns, and changes in oxygen concentration, especially near the sea floor, influenced by respiration of organic matter, decomposition of organic matter, lack of oxygen supply, and temperature. Studies on salinity and temperature, among other variables, are frequently carried out in the North Sea area. For example, the study carried out on the oceanographic event known as "Great Salinity Anomaly" (GSA). During the 1960s and 1970s, a significant change in salinity patterns was detected on the surface of the North Atlantic Ocean, particularly in the North Sea and Norwegian Sea regions. This change was characterized by a large accumulation of fresh water on the ocean surface, resulting in a sharp decrease in water salinity, as a consequence of the accelerated melting of sea ice and glaciers caused by climate change [2] and more recently [5]. Monitoring seawater quality helps identify and mitigate environmental threats, ensuring the preservation of marine resources, the prevention of environmental disasters and the support of sustainable economic practices, essential for the well-being of coastal communities and conservation of marine ecosystems. Statistical geospatial knowledge of water analysis variables in the North Sea is still motivated by a number of critical reasons, including those mentioned above. In addition to monitoring water quality for public health and the sustainability of marine life, the constant monitoring of these variables assesses the impacts of climate change and human activity in the region, such as pollution and the exploitation of natural resources. They also provide essential data for making informed political and economic decisions aimed at the environmental sustainability and economic development of the region [8,10]. The TRAM project (Tracing origin and distribution of geogenic and anthropogenic dissolved and particulate critical high-technology metals in the southern North Sea), led by Dr. Andrea Koschinsky from Jacobs College in Germany, was a research cruise, the RV Meteor, which took place in 2020 and aimed to study anthropogenic inputs of emerging critical metallic contaminants, from the German rivers Elbe, Weser and Sem to the ocean (TRAM, 2024). On its trip, the work area comprised the estuaries of these rivers, from the freshwater end member to the seawater end member along the salinity gradient, as well as the dispersion of the plume along the prevailing currents, mostly in heading east.

This work aims to use multivariate and geospatial statistical techniques to obtain indicators that allow qualifying the variables obtained from RV Meteor

and looking for patterns of similarity between them in the studied area. To this end, spatial data sets of different variables were analyzed in an attempt to locate specific patterns and characterize the oceanographic profile in the region.

This paper is structured as follows: Sect. 2 provides a description of the material and methods used. In Sect. 3, the results of the exploratory spatial data analysis as well as the application of Principal Components Analysis to data of the North Sea to dimension reduction and spatial correlation analysis with Moran test. Conclusions and some remarks are presented in Sect. 4.

## 2 Material and Methods

### 2.1 Study Area and Data Sources

Data is from TRAM project which aims to study anthropogenic inputs of emerging critical metallic contaminants, such as rare earth elements, Sc, Ga, Ge, Pt, Zr, Ti, Mo and V, from the German rivers Elbe, Weser and Sem to the ocean. The scientific cruise *RV Meteor* collected data during a week of december 2020, based on two autonomous thermosalinography systems (thermosalinograph - TSG), named TSG1 and TSG2, consisting of the use of the TSG model SBE21 together with an SBE38 thermometer. The systems worked independently of each other throughout the cruise, ensuring cross-validation of data and redundancy in collection. Observations were taken every 5 s at TSG1 and 10 s at TSG2. In the data not yet processed from the sensors, averages of the minute-by-minute measurements were taken, discarding values that exceed 2 times the value of the standard deviation. Finally, the group chose the TSG1 dataset to publish the results, with salinity and temperature values being very close for both sensors as demonstrated in the time series in Fig. 1. For specific technical details of the sensors and data quality control steps, see the full data processing report provided by the research team [9], which was used as a reference for pre-processing data in the present study.

Although the previous study [9] only considered salinity and temperature data, the characterization of the area proposed by the present study considered the value of the other variables taken in each observation by the same sensor. The raw TSG1 data was then made available by the team for 11 variables (Table 1), including date and time of the observation and coordinates (latitude and longitude). The variables calculated from the absorptions at wavelengths can be seen in the spectrum in Fig. 2.

The analyzed data set was divided into three zones so that the waters in each zone had the potential to show different profiles in the studied variables. For this, the official German division presented in Fig. 3 (Maritime Borders of the Federal Republic of Germany, 2019) was used. The considered zones were Internal Waters (IW) - referring to fresh waters and nearby slopes, Territorial Sea (TS) - being the German part of the North Sea, which covers around 22km of the coast, and North Sea (NS) - being international waters. It is expected that each zone will present differences in at least some variables, such as salinity and temperature. It is also necessary to mention that the zones are not being considered as areas

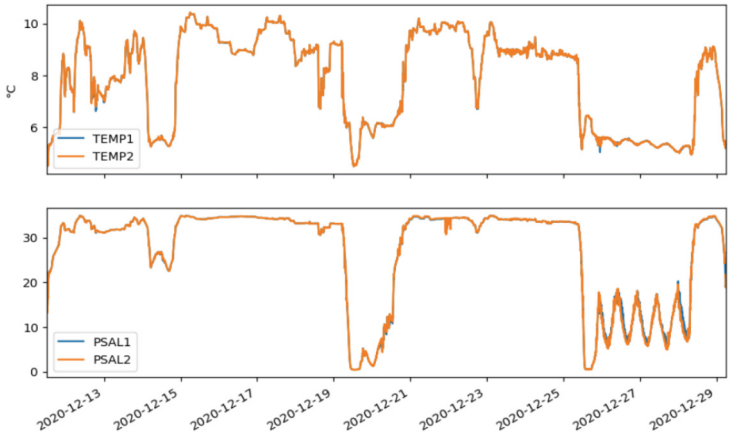


**Fig. 1.** Time series of Temperature (TEMP1, TEMP2) and Salinity (PSAL1, PSAL2) measured by two sensors TSG1 and TSG2 respectively.

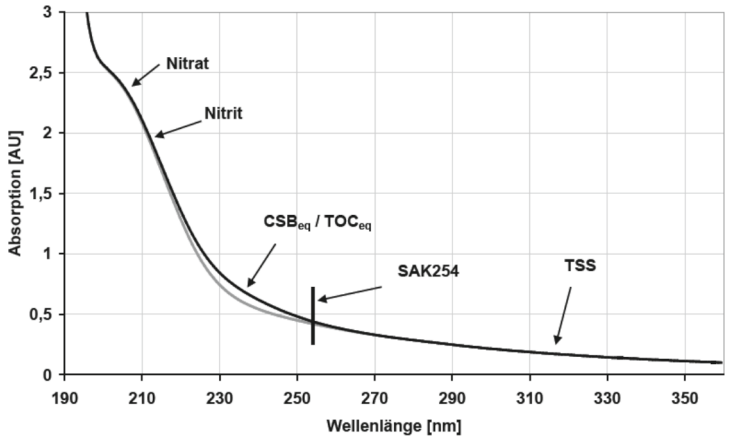


**Fig. 2.** Absorption Regions of the Different Calculated Variables (TriOS GmbH, 2017).
Note: CSBeq - Chemical Oxigen Demand; TOCeq - Total Organic Carbon; SAK254 - Specific Absorption of Carbon at 254 nm; TSS - Total Suspended solids.

(polygons) in spatial statistics tests, but are only being analyzed separately through the points collected within each region. Figure 3 also shows the cruise route during the considered period, represented by the gray dotted line, as well as points in each region colored according to their location.

To make the data suitable for subsequent analysis, pre-processing was carried out following the SeaDataNet Data Quality Control protocol [6], which will be described in the next topic.

**Table 1.** Description of water quality variables under study.

| Variable | Acronym | Measurement unit | Observations |
|---|---|---|---|
| Salinity | | PSU (Practical Salinity Units) | Estimated within the interior of the TSG based on conductivity and interior temperature; external temperature, measured by the SBE38 thermometer |
| Temperature | TempExtern | °C | |
| Nitrate | N.NO3 | mg/L | Concentration of nitrate ions (NO3-) in the water sample. Nitrate analysis is important because high levels of this ion in water can indicate pollution or contamination, which can have harmful effects on aquatic ecosystems and pose risks to human health (World Heath Organization, 2016) |
| Dissolved Organic Carbon | DOCeq | mg/L | Assess the quality of water |
| Absorption at each wavelength (210 nm, 254 nm and 360 nm) | Abs210Abs254,Abs360 | AU (Absorption Units) | |
| Specific Absorption of Carbon at 254nm | SAC254 | AU (Absorption Units) | Useful indicator of organic water quality and can be used to estimate the amount of natural organic matter, including humic and fulvic acids, present in the water sample |

### 2.2 Pre-processing: Data Validation and Imputation

The raw data contained around 17 thousand observations that were obtained over a period of 7 d, and the first step of pre-processing was to evaluate whether there were missing data. As the focus of the expedition was only on salinity and temperature data, it was possible that the other variables could be incomplete. The variables N.NO3, DOCeq and Abs210 were the most compromised, with around 31.27%, 28.93% and 22.85% of missing data. Most of these missing spots in the spectroscopy were found between Glückstadt and Wedel, on the River Elbe, near Hamburg. The points with missing values for N.NO3 reading were removed, thus having around 0.78% of the DOCeq points missing and 10.8% of the Abs210 points.

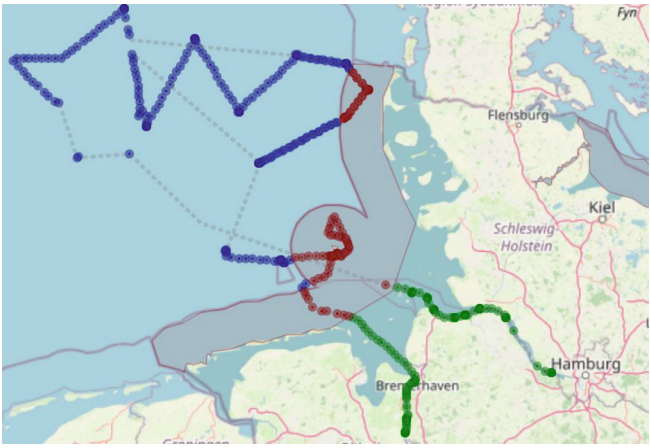


**Fig. 3.** Zones: Internal Waters (Green), Territorial Sea (Red) and North Sea (blue). (Color figure online)

For the remaining incomplete points, spatial interpolation of the data was carried out. The method chosen for spatial interpolation was inverse distance weighting (IDW). In this method, values at unsampled locations are estimated as the weighted average of values at the rest of the locations with weights inversely proportional to the distance between the unsampled and sampled locations. The calculation is carried out through $\hat{Z}(s_0) = \sum_{i=1}^{n} Z(S_i) w_i$, where $\hat{Z}(s_0)$ is the predicted value at $(s_0)$, $n$ is the number of sampled locations (about 7700, depending on the variable), $Z(S_i)$ is the value at location $(S_i)$ [7]. The weights, $w_i$, are given by $w_i = \frac{d_i^{-\beta}}{\sum_{i=1}^{n} d_i^{-\beta}}$, being $d_i$ the distance between locations $s_i$ and $s_0$ and $\beta$ the power of distance $d_i$ which determines the degree to which closer distances are preferred over more distant locations. In this study was considered $\beta = 1$.

The next pre-processing step was analyzing data consistency in relation to sequential measurements, that is, the differences between sequential measurements, where one measurement is quite different from adjacent ones, presenting a peak in magnitude and gradient. This treatment is guided by SeaDataNet's Data Quality Control Procedures [6].To test the magnitude between sequential measurements the Spike test was performed [11].

### 2.3 Exploratory Spatial Data Analysis

In the analysis of spatial processes, the interaction between heterogeneity and spatial dependence presents significant challenges in the specification of spatial models, making the specification process time-consuming and error-prone, potentially leading to inappropriate models. An effective tool for defining such models is Exploratory Spatial Data Analysis (ESDA), which is essential to better understand the data before modeling. This preliminary approach helps ensure

more accurate modeling, following the recommended practice of exploring the data before proceeding with more sophisticated analyzes [1].

### 2.4 Spatial Autocorrelation: Moran's *I*

Spatial autocorrelation gives us an idea of the degree to which a set of features tends to be clustered together or evenly dispersed over the Earth's surface. Global Moran's $I$ is evaluated by measuring the covariance between attributes at each place and near sites toward the overall mean. If both neighboring values are above or below the mean (similar high-high - HH - or low-low - LL - values), the product is positive, reflecting the presence of a similar spatial autocorrelation. Otherwise, the product of the two mean deviations will be negative (unrelated high-low - HL - and low-high - LH - values), indicating a non-positive situation. In a certain way, this index is the modified Pearson coefficient for a single variable. Global Moran $I$ is given by $I = \frac{N}{W} \times \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$,where $N$ is the number of spatial observations, $x_i$ and $x_j$ are the values of variable in locations $i$ and $j$, $\bar{x}$ is the mean, $w_{ij}$ is the $ij$-th element of the spatial weight matrix and $W$ is the sum of spatial weights. To calculate $W$ the inverse distance of nearest neighbors method was considered, $w_{ij}(k) = \begin{cases} 0, & if \quad i = j \\ \frac{1}{d_{ij}}, & if\ d_{ij} \leq d_i \\ 0, & if\ d_{ij} > d_i \end{cases}$, where $d_{ij}$ is the distance between the point and the nearest neighbor according to a cutoff distance $d_i(k)$[7][12]. The distance $d_{ij}$ was determined based on the great circle distance, which represents the minimum distance between two points on a path on the spherical surface [11]. This distance is calculated by $d_{ij} = Rcos^{-1}\left[sen\theta_i sen\theta_j + cos\theta_i \theta_j cos\left(k_i - k_j\right)\right]$. Where $R$ is the radius of the Earth around the Equator (6378Km); $\theta$ and $k$ are the latitude and longitude, respectively. To test the significance of Global Moran's $I$ under the null hypothesis of no spatial autocorrelation, the statistic is given by $z = \frac{I - E(I)}{Var(I)^{1/2}}$, where $E[I] = \frac{-1}{n-1}$, $Var[I] = \frac{n^2(n-1)S_1 - n(n-1)S_2 - 2S_0^2}{(n+1)(n-1)^2 S_0^2}$, $S_0 = \sum_{i \neq j} w_{ij}$, $S_1 = \frac{1}{2}\sum_{i \neq j}(w_{ij} + w_{ji})^2$ and $S_2 = \sum_k \left(\sum_j w_{kj} + \sum_i w_{ik}\right)^2$. This statistic follows asymptotically a standard normal distribution for a sufficiently large number of regions. Since there are only three regions in this study, an alternative approach to judge significance was used, namely Monte Carlo randomization.

The local Moran $I$ is calculated when there is interest in evaluating the local similarity between the value of each area (region) and that of neighboring areas. Local Indicators of Spatial Association (LISA), such as Local Moran's $I$ is a decomposition of Global Moran $I$ into a set of local statistics, given by $I_i = \frac{n(x_i - \bar{x})}{\sum_j (x_j - \bar{x})^2} \sum_j w_{ij}\left(x_i - \bar{x}\right)$.

### 2.5 Principal Components Analysis (PCA)

The central idea of PCA is dimensionality reduction for a data set that consists of a considerable number $p$ of correlated variables, while seeking to retain as

much variability in data as possible. The original set is transformed into a set of new $p$ variables, the Principal Components (PC), after obtaining the eigenvalues and eigenvectors of the covariance matrix [3]. The $i$-th PC is the weighted linear combination of the original variables associated with the $i$-th eigenvector of the covariance matrix, and is given by $Y_i = a_{i1}X_1 + a_{i2}X_2 + ... + a_{ip}X_p$, where $a_{ij}$ are the elements of the $i$-th eigenvector, also called loadings, which describe how much each variable contributes to a particular principal component. Large loadings (positive or negative) indicate that a particular variable strongly relates to a particular principal component. The sign of a loading indicates whether a variable and a principal component are positively or negatively correlated. The coefficients $a_{ij}, i, j = 1, ...p$ are chosen to satisfy three requirements: **i)** variance of component 1 ($PC_1$) is large as possible, e.g. $Var\,[PC_1] \geq Var\,[PC_2] \geq ... \geq Var\,[PC_p]$; **ii)** the principal components are uncorrelated; **iii)** for any principal component $a_{i1}^2 + a_{i2}^2 + ... + a_{ip}^2 = 1$.

In the present study, considering the high number of variables and different measurement scales, PCA was applied on the correlation matrix after standardization of data. The components were extracted forcing independence using the Varimax rotation with Kaiser normalization to ensure orthogonality. To evaluate the quality of the PCA, the Kaiser-Meyer-Olkin statistic (KMO), Bartlett test for sphericity (uncorrelated variables) and square of the cosine ($\cos^2$) of the variables were used:

- $KMO = \frac{\sum\sum_{j\neq k} r_{jk}^2}{\sum\sum_{j\neq k} r_{jk}^2 + \sum\sum_{j\neq k} p_{jk}^2}$, where $r_{jk}$ is the simple correlation coefficient between variables $X_j$ and $X_k$ and $p_{jk}$ is the partial correlation coefficient between $X_j$ and $X_k$, given the other $X_s$ variables. KMO varies between 0 and 1, and higher values indicate a better PCA fit to the data.

- Bartlett statistic $\chi^2 = -\left[(n-1) - \frac{2p+5}{6}\right] ln\,|R|$, that follows a $\chi^2$ distribution, with $\frac{p(p-1)}{2}$ degrees of freedom, where $n$ is the sample dimension, $p$ is the number of variables and $|R|$ is the correlation's matrix determinant.

- $cos_{ij}^2 = \frac{a_{ij}^2}{\sum_{k=1}^{p} a_{jk}^2}$, where $a$ represents the observation's score, $i$ represents the component, $j$ represents the variable and $p$ is the number of components [3]. $cos_{ij}^2$, varies in $[0; 1]$ and higher values indicate that component $i$ retains a high proportion of the variance of the variable $j$.

Keiser's criterion was used to decide how many PC to retain, that is the components for which the eigenvalue is above 1.

# 3 Results

## 3.1 ESDA

Preliminary analysis through histograms for all variables after standardization made it possible to obtain a representation of distributions for comparison purposes. A similar pattern is observed between the TS and NS zones, with the greatest variability occurring the IW zone (Fig. 4). Once the presence of outliers was detected in all variables, Box-Cox transformations were performed to normalize the data for the remaining analyses.

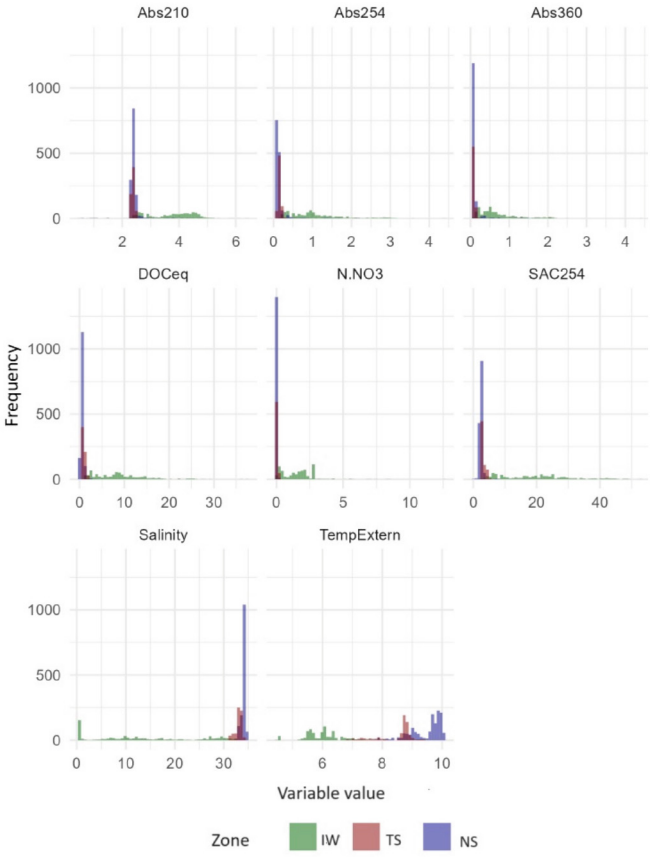


**Fig. 4.** Histograms for all variables by zone: IW-Internal waters, TS-Territorial Sea, NS-North Sea

In the analysis by zones, we noticed that temperature and salinity have a moderate correlation in IW, but significantly strong for the other two regions. This may occur because both vary differently as one moves towards the North

Sea, or they may also have different variation in waters further away from the sea, entering to the mainland. They still present a negative correlation varying from moderate to strong intensity for all other variables. N.NO3 and Abs210 have a correlation of similar magnitude with the other variables, and moderate between them. The absorption variables showed a strongly positive correlation between them, which may be influenced by the turbidity of the water (TriOS GmbH, 2017) (Fig. 5).

The NS region showed weak and/or non-statistically significant correlations for some pairs of variables. This result can be explained by the compression of the data in that area. N.NO3, DOCeq and SAC are considerably compressed and in their small variation no linear correlation between them is observed. Still in the NS, Salinity and TempExtern present a strong positive correlation for this area, demonstrating a linear correlation significantly stronger than in the IW area. The global correlogram indicated that, when considering the complete set of data, the variables present a strong correlation with each other (Fig. 5).

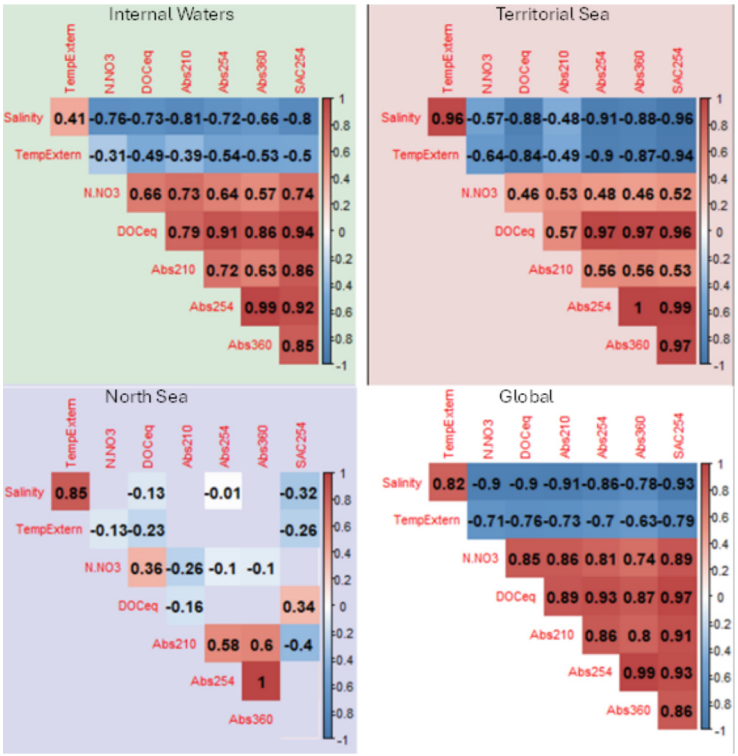


**Fig. 5.** Correlogram by Zone and Global.

From this initial analysis of the global data, we can then conclude that we mainly have two groups of variables that strongly positively correlate with each

other: the group with Salinity and External Temperature (TempExtern) and the group with all other variables. We also see that between these two groups there is a strongly negative correlation.

After analyzing the correlograms, it was observed that the variables had a high correlation structure that PCA can explore.

### 3.2 PCA: By Region and Globally

In the IW zone, the loadings demonstrate that no variable or group of variables is mostly loaded in PC1, although it contributes around 74.5% of the total variability of the original data. In PC2, the loading of the TempExtern ( 0.7) stands out, followed by N.NO3 ( 0.45). The first two PCs together represent 85% of the variability of the original variables. In the TS zone, loadings on PC1 demonstrate similar behavior to IW. PC2 represent more effectively the variables N.NO3 ( 0.7) and Abs210 ( 0.6). The first two components preserve 90% of the variability of the original variables in that zone. In NS zone, the components extracted reflects also the correlogram analysis. It was seen that most of the variables did not present a significantly strong linear correlation and it is known that this characteristic reduces the effectiveness of PCA. In this case, the % of the variance explained is now more distributed between the two chosen components and the total variability explained was around 60%. (Fig. 6).

| | Internal Water | | Territorial Sea | | North Sea | |
|---|---|---|---|---|---|---|
| | **PC1** | **PC2** | **PC1** | **PC2** | **PC1** | **PC2** |
| *Salinity* | -0.3526 | -0.2834 | -0.3772 | 0.11432 | 0.20344 | -0.4937 |
| *TempExtern* | -0.2371 | 0.71746 | -0.3759 | 0.01752 | 0.30041 | -0.4558 |
| *N.NO3* | 0.32304 | 0.44178 | 0.24799 | 0.70834 | -0.1808 | 0.13156 |
| *DOCeq* | 0.38411 | -0.0428 | 0.37805 | -0.1823 | -0.0476 | 0.40351 |
| *Abs210* | 0.35595 | 0.30206 | 0.25221 | 0.5962 | 0.48835 | 0.01666 |
| *Abs254* | 0.38692 | -0.1942 | 0.38763 | -0.1816 | 0.53044 | 0.31932 |
| *Abs360* | 0.36518 | -0.2789 | 0.38144 | -0.1862 | 0.54275 | 0.29665 |
| *SAC254* | 0.39722 | 0.0338 | 0.39066 | -0.1689 | -0.1378 | 0.42204 |

| | Internal Water | | Territorial Sea | | North Sea | |
|---|---|---|---|---|---|---|
| | **PC1** | **PC2** | **PC1** | **PC2** | **PC1** | **PC2** |
| Variance | 5.964 | 0.843 | 6.353 | 0.852 | 2.613 | 2.17 |
| % of Variance | 74.545 | 10.542 | 79.408 | 10.644 | 32.664 | 27.119 |
| Cumulative % of Variance | 74.545 | **85.087** | 79.408 | **90.052** | 32.664 | **59.784** |

**Fig. 6.** Heatmap - PCA Loadings and explained variance by Zone.

Regarding to the global, PC1 presents the same separation of variables found in the correlogram. This strong correlation also justifies the large portion of the variability explained in the first PC, 86.4%. PC2 was characterized by the variables TempExtern, Abs254 and Abs360, but this represented only 5.9% of the percentage of the total variance (Fig. 7).

| | Global | |
|---|---|---|
| | PC1 | PC2 |
| *Salinity* | -0.3631 | 0.2442 |
| *TempExtern* | -0.3122 | 0.65956 |
| *N.NO3* | 0.34589 | -0.1737 |
| *DOCeq* | 0.36757 | 0.09276 |
| *Abs210* | 0.35708 | -0.0467 |
| *Abs254* | 0.36301 | 0.39596 |
| *Abs360* | 0.34225 | 0.55452 |
| *SAC254* | 0.37355 | 0.01002 |

| | Global | |
|---|---|---|
| | PC1 | PC2 |
| Variance | 6.912 | 0.47 |
| % of Variance | 86.402 | 5.876 |
| Cumulative % of Variance | 86.402 | **92.278** |

**Fig. 7.** Heatmap - PCA Loadings and explained variance for global region.

In Fig. 8, the biplot of IW and TS show the negative correlations of Salinity and TempExtern with the other variables. In both graphs (IW and TS), is also observed that most of the variables were close to the PC1 axis. The $cos^2$ values indicated by the colors of the vectors illustrates the adequacy of only 2 components. The NS zone biplot emphasized that the 2 components do not correctly represent the variability of N.NO3, DOCeq and SAC254 data. The magnitude of their vectors was smaller and consequently the $cos^2$ values were low. The biplot of the Global clearly demonstrated the strong linear correlations between the variables. The vectors presented are located around the PC1 axis, confirming their efficiency in representing the variability of the original data, with little loss of this information ( 86%). The $cos^2$ values support this result, having values around 0.9 for most variables.

### 3.3 Global and Local Moran's $I-$statistic

To define the weight matrix, the arrangement chosen established a temporal spacing of 60 min and $k = 5$ neighbors, as shown in Fig. . Such an arrangement with 49 points allowed broad coverage of the navigated area, avoiding dense agglomerations of points in a single region, at the same time that connections are made without leaving 'islands'. Due to the number of points, the computational load of calculating the weight matrix was not extremely onerous in terms of time and machine consumption. Furthermore, consider a large number of close neighbors can bias the results towards more uniform predictions due to the potential dilution effect that may occur between the many points, reducing local sensitivity. The arrangement obtained is represented in Fig. 9.

Considering the weight matrix, the Moran index was calculated and results are presented in Table 2. All $I$ values represented positive spatial autocorrelation. Taking into account that the p-values were lower than 0.05, the hypothesis of negative spatial autocorrelation or absence of autocorrelation is rejected, and therefore it is concluded that the data provide evidence for the existence of positive spatial autocorrelation.

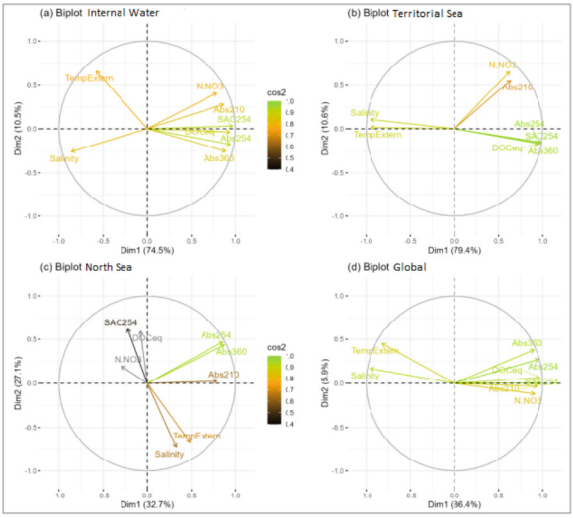


**Fig. 8.** Biplot by zone and global with $\cos^2$.

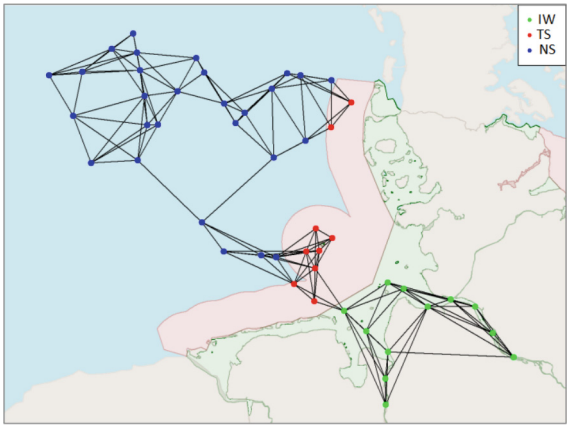


**Fig. 9.** Arrangement to calculate the Weight Matrix, $W$.

The variable with the highest value was TempExtern which, according to histograms of the variable distributions by zone (Fig. 4), indicated a gradual increase towards the NS zone. The second highest value was Salinity which, had an increase in the same direction. However, the variability in this variable is large for the TS and NS zones, remaining reasonably constant. Therefore, the variation of these two variables is constant and increasing towards the MN zone. The reduced number of extreme values observed for these two variables is also a potential factor that contributes to a more prominent spatial autocorrelation. For

**Table 2.** Moran's $I$ per variable and for PC1 and PC2.

| | Moran's $I$ | z-score | p-value |
|---|---|---|---|
| Salinity | 0.65463 | 7.2445 | 2.17E-13 |
| TempExtern | 0.80615 | 8.5925 | 4.25E-18 |
| N.NO3 | 0.50486 | 5.7203 | 5.32E-09 |
| DOCeq | 0.39864 | 4.9059 | 4.65E-07 |
| Abs210 | 0.36964 | 4.1697 | 1.52E-05 |
| Abs254 | 0.45485 | 5.3906 | 3.51E-08 |
| Abs360 | 0.42591 | 4.9868 | 3.07E-07 |
| SAC254 | 0.41866 | 5.0087 | 2.74E-07 |
| **PC1** | 0.55069 | 6.3064 | 1.43E-10 |
| **PC2** | 0.57698 | 6.4722 | 4.83E-11 |

the remaining variables from the original data set, Moran's $I$ reveals moderate values (Fig. 10).

Taking PC1 and PC2, a moderate value of Global Moran's $I$ for both was observed (Table 2). As these variables represent the variability of the data set as a whole, it is interesting to see that their values were intermediate to the values of the other variables. The data set, as a whole, presents a significant spatial correlation, especially considering the results of PC1 due to its high loading of originals data variability. Thus, the use of PCs in this analysis helps in the interpretation of the joint result of the spatial autocorrelation of the variables studied, indicating that, when considered together, they are not only strongly correlated (see Fig. 5), but they also present positive spatial autocorrelation.

Next, the local Moran $I$ (LISA) values were analyzed. Remember that the objective of LISA is to indicate statistically significant spatial clusters for each variable. From the analysis of the results presented in Fig. , there is an absence of significant spatial autocorrelation in the TS and NS zones for most variables. It was also found that observations from the IW zone showed positive spatial autocorrelation for most variables. This profile was reflected in PC1, a component that explained most of the variability in the original data in the PCA test.

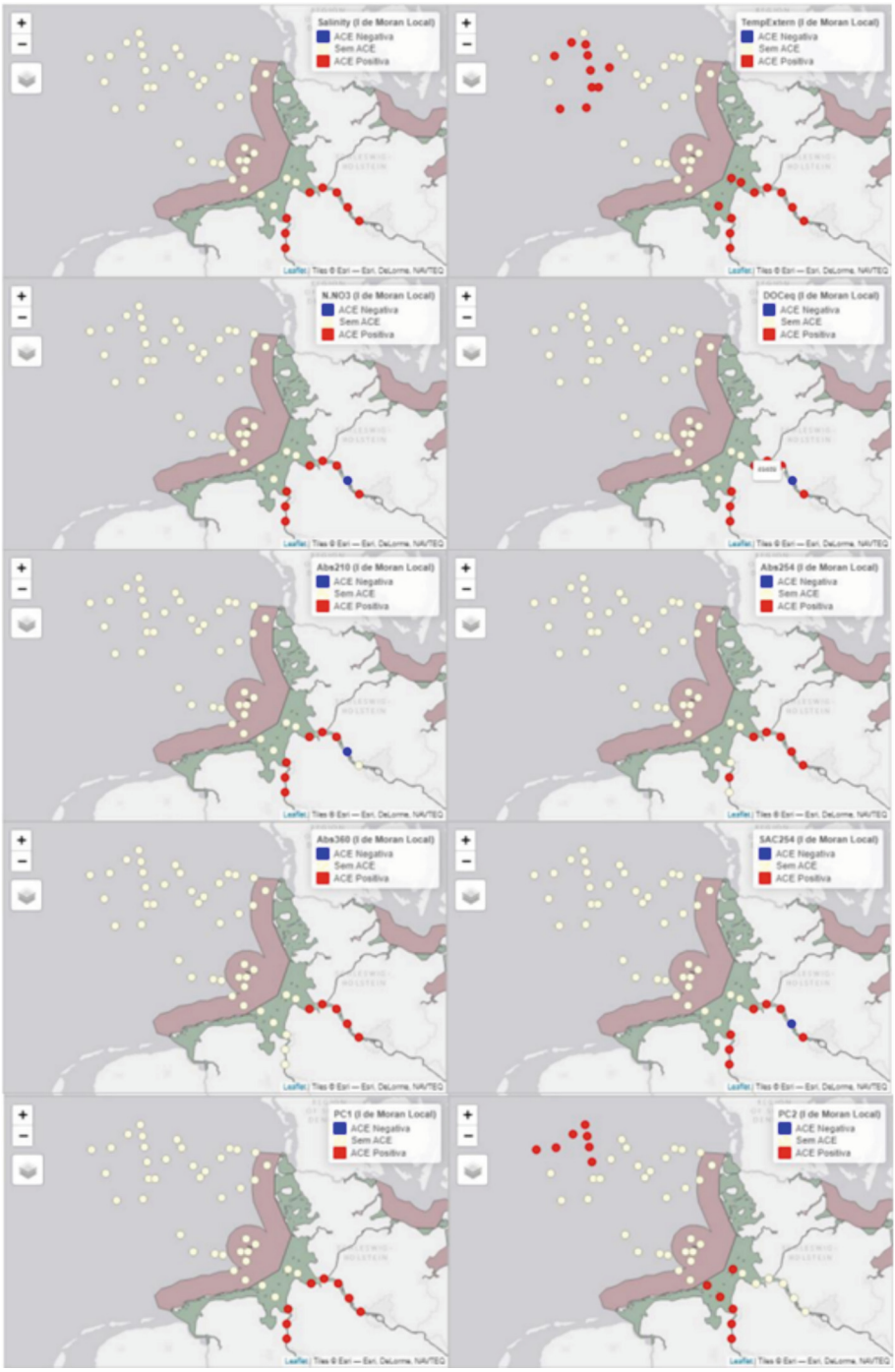


**Fig. 10.** Local Moran's $I$ for all variables and principal components PC1 and PC2. Blue dots represent Negative spatial autocorrelation, white dots are no spatial autocorrelation and red dots are positive spatial autocorrelation. (Color figure online)

## 4 Comments and Concluding Remarks

Exploratory Spatial Data Analysis made by zone revealed that most of the variables were flattened into a small range in the Territorial Sea and North Sea (south) zones. Salinity and External temperature differed from the other variables but were similar to each other. The Internal Waters zone shown the greatest variance for all the variables. ESDA also indicated a possible spatial autocorrelation from the Internal Waters in direction to North Sea. As such, spatial autocorrelations under global Moran's $I$ indicate that they are mostly located in the Internal Waters. LL clusters were found for Salinity and External temperature in the Internal Waters region while HH clusters were found in the same zone for the other original variables. PCA resume data into two principal components. PC1 point to clusters similar to the original variables as a whole and PC2 similar to External Temperature. The same point was determined as a LH spatial outlier for 4 variables, indicating that there may be a variation of interest in that zone, as there is a drop in the value of some of the variables in relation to their near neighbors.

This study has some limitations, such as the data collection method. Data was collected during the voyage along the cruise route. Thus, there aren't simultaneous observations at the different points, which potentially influenced the results, due to tides and/or other local maritime and meteorological dynamics. Another limitation concerns the period of data collection, which is too short to apply more robust spatio-temporal methods that would be convenient for understanding the changes in the variables considered over a more significant period. Seasonal imbalances in the region, such as the aforementioned Great Salinity Anomaly, could take a considerable amount of time to be realized, perhaps years, if not decades.

In the future, it is intended to obtain data from the same region for a longer period and different seasons which will allow to confirm, or not, the spatial autocorrelation found, identify seasonal effects, and ultimately implement spatial or even spatio-temporal modelling.

**Acknowledgments.** This work is partially financed by national funds through FCT-Fundaąão para a Ciência e a Tecnologia under the projects:
UIDP/05608/2020. DOI 10.54499/UIDP/05608/2020
(https://doi.org/10.54499/UIDP/05608/2020),
UIDB/05608/2020. DOI 10.54499/UIDB/05608/2020
(https://doi.org/10.54499/UIDB/05608/2020) and
UIDB/00006/2020. DOI: 10.54499/UIDB/00006/2020
(https://doi.org/10.54499/UIDB/00006/2020).

The authors thank the three referees for their constructive comments and valuable suggestions, which led to improvements to this work and will contribute to future work. The authors thank to Dr. Andrea Koschin from Jacobs College in Germany, coordinator of the TRAM project who provided the data.

## References


1. Anselin, L.: Spatial Econometrics: Methods and Models. Studies in Operational Regional Science. Springer Netherlands (1988). https://books.google.pt/books?id=3dPIXClv4YYC
2. Belkin, I.M., Levitus, S., Antonov, J., Malmberg, S.A.: "Great salinity anomalies" in the north atlantic. Prog. Oceanogr. **41**(1), 1–68 (1998)
3. Hair, J., Babin, B., Black, W., Anderson, R.: Multivariate data analysis. Cengage (2019). https://books.google.pt/books?id=0R9ZswEACAAJ
4. Huthnance, J., et al.: Recent change—North Sea. In: Quante, M., Colijn, F. (eds.) North Sea Region Climate Change Assessment. RCS, pp. 85–136. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-39745-0_3
5. Kim, W.M., Yeager, S., Danabasoglu, G.: Revisiting the causal connection between the great salinityanomaly of the 1970s and the shutdown of labrador sea deep convection. J. Climate **34**(2), 675 – 696 (2021). https://doi.org/10.1175/JCLI-D-20-0327.1, https://journals.ametsoc.org/view/journals/clim/34/2/JCLI-D-20-0327.1.xml
6. Management, P.E.I.F.O..M.D.: Seadatanet.data quality control (2019). https://www.seadatanet.org/Standards/Data-Quality-Control, (Accessed 22 Mar 2023)
7. Moraga, P.: Spatial Statistics for Data Science: Theory and Practice with R. Chapman & Hall/CRC Data Science Series (2023)
8. Quante, M., Colijns, F.: North Sea Region Climate Change Assessment. Springer International Publishing (2016)
9. Schlundt, M.: Continuous thermosalinograph oceanography along rv meteor cruise m169-data processing report (2021). https://doi.org/10.1594/PANGAEA.938474
10. Walday, M., Kroglund, T.: The north sea. European Environ. Agency **32** (2008)
11. Wong, A., Keeley, R., Carval, T., Team, A.D.M.: Argo quality control manual for ctd and trajectory data (2024). https://doi.org/10.13155/33951
12. Zhou, X., Lin, H.: Spatial Weights Matrix. Encyclopedia of GIS. Springer, Boston (2008)

# Author Queries

**Chapter 12**

| Query Refs. | Details Required | Author's response |
|---|---|---|
| AQ1 | As per Springer style, both city and country names must be present in the affiliations. Accordingly, we have inserted the city name in the affiliation. Please check and confirm if the inserted city name is correct. If not, please provide us with the correct city name. | correct information |