NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE

PEDRO ALEXANDRE ALDONSO CALDEIRÃO

B.Sc. in Computer Science and Engineering

WIND STRESS COUPLED CLUSTERING-CLASSIFICATION FOR SEA SURFACE TEMPERATURE UPWELLING ANALYSIS

MASTER IN COMPUTER SCIENCE AND ENGINEERING

NOVA University Lisbon
September, 2024

NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

DEPARTMENT OF
COMPUTER SCIENCE

WIND STRESS COUPLED CLUSTERING-CLASSIFICATION FOR SEA SURFACE TEMPERATURE UPWELLING ANALYSIS

PEDRO ALEXANDRE ALDONSO CALDEIRÃO

B.Sc. in Computer Science and Engineering

Adviser: Susana Maria Nascimento
Assistant Professor, NOVA University Lisbon

Co-adviser: Paulo Relvas
Assistant Professor, Campus de Gambelas, Centro de Ciências do Mar (CCMAR), Universidade do Algarve

Examination Committee

Chair: Miguel Jorge Tavares Pessoa Monteiro
Assistant Professor, NOVA University Lisbon

Rapporteur: Pedro Lopes da Silva Mariano
Assistant Researcher, ISCTE

MASTER IN COMPUTER SCIENCE AND ENGINEERING

NOVA University Lisbon
September, 2024

Wind Stress Coupled Clustering-Classification For Sea Surface Temperature Upwelling Analysis

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v7.1.3) [1].

Acknowledgements

Gostaria de agradecer à minha orientadora **Susana Nascimento**, cujo apoio e ajuda foram fundamentais para a realização desta tese. Ao professor Paulo Relvas por ter disponibilizado os dados usados para o estudo feito e por esclarecer conceitos que não são da minha área. Ao professor Boris Mirkin pela ajuda prestada nas reuniões feitas com todo o grupo.

Agradeço à minha mãe **Virgínia**, ao meu pai **Carlos** e ao meu irmão **André** por me motivarem e darem força ao longo dos 5 anos do curso.

Agradeço também aos meus amigos mais antigos **Bea**, **Duarte** e **Gonçalo** por nunca terem desistido de mim, sempre a motivar-me especialmente nos momentos mais baixos enquanto escrevia esta dissertação e realizava o estudo prático.

Agradeço também às pessoas que conheci e convivi no ATL: **Paula**, **Ana**, **Hugo** e ao **Tio Luís**, e muitos outros que me ajudaram sempre tanto no meu percurso escolar como pessoal antes embarcar nestes 5 anos de Engenharia Informática.

Por fim, agradeço a todos os meus amigos que conheci na universidade durantes estes 5 anos e por todos os momentos que passei com eles sendo em aulas, projetos, festas e outros eventos.

Abstract

The scientific community continuously studies oceanographic events, crucial to understanding our planet's natural systems and interconnections. Upwelling is one such event, where deep, cold waters rise to replace warmer waters moved offshore by coastline-parallel blowing winds.

A recently developed framework, the core-shell clustering automatically segments coastal upwelling regions and models Upwelling Stability Period (USP) during the upwelling seasons using Sea Surface Temperature (SST) maps derived from MODIS-Aqua satellite imagery. This allows long-term spatiotemporal analysis of coastal upwelling.

Upwelling is a wind-driven event, so studying the relation between wind and upwelling patterns leads to a better understanding of this phenomenon. However, recent studies show that there is a lack of correlation between wind and upwelling patterns.

The main goal of this dissertation is two-fold: to investigate whether wind patterns constructed from clustered ocean wind data are associated with USP's derived from core-shell clustering; to explore if these USP's can be predicted from wind data.

We explore a clustering-classification approach applied to Wind Stress Anomaly (WSA) data. We first construct WSA maps from wind reanalysis products to then apply an unsupervised clustering algorithm to these maps, extracting features from the resulting clusters to build new WSA datasets. These datasets are then labelled with USP taking advantage of the trapezoidal membership function and popular defuzzification functions. The labeled WSA datasets are used as input to state-of-the-art classifiers, Random Forest (RF), Ordinal Forest (OF) and K-Nearest Neighbors (K-NN), to predict USP's.

The experimental study using data from three representative upwelling seasons of north and south Morocco geographic regions show that the RF classifier proves to be an adequate choice for predicting the USP as it presents the best tradeoff between classification quality and computational fit times. The OF classifier, although having the highest test set classification quality, suffers from an excessively high fit time.

Keywords: unsupervised clustering, ensemble classifiers, defuzzification functions, wind stress anomaly data, coastal upwelling

Resumo

A comunidade científica estuda continuamente os fenómenos oceanográficos, cruciais para compreender os sistemas naturais e as interligações do nosso planeta. O afloramento costeiro é um desses eventos, em que águas profundas e frias sobem para substituir águas mais quentes deslocadas para lá da costa por ventos a soprar paralelos à costa.

Uma *framework* recentemente desenvolvida, o *core-shell clustering* segmenta automaticamente as regiões de afloramento costeiro e modela períodos de estabilidade (PE) durante as épocas de afloramento costeiro, ao usar mapas de SST derivados de imagens de satélite MODIS-Aqua. Isto permite que seja feita uma análise espácio-temporal a longo prazo do afloramento costeiro.

O afloramento costeiro é um fenómeno provocado pelo vento, pelo que, o estudo da relação entre o vento e os padrões de afloramento costeiro leva a uma melhor compreensão deste fenómeno. No entanto, estudos recentes mostram que existe uma falta de correlação entre os padrões de vento e os padrões de afloramento costeiro.

O objetivo princial desta dissertação é duplo: investigar se os padrões de vento construídos a partir de dados de vento oceânico segmentados estão associados a PE derivados do *core-shell clustering*; explorar se estes PE podem ser previstos a partir de dados de vento.

Exploramos uma abordagem de *clustering* e classificação aplicada a dados de WSA. Começamos por construir mapas de WSA a partir de produtos de reanálise do vento para depois aplicar um algoritmo de *clustering* não supervisionado a esses mapas, extraindo *features* desses *clusters* para construir novos conjuntos de dados WSA. Estes conjuntos de dados são depois rotulados com o PE tirando partido de função de pertença trapezoidal e de funções de defuzzificação populares. Os conjuntos de dados WSA rotulados são utilizados como entrada para classificadores topo de gama, RF, OF e K-NN, para prever PE.

O estudo experimental que usou dados de três épocas representativas de afloramento costeiro nas regiões geográficas de norte e sul de Marrocos mostra que os classificadores RF provam ser uma escolha adequada para prever o PE, uma vez que apresentam a melhor relação entre a qualidade da classificação e o tempo de computação. Os classificadores OF, embora tenham a melhor qualidade de classificação do conjunto de teste, têm um tempo

de computação excessivamente elevado.

Palavras-chave: *clustering* não-supervisionado, classificadores *ensemble*, *funções de desfuzzificação*, dados de *wind stress anomalies*, afloramento costeiro

Contents

List of Figures

List of Tables

Acronyms

ARIMA Autoregressive Integrated Moving Average *(p. 13)*
ASCAT Advanced SCATterometer *(pp. 8, 10, 11, 20)*

CART Classification and Regression Trees *(p. 27)*
CCA Canonical Correlation Analysis *(pp. 6, 7)*
CCMP Cross-Calibrated Multi Platform *(pp. 8, 9, 11, 20)*
CCUS Canary current upwelling system *(pp. 2, 3, 8, 82, 89)*
Chl-a Chlorophyll-a *(pp. 1, 6, 8, 10)*
CLARA Clustering large applications algorithm *(pp. 11, 12)*
CMEMS Copernicus Marine Environmental Monitoring Service *(p. 6)*
COADS Comprehensive Ocean-Atmosphere Data Set *(pp. 7, 11)*
COG Center of Gravity *(pp. xii, 43–46, 57–62, 66, 68, 69, 96–98)*

DBSCAN Density-Based Spatial Clustering of Applications with Noise *(pp. 5, 13, 22, 24)*
DIANA Divisive Analysis Clustering Algorithm *(p. 12)*
DPCA Dynamic Principal Component Analysis *(p. 13)*

EA East Atlantic *(p. 8)*
ECPM East Coast of Peninsular Malaysia *(p. 9)*
ECWMF European Centre for Medium-Range Weather Forecasts *(pp. 9, 20)*
EMC East Madagascar Current *(p. 9)*
ENSO El Niño Southern Oscillation *(pp. 9, 10)*
EOF Empirical Orthogonal Function *(p. 10)*

FADDIS Fuzzy Additive Spectral Clustering *(pp. 87, 89)*
FCM Fuzzy C-means *(pp. 5, 9, 12, 14, 27, 86)*
FCPM Fuzzy clustering with proportional membership *(pp. 27, 84–86, 89)*

GAS Generalized Autoregressive Score *(p. 13)*

IAP Iterative Anomalous Pattern *(pp. ix, x, xii, 3, 4, 22, 24, 26, 27, 32, 36–40, 51, 53, 54, 56)*
ICM Interdisciplinary Center for Mathematical and Computational Modelling *(p. 7)*
IPO Interdecadal Pacific Oscillation *(p. 10)*

K-NN K-Nearest Neighbors *(pp. iii, iv, ix, xii, 4, 30, 31, 49, 50, 57, 61–68, 95)*

LOM Largest of Maximum *(pp. 43, 46, 57–60, 68, 96–98)*
LOOCV Leave One Out Cross-Validation *(pp. xii, 31, 49, 50, 60–62)*

MOM Middle of Maxima *(pp. 43, 46, 57–62, 68, 96–98)*
MuSTC Multi-Stage Spatio-Temporal Clustering Method *(p. 11)*

NAO North Atlantic Oscillation *(p. 8)*
NM North-Morocco *(pp. x–xii, 26, 33, 34, 44–46, 51, 52, 55–62, 66, 68, 69, 92–97)*
NOAA National Oceanic and Atmospheric Administration *(pp. 8, 10, 11)*

OF Ordinal Forest *(pp. iii, iv, xii, 4, 29, 30, 49, 57, 59, 60, 62–68)*

PAM Partitioning Around Medoids *(p. 11)*
PUS Pacific Upwelling System *(p. 10)*

QuikSCAT Quick Scatterometer *(pp. 6, 8, 10, 11)*

RF Random Forest *(pp. iii, iv, ix, xii, 4, 27–29, 31, 48, 57–60, 62–69)*

S-STSEC Sequential Self Tunning Seeded Expanding Cluster *(pp. ix, 1, 22–24)*
SEAS South Eastern Arabian Sea *(p. 9)*
SM South-Morocco *(pp. ix–xii, 2, 26, 33–40, 42–46, 51–53, 55–58, 60, 62, 66, 68, 69, 92–95, 97, 98)*
SMI Swedish Meteorological Institute *(p. 7)*
SOM Smallest of Maximum *(pp. 43, 46, 57–59, 61, 68, 96–98)*
SST Sea Surface Temperature *(pp. iii, iv, x, 1, 3, 6–10, 22–25, 27, 34, 35, 38–42, 51–54, 68, 69, 82, 89)*

TPI Topographic Position Index *(p. 10)*

UI Upwelling Index *(pp. 7, 11)*

USP Upwelling Stability Period *(pp. iii, ix, x, xii, 2–4, 22–24, 26, 32, 40–44, 46, 47, 51–57, 63–66, 68, 69)*

WSA Wind Stress Anomaly *(pp. iii, iv, x–xii, 3, 4, 32, 35–40, 42, 44–46, 51–56, 68, 92–95)*

1 Introduction

1.1 The problem and its importance

The scientific community continuously studies oceanographic events, which are crucial to understanding our planet's natural systems and interconnections. Upwelling is one such event, where deep, cold waters rise to replace warmer waters moved offshore by coastline-parallel blowing winds. This enriches water nutrient mixing, leading to large populations of marine animals coming into the affected region. Investigating and interpreting these affected regions is critical to numerous application areas such as fisheries and studies about ocean dynamics. This analysis is aided by regular and continuous global observation of Sea Surface Temperature (SST) through sattelite remote sensing which have provided almost 40 years of data. Figure 1.1 presents the phenomenon dynamics.

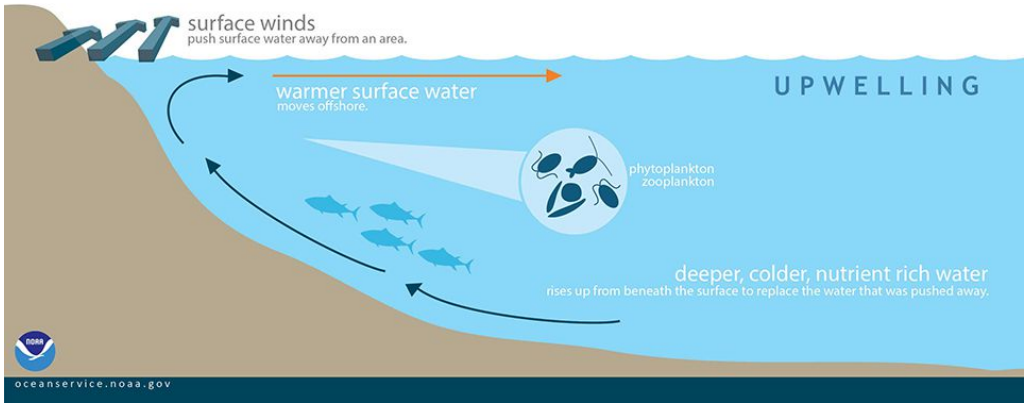


Figure 1.1: Upwelling diagram representing the upwelling phenomenon dynamics. Image taken from [2]

To track and segment the phenomenon several aprroaches have developed such as binary classification methods which focus on the SST of each pixel in a SST image [3] or even fusion methods utilizing SST and also Chlorophyll-a (Chl-a) sea surface images [4].

A novel spatiotemporal framework was developed [5, 6] for the automatic segmentation of upwelling and its spatiotemporal analysis from SST data derived from MODIS-Aqua satellite images. This framework extends the Sequential Self Tunning Seeded Expanding

Cluster (S-STSEC) algorithm, a spatial clustering algorithm proposed in [7] for temporal analysis. The clustering framework unsupervisedly finds temporal intervals where coastal upwelling stays relatively stable during an upwelling season, designating these intervals as Upwelling Stability Period (USP). From each of the USPs defined in an upwelling season, it is built a cluster segmentation which employs the core-shell structure. This cluster is characterized by a constant part, a core, where upwelling is continously happening during that USP. The variable part of this core-shell cluster, the shell, characterizes the variable part of the upwelling region, changing over the period of the defined USP, thus representing the dynamic characteristic of the event. This framework was successfully applied to 16 years of data in the Canary current upwelling system (CCUS), covering three distinct geographic regions: Portuguese coast, North and South Morocco, unsupervisedly defining USP throughout the upwelling seasons.

Although several upwelling detection and monitoring techniques exist, studies on the inter-relation between wind and upwelling patterns are scarce in the literature. As the wind is a primary driver of coastal upwelling, by studying and understanding its patterns and how they relate to the event, we can get a deeper knowledge of the mechanisms that drive it.

Wind is the primary driver of coastal upwelling as the event starts when wind blows parallel to the coast. From this interaction, the water molecules displaced in the sea surface layer are moved offshore and to replace this lost matter deeper and colder waters have to emerge, therefore triggering upwelling.

Figure 1.2 ilustrates the impact of wind on the upwelling phenomenon, showcasing its variation across the year 2007 for the South-Morocco (SM) geographic region from April to October.

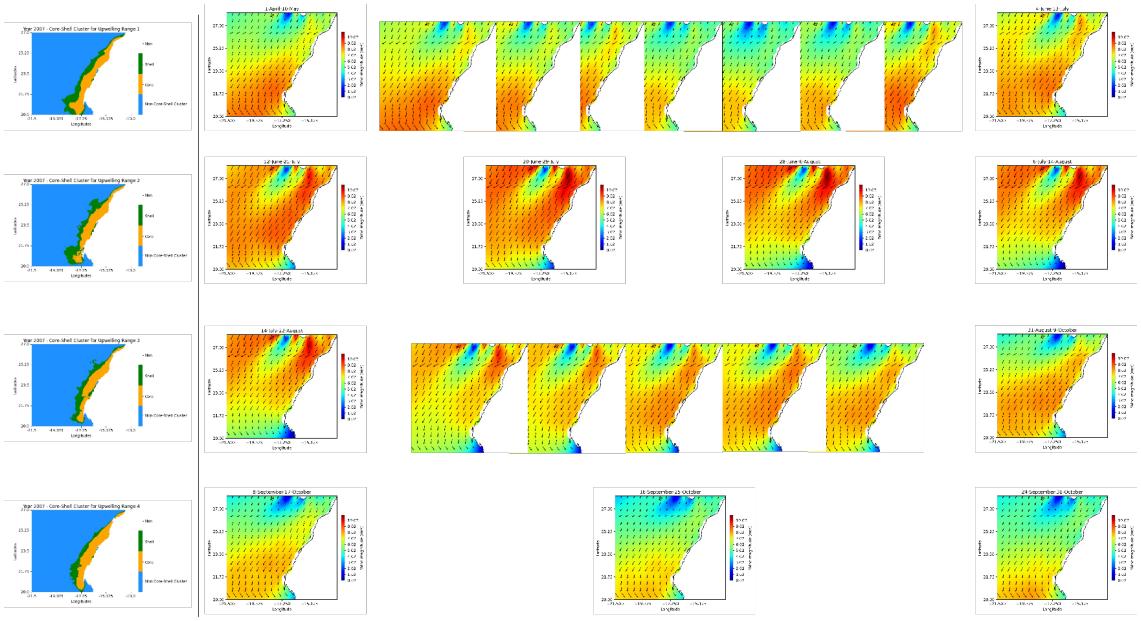


Figure 1.2: Wind map sequences for the year 2007 for the south Morocco geographic region-April to October

On the left, it is present the core-shell upwelling SST segmentation of the four USP defined by the core-shell clustering framework for this upwelling season, where in orange is colored the core of the cluster and in green the shell of the cluster. On the right is present the wind speed map segmentations corresponding to the SST instants that compose each USP. The figure displays an interesting trend, where distinct wind speeds are evident within in each USP, and wind patterns shift when transitioning from one USP to another. We can also see that the intermideate USP correspond to the higher wind speed intensities, characterized by the orange and red colors, as expected. These intermideate USP are inserted during summer season, a period where upwelling is more intense in the area.

The main goal of this dissertation is to analyze the wind fields as the primary force for coastal upwelling across the CCUS in their relation to upwelling patterns using long yearly wind reanalysis products.

The developed work was guided to answer the following questions:

(Q1) Are wind speed and Wind Stress Anomaly (WSA) patterns consistent with the piece-wise constancy of coastal upwelling?

(Q2) Can the USP defined by the core-shell clustering be predicted by WSA data?

To investigate these questions, we will employ a framework which preprocesses the collected wind data on a daily scale. Subsquently, we will segment the daily maps in a unsupervised manner, extracting WSA features from the clusters for the purpose of constructing yearly datasets. Following this, the aforementioned datasets will be labelled according to the USP found by the core-shell clustering framework, serving as ground truth. Finally we will use three classifier models with the purpose of predicting the USP of a daily wind map. Each daily wind map USP is going to be modeled according to a trapezoidal fuzzy membership function and defuzzified with popular defuzzfication functions.

The study will be conducted in the CCUS, in the North (28°N-36°N;5.5°W-16°W) and South (20°N-27°N;13°W-21°W) Morocco geographic regions, ignoring the strip between 27°N and 30°N due to the presence of the canary islands, known to induce significant anomalies in upwelling patterns [5].

1.2 Objectives and main contributions

The main objective of this dissertation is to explore if WSA segmented from ocean wind data relate to the USP modeled by the core-shell clustering framework and if a USP can be predicted by wind data.

As there is a lack of work analyzing wind fields in upwelling using prediction methods, we propose to unsupervisedly segment wind data with the Iterative Anomalous Pattern (IAP) algorithm. From the obtained WSA cluster, we will extract features to build WSA datasets and label them according to their USP modeled by a trapezoid fuzzy membership

function and defuzzifying them with popular defuzzifcation functions. These constructed datasets will serve as input to the chosen classifier models to then answer if an USP can be predicted with wind data.

The main contributions of this dissertation are:

1. Construction of WSA maps from wind reanalysis products;
2. Unsupervised clustering of WSA maps with an algorithm that automatically fine-tunes the number of clusters to retrieve from data;
3. Construction of new WSA data from WSA clusters;
4. Labeling of WSA data with USP (derived from core-shell clustering) for which the matching of temporal resolution of WSA into the USP is modeled by popular fuzzy membership functions;
5. Application of state of the art ensemble classifiers-Random Forest (RF), Ordinal Forest (OF), and K-Nearest Neighbors (K-NN)-to predict USP from the derived WSA data;
6. Application of the developed experimental clustering-classification framework to datasets as representative samples for an exploratory work;

1.3 Organization of the document

The document is organized as follows. Chapter 2 reviews the clustering techniques used in the analysis of geo-oceanographic phenomena, studies carried out in the spatiotemporal analysis of upwelling using wind data, and clustering methods applied to wind data. Several studies will be presented for each section along with the techniques and protocols used for their purpose. Chapter 3, presents the role of wind in coastal upwelling, going deeper into the interactions and factors that act to trigger such an event. Also in this chapter, it is presented an overview of what possible wind field datasets can be obtained and their respective sources. Chapter 4 reviews and describes the algorithms and frameworks used for the experimental study: core-shell clustering framework, IAP clustering algorithm, RF, OF, and K-NN algorithms. Chapter 5 presents the proposed experimental methodology to answer the main questions of the dissertation. In Chapter 6, the findings of the experimental study are presented and discussed, focusing mainly on the comparison of the defuzzification functions utilized and model by model comparisons for our classification task. In Chapter 7 are presented the conclusion of this dissertation along with possible extensions of the experimental study made.

2

State of the art

2.1 Clustering in the analysis of geo-oceanographic phenomena

The study of coastal upwelling can be interpreted as a spatiotemporal clustering problem. Clustering approaches have also been used to analyze other geo-oceanographic phenomena. These approaches can be categorized into five groups: trajectory, generative model, hotspot discovery, correlation study, and time-stable clustering.

The work in [8] developed a process-oriented clustering method to cluster complex trajectories of dynamic geographic phenomena, such as eddies or storm events. The process started with a process-oriented representation of the trajectories. Then a hierarchical similarity measure was applied. Finally, a density-based trajectory clustering algorithm was used to produce the desired clusters and patterns. The framework presented in [9] utilized hierarchical clustering to mine spatiotemporal periodic patterns from moving object data. The objects' trajectories were divided into multiple line segments at points where the object made sharp turns. Similar segments were then grouped using an extended version of the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm, based on line segments. The algorithm generated a hierarchy of reference points using hierarchical clustering. This method produced more reference points than traditional approaches.

The work presented in [10] proposed using a Gaussian density function to model clusters which can detect anomalous patterns deviating from the Gaussian. The team in [11] compared temperature and salinity profiles with previous knowledge of the ocean structure. The resulting clusters, consisting of 200 layers formed spatially contiguous regions. The study applied a Gaussian mixture model.

In the study of hotspot discovery, in [12] it was applied a spatiotemporal version of an extended Fuzzy C-means (FCM) algorithm for hotspot detection and prediction. The algorithm defined cluster centers using hyperspheres, which were then applied to forecast earthquake hotspots. The results were satisfactory and the reliability was comparable to the ST-DBSCAN algorithm.

The work in [13] conducted a correlation study by performing spatiotemporal clustering

in seismic data. The study aimed to identify potential correlations between energy release rates and the time interval between large earthquakes using a deep neural network.

It was introduced in [14] a new concept in spatiotemporal clustering: time-stable clustering. A cluster is defined as a set of unalterable "core points", i.e., points that never change cluster memberships during a given period. The purpose of this method is to extract dynamic clusters (clusters whose size, shape, and location can change with time) from a continuous spatiotemporal field such as a body of water. The method consists of four steps: specifying the cluster objectives, discovering stable clusters, refining each cluster for each time point, and then processing them. This method is more robust to noise and missing data and could capture the dynamics of water bodies over time.

2.2 Spatio-temporal analysis of coastal upwelling with winds

Several studies of upwelling have been carried out around the world. To achieve this, the researchers use two main types of data: Sea Surface Temperature (SST) and wind data. The objectives can range from developing a new workflow/operational system to detect upwelling to studying the spatiotemporal variability of upwelling in the geographic region of interest. In the next subsections, organized by geographic regions, we present an overview of approaches for spatiotemporal analysis of upwelling, focusing on the exploration of winds. We will highlight the following aspects:

(i) Objectives of the study

(ii) Adopted method(s)

(iii) Geographic region of study

(iv) Wind data characterization

2.2.1 Europe

In the European continent, the work proposed in [15] aimed to analyze trends in SST, Chlorophyll-a (Chl-a), Productivity, and wind stress in upwelling areas in the Upper Eastern North Atlantic Subtropical Gyre (20ºN-45ºN; 30ºW-5ºW). The wind data was obtained from the Copernicus Marine Environmental Monitoring Service (CMEMS) product, based on reprocessed time series of surface wind analyses from several satellite sensors. The wind components used were direction and intensity. To analyze the data, the author and his team used a regression-based approach to assess the trends of the variables in question. The team concluded that the interplay of wind stress and stratification regulates the intensity of the phenomenon. The study presented in [16] analyzed the spatiotemporal variability and other upwelling characteristics with a regression-based approach, specifically Canonical Correlation Analysis (CCA) in the West Iberian Peninsula. The data utilized came from the Quick Scatterometer (QuikSCAT) satellite product and the

Comprehensive Ocean-Atmosphere Data Set (COADS) product. Wind speed and direction were the components used. The researchers concluded that CCA can be used to identify linear relationships between sea surface temperature and wind patterns.

In [17] it was analyzed the spatiotemporal variability on the Portuguese coast by exploring several indices. These included Upwelling Index $(UI)_{SST}$, which is a simple mathematical difference between the temperature offshore and the temperature inshore, and UI_{WIND}. The latter index does not have a consensus within the community; the study considered Ekman transport as UI_{WIND}, although it can change according to the study and the author. Data were obtained from the ERA5 product, a mixture of observations and reanalysis of wind data. Wind direction and intensity were the components used. Overall, the study concluded that the wind-based index is more reliable than the SST index, and upwelling-promoting conditions are more intense on the west than on the south coast.

In the Baltic Sea (54º-66ºN;10ª-33ºE), the work in [3] developed an operational system for automatically detecting coastal upwelling events, assessing their frequency and location. The system is based on a binary classification method, that detects an event if the temperature at a given pixel differs from the corresponding mean by more than a threshold of 2ºC. The analysis used daily mean values of wind direction and intensity, provided by the Interdisciplinary Center for Mathematical and Computational Modelling (ICM). The results were consistent with previous studies, indicating that the areas of the Swedish coast, the Gulf of Finland, the Polish coast, and the Bay of Bothnia were the most active areas for upwelling. The study conducted in [18] also studied the frequency of upwelling and analyzed the phenomenon with statistical analysis while simultaneously detecting it. The data for this study were modeled data provided by the Swedish Meteorological Institute (SMI) with the wind components used being the direction and intensity. Two methods were used to detect upwelling: the first was a visual method in which a horizontal grid was superimposed on an SST map, and the second, a binary classification method similar to the one proposed by [3]. The author and his team concluded that this event is common in the Baltic Sea with regional and seasonal variations in frequency and location. The work presented in [19] utilized the binary classification method introduced in [18] to map upwelling and analyze spatiotemporal variability and characteristics of the event in the study region. Wind data was provided by the ERA5 product and its direction and intensity were used. The study concluded coastal upwelling in the Baltic Sea is primarily driven by Ekman transport, induced by the wind blowing along the coastline.

Another study in [20] analyzed upwelling triggered by specific events such as a storm. The study was conducted in the Balearic islands (39.5ºN;3.0ºE), and explored various indices, including UI_{SST}, $UI_{Ekmantransport}$, $UI_{Totaltransport}$, and also the alongshore velocity. The EU Copernicus Marine Service product provided the data, and the wind intensity and direction were utilized. The team verified that this event had the most significant impact on SST and alongshore velocity in a 9-year time series. However, it did not have the most significant impact on cross-shore transport. The team also suggested that the event might have impacted local primary productivity and the marine ecosystem.

The work presented in [21] analyzed the vertical structure of the ocean during upwelling events and the influence of patterns in the Canary current upwelling system (CCUS) (25ºN-35ºN). The ERA5 and National Oceanic and Atmospheric Administration (NOAA) products provided the data, and the wind components used were the direction and intensity. A linear regression model was used and a correlation between upwelling indices and climate patterns was made. Overall, the team concluded that upwelling in this region is stable and seasonal, being influenced by the North Atlantic Oscillation (NAO) and the East Atlantic (EA) climate patterns. It was also suggested that, under certain conditions, the coupled NAO and EA phases can generate extremes in upwelling.

2.2.2 Africa

Moving to the African continent, the study in [4] developed a new upwelling detection method and analyzed its spatiotemporal variability in Northwest Africa (21ºN-36ºN;4ºW-19ºW). The data used in the study were obtained from the Cross-Calibrated Multi Platform (CCMP) and the NOAA's product, NCEP, with wind intensity and direction being analyzed. To detect upwelling the team utilized a classification method combined with a voting mechanism. A pixel was classified as belonging to an upwelling class if the majority agreed. To study the spatiotemporal variability, several indices were explored, including the cross-shore Ekman transport, Chl-a index, and a SST index, which indicates the upwelling intensity. It was concluded that upwelling in the region can be divided into three regions and the proposed method performed better than other popular fusion methods.

The team of researchers in [22] shared the same objectives as the previous authors and studied the same area. The researchers detected upwelling using a clustering and merging approach and explored indices. Wind data from QuikSCAT and Advanced SCATterometer (ASCAT) satellites was used for this study. To detect upwelling, the authors employed a particle swarm clustering algorithm, identifying upwelling water as the cluster with the lowest centroid and the smallest cardinality. The study merged connected regions to form a larger region. The indices examined in the study were the SST index, cross-shore Ekman transport, NAO index, and Chl-a index. The study concluded that upwelling dynamics are influenced by seasonal and interannual variations of the trade winds, mesoscale processes, and NAO.

A new upwelling index was developed and studied in [23] in the same region as the other two authors with wind data from the Copernicus product. This new index was compared to two existing indices, the Ekman transport and the SST index. The new index provided a more accurate and comprehensive representation of upwelling dynamics, capturing submesoscale oscillations not detected by the wind-based index. The work conducted in [24] investigated upwelling in Madagascar (18.7669°S; 46.8691°E) using SST indices and also the cross-shore Ekman transport, with wind data from the MetOp-A product, derived from the ASCAT satellite. The results showed that coastal upwelling in

the south of Madagascar occurs in two distinct cores, with wind and the East Madagascar Current (EMC) contributing to the generation of upwelling in the two of them.

2.2.3 Asia

Moving on to the Asian continent, in [25] the phenomenon was studied in the Caspian Sea (42.0°N;50.5°E) to evaluate the potential of Sentinel-3 to detect it, while also studying its spatiotemporal variability. The authors obtained wind data from the CCMP dataset. To evaluate the performance of the satellite, the authors employed a specific workflow pipeline. The team preprocessed the data and delineated potential upwelling dates based on wind speed and spatial patterns of upwelling with SST. Finally, the FCM clustering algorithm was applied to detect favorable upwelling cells. To study the variability, SST and wind-based indices were used. It was discovered that upwelling occurs in three cells, influenced by a combination of factors including wind, the topography of the area, and eddies. The study in [26] evaluated the influence of wind and bottom topography in the same region using the Princeton Ocean Model adapted to the area. Wind data used in this study was from the European Centre for Medium-Range Weather Forecasts (ECWMF) product. The study concluded that the wind field and the bottom topography played different roles in the formation of upwelling. Wind-driven upwelling was found to be more intense and frequent in the middle basin than in the southern basin.

It was conducted a study on the spatiotemporal variability of upwelling on the Southwest coast of India in [27]. The study utilized wind data from the CCMP dataset and explored several indices and indicators using linear regression to evaluate their trends. The results showed that winds and remote forcing drive coastal upwelling in the South Eastern Arabian Sea (SEAS). The trend analysis revealed a slight increase in upwelling intensity in the southern part and a decrease in the northern part. The study in [28] compared several indices to analyze the spatiotemporal variability of upwelling in the southwestern region of the South China Sea (Equator-25ºN; 99ºE-122ºE). The wind data was obtained from the ECWMF product and the indices explored were the Ekman transport, Ekman pumping, and SST index. The study revealed that the coastline of the East Coast of Peninsular Malaysia (ECPM) is wind-driven and influenced by the advection of cooler water from a strait. The interannual variability is related to the El Niño Southern Oscillation (ENSO) effect.

Upwelling in the west central part of the South China Sea was detected and studied in [29] using wind data from the Copernicus product. The phenomenon was identified using a binary classification method that required the difference between the mean SST and latitudinal mean SST to exceed a threshold. Spatiotemporal analysis was conducted by exploring indices and using correlation analysis to examine the lagged response of upwelling to the local wind field. This area can be divided into three sub-regions of upwelling. The spatial and temporal variability of this area is significantly impacted by the characteristics of the wind field.

In eastern Taiwan, the work in [30] mapped and studied the phenomenon using wind data from the CFSv2, a product from NOAA. The team utilized a new SST index called Topographic Position Index (TPI) and the Ekman transport that served as the wind upwelling index. To identify upwelling, a binary classification method was used, choosing areas of negative TPI using a threshold and the same method but for sea surface temperature. The study analyzed spatial and temporal variability using a time series of upwelling favorable wind events and correlated the number of days with favorable winds to satellite data. The results indicate that TPI is effective for mapping coastal upwelling. The study suggests that Ekman transport and pumping are the primary driving forces behind the phenomenon in the area.

2.2.4 America

On the American continent, in [31] upwelling fronts were identified and examined in the central Chile region (36.5ºS-37ºS) using an SST-based method. The team also used wind data from QuikSCAT and ASCAT satellites, as well as the Ekman transport as a wind-upwelling index. The fronts were identified employing a specific framework of binary classifications and were analyzed with a linear correlation between their characteristics. The team concluded that front characteristics vary seasonally and interannually. The work proposed in [32] investigated the role of the wind stress field on the North Humboldt Upwelling System (4ºS-19ºS) using wind data from the ERA5 product. The study employed an Empirical Orthogonal Function (EOF) to analyze the co-variability of wind-field and SST and linearly correlating several indices. The co-variability between the two fields across the Pacific Upwelling System (PUS) was found to be complex and asymmetric. The wind field over the system exhibits strong fluctuations on several scales, closely linked to the ENSO and the Interdecadal Pacific Oscillation (IPO). The study conducted in [33] used data from the National Data Buoy Center, where upwelling was examined in a small embayment on the Central California Coast. The phenomenon was investigated by several indices, including the Large and Pond wind upwelling index, coefficient of variation, standard upwelling SST indices, and the Chl-a concentration. Five upwelling seasons were identified and temperature and chlorophyll patterns are influenced by the phenomenon's seasonality.

2.2.5 Other regions

In some other regions of the globe, in the Australian South-Eastern Coast (132.5ºE-154ºE; 28ºS-44ºS), an upwelling mapping technique was employed in [34], using the TPI, and the Ekman transport. The data was obtained from the Australian Bureau of Meteorology. The study identified upwelling with a specific framework of binary classifications and visual selection. The spatiotemporal variability of upwelling characteristics was analyzed using linear correlation analysis. In a more scaled study, in [35], were only used SST-based indices to uncover the regionality of Global SST to help future studies of the upwelling

phenomenon. It was developed a model called the Multi-Stage Spatio-Temporal Clustering Method (MuSTC), and the data was provided by the NCEI, a NOAA product. The method proved to be useful in uncovering the regionality of the global sea surface temperature.

2.2.6 Summary

With this overview, we can summarize the aspects mentioned in the beginning.

(i) Detecting and analyzing upwelling are the main goals

(ii) Using binary approaches to detect upwelling is widely used by the community

(iii) Upwelling studies are mainly done by exploring indices

(iv) There isn't a consensus in the community regarding some indices like the UI_{WIND}, for this, the researchers adapt it to the geographical region being studied

(v) Apart from data given by meteorological institutes, the most popular product datasets used are the *COADS*, *ERA5*, *QuikSCAT-ASCAT*, *and CCMP*

2.3 Clustering wind data

Several studies have been conducted in various regions using wind data, similar to the study of upwelling. A particular objective of these studies can be to discover the potential energy that wind can generate. The most commonly used method is a clustering algorithmic approach, due to the absence of ground-truth data. This makes it both a reason and an advantage to use, as the results obtained can be interpreted. This approach offers additional benefits, such as identifying structures in the data and reducing computation time compared to other supervised methods. This section will focus on the objective of the study, the algorithm(s) used, and the entities clustered.

2.3.1 Introduction

In [36] it comprised a collection of clustering algorithms used in a comparative study of wind speed clustering. The study concludes that the Linkage-Ward method is the more accurate, despite its higher computation requirements due to complex calculations, when compared to other methods such as K-Means. Clustering algorithms were reviewed in [37] to identify the temporal wind speed profiles in a specific region of South Africa. The studied algorithms were categorized into three groups: *Partioning algorithms*, *Hierarchical algorithms*, and *Advanced algorithms*. They were validated using several methods, including the Silhouette Coefficient, the number of incorrect cluster assignments, the Calinski-Harabasz index, the average distance between clusters, and the Dunn index. The *Partitioning algorithms* tested were K-Means, Partitioning Around Medoids (PAM), and the Clustering large applications algorithm (CLARA). In the *Hierarchical algorithms* section, the study

tested the agglomerative algorithm and the Divisive Analysis Clustering Algorithm (DIANA). Only one algorithm, the FCM, was tested in the last section of *Advanced algorithms*. The study concluded that the CLARA approach was the most suitable for achieving its objective.

2.3.2 K-Means application

Moving on to more concrete approaches, the K-Means algorithm was used in [38] to identify the most prominent cities regarding monthly average wind speed in a group of 75 cities. The algorithm underwent testing with four distance measures, with the squared Euclidean distance measure ranking highest among City-Block, Cosine, and Pearson Correlation. This algorithm was also used in [39] to identify relationships between winds at turbine height and climate oscillations. K-Means was applied to wind data (zonal and meridional components) up to 80m in height, with the features being its speed and direction. The purpose of this study was to develop a method that could predict the impacts of climate change on wind resources. The work proposed in [40] utilized K-Means to classify synoptic and local-scale wind patterns in a coastal area of the Tyrrhenian Sea in Italy. Daily wind profiles were clustered, with features such as zonal and meridional components. The study analyzed the wind intensity and direction data and identified three clusters: The northeast cluster, the Breeze cluster, and the Southeast cluster. These clusters were then thoroughly analyzed.

2.3.3 Hierarchical Clustering application

The work presented in [41] characterized vertical wind speed profiles using Ward's Agglomerative Clustering algorithm. The wind profiles were based on the wind speed and direction at various heights. The researchers yielded satisfactory results in identifying the most probable wind vector patterns for different sample times and heights. Hierarchical Clustering was used in [42] to study the impact of winds on evaporation in Eğirdir Lake, Turkey. Monthly evaporation losses were clustered with monthly wind speed and wind blow number. The study concluded that the algorithm, coupled with the single-link method, produced satisfactory results. Monthly wind speed and insolation data were analyzed in [43] in Turkey to determine the potential for installing wind and solar power plants. The study utilized the Hierarchical Agglomerative Clustering with several different metrics and developed a graphical model. The results identified the cities with the highest potential for renewable energy power plants. In [44], hierarchical clustering was employed with average linkage to gain knowledge about wind characteristics in La Plata, Argentina. Hourly wind roses were clustered representing the prevailing winds for different periods of the day and seasons, with the features being wind direction frequencies. The work proposed in [45] utilized agglomerative hierarchical clustering to group time series of wind speed in different regions of Iran. The study identified a turning point of significant wind speed around 1990.

2.3.4 Hybrid Methods

The framework employed in [46] used a hybrid clustering-statistical method to forecast year-ahead wind speed. Daily wind speed observations were clustered. The researcher's approach involved utilizing the DBSCAN algorithm to prune wind speed data from the original dataset. Next, the statistical model of Autoregressive Integrated Moving Average (ARIMA) statistical model was used to project future wind speed. The study concluded that this hybrid model outperforms using only the ARIMA model. The work in [47] used trend-based time series data clustering for wind speed forecasting. The technique involved finding clusters of time series data with identical components. Similar to the previously mentioned work[46], statistical methods, such as the ARIMA and the Generalized Autoregressive Score (GAS) were applied to each cluster. The study confirms that incorporating a clustering method before applying statistical methods enhances the accuracy of wind forecasting models.

In Cartagena, Spain, in [48] cluster analysis was used to classify winds for the development of predictive statistical models on atmospheric pollution. The study conducted a cluster analysis in two parts. The first part utilized hierarchical cluster analysis with average linkage methodology, while the second part employed K-Means for wind speed and direction classification. The study identified five wind patterns with different predominant wind directions. The researchers concluded that this would enable them to develop predictive regression models for each cluster. Also on the topic of wind speed forecasting, a combination of Spectral Clustering and an Echo State Network was employed in [49]. Spectral Clustering was utilized to select similar data from the historical data to form training and validation sets, while the Echo State Network was used to make the prediction. The study concluded that the model performed well and was more accurate than other traditional models.

The work presented in [50] used cluster analysis to uncover synoptic wind regimes in California. The cluster approach employed was a hybrid of the cluster algorithms K-Means or Hierarchical Clustering. Initially, a similar iterative process to k-means was used to create partitions into k clusters, each of which was represented by a Dynamic Principal Component Analysis (DPCA) model. The solutions are aggregated into a single distance matrix and Hierarchical Clustering is used to form a dendrogram and choose the final partition. Overall, four clusters were identified, with two of them having enhanced ventilation, and the other two capturing distinct meteorological patterns.

2.3.5 Other methods

The Linkage-Ward method was used in [51] in an Iranian case study to cluster wind speeds. The average values of wind speed up to 40m in height and over 10-minute intervals were clustered. The purpose was to meet the increasing electricity demand, particularly from renewable sources such as wind, and to take advantage of Iran's potential for wind turbine installation. This study compared the algorithm with K-Means. The results showed

that the Linkage-Ward had a lower relative error, but had a higher execution time than K-Means. The work proposed in [52] utilized a modified version of FCM in a fuzzy model to predict wind farm power generation. Wind speed, air temperature, and wind power were the entities clustered. The optimal fuzzy rules for the model were determined using the clustering algorithm. Along with other meteorological variables, wind speed data was used.

2.3.6 Summary

Based on the literature, it appears that K-Means and Agglomerative clustering are widely used. The entities being clustered are study-dependent, with daily and monthly wind averages the most frequently used. The features clustered usually include the wind speed and its direction. From the objectives, two stand out: Identifying wind profiles and from these profiles obtained using them as input to predictive models.

3

Wind data

This chapter focuses on wind data characterization, from the role wind has in coastal upwelling to the specifications of the products that collect it. Section 3.1 provides a detailed description of how wind contributes to upwelling. Section 3.2 describes several wind data extraction products and the process of obtaining wind data from the products described.

3.1 The role of the winds in coastal upwelling

3.1.1 Wind component

To initiate upwelling, the wind must blow parallel to the coastline at a specific speed to trigger such an event. However, several other factors such as the coastline geometry [18] and the bottom topography [26] of the area can also influence upwelling, potentially allowing for non-parallel wind directions. The wind itself, as several studies present in chapter 4, has two components, u and v. The u component, parallel to the x-axis (longitude), represents the eastward direction of the wind, being positive in a west-to-east flow, and negative otherwise. The v component is parallel to the y-axis (latitude) and represents the northward flow, being positive in a south-to-north flow. To determine the component that contributes to coastal upwelling formation, the *alongshore component*, a transformation has to be done, which will be explained in a later phase of the document.

3.1.2 Coriolis Effect

The Coriolis Effect is a phenomenon that illustrates how objects not fixed to the Earth's surface deviate from their intended path as they move across the planet [53].

Generalizing, the effect can be understood as the impact of an inertial force on moving objects within a rotating non-inertial frame relative to an inertial frame.

To demonstrate this, the analogy made in *What is the Coriolis Effect?* [54] is simple. Let there be 3 trains, 2 red trains, and 1 blue train which are on different latitudes of the globe. Of the 2 red trains, one is located in the northern topic and the other is located in the

southern topic, the blue train is set in the Equator. Due to the Earth's rotation, the blue train is moving faster than the other trains, although, from a bird's eye perspective they appear to go at the same speed. If one wanted to kick a football from the blue train in the direction of a goal present in each red train, the ball would be deflected to the right of the northern topic goal and left of the southern topic goal (facing the goal). In this example, the blue train is considered the non-inertial reference frame, because it is rotating relative to the inertial frame, the Earth, while the moving objects are the observer and the football. This is illustrated in figure 3.1, with the red line describing the trajectory of the ball.

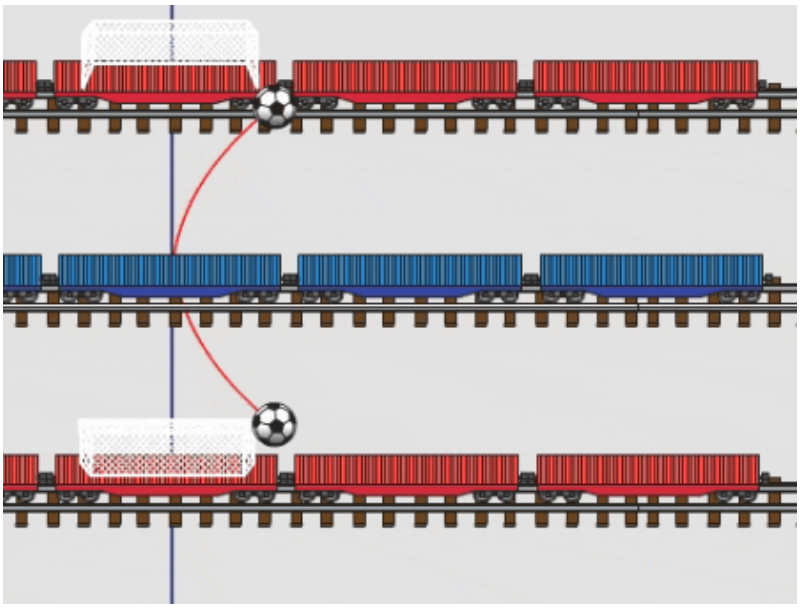


Figure 3.1: Ilustration of the Coriolis effect. Image taken from [54]

Applied to weather patterns, this effect will lead to curved paths as the Earth rotates on its axis, deflecting the atmosphere to the right in the northern hemisphere and left in the south hemisphere. The formation of phenomena such as cyclones and trade winds are also an effect of this. In the upwelling phenomenon, when the wind blows parallel to the coastline, the Coriolis Effect can move the water at the right angles to trigger this event as detailed in [55].

3.1.3 Ekman Spiral

To explain the Ekman Spiral, we first start with an example where the Coriolis effect is not present. Given a stack of paper, this will be displaced if we apply a dragging force to the top sheet. However, the top sheet is not the only one being displaced, the second sheet will be dragged by the displacement of the first sheet, although it will not move as much. This continues until there isn't enough force to move a sheet of paper. Figure 3.2 represents this.

Applying the Coriolis Effect to the previous example, each sheet of paper will be turned at a specific angle, one more accentuated than the other, thus making a spiral. In the

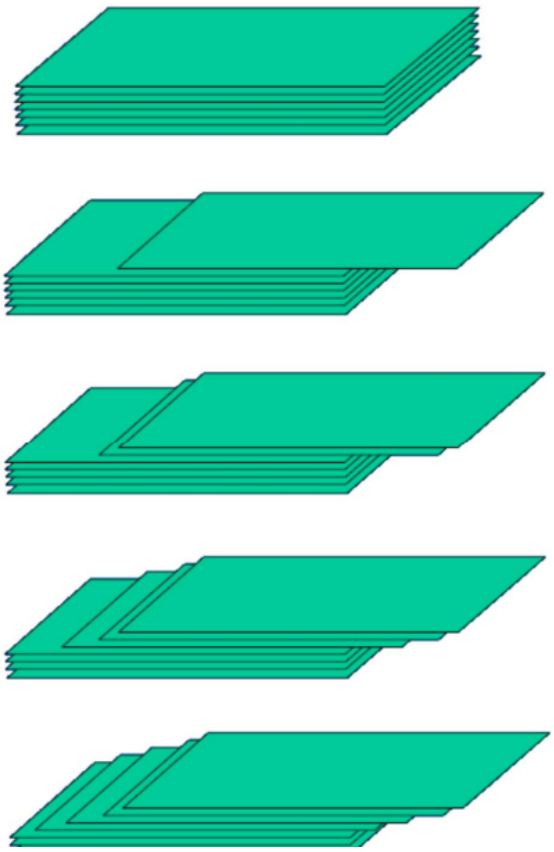


Figure 3.2: Ilustration of the Ekman Spiral without the Coriolis Effect. Figure provided by Paulo Relvas.

problem context, as the depth increases, the velocity vector will decrease exponentially in intensity rotating to the right in the northern hemisphere (left in the southern hemisphere). Due to this decreased velocity intensity and increased rotation in each vector, their edges will form a logarithmic spiral. This is shown in figure 3.3:

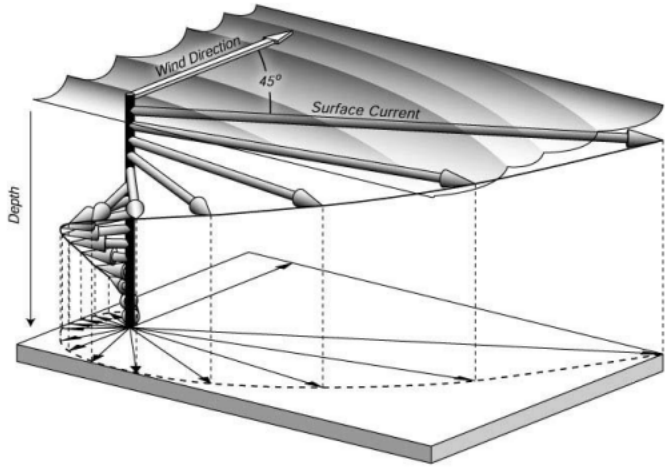


Figure 3.3: Ekman Spiral. Figure provided by Paulo Relvas.

where the arrow length is proportional to the current intensity and their direction is the same as the current.

3.1.4 Ekman Transport

The Ekman Transport is described as the total movement of water per unit of time, at a 90° angle to the direction of the wind (to the right in the northern and the left in the southern hemisphere) as a result of the balance of Coriolis and drag forces [56]. Figure 3.4 represents a bird's eye view of the Ekman transport:

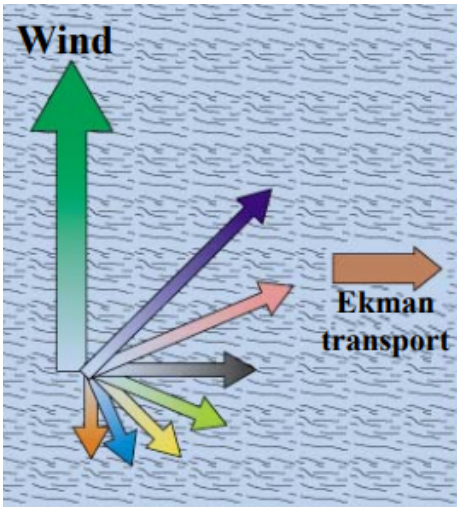


Figure 3.4: Ekman Transport from a top view. Figure provided by Paulo Relvas.

The phenomenon's direction is crucial for upwelling as it depends on the movement of water away from the shore. Conversely, downwelling occurs when water moves towards the shore, as figure 3.5 represents:

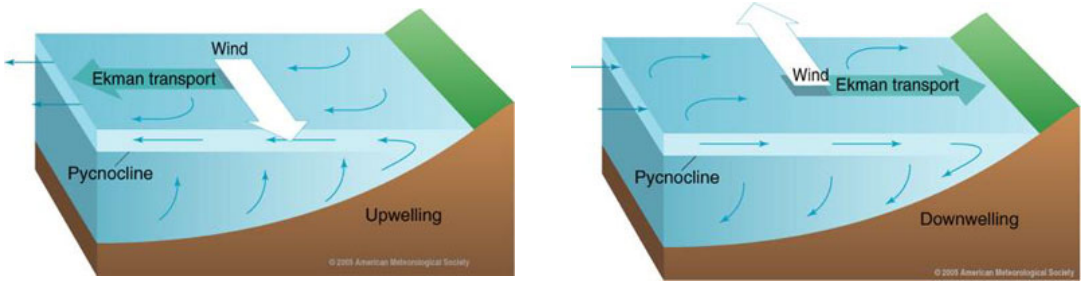


Figure 3.5: Upwelling and downwelling occurrences. Image taken from [57]

3.1.5 Wind-stress

Wind stress is a frictional force, per unit of area, that results from the wind blowing over the sea surface. The wind speed, roughness of the sea surface, and prevailing atmospheric conditions are the main factors that influence wind stress. In upwelling, the alongshore wind stress, i.e., the wind stress parallel to the coastline is one of the factors responsible

for inducing it [58]. This force is proportional to the square of the wind speed and is given by the following formula:

$$\tau = c * W^2 \tag{3.1}$$

where, c is the coefficient of friction and W is the wind speed.

Wind stress curl originates from the wind stress shear, or, spatial variations of the wind stress vector (τ_x and τ_y). Specifically, the horizontal stress shear. Wind stress curl can be positive or negative, being indicated by the direction of its vorticity. The direction of the vorticity of the wind stress curl can be determined by applying the right-hand rule, with the thumb pointing up for positive and down for negative. The usual wind shear pattern for a positive vorticity is an increasing intensity away from the shore, and the opposite for a negative vorticity. Figure 3.6 represents this:

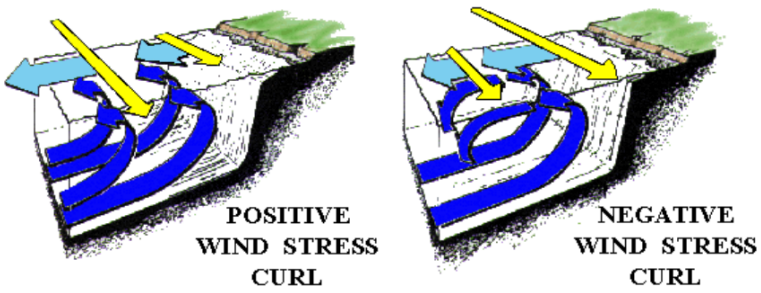


Figure 3.6: Upwelling events induced by variable cross shore intensity winds. Figure provided by Paulo Relvas.

Here, the yellow arrows represent the wind stress vector, the light blue arrows represent the Ekman Transport vector, and the darker blue arrows represent the upwelling-downwelling vector. As seen, the more intense the wind, the more intense the Ekman Transport, leading to convergent or divergent zones, characterized by negative and positive wind stress curl, respectively. In a convergent zone, water is pushed down, suppressing water mixing, while in a divergent zone, water is pulled up, causing upwelling. However, it should be noted that the actual impact of the wind-stress curl on upwelling is highly complex and dependent on several other factors. Different wind stress patterns will lead to different wind stress curl patterns, being important factors in shaping the upwelling pattern.

3.2 Products

The assessment and analysis of atmospheric conditions are crucial for various applications, including weather forecasting climate studies, and phenomena like the one being studied in this dissertation. Multiple products are available to provide this information, each with its strengths and limitations.

The products are divided into two main groups: satellite observation and reanalysis products. The main difference between these types is the use of a combination of observation data and prediction models by the second type.

Although the resulting dataset for the product contains several variables, only the wind condition variables will be used. Specifically, the wind speed and wind direction at surface level, 10 meters, will be used.

3.2.1 ASCAT Datasets

The Advanced SCATterometer (ASCAT)'s primary objective is to measure wind speed and direction at the ocean surface [59]. Secondary objectives include measuring sea-ice type and soil moisture. ASCAT's observations provide a wealth of data products [60]:

- ASCAT Ocean Surface Wind Vectors data of 50 km resolution
- ASCAT Ocean Surface Wind Vectors data of 25 km resolution
- ASCAT Storm data
- ASCAT Global Ambiguity
- ASCAT Ice Data

The first two products are the most relevant for upwelling studies since the other type of data is not utilized. The wind components are measured in meters per second.

3.2.2 ERA5 Datasets

The ERA5 product is the fifth-generation atmospheric reanalysis of the global climate by European Centre for Medium-Range Weather Forecasts (ECWMF). Its predecessors are ERA-15, ERA-40, and ERA-Interim (from older to newer). The data is characterized by a high resolution of $0.25° \times 0.25°$. Additionally, due to its reanalysis nature, it introduces estimates of uncertainty in the data [61]. One possible way to obtain a dataset of this product is through [62] where is present an example entry: *"ERA5 hourly data on single levels from 1940 to present"*. The entry describes gridded data projected on a regular latitude grid, with possible resolutions of $0.25° \times 0.25°$, $0.5° \times 0.5°$, or $1° \times 1°$. The dataset covers the period from 1940 to the present with an hourly resolution. It contains various variables, but for this work, the most important is the northward and eastward components (u and v) of the wind at 10 meters, all measured in meters per second. The base product is sufficient for the intended work, although variations exist, such as the ERA5-Land.

3.2.3 CCMP Datasets

The Cross-Calibrated Multi Platform (CCMP) dataset provides a gridded analysis of wind vectors. The input data is derived from inter-calibrated satellite data and in-situ data

measurements. The dataset has high spatial and temporal resolutions ($0.25° \times 0.25°$ and 6-hourly) and a long data record, spanning from 1987 to the present, making it useful for studies of interannual variability. However, it should be noted that the dataset has limitations, including low accuracy in rainy conditions and unsuitability for trend studies [63]. The latest version, V3.1, covers a global region with temporal coverage from January 1993 to December 2019 [64]. The Northernmost Latitude is 80º, the Southernmost Latitude is -80º, the Westernmost Longitude is 0º, and the Easternmost Longitude is 360º. There are two types of datasets, the 4x Daily (6 hourly) and the monthly dataset. The first dataset includes the meridional and zonal wind components, as well as the wind speed measured at 10 meters, all in meters per second. Similarly, the second dataset contains monthly wind averages and anomalies, also measured in meters per second.

4

Background knowledge

This chapter is dedicated to the introduction and description of the various frameworks and algorithms to utilize during the experimental study. Each section is structured in a similar manner, beggining with a concise overview of the technology in question, followed by a more comprehensive explanation of the framework or algorithm's functionality. Finally, each section presents case studies in which the technology under discussion was employed, both inside and outside the application's domain.

4.1 The core-shell clustering framework

The core-shell clustering framework is an automatic detection and tracking framework for upwelling Sea Surface Temperature (SST) regions, that employs the concept of a core-shell structure. Developed in [6], this framework is designed to address the limitations of other Density-Based Spatial Clustering of Applications with Noise (DBSCAN) approaches, which often require significant user input and the selection of appropriate of parameters [65].

The detailed workflow is ilustrated in Figure 4.1, and below are briefly described the main steps composing it:

1. Preprocessing M SST grids from an upwelling season utilizing a preprocessing pipeline outputting N preprocessed SST grids.
2. These preprocessed SST grids serve then as input for the Sequential Self Tunning Seeded Expanding Cluster (S-STSEC) algorithm, unsupervisingly identifying upwelling and non-upwelling regions, originating a bipartition map.
3. From the bipartition maps 4 features are extracted: the total upwelling area, the average SST temperature and the latitudes of the northernmost and the southernmost regions, creating a time series.
4. With the time series constructed, these are unsupervisedly grouped with the help of the Iterative Anomalous Pattern (IAP) algorithm, grouping similar SST instants. These groups of instants are designated Upwelling Stability Period (USP).

5. From the derived USP, the T consecutive SST instants present in each collection are used as input for the core-shell clustering algorithm. This algorithm models a core-shell spatial cluster structure, whoose core defines the constant part of the upwelling region, and the shell, represents the dynamic part of the upwelling region.

6. Features are then extracted from these core-shell cluster time-series, allowing for inter-anual analysis and coastal upwelling trend studies across the world.

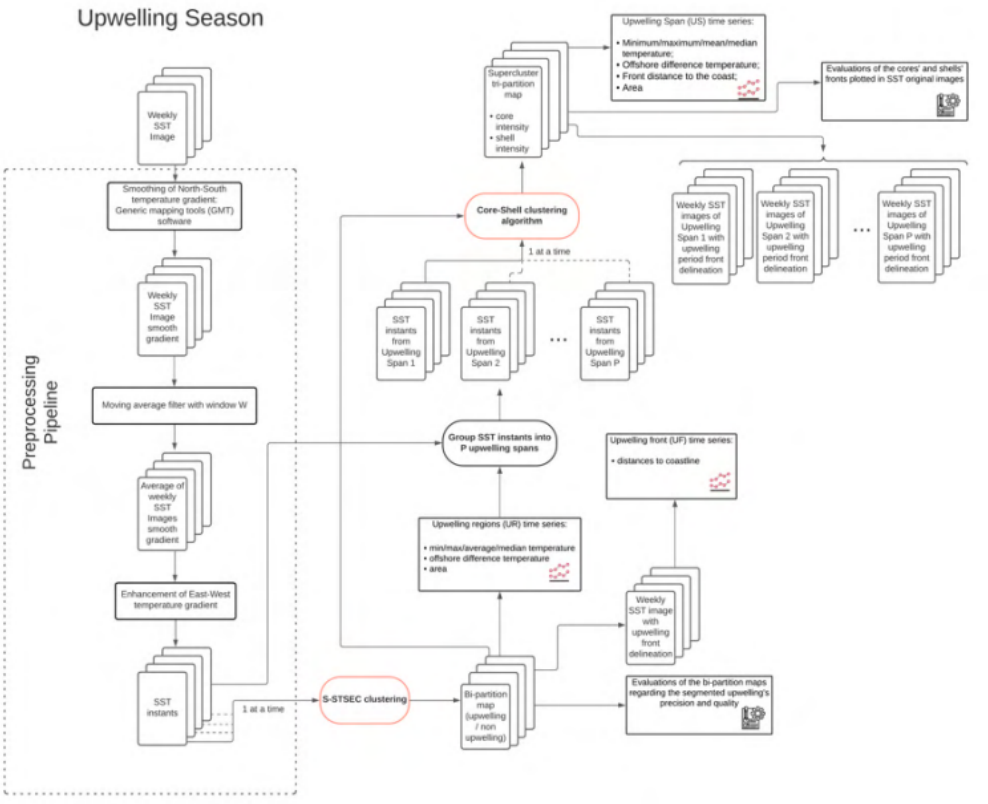


Figure 4.1: Full workflow pipeline. Image taken from [6]

The figure above presents the core-shell clustering framework workflow, from the preprocessing of the SST grids to the segmentation of those grids by the core-shell clustering algorithm. In the preprocessing stage, the SST grids undergo 3 steps. Initially, the North-South gradient is first smoothed out producing better S-STSEC segmentations [6]. Subsequently, a mooving average window is applied, creating a much smoother collection of SST grids, now designated SST instants. Finally, the preprocessing pipeline addresses the issue of over-segmentation when employing the S-STSEC algorithm.

These SST instants are then used as input to the S-STSEC algorithm, which initially delineates the "upwelling front", separating oceanic from coastal waters, outputting a binary grid that defines the upwelling and non-upwelling regions with the help of the seeded region growing technique.

The features mentioned in step 3 of the aforementioned steps are then extracted from the bipartition grids, serving now as input for the IAP to divide the upwelling season into shorter periods in which upwelling remained relatively stable, or USP. This is done by extracting four features: average temperature of the upwelling area, the total upwelling area and the southern most and the northern most latitudes of the upwelling area. So, an USP is defined as a collection of SST grids that shared similar values for these four upwelling attributes.

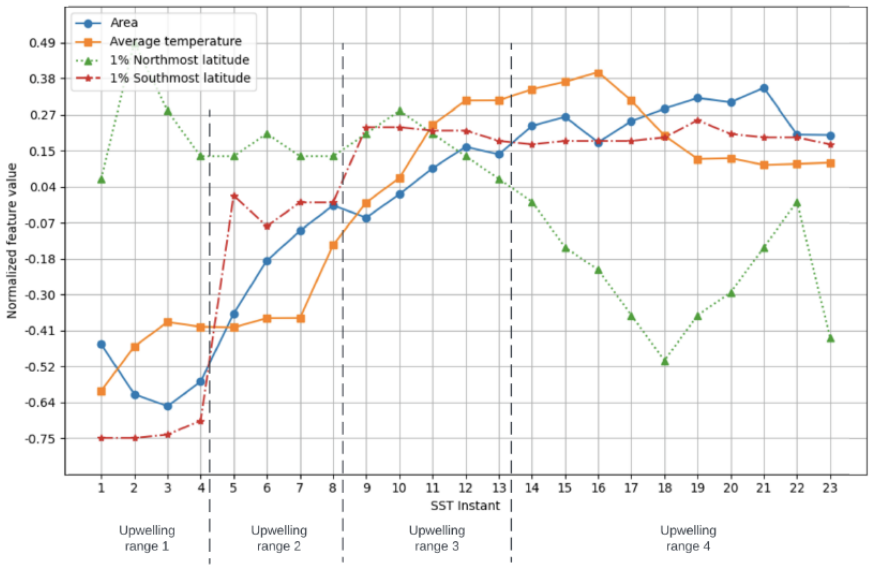


Figure 4.2: USP obtained by applying IAP to S-STSEC segmentations. Image taken from [6]

In Figure 4.2, are presented the results of applying IAP to the SST instants obtained from the S-STSEC segmentations, totalling in 4 USP. The geographic region applied was Portugal during an upwelling season and the year is 2019. We can see an increase in the average temperature (orange), area (blue) and the southern most latitude until USP 3, stabilising along the final instants. We can also observe that the northern most latitude decreases over the year.

Figure 4.3 presents the original SST instants of the Portugal geographic region and the corresponding core-shell segmentations. The core-shell segmentations present the upwelling (orange and green) and non-upwelling (blue and white) regions. The instants ilustrated, all are part of one common USP, with this being USP 3.

The core-shell clustering algorithm is the crucial part of this framework, as it is the final step taken before the final results are outputted, integrating the segmentation results of the S-STSEC algorithm. The algorithm differentiates from other technologies such as the DBSCAN by using the core-shell structure concept instead of collecting dense proportions of data just like the former does. The model proposed for this algorithm was proposed in [66] being described as follows:

Let a preprocessed SST grid be defined as $A^t(I, J) = (a^t_{ij})$, with temperature a^t_{ij} at

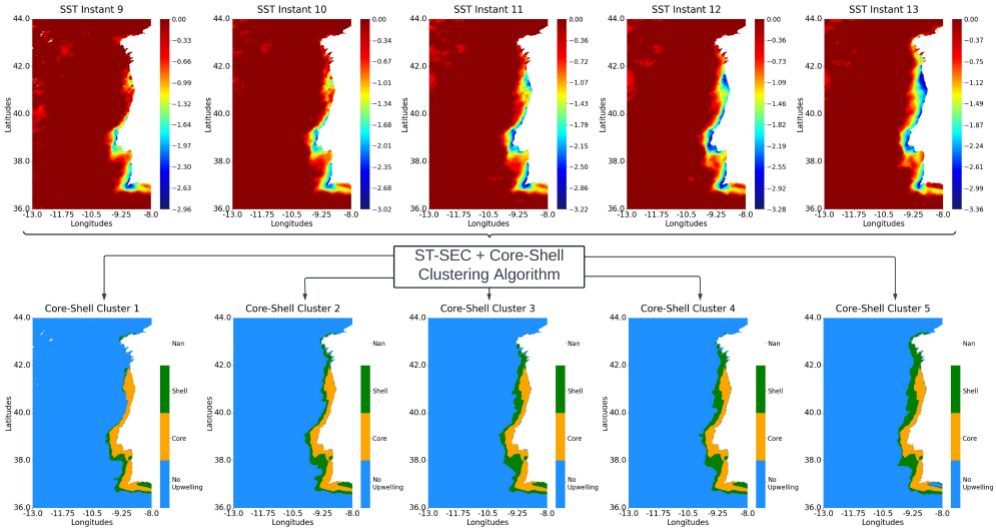


Figure 4.3: Core-shell example output. Image taken from [6]

longitude i ($i = 1, 2, \ldots, I$) and latitude j ($j = 1, 2, \ldots, J$) at a given period t ($t = 1, 2, \ldots, T$). A core-shell cluster is composed by two non-overlapping binary sets, $R \cup S^t$ with $r_{ij} \in R$ being the core and $s_{ij} \in S^t$ the shell at period t, such that at any given period t, $r_{ij} \times s_{ij}^t = 0$. Let the shells S^t be characterized by their intensity values, λ^t. The core's intensity should always be greater than of a shell, so the core's intensity is described as $\lambda^t + \mu^t$, with $\mu^t > 0$. With this the SST at point ij is defined as:

$$a_{ij}^t = (\lambda^t + \mu^t) r_{ij} + \lambda^t s_{ij}^t + e_{ij}^t, \tag{4.1}$$

where the residual values e_{ij}^t should be minimized according to the least squares criterion

$$\Delta = \sum_{t=1}^{T} \sum_{i=1}^{I} \sum_{j=1}^{J} (a_{ij}^t - (\lambda^t + \mu^t) r_{ij} + \lambda^t s_{ij}^t)^2. \tag{4.2}$$

Applying the first order derivation to the previous equation in order of λ^t and $\lambda^t + \mu^t$, we get the shell and core intensities, respectively

$$\lambda^t = \frac{\sum_{i,j} a_{ij}^t s_{ij}^t}{\sum_{i,j} s_{ij}^t}. \tag{4.3}$$

$$\lambda^t + \mu^t = \frac{\sum_{i,j} a_{ij}^t rij}{\sum_{i,j} rij}. \tag{4.4}$$

Going back to equation 4.1, we can now substitute the intensity values with the new formulas getting

$$\Delta = \sum_{t=1}^{T}\sum_{i=1}^{I}\sum_{j=1}^{J}(a_{ij}^{t})^2 - \sum_{t}((\lambda^t + \mu^t)^2 \times |R| + (\lambda^{t^2} \times |S|^t)), \tag{4.5}$$

where $|R|$ is the number of data points in the core and $|S|^t$ the number of points in the shell at period t.

Criterion 4.5 can be written as

$$\Delta = D - G, \tag{4.6}$$

where D is the total data scatter and G the contribution of the core-shell cluster to the data scatter.

As D is constant, minimizing the least squares criterion 4.5 is equivalent of maximizing G.

This framework has been previously applied in [6, 65] in the portuguese coast and also applied in [5, 66] to North-Morocco (NM) and South-Morocco (SM) coast, to explore the upwelling dynamics in the region. Also, it was developed an extension of this framework in [67] where the algorithm's limitations were tested by exploring different clustering algorithms for segmenting time series data to define USP.

4.2 The Iterative Anomalous Pattern

The IAP cluster algorithm by Mirkin [68] is a unsupervised learning algorithm based on the divide and conquer method. Consider an entity-to-feature data matrix X. The method iteratively extracts clusters from a tabular standardized dataset Y, obtained by shifting the origin of X to the grand mean $\overline{x}$ and then rescaling the features according to their respective ranges. This feature vector $\overline{x}$ is then taken as the first reference point, and the farthest point away from it is taken as the seed point. Follwing this, a cluster C_t is constructed, defined as the set of data points closer to the seed point than to $\overline{x}$. After this process, cluster's C_t seed is defined as its center of gravity. The method then reiterates over the residual data $Y_{t+1} = Y_t - C_t$ until one of the stop conditions are met:

(S1) All entities are clustered, meaning Y_{t+1} is empty

(S2) The overall cumulative cluster contribution reaches a predefined threshold τ

(S3) The cluster's contribution is too small

(S4) The number of clusters reached a certain predefined value, K^*

As dataset Y can be considered a matrix of N rows and D columns, the total data scatter of the data points, $T(Y)$ is defined as [69]:

$$T(Y) = \sum_{i=1}^{N} \sum_{h=1}^{D} y_{ih}^2, \tag{4.7}$$

where y_{ih}^2 is the squared value of data point i at feature h.

In [68] is demonstrated that the data scatter $T(Y)$ can be decomposed into two distinct parts: an explained part due to the retrieved cluster structure and an unexplained one, corresponding to the K-means criterion. Thus, the individual contribution of cluster C_t, $W((C, v))$ is defined as:

$$W((C, v)) = \frac{n \sum_{h=1}^{D} v_h^2}{T(Y)} = \frac{n \sum_{h=1}^{D} v_h^2}{\sum_{i=1}^{N} \sum_{h=1}^{D} y_{ih}^2}, \tag{4.8}$$

where n is cluster's C_n cardinality and v_h^2 the squared value of prototype v component at feature h.

In [69], was made a comparative study with the furthest sum initialization method where the authors utilized several different clustering algorithms such as the Fuzzy clustering with proportional membership (FCPM) and archetypal analysis, analyzing different aspects which include the number of iterations and the convergence of the cluster's contribution on several different datasets. Overall, the authors concluded that using this initialization method was a good modelling strategy to determine the number of clusters to extract. IAP was also utilized as an initial setup to the Fuzzy C-means (FCM) algorithm in [70], in which a tool was developed to automatically delineate SST upwelling areas using the combination of the two algorithms. On the works of [6], this algorithm partakes in a crucial part of the framework developed. Here, the method was to segment time series data, creating several upwelling spans which would be later applied in the author's framework.

4.3 Random forest classifier

The Random Forest (RF) algorithm [71] is a supervised learning ensemble method that can be utilized in classification and regression problems. This algorithm is inserted the ensemble learning category as it combines the outputs of multiple individual models, in this case decision trees, to improve prediction quality. In classification problems, a majority vote is taken to classify the point in question. For regression, the average of the decision trees' output is computed. A RF is then composed by a set of decision trees, where each tree is built using the Classification and Regression Trees (CART) algorithm. The CART algorithm's unstability [72] is overcome when using an RF model, as the latter is composed of multiple trees. Thus, a RF model is more robust than a single decision tree. Figure 4.4 ilustrates the main idea of the RF algorithm.

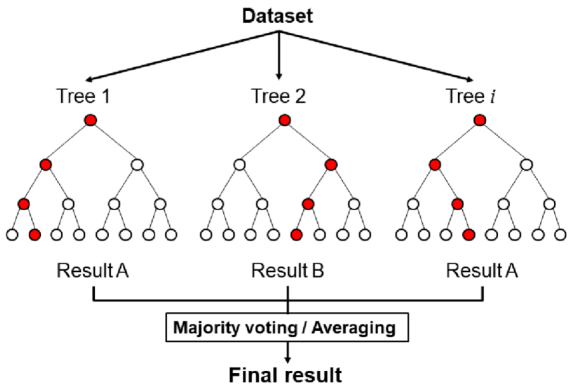


Figure 4.4: Schematic representation of the RF classifier. Image taken from [73]

The randomness characteristic of a RF model is a consequence of the manner in which each tree is constructed. Summarizing the framework described in [72], each tree is built with resource to the bootstrapping aggregation technique. This technique involves the use of a random subset with replacement of the original data for the purpose of training a tree. This technique is also referred to as bagging, which serves to reduce the variance associated with each individual tree. Next, the features that will be available for growth of the tree are determined, and the tree itself is trained on each bootstrap sample Applying this metodology to every tree outputs a RF model. As it is a set of trees, the RF should be composed by diverse and low-correlated trees, which is already achieved by the usage of the bagging technique. However, this composition can also be improved by applying feature selection techniques. As previously outlined in [72], several approaches may be employed, such as using the totality of available features during the tree building phase, randomly draw a subset of features at every splitting node and use it as input to determine the optimal split at the node itself, with the possibility of the former 2 being combined.

The RF with its default hyperparameters has been demonstrated to achieve good results, as several reports suggest [74, 75]. However, it is imperative to conduct fine-tuning to enhance its performance. The hyperparameters present in a RF model are vastly diverse, controlling each tree structure or even its randomness. These hyperparameters include the maximum depth of the decision tree, defined as the longest path from the root to a leaf node; the minimum samples required to split an internal node, thus also controlling the tree's depth; and the splitting rule utilized, where for classification it is commonly used the Gini impurity criterion or the Entropy criterion. The randomness of the tree can be controlled with resource to parameters such as *mtry* parameter which is the number of features selected for each split. With the sample size parameter, the randomness can also be controlled as this hyperparameter corresponds to the number of observations drawn for each tree. Another crucial hyperparameter for the RF model is the number of trees which balances the tradeoff between the RF performance and computational cost.

The RF model is a flexible algorithm as it can handle both regression and classification problems. Additionally, it exhibits a high capacity for handling high-dimensional data,

not being sensitive to noisy and missing data. The model presents a reduced risk of overfitting due to the fact that with more decision trees, specially lowly correlated with each other, the variance and prediction error are reduced. This model is also fairly simple to evaluate as it provides the variable importance to it. However, the results outputted by a RF are of harder interpretation when compared to a decision tree, and its computational fit time is also greater than of a decision tree given the fact that the model is an ensemble of decision trees.

Nowadays, the RF model is employed in a number of fields such as finance, healthcare and e-commerce. In the financial sector, this algorithm is utilized for the detection of fraud and the assessment of customers with high credit risk [76]. In healthcare, this algorithm was used in the domain of bioinformatics for the analysis of gene expression and the classification of disease samples, as well as the identification of biomarkers [77]. In e-commerce, this model can be used for recommendation engines based on consumer data [78].

4.4 Ordinal Forest classifier

The Ordinal Forest (OF) classifier is a relatively recent random forest-based prediction method for ordinal data [79]. The method was introduced to capitalise on the ordinal nature of response variables, typically treated as nominal, and also due to the limited availability of prediction methods designed to the specific characteristics of such variables. In contrast to the conventional approach of treating ordinal response variables as nominal, this method considers them as continuous variables, thereby underlying their ordinal behaviour. As a result, this method shares similarities with a regression forest, which is also reffered to as naive OF in this context.

The OF algorithm initiates the process by transforming the class values into score values, which are then optimized with the objective of maximizing the out-of-bag performance [79]. From these score values, it is created a candidate score set $\{s_1, \ldots, s_J\}$, with $1 \ldots J$ representing the ordinal response variable values. An OF model is constructed as a regression forest utilizing this generated score set, according to the measure of the out-of-bag function, also named performance function. Subsequently, several additional randomly generated candidate score sets undergo the aforementioned steps. The final score set is then computed as the summary of the candidate score sets that exhibited the highest out-of-bag function. Finally, an OF is constructed with this optimized score set.

Given its novelty, this method is not yet widely available in many machine learning packages. However, a detailed explanation of this model's hyperparameters is provided in [80]. As it is closely related to a regression forest, the actual construction and associated hyperparameters are not included. Nevertheless, there are certain hyperparameters which ought to be fine-tuned to achieve optimal performance from the model. One such parameter is the ntreefinal parameter which corresponds to the number of trees in the final OF. Additionally, the nsets parameter, which denotes the number of score sets to

be attempted prior to the optimization of the score set, and the nbest, which relates to the number of score sets used to compute the optimized score set, can also be adjusted. Moreover, the performance function can also be fine-tuned as this algorithm offers a range of alternative functions.

In a study conducted by [79], the OF classifier outperformed its competitors in terms of prediction performance utilizing five distinct datasets already used in a preceding study [81]. Furthermore, the model enables the analysis of covariate importance, facilitating the discrimination between influential and noise covariates. The algorithm is also suitable for low and high-dimensional data, proving to be versatile for various applications. However, this algorithm may be computationally expensive and infeasible when used with high-dimensional data, particularly when employing conditional inference trees. OF were subject to a study in [82], in which they were extended to respect the data's ordinal scale without assigning artificial scores.

4.5 K-NN classifier

The K-Nearest Neighbors (K-NN) algorithm is a supervised machine learning method which can either be used for classification or regression. The algorithm is based on a straightforward concept: classification is achieved by identifying the nearest neighbors to a given query and using them to determine its class. This classifier is included in the lazy learning algorithm category, meaning that computation is only performed when the system is queried. As the training samples are only needed at runtime, this algorithm can also be referred to as Memory-Based Classification [83]. Figure 4.5 ilustrates the main idea of the K-NN algorithm.

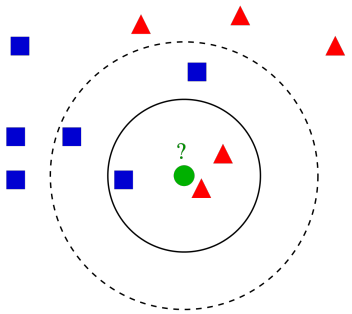


Figure 4.5: K-NN representation for $K = 3$ and $K = 5$. Image taken from [84]. For K=3, test sample (green circle) would be classified as a red triangle and for K=5 as a blue square.

The K-NN classification comprises of two phases: the computation of the K nearest neighbors and the subsquent classification of data points using those neighbors. In the initial phase, a variety of distance metrics may be employed including the Euclidean or the Manhathan distance metrics. Alternatively, similarity metrics like the cosine similarity or the correlation, can also be utilized, particularly in suitable contexts [83]. Furthermore,

distance weighted voting can be incorporated in the second phase, whereby the K nearest neigbors are assigned greater weight by inversing their distance to the query point.

A downside of K-NN is its computional time, which degrades as the dataset scales up. Additionally, the choice of distance metric can significantly impact the algorithm's efficiency, which is the case when using the Earth Mover's Distance (EMD) [85]. To improve the runtime of this algorithm, several structures can be employed as strategies including the use of Kd-Trees and Ball trees instead of the algorithm's naive approach of brute forcing through the dataset [83].

To fine-tune K-NN, the crucial hyperparameter for the matter is the desired number of nearest neighbors, K. This number can be predicted through several methods such as cross-validation, using the Leave One Out Cross-Validation (LOOCV) variant for the matter, and the elbow method, which involves plotting the error rate against different K values and selecting the optimal value based on the point at which error rate begins to level off [86]. Furthermore, evolutionary algorithms can also be employed to predict the appropriate number of K for the problem at hand [87]. It should be noted that other hyperparameters may also exert an influence on the algorithm's performance including the distance measure employed, as previously discussed. The Minkowski distance measure allows for the calculation of multiple distance measurements by varying the value of p. When p is set to 1, the resulting distance measurement is the Manhatthan distance, while p = 2 leads to the Euclidean distance. As p increases, i.e., $p \mapsto \infty$ the distance measurement approaches the Chebyshev distance [88], an extreme case of the latter.

Overall, the K-NN algorithm is characterized by a transparent process and an easily analyzable output, which are benefitial in a variety of contexts. Its simplicity makes it an appropriate algorithm for many classification tasks. However, as the computation is performed at runtime, K-NN may not perform optimally with large datasets. Additionally, this algorithm is highly sensitive to redundant features, which can negatively impact its performance.

In the field of machine learning, this algorithm is not only employed for classification and regression tasks; it can also be utilized for data preprocessing purposes by assigning a value to missing values in datasets, as this is a common occurence [89]. Similarly to a RF, it can be used as a recommendation engine [90], classifying users into specific groups based on their behaviour and subsequently recommending additional content. In the financial sector, K-NN was used to assess the potential risk a bank might encounter when extending loans to an organization or an individual, employing a weighted version of the algorithm to generate a credit score [91]. The algorithm has also been applied for pattern recognition in text and also digital classification [92].

5

Proposed Experimental Methodology

5.1 Introduction

The proposed methodology has been designed to answer the main questions of this dissertation: Are Wind Stress Anomaly (WSA) consistent with Upwelling Stability Period (USP) found by the core-shell clustering framework?; Can the USP be predicted by WSA data?

According to this we developed an experimental methodology with the following main steps:

- Extraction, preprocessing and feature construction of WSA maps from wind reanalysis products;
- Segmentation of WSA maps through an unsupervised clustering algorithm that automatically fine tunes the number of clusters to retrieve from data;
- Construction of new WSA datasets from the wind stress anomaly clusters, and labelling[1] of the WSA data with the USP values;
- We apply state of the art fuzzy membership functions to map the daily WSA data into the USP (time-ranges).
- Aplication of state of the art classifiers to predict USP from the labelled WSA datasets;

The chapter is organized as follows: Section 5.2 details the preprocessing steps taken since the collection of data until the building of datasets which will serve as input for the Iterative Anomalous Pattern (IAP) clustering algorithm. Section 5.3 explains the experimental setup of the clustering algorithm, along with the decisions taken to fine tune it. Section 5.4 ilustrates the steps present in the visualization procedure and section 5.5 the actions taken to build and label new WSA datasets from the segmented clusters. These datasets will serve as input for the classifier models whoose building process is detailed in section 5.6.

[1] a label is a category (USP value) that allows to differentiate our data

5.2 Wind data: collection, preprocessing and feature construction

5.2.1 Data collection

For the studies, the product chosen is the ERA5 dataset from the Copernicus datastore [93]. The region of analysis will be the North-Morocco (NM) (28°N-36°N;5.5°W-16°W) geographic region and the South-Morocco (SM) (South coast: 20°N-27°N;13°W-21°W), the same explored geographic regions as in [5]. The product type is Reanalysis. The dataset includes the hourly records of the 10m u-component and 10m v-component of wind (measured in meters per second) from 2004 to 2019. The file format is *.grib*.

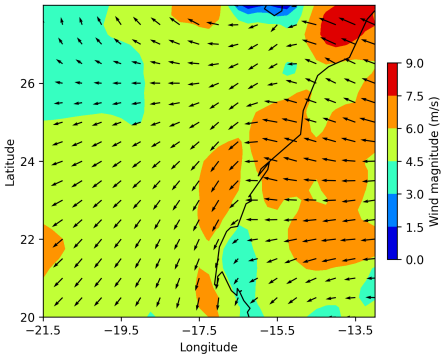


Figure 5.1: Wind plot from 00:00 UTC 01/01/2019-South Morroco

In Figure 5.1 is plotted the first timestamp of the SM geographic region for year 2019, where the black arrows represent the speed vector (u and v components combined), and the color the wind magnitude.

5.2.2 Data preprocessing

After the wind maps collection they enter a preprocessing pipeline organized in 5 steps:

1. **Data cleaning:** To eliminate unnecessary data, the grid position values corresponding to inland Morocco were replaced by NaN (not a number) values as they do not contribute to the phenomenon under study.

2. **Data aggregation:** Given the hourly nature of the datasets, they must be aggregated by averaging them over a period of 24 consecutive hourly wind maps. This process transforms the hourly dataset into a daily wind map dataset.

3. **Computing the wind component parallel to the coast:** In this stage the datasets refering to the u and v components are merged, computing the *alongshore component* with the formula [94]:

$$u \cdot \cos(-\theta) - v \cdot \sin(-\theta), \tag{5.1}$$

where u and v are the respective wind components and θ the average coastline difference in the whole geographic region.

In our study, $\theta = 55°$, the same used in [95], the negative value comes from the coastline orientation as it is in the opposite direction. Using this formula we can compute the actual wind speed contributing to upwelling. To support the visualization, this *alongshore component* is decomposed back to the u and v components which is done with the help of a function from the metpy library [96]. As the the function takes as input the wind direction, we need to compute the required coastline angle. Due to the irregularities of the Moroccan coast, this angle tends to differ. To then obtain more accurate results, we determine the coastline angles in certain latitude ranges that will have similar coastline angles. Using Google Earth's tools [97] we determined the following angles for the South Moroccan coast:

Starting Latitude	Ending Latitude	Coastline Angle
20°N	20.5°N	144°
20.75°N	22.25°N	190°
22.5°N	24.5°N	210°
24.75°N	26°N	200°
26.25°N	26.75°N	240°
27°N	28°N	204°

Table 5.1: Coastline angles for the SM geographic region coast

The angles used for the NM geographic region coast are present in Table B.1 of appendix B.

From this point forward, the datasets to be worked with will be referent to this component.

4. **Data interpolation:** Due to different spatial resolutions between Sea Surface Temperature (SST) grids and wind maps, the latter must to be interpolated to the spatial resolution of the former. This is achieved through the use of two Python libraries, NumPy [98] and SciPy [99]. First, a new linear space is created with NumPy's function *linspace*, to obtain the same interval between grid points in the wind dataset as the core-shell upwelling SST segmentations one, while maintaining the original range of longitude and latitude. Subsequently, the actual data is interpolated to this newly created linear space with the help of SciPy's function *griddata*. Due to inherent difficulties in interpolating NaN values within the original dataset, these values are transformed into a sentinel value and replaced by NaN values again when interpolated.

5. **Wind stress anomaly computation** As described in section 3.1.5, wind stress is the frictional force, per unit, of area, acting on the sea surface as a result of the wind blowing over it. Wind stress is defined by equation 3.1. To avoid excessive computations, the equation can be simplified to:

$$\tau = w^2, \tag{5.2}$$

as c can be considered as a constant in our context according to the domain expert.

Given previous experiments of clustering wind intensities, clustering their square would lead to similar results. It was then decided to cluster wind stress anomalies, given by the formula:

$$\tau = (\overline{w} - w_{ij})^2, \tag{5.3}$$

where $\overline{w}$ is the average wind speed of the current wind map W and w_{ij} the wind speed at grid point ij. Wind map W is defined as $W(I, J) = (w_{ij})$ with wind speed values w_{ij} at point ij, where i ($i = 1, 2, \ldots, I$) is the longitude and j ($j = 1, 2, \ldots, J$) the latitude and $\overline{w}$ the average of the $|I| \times |J|$ w_{ij} values. This WSA will reach its highest values when the wind speed intensity at point ij is significantly higher or lower than the grids's average wind speed. So, a WSA can be defined as "how much does the actual wind speed intensity differs from the average wind speed in the region of analysis".

6. **Moving average procedure:** This procedure is applied to set the temporal scale of the wind map segmentations as equal to the core-shell upwelling SST segmentations. The procedure starts by aggregating the daily wind maps into weekly wind maps considering an 8 day week [6], then applying a sliding window technique to them. Applying this will bring benefits, such as flexibility as we can change the window size and the possibility of detecting local trends and anomalies.

   A window with size W ranging from 2 to 25 was studied to determine which size kept the better temporal resolution in the data, while also smoothing it. As expected, a bigger window size will bring lower variability between the wind maps, implying a loss of temporal resolution. A smaller window size, although preserving temporal resolution, does not smooth the data with enough quality. After observing this in the study, is chosen $W = 5$, the same used in [6]. In Figure 5.2 it is plotted the overall mean average wind speed values, along with their correspondent standard deviations, for the chosen size.

   It was also taken into consideration maximizing the *alongshore component* however, the visualized segmentations by averaging instead of maximizing when applying the sliding window technique were superior and a better picture of the spatial evolution of the *alongshore component* was obtained.

Figure 5.3 ilustrates the preprocessing pipeline steps apllied to the first timestamp of the daily wind dataset of the SM geographic region for 2019, outputting a daily WSA map:

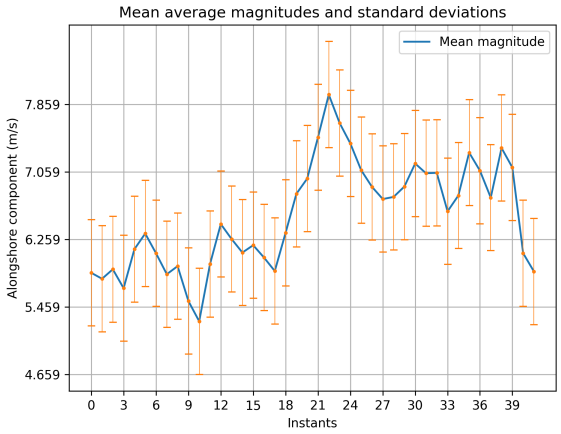


Figure 5.2: Mean and standard deviations results of the sliding window study for average weekly values

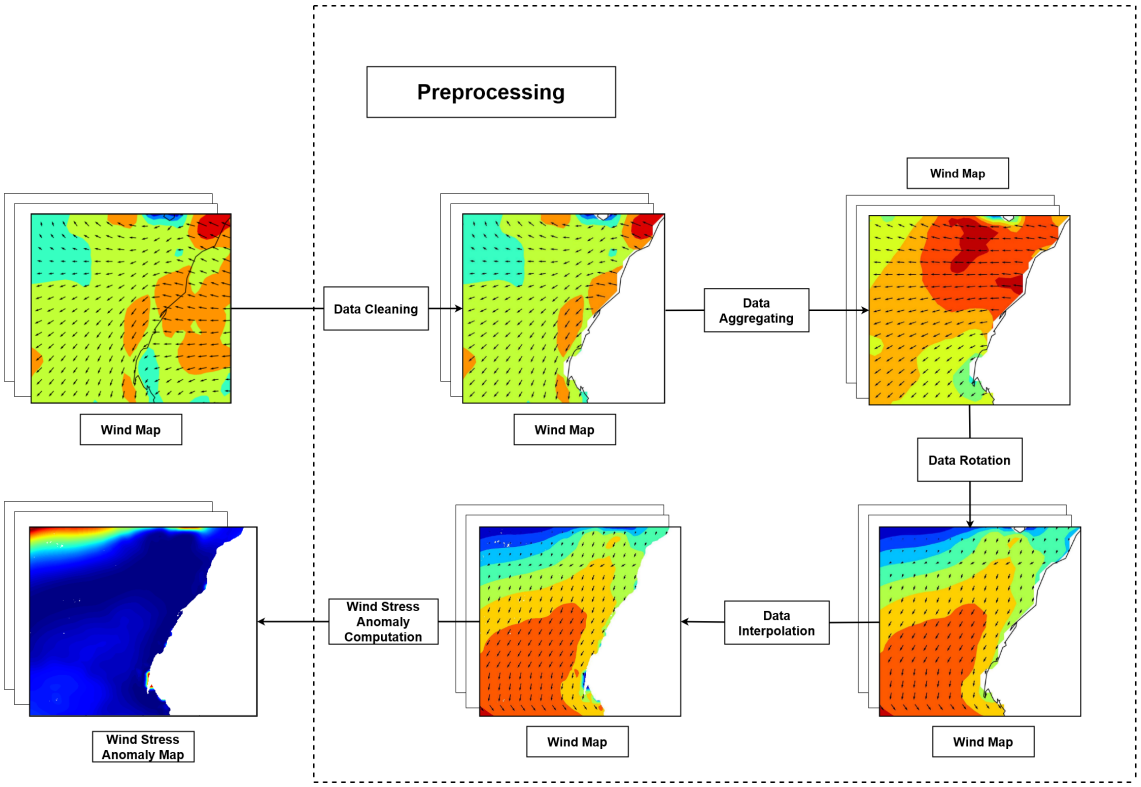


Figure 5.3: Visualized steps of the preprocessing pipeline-First timestamp-2019-SM geographic region

5.2.3 Wind stress anomaly maps

A WSA map is the final result of applying the full preprocess pipeline described in section 5.2.2. This WSA map is defined as a spatial grid $G(I, J) = (g_{ij})$ with WSA values g_{ij} at point ij, where i ($i = 1, 2, \ldots, I$) is the longitude and j ($j = 1, 2, \ldots, J$) the latitude. Each of these maps will then serve as input to the IAP clustering algorithm to then extract features from the obtained WSA cluster segmentations in order to build the datasets required

for the experimental study. As a first step, each WSA map is going to be normalized by shifting the origin of the data to its grand mean and scaled to the WSA value range to facilitate the algorithm's computation, done with the following formula:

$$a_{ij} = \frac{g_{ij} - mean_G}{max_G - min_G}, \tag{5.4}$$

where a_{ij} is the normalized WSA value at point ij, g_{ij} the original value at point ij, $mean_G$ the average value of the $|I| \times |J|$ g_{ij} values ($i = 1, \ldots I; j = 1, \ldots, J$), and max_G and min_G the maximum and minimum values of the $|I| \times |J|$ g_{ij} values ($i = 1, \ldots I; j = 1, \ldots, J$), respectively.

Figure 5.4 ilustrates on the left the original WSA map and to the right the result of applying normalization equation 5.4. From this figure, the normalized WSA map presents a much clear visualization of the behavior of the WSA value by showing different shapes given by different shades of blue, otherwise impossible to dettect with the original WSA map.

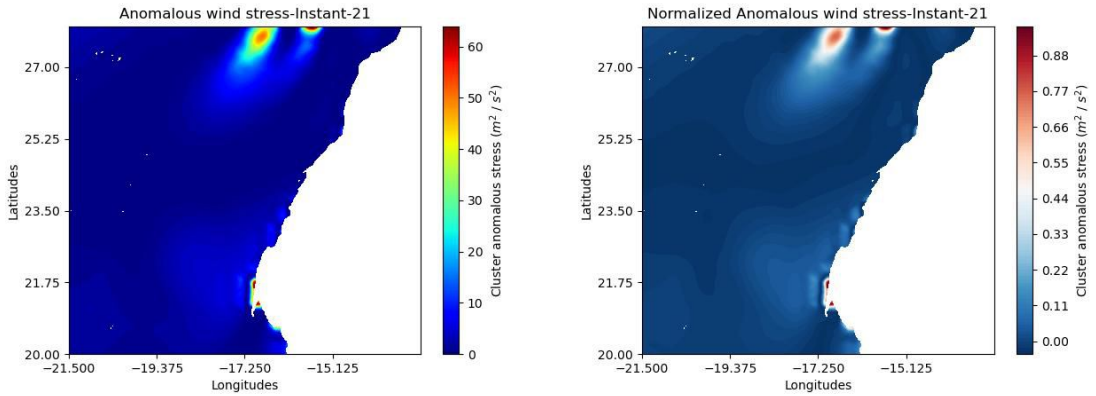


Figure 5.4: Wind stress anomaly map (left) and corresponding map normalized (right)-2019-SM geographic region-First timestamp

5.3 Wind stress anomaly maps segmentation through Anomalous Clustering

As described in section 4.2, the IAP algorithm takes as input an entity-to-feature matrix with N rows and D columns, where N is the number of rows or entities and D the number of features. Since we pretend to segment daily WSA maps, where each one is a spatial grid G, as defined in 5.2.3, we need to transform these spatial grids into the feature-to-entity matrices required. Given this, the experimental setup of applying IAP is composed of two phases described below:

- **Data flattening**: Grid G has to be reshaped to the required entity-to-feature matrix. As we are only clustering one feature, the WSA value, this reshape operation can be

described as a flattening operation. With this procedure, spatial grid G is reshaped to a column vector I, with the shape $n \times 1$, where n is the $|I| \times |J|$ number of WSA values of grid G.

- **Fine-tuning the stop condition**: To fine-tune the IAP stop condition, we consider different values of K=2,3, then analyzing the clustered segmentations and deciding which K number of clusters better segments the data, thus utilizing stop condition (S4) of the ones described in section 4.2.

Following the IAP's execution, the remaining WSA points are assigned to the cluster whoose distance to its prototype is minimum.

With the clusters obtained, these are ordered according to their prototypes (average WSA value) in ascending order designated by cluster 1, 2 and 3, with a higher cluster designation meaning a greater prototype.

5.4 Visualization of clustered wind stress anomalies

The visualization of the clustered WSA maps follows a mapping procedure so that the final result is the clustered WSA values mapped over a spatial grid (longitude × latitude).

From the input dataset, column vector I is reshaped back to grid's G original shape. Each grid point ij is then assigned its corresponding WSA value g_{ij} and cluster label L= 1,2,3.

A unique color is assigned to each WSA cluster label L and grid point ij is colored according to its assigned cluster label corresponding color, color$_L$. This mapping process allows for a clear visualization of the clustered results, where each label L is visualized by a distinct color and the cluster's spatial location can be easily interpreted.

Figure 5.5 ilustrates the result of applying this mapping procedure:

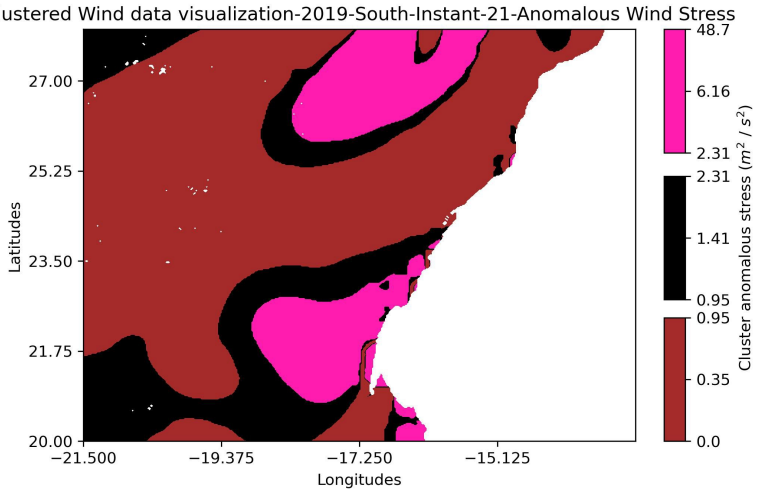


Figure 5.5: SST instant 21 (9th June-18th July) corresponding WSA map IAP segmentation visualization-2019-SM geographic region

From the figure above, cluster label 1 is colored in brown, cluster label 2 in black and label 3 in pink. The color scales are ordered according to the WSA cluster's prototypes.

We also want to visualize the clustered WSA maps over the core-shell upwelling SST segmentation maps. The purpose of this is to inspect the WSA clusters over the core-shell upwelling SST segmentations as a ground-truth. For that, we apply the moving average procedure described in section 5.2.2 to the daily WSA maps to be transformed in weekly average maps just like the SST instants.

Visualizing now the segmented results over the upwelling region SST segmentation maps involves the WSA clusters being visualized in a different spatial grid, U, with same shape as the original grid G. To avoid color mixing, as each grid point ij of grid U is already assigned a color, it is plotted the original wind speed vector over this spatial grid, colored according to the WSA clusters' color map defined in the mapping procedure. This step is done with help of Python's library matplotlib [100] function quiver.

Figure 5.6 ilustrates the result of this visualization process over a core-shell upwelling segmentation SST map. The upwelling core is in orange and the shell in green. The arrows represent the original wind speed whoose colors correspond to the colored label clusters (first color scale on the right).

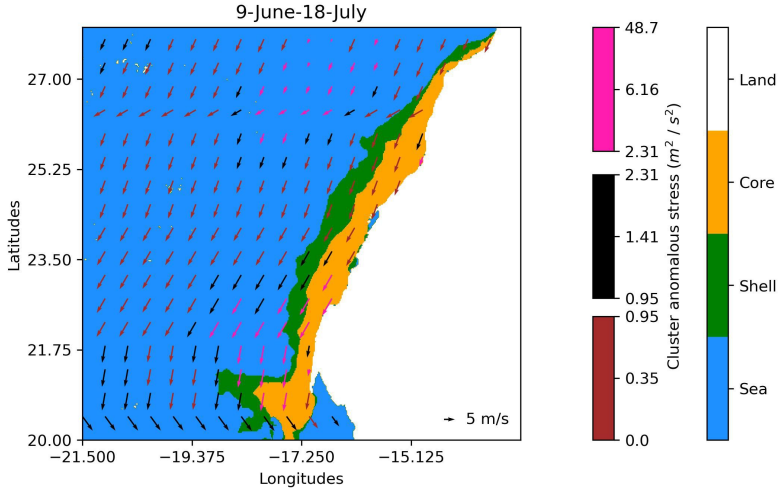


Figure 5.6: SST instant 21 (9th June-18th July) corresponding WSA map IAP segmentation visualized over SST instant 21 core-shell upwelling SST segmentation-2019-SM geographic region

5.5 Construction of labeled wind data sets

5.5.1 Feature extraction from clustered wind stress anomaly data

From the each segmented daily WSA maps are extracted two features from each of the three clusters: the cluster's average WSA and maximum WSA, both in m^2/s^2 in a total of 6 features. With this we construct an entity-to-feature data matrix considering these

6 features. Consider as an example the clustered WSA map in Figure 5.6, the obtained features are marked on the first color scale on the right.

To force the temporal order of the new WSA data it is added an additional feature to each daily WSA map that is the number of the corresponding month.

Clustering the daily WSA maps of an year and geographic region, leads to a construction of a dataset with the shape of 365×7. So, the constructed clustered WSA datasets are construced as an entity-to-feature 365×7 data matrix for each geographic region. The entities of this *region-year* dataset are the daily clustered WSA maps characterized by the extracted cluster features and the map's corresponding month. In Table 5.2 it is presented in a tabular manner a representative sample of this *region-year* dataset.

Average WSA Cluster 1	**Maximum WSA Cluster 1**	**Average WSA Cluster 2**	**Maximum WSA Cluster 2**	**Average WSA Cluster 3**	**Maximum WSA Cluster 3**	**Month**
0.34	0.96	1.38	3.11	4.93	16.13	1
...	...	...	...	...	...	...
0.23	0.70	1.37	2.63	5.89	44.48	2
...	...	...	...	...	...	...
1.13	3.75	5.85	9.18	14.39	37.8	3
...	...	...	...	...	...	...
0.25	0.66	1.14	2.66	5.44	32.68	12

Table 5.2: Ilustration of a sample of the dataset obtained after applying IAP to daily WSA maps of the SM geographic region of 2019 and extracting the desired features

This process is then applied to each region and year combination for the intended study, totalling in 6 datasets to be labbeled. These datasets will be named according to their respective region and year with the format *region-year*.

5.5.2 Labelling wind stress anomaly data set with upwelling stability period

It is now pretended to label the daily *region-year* clustered WSA datasets with the corresponding USP. As described in section 4.1, an USP is a sequence of SST instants characterizing a period of upwelling stability whoose core-shell cluster represents the constant upwelling region (the core) and the shell the variable one.

In Figure 5.7, the USP are colored in green, the SST instants in orange and the weeks in blue. As observed in the figure, there is an intersection between USP, making it impossible to accurately label the daily WSA maps present in this interval intersection according to the USP.

Take day 180 (28th June) as an example, this daily WSA map not only belongs to SST instants 20 to 23, but also belongs to USP 2 and USP 3. So to label each entry of the *region-year* WSA clustered dataset it is employed a fuzzy memebrship function.

The fuzzy membership function to be used in this case is a trapezoidal function of the general format [101]:

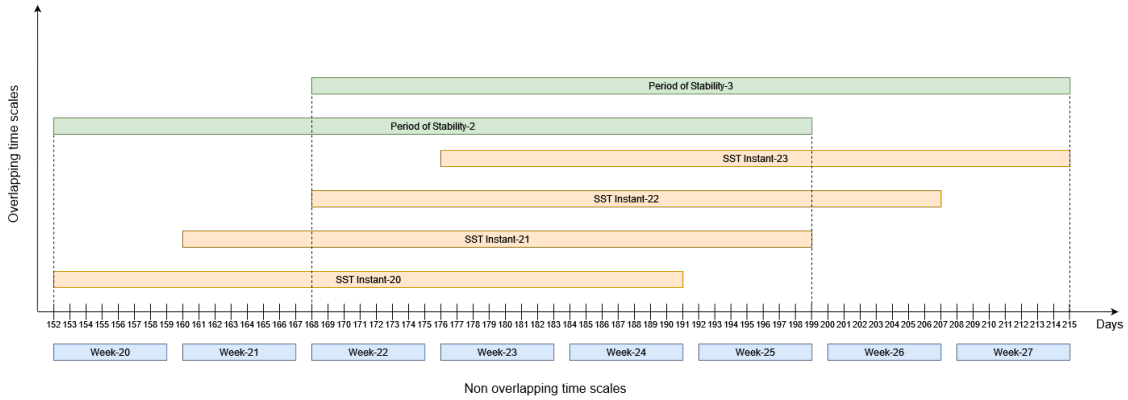


Figure 5.7: Overlapping days between instants

$$\mu(x) = \begin{cases} 0 & \text{if } x < a \\ (x-a)/(b-a) & \text{if } a \leq x < b \\ 1 & \text{if } b \leq x < c \\ (d-x)/(d-c) & \text{if } c \leq x < d \\ 0 & \text{if } x > d \end{cases}$$

where a,b,c,d are the parameters of the function.

The fundamental concepts of a trapezoidal function are as follows [102]:

- Support: Set of elements with non-zero degree of membership:$[a-d]$.
- Core: Set of elements with degree of membership of 1:$[b-c]$.
- α-Cut: Set of elements with degree of membership greater then α.
- Height: the maximum degree of memebership, 1 in our case.

5.5.2.1 Computing the fuzzy membership function parameters

The discrepancies in the USP observed across regions and years will result in disparate membership functions. In light of these challenges, it is necessary to compute the four parameters of each fuzzy membership function.

To this end, a simple mapping procedure was developed. This procedure starts by computing the support of the membership function, parameters a and d, by mapping the days belonging to each SST instant. This is achieved by decomposing the SST instants into the weeks used for computation and then decomposing the weeks in days. Subsequently, the first day of the first week and the last day of the last week are retrieved.

Let us take as an example SST instant 1:

SST Instant 1: $\mapsto$ weeks: 1 to 5 $\mapsto$ days: 1 to 40 (1st January - 9th February)

The days are then mapped to actual calendar days and then to the corresponding daily WSA map, but for the sake of the example those steps are left out.

To define the duration of a USP in days, we take the first and last instant and apply the procedure to compute the first day of the first week (belonging to the first instant) and the last day of the last week (belonging to the last instant).

An example with the first USP for the geographic region of SM for 2019, ranging from SST instant 1 to 14:

SST Instant 1: $\mapsto$ weeks: 1 to 5 $\mapsto$ days: 1 to 40 (1st January - 9th February)

SST Instant 14: $\mapsto$ weeks: 14 to 18 $\mapsto$ days: 105 to 144 (14th April - 23rd May)

Support of USP: 1 – 144 (1st January - 23rd May)

To compute the core of the membership function, parameters b and c, some aspects of set theory are employed, eventhough its sequences being worked. The core of the membership function is computed by applying the set difference between the membership functions' supports from the current USP and the next USP. Let us consider the next USP of the same region and year, ranging from SST instants 15 to 21. In this case its membership function support extends from day 113 to day 200.

Applying the difference to the two support sets:

Support of the current USP: 1 to 144 (1st January - 9th February)

Support of the next USP: 113 to 200 (22nd April - 28th July)

Core of the current USP: 1 to 144 – 113 to 200 $\mapsto$ 1 to 113 (1st January - 22nd April)

With this, the membership function for the first USP regarding the SM geographic region for year 2019 is defined with the following parameter set: $\{a = 1, b = 1, c = 113, d = 144\}$

In Figure 5.8 below is visualized the defined function.

As this USP is the first one, its membership function's core and support start on the same day. The vertical dashed line indicates that same start.

Applying this procedure to the complete year of 2019, we will get the set of trapezoids represented in Figure 5.9. Where the dashed lines mark the start and the end of the membership function core of the respective USP. Between the end of the support set (descending slope) of one USP and the beggining of the support set (ascending slope) next USP there is an overlapping interval between fuzzy membership functions. The USP assignement to the days contained in this overlapping interval is going to be explained in section 5.5.2.2.

5.5.2.2 Constructing wind stress anomaly data sets labelled with upwelling stability period

With the mapping procedure in place, it is crucial to defuzzify the trapezoidal membership functions and then construct the dataset. To this intent, the most popular defuzzification

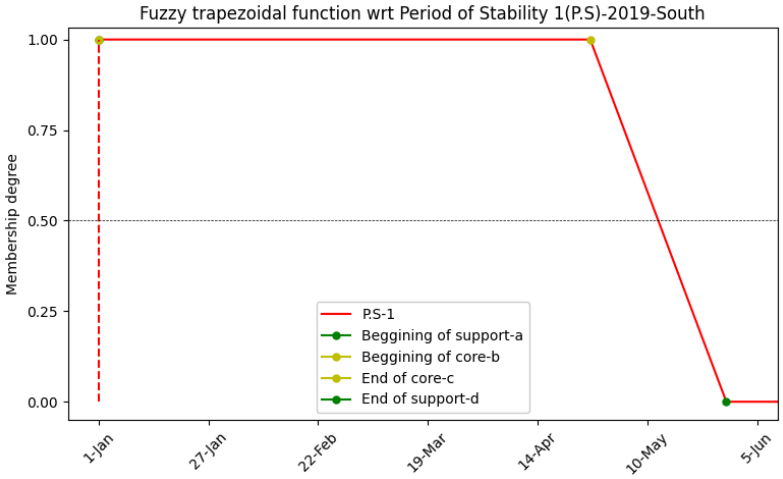


Figure 5.8: Fuzzy trapezoidal membership function w.r.t USP-1-2019-SM geographic region

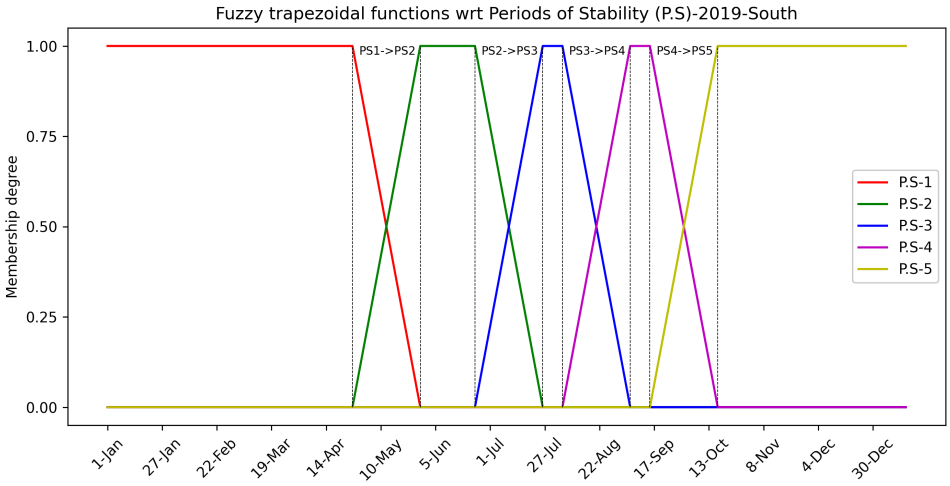


Figure 5.9: Fuzzy trapezoidal functions w.r.t USP-2019-South

functions will be tested [103]: Center of Gravity (COG), Middle of Maxima (MOM), Largest of Maximum (LOM), Smallest of Maximum (SOM).

Center of gravity: Computes the center of gravity under the membership function and the output is selected as the deffuzification value.

$$\text{COG}(\mu(x)) = \frac{\sum_{x_{min}}^{x_{max}} x \cdot \mu(x)}{\sum_{x_{min}}^{x_{max}} \mu(x)}, \tag{5.5}$$

where $\mu(x)$ represents the membership function.

Mean of Maximum: The middle element from the membership function's core is selected as the deffuzification value.

$$\text{MOM}(\mu(x)) = \frac{\sum_{x_i \in \text{core}} x_i}{|\text{core}|}, \tag{5.6}$$

Smallest of Maximum: The smallest element from core of the the membership function is selected as the deffuzification value.

$$\text{SOM}(\mu(x)) = \min \text{core}(\mu(x)), \tag{5.7}$$

where $\mu(x)$ represents the membership function.

Largest of Maximum: The largest element from the membership function core is selected as the deffuzification value.

$$\text{LOM}(\mu(x)) = \max \text{core}(\mu(x)), \tag{5.8}$$

where $\mu(x)$ represents the membership function.

Let us go into more detail of a USP trapezoidal membership function, Figure 5.10 shows the output of all of these functions applied to USP 2 of 2019 for the SM geographic region.

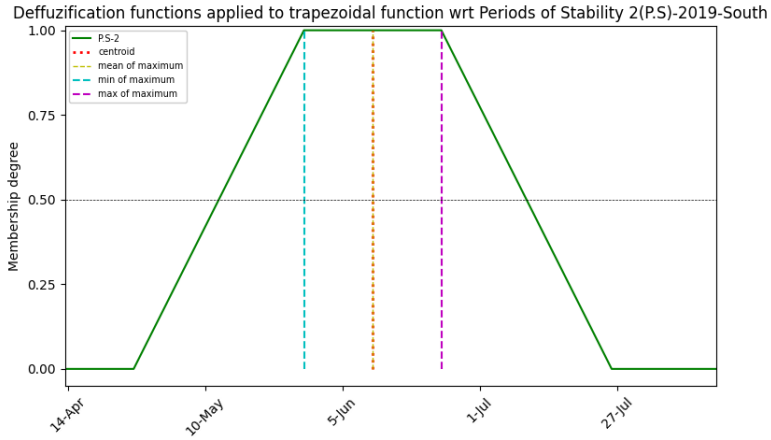


Figure 5.10: Deffuzification functions applied to USP-2-2019-SM geographic region

Consider now the COG function (centroid) and let us apply to the rest of the current year's USP, represented by Figure 5.11.

From Figure 5.11, each USP gets computed a *center of gravity*. To now construct the dataset, each daily WSA map gets classified according to the *center of gravity* is closer to. After defuzzifying the corresponding year and region with each of the defuzzification functions, four datasets are constructed which will be used as input for the classifier models of the experimental study and evaluated to determine which defuzzification function is most appropriate for the task at hand. Table 5.3 is a sample of the dataset generated by applying the COG defuzzification function:

Let us inspect the correlation each feature has with the USP class attribute, Figure 5.12 presents this in the form of correlation matrices for the geographic regions of NM and SM:

From a brief observation, we can see different behaviors between the NM and SM geographic regions, which perpetuate the fact that these two geographic regions should be looked at separately. An example of this is feature `Max_Wind_Stress_C3`, which in the geographic region of NM is positively correlated with the features: `Avg_Wind_Stress_C1`,

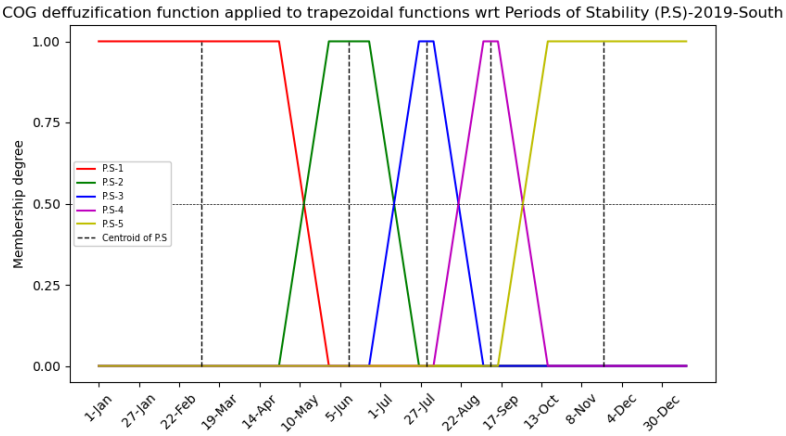


Figure 5.11: Center of gravity deffuzification function applied to 2019-SM geographic region

Avgerage WSA Cluster 1	Maximum WSA Cluster 1	Avgerage WSA Cluster 2	Maximum WSA Cluster 2	Avgerage WSA Cluster 3	Maximum WSA Cluster 3	Month	Class
0.29	0.99	1.69	2.67	8.73	72.72	4	1
0.43	1.49	2.79	4.91	10.94	48.67	4	1
0.36	1.60	2.50	3.62	7.17	36.98	4	2
0.52	1.62	2.65	4.97	12.19	28.59	4	2
0.50	2.10	4.36	12.72	21.69	30.04	4	2

Table 5.3: Ilustration after applying the COG deffuzification function to 2019-SM geographic region

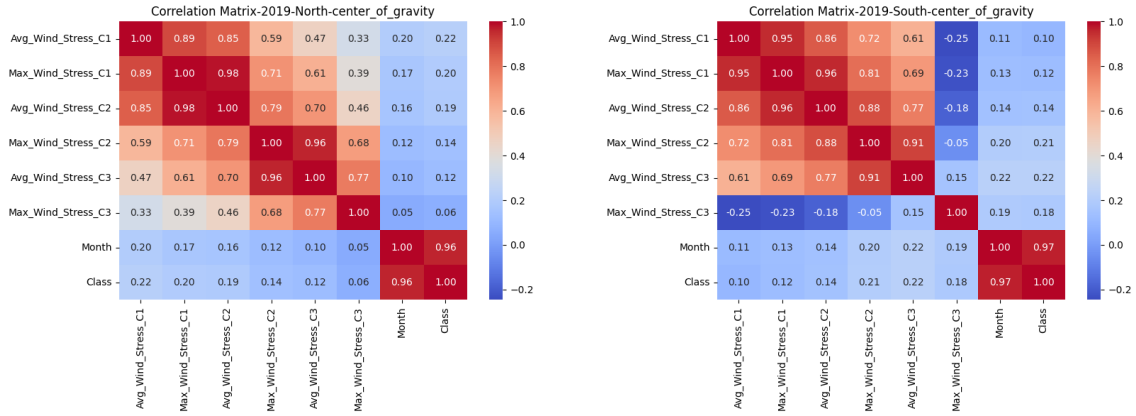


Figure 5.12: Center of gravity feature-class correlation matrices for NM and SM geographic regions-2019

`Max_Wind_Stress_C1`, `Avg_Wind_Stress_C2`, `Max_Wind_Stress_C2` and in SM it is negatively correlated with the same features.

A common trend observable in the correlation matrices is that none of the features are correlated with the class attribute, with the exception of the feature `Month`. This feature exhibits this behavior due to its data type, as temporally, the month of the year is presented

in ascending order (1 < 2< 3 < 4 . . .), as well as the USP class (1<2<3 . . .). Therefore, this temporal feature helps to guarantee the sequence of the data.

The four defuzzification functions will be applied to the previous 6 clustered WSA datasets, generating in total 24 new labeled datasets (12 per geographic region). The naming of these datasets will follow the structure:

<geographic_region>-<year>-<defuzzification_function>,

where the geographic region contians the values: NM and SM; the year is one the three years utilized for the study: 2007,2015, and 2019; the defuzzification_function: COG, LOM, MOM, SOM.

5.6 Random Forest/Ordinal Forest/K-NN experimental setup

To predict USP, independently of the model tested, each classifier will follow a general model building protocol composed by three stages, visualized in Figure 5.13 and described below.

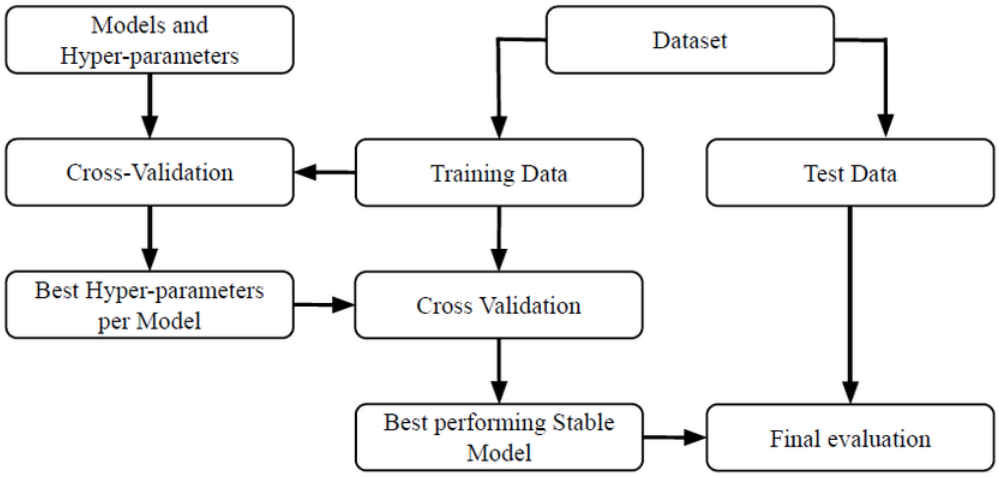


Figure 5.13: Model building protocol workflow. Image taken from [104]

- **Data split:** Let X be a 365 × 7 entity-to-feature data matrix with $N = 365$ rows and $V = 7$ columns. When splitting X, it will be splitted following a 80% training and 20% testing split, stratified by the class attribute USP in order to maintain the original USP class proportions. After this split, both training and testing datasets are standardized following range normalization, where the origin of each feature,v, is shifted to the feature's average and then scaled to its range. This is described by equation 5.9.

$$y_{nv} = \frac{x_{nv} - mean_v}{max_v - min_v}, \tag{5.9}$$

where x_{nv} is the value of X at row n and feature v, $mean_v$, max_v, and min_v are feature's v average, maximum and minimum values, respectively.

Following the normalization of the training set, we are going to keep its original mean, maximum, and minimum values for each feature and apply the previous normalization equation to the test set with these values.

- **Train-Validation:** During this phase, the optimal model for each classifier will be determined according to the dataset used through a process of training and fine-tuning. In order to achieve this, a form of K-fold cross-validation will be employed. K-fold cross-validation [105] entails the partitioning of the input data into K specified groups or folds. From those folds, one is taken as the testing set of that iteration while the others are taken as training. The model under evaluation will be fitted to the training folds and evaluated on the testing folds, with these being distinct in each iteration. The final score for each model will be determined by computing the mean of the scores obtained across all iterations. The model with the highest overall score will be deemed the optimal one. Given the classification task at hand and the possibility of having an imbalanced dataset, stratified K-fold cross-validation is going to be utilized to ensure that every fold maintains the original USP class proportions.

  During this cross-validation cycle, the model's parameters are fine-tuned to ensure that no overfitting occurs. It is applied technique named *Grid Search*. *Grid Search* is a valuable tool for hyperparameter tuning due to its comprehensive approach to testing a predefined set of hyperparameters and identifying the optimal combination by applying cross-validation to each model tested. A significant benefit of this function is its time efficiency, which is achieved through the reduction in time when compared to manual fine tuning, despite its exhaustive search [106]. This leads to an overall improvement in model accuracy and robustness, as well as a reduction of overfitting.

- **Model evaluation:** With the 20% of data reserved for testing, the optimal model is tested in order to evaluate the model's behavior on independent data. This phase will also serve as an evaluation phase for each defuzzification function utilized.

To evaluate the models performances, four different metrics will be used: Balanced accuracy, Precision, Recall, and the F1 score [107]. Balanced accuracy will give us an accurate measurement as in some cases, the dataset might be imbalanced, so taking advantage of this metric is appropriate for this problem. Regarding the precision and recall metrics, we are going to compute the macro average of both. As the F1 score generally evaluates the quality of predictions by being an harmonic mean between precision and recall [108], this metric will be the one with the most relevance when determining the optimal model both during the fine-tuning and the testing phases.

5.6.1 Random Forest

The Random Forest (RF) algorithm offers a wide range of user set hyperparameters which can control the individual structure of a tree, the structure and size of the forest or even its randomness. Thus the process of building an optimal RF model is achieved by obtaining an optimal tradeoff between low correlated but strong trees [74].

From the available hyperparameters, we will focus on fine-tuning only the ones controlling the trees and overall forest structure. Given our low-dimensionality data, we decided to not fine-tune parameters that directly relate to the trees' randomness. The subset of these hyperparameters is defined as follows:

- **Number of estimators**: Number of trees in the forest. The number of trees of a forest should be set sufficiently high in order to maximize the performance gain. Although theoretically it is benefitial to increase the number of trees [109], a tradeoff between model complexity and performance has to be achieved, due to the decreasing improvement obtained from adding more trees to a forest [74].

- **Max depth**: Maximum depth of the tree. This parameter can be defined as the longest path from the root node to a leaf node in the tree, limiting the growth of each tree. It is verified that as this parameter increases the models tend to lose their generalizing capacity [110] so it is ideal to find the optimal value so that this overfitting case does not happen.

- **Minimum samples split**: Minimum number of samples required to split an internal node. The trees' depth is also controlled with this parameter as it controls the number of observations needed to split a tree's internal node. A low enoguh value might lead to model overfitt, as more splits are made and the resulting nodes become purer, while, a bigger value leads to less splits being made and thus to model underfitting.

- **Minimum samples leaf**: Minimum number of observations to consider a node a terminal or leaf node. This hyperparameter goes hand-in-hand with the max depth parameter, as it also controls the depth of the built tree. Setting this parameter with a lower value leads to deeper trees, as more splits are made until reaching a terminal node. Setting it higher will not only reduce the trees depth but also reduces computational time as less splits are done.

The values to be tested for every hyperparameter are presented in Table 5.4.

Hyperarameter	Values
number of estimators	$300, 400, \ldots, 1000$
max depth	$10, 15, 20$
min samples split	$2, 3, \ldots, 7$
min samples leaf	$1, 3, 5, 7, 9$

Table 5.4: Hyperparameters to be fine-tuned

In total are evaluated 1440 hyperparameter combinations with 10-fold cross-validation, totalling 14400 fit operations. Once the optimal hyperparameters have been identified within this set, a new model is constructed using them. Training is then conducted using the entire training dataset, and the model is subsequently evaluated using the test set.

5.6.2 Ordinal forest

As it is a recently developed algorithm, it is not greatly optimized and it is not yet included in many machine learning packages, so we are going to change to R programming language for this model. To overcome this challenge, some changes had to be made to the train-validation phase. The grid space tested has to be drastically smaller given the high default values for some of the hyperparameters tested, which weight heavily on computational fit times. Also, it was utilized 5-fold cross-validation, instead of 10-fold, again due to the high computational time needed to train an Ordinal Forest (OF) model.

The hyperparameters to fine-tune are the following [111]:

- **ntreeperdiv**: Number of trees used in the smaller regression forests constructed for each of the *nsets* different socres being tested.
- **ntreefinal**: Number of trees in the OF.
- **npermtrial**: Number of permutations of the class width ordering to try for the second to the nsetsth score set tried prior to the calculation of the optimized score set.
- **nbest**: Number of best score sets used to calculate the optimized score set

Hyperarameter	Values
ntreeperdiv	25, 50, 100
ntreefinal	2500, 5000
npermtrial	500, 600, 700
nbest	5, 10

Table 5.5: Hyperparameters to be fine-tuned for the Ordinal Forest

The combination of these hyperparameters leads to 36 distinct OF models with 5-fold cross validation totalling 180 fit operations.

5.6.3 K-NN

For the K-Nearest Neighbors (K-NN) algorithm, there is also some changes needed to be done in the train-validation phase. For the cross-validation phase we tested three forms of cross-validation: 5-fold, 10-fold, and Leave One Out Cross-Validation (LOOCV). The difference between LOOCV [105] and the other forms of cross-validation is that this method fits the entire dataset into the algorithm, leaving only one sample for testing,

applying this for n iterations, with n being the number of observations in the dataset. From the study made, we decided to use LOOCV, as most of the times this form of cross-validation presented the lowest error but also due to the small number of rows in our datasets, so the computational time increase was not a problem. For each neighborhood number, K, the algorithm is run 15 times where then is averaged the LOOCV and the test error. The test error is computed by applying the procedure of the evaluation phase, being considered a preliminary test error. The optimal K is calculated by computing the minimum absolute difference between the two computed errors and in the case of draws, the selected K is the smallest one, as it is the one which outputs the simpler model.

The grid space used was the following:

Hyperarameter	**Values**
K	$2, 3, \ldots, 26$

Table 5.6: Hyperparameters to be fine-tuned for K-NN

The number of fits performed in the fine-tuning process is far greater than the other models, however, given the simplicity of the algorithm and due to its lazy learning nature [83] this algorithm becomes the fastest one out of the three to train, validate and fine-tune.

6

Experimental Study

This chapter is dedicated to the presentation and discussion the results of the experimental study. Sections 6.1 pertain to the data utilized in the study itself, delineating the source and overall characteristics of said data. Section 6.2 presents the results of applying the Iterative Anomalous Pattern (IAP) algorithm to the Wind Stress Anomaly (WSA) maps and the datasets obtained by applying it to daily maps. Section 6.3 presents a preliminary visual exploration of the datasets constructed in the previous section with regard to Upwelling Stability Period (USP) class proportions. Sections 6.4 to 6.6 present the experimental results of the study, initially by comparing the performance of each classifier with regard to each defuzzification function and subsequently by comparing the most effective classifiers with one another. The final section outlines the conclusions derived from this study.

6.1 Data type collections

In this study we are going to deal with two types of data: Wind maps collected from the ERA5 [93] datastore and Sea Surface Temperature (SST) grids obtained from a previous project [5]. These two data collections comprise the same 16 years (2004-2019) of the same geographic regions (North-Morocco (NM) and South-Morocco (SM)). Each individual wind map contains the u and v wind component in meters per second, with a spatial resolution of 25km and is of hourly nature. Each SST grid contains the SST measured in degrees celsius, with a spatial resolution of 2km and a weekly temporal resolution. Table 6.1 summarises the data collected for the study.

Data Collection	Source	Geographic Regions (Morocco)	Years	Variables	Spatial resolution	Temporal resolution
Wind maps	ERA 5 [93]	North:28°N-36°N;5.5°W-16°W South:20°N-27°N;13°W-21°W	2004-2019	u wind component (m/s) v wind component (m/s)	25km	Hourly
SST grids	Previous project [5]	North:28°N-36°N;5.5°W-16°W South:20°N-27°N:13°W-21°W	2004-2019	Sea Surface Temperature (°C)	2km	Weekly

Table 6.1: Data collections used in the experimental study

For the study at hand, we are going to use three years of data for both Morrocan geographic regions: 2007, 2015 and 2019, as these years already proved to be representative

data samples for the study of coastal upwelling [5, 6]. In total, per region and year, there are 8760 hourly wind maps and 46 weekly SST grids.

Figure 6.1 presents the first time stamp for both data type collections for the SM geographic region for 2019:

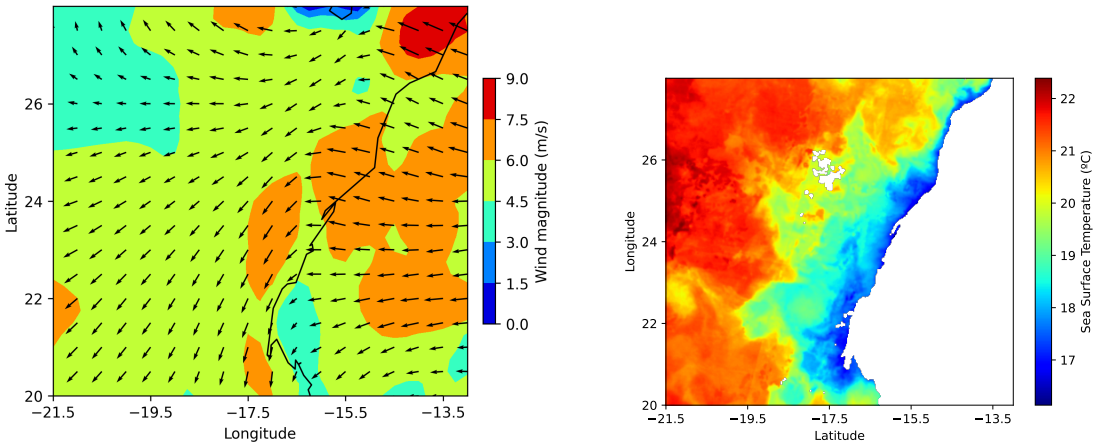


Figure 6.1: First time stamp for the wind map collections (left) and SST grids (right)-2019-SM geographic region

The wind maps are subjected to the preprocessing pipeline described in section 5.2.2 whoose steps are ilustrated in Figure 5.3. Applying the preprocessing pipeline to the 8760 wind maps leads to the construction of 365 new wind maps per region and year. These new wind maps are named WSA maps.

The SST grids are subjected to the core-shell clustering framework to then get the corresponding core-shell upwelling SST segmentations and also get the number of USP for each region and year. Tables 6.2 and 6.3 below presents the number of USP for each region and year along with the first and last SST instant of the USP. We can see that the SM geographic region presents more USPs than the NM geographic region.

	USP 1	USP 2	USP 3	USP 4
2007	1-11	12-22	23-42	-
2015	1-8	9-21	22-25	26-42
2019	1-11	12-20	21-39	40-42

Table 6.2: USP determined by the core shell framework for the NM geographical region for each year

	USP-1	USP-2	USP-3	USP-4	USP-5	USP-6
2007	1-18	19-23	24-32	33-42	-	-
2015	1-12	13-15	16-22	23-30	31-36	37-42
2019	1-14	15-21	22-26	27-31	32-42	-

Table 6.3: USP determined by the core shell framework for the SM geographical region for each year

6.2 Clustering wind stress anomaly data

After building the collection of WSA maps (365 per year and region), we need to determine how many clusters we intend to extract. It is applied IAP and fine-tuned its stop condition as described in section 5.3, where we first apply it to the WSA maps with the same temporal scale as the core-shell SST upwelling segmentations.

Let us take as an example the SM geographic region of 2019 whoose core-shell clustering framework result determines 5 USPs (see Table 6.3). Figure 6.2 presents on the left column the sequence of SST instants 20,21 and 22 with SST instant 21 (the last of USP 2) the transition instant between USP 2 and 3 of the aforementioned year. From left to right it is shown the SST instant, the corresponding SST upwelling core-shell segmentation, and the WSA segmentation maps for 2 and 3 clusters, respectively.

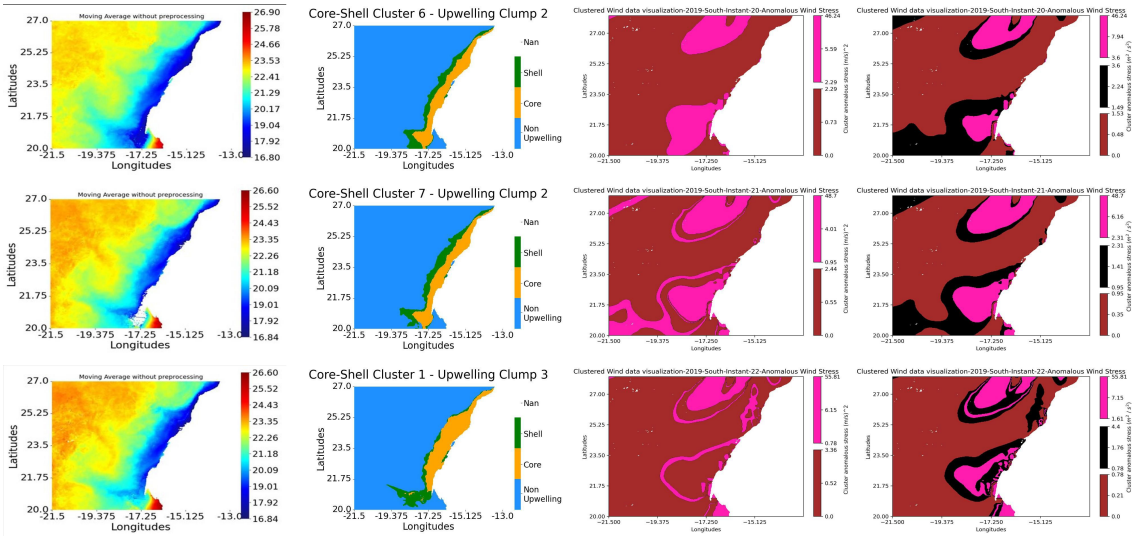


Figure 6.2: Visualized SST instants and core-shell SST segmentations along with the corresponding WSA maps segmentations for 2 and 3 clusters-SST instants 20 to 22-2019-SM geographic region.

The three clustered instants shown allow the observation of a change in the WSA segmented areas for 3 clusters. From the final SST instants of USP 2 corresponding WSA maps, instant 20 to 21, there is a gradual spatial change in which the pink cluster enlarges and the brown cluster in the southernmost part seems to be closing in the area slightly up north. In the following instant, SST instant 22, there is an abrupt spatial change where not only the brown cluster closed in, filling a vast majority of the segmented WSA map but is also present a segment up north of the black cluster. This sudden change coincides with the start of the new USP.

In the WSA segmentation maps for 2 clusters, this observation is not as clear, as the brown cluster does not present the same evolving behavior. For instance, in the WSA segmented map for 2 clusters corresponding to SST instant 20, the brown cluster is filling the majority of the map. In the following SST instant, it presents a similar spatial location

to the brown cluster of the 3 cluster segmentation and in the last SST instant it fills again the majority of the map. Also, in the WSA map corresponding to SST instant 22, the brown cluster suddenly takes over the southernmost part of the pink cluster area, which in the previous WSA maps corresponding to the previous SST instants was not showing signs of happening. Thus a two-cluster segmentation approach does not output as much information as a three-cluster segmentation approach.

This type of observation was consistent on the two geographic regions along the years. Additionally, by segmenting WSAs with 3 clusters we extract 6 features (average and maximum WSA of each cluster), advantageous to build the classification models. Therefore a three-cluster segmentation approach may offer a more precise delineation of WSAs.

Figure 6.3 ilustrates these changes, in the IAP and core-shell SST segmentations.

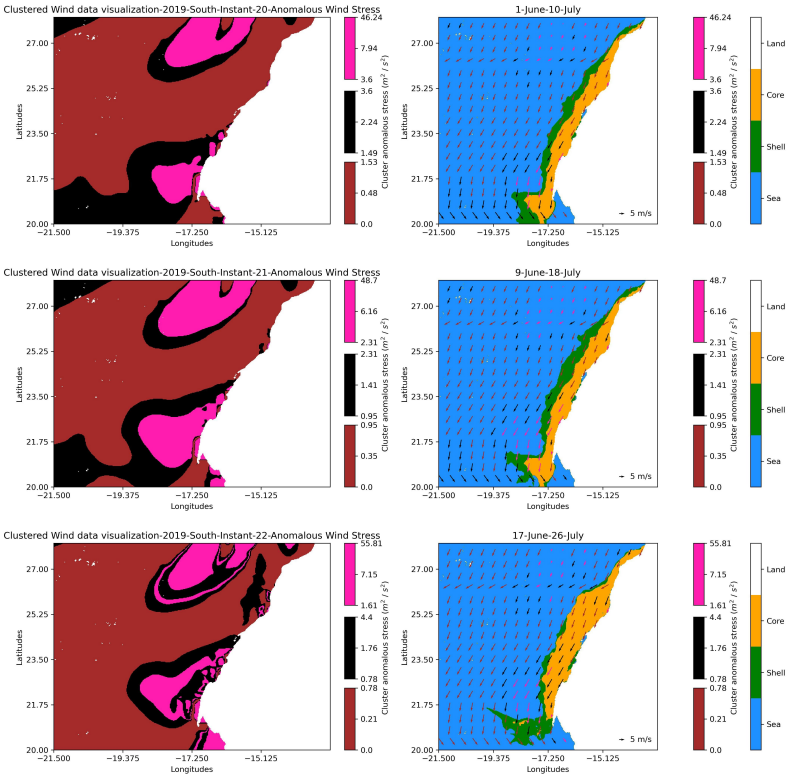


Figure 6.3: Spatial evolution of segmented IAP areas and core-shell segmentations

In the core-shell upwelling SST segmentations (right), the core/shell is visualized in orange/green. From SST instant 20 to 21, we can see the shell in the southernmost part moving slightly upwards. From SST instant 21 to 22, the USP has changed, shown by the changing of the orange area of the core-shell SST segmentations. We can see that in this SST instant, the pink cluster of the WSA map segmentation has become more divided between the black cluster, a good indicator of changing USP. These interesting spatial changes observed show how WSA patterns can be consistent with the USP.

The graphics in Figures 6.4 and 6.5 present the WSA averages for each of the three

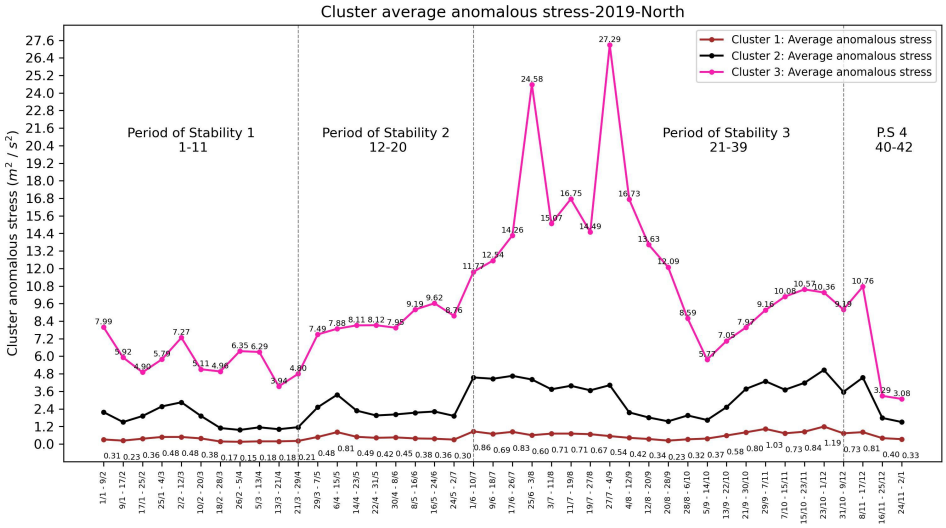


Figure 6.4: Evolution of the average cluster WSA in 2019-NM geographic region

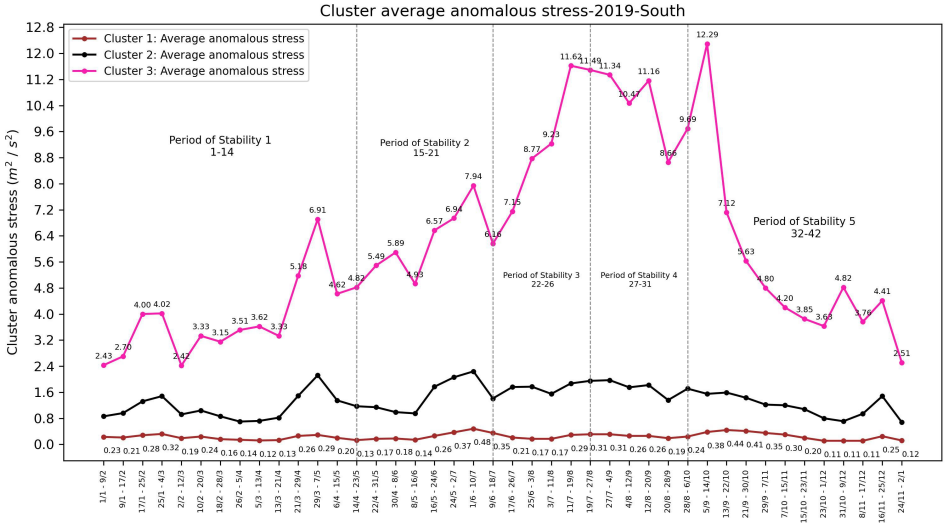


Figure 6.5: Evolution of the average cluster WSA in 2019-SM geographic region

clusters along the upwelling season of 2019. The USP described in the figures were determined with the use of the core-shell clustering framework, which be seen in Tables 6.2 and 6.3. We can see that the NM geographic region has less USPs than the SM geographic region, however is characterized by a longer USP, in this case USP 3. The trends observed in the average WSA of clusters 1 and 2 are quite similar to each other in both geographic regions. Cluster's 3 average WSA value evolution presents a similar trend of increasing until the end of the summer season and then decreasing until the end of the year. However, we can see that in the NM geographic region that drop happens way sooner (27th June - 4th September) than in the SM geographic region (5th September - 14th October). Coincidently, this rapid decreasing tendency happens when its reached the cluster's 3 global maximum

WSA average value.

The same tendency patterns can be obserseved for the other regions whoose graphics are presented in Figures B.6 to B.9 in section B.4 of appendix B. These results are promissing about a relation between WSA and the USP found by the core-shell clustering.

Changing now the temporal resolution to daily WSA maps, we apply IAP to extract 3 clusters from the data in order to build datasets similar to the one described in table 5.2 whoose labelling of the USP will follow the procedure described in section 5.5.2, resulting in the construction of 4 clustered WSA datasets per region and year.

6.3 Labelled wind stress anomaly data: brief analysis

We apply the procedure for labelling the clustered WSA data with the USP (see subsection 5.5.2). From this are produced 24 labbeled WSA data collections following the naming structure specified in subsection 5.5.2.2 as shown in the first column of table 6.4. Each dataset has in total 365 samples as they correspond to an year and are of daily nature.

Prior to commencing the construction of any models, it is essential to conduct a preliminary examination of the datasets to analyze the data cases distribution per USP class, or the USP class proportions. The USP class distributions for each dataset utilized in each year and region are presented in Table 6.4. This allows for an understanding of the behavior of each defuzzification function employed.

Dataset	USP Class proportions (%)					
	1	2	3	4	5	6
NM-2007-COG	27.40	30.96	41.64	-	-	-
NM-2007-MOM	26.30	33.15	40.55	-	-	-
NM-2007-LOM	36.44	37.81	25.75	-	-	-
NM-2007-SOM	16.44	28.49	**55.07**	-	-	-
NM-2015-COG	23.56	21.92	21.64	32.88	-	-
NM-2015-MOM	22.47	23.01	22.74	31.78	-	-
NM-2015-LOM	**32.05**	18.63	26.85	22.47	-	-
NM-2015-SOM	13.15	27.40	18.63	**40.82**	-	-
NM-2019-COG	26.30	27.40	28.22	18.08	-	-
NM-2019-MOM	25.21	28.49	29.32	16.99	-	-
NM-2019-LOM	**34.25**	30.68	27.95	7.12	-	-
NM-2019-SOM	16.44	26.30	30.68	26.58	-	-

Dataset	USP Class proportions (%)					
	1	2	3	4	5	6
SM-2007-COG	**35.62**	21.37	18.90	24.11	-	-
SM-2007-MOM	34.52	22.47	20.00	23.01	-	-
SM-2007-LOM	**45.21**	15.34	24.66	14.79	-	-
SM-2007-SOM	24.11	29.59	15.34	**30.96**	-	-
SM-2015-COG	24.66	14.79	13.70	15.89	15.07	15.89
SM-2015-MOM	23.56	15.89	13.70	15.89	16.16	14.79
SM-2015-LOM	**30.68**	10.96	15.62	15.34	16.99	10.41
SM-2015-SOM	16.71	20.55	12.05	16.44	15.34	18.90
SM-2019-COG	**30.14**	19.18	12.05	15.07	23.56	-
SM-2019-MOM	**29.04**	20.27	12.05	16.16	22.47	-
SM-2019-LOM	**38.63**	13.15	10.96	21.37	25.89	-
SM-2019-SOM	19.73	27.40	13.15	10.96	**28.77**	-

Table 6.4: Distribution of USP class proportions for WSA labbeled data collections (NM and SM geographic regions)

As evidenced in Table 6.4, the datasets under consideration do not exhibit a significant imbalance in their USP class distributions, with the cases of class imbalancing being highlighted in boldface according to the majority class. However, it is important to note that in certain datasets, this imbalance may require consideration when evaluating the performance of a model constructed from them. The most notable instance of this is observed in USP 4 of the NM-2019-LOM dataset, which represents a mere 7.12% of the entire dataset. In certain instances, other USP classes have also demonstrated a similar trend, with class proportion levels reaching just over 10%.USP 3 of SM-2015-SOM is an example of these instances. We can also see that functions which output a central

defuzzification value of the membership function, like the Center of Gravity (COG) and Middle of Maxima (MOM) tend to generate more balanced datasets, like the SM-2015 dataset for both of the funtions. On the other hand, functions such as the Smallest of Maximum (SOM) and Largest of Maximum (LOM), which output a defuzzification value at the beggining, SOM, or the end, LOM of the core of the membership function tend to generate more imbalanced datasets, like the previous mentioned case of the NM-2019-LOM or the NM-2007-SOM dataset, with latter having a USP class representing more than half of the entire dataset. Futhermore, we can observe that in the case of the NM geographic region, the last USP exhibits the highest USP class proportion, while in geographic region of SM, this is observed in the first USP.

6.4 Predicting upwelling stability period from wind stress anomalies

In this section we present and discuss the results of applying the Random Forest (RF), Ordinal Forest (OF), and K-Nearest Neighbors (K-NN) classifiers following the experimental setup protocol described in section 5.6 to the 24 labbeled datasets. Particular attention is given to analyze the consistency of each of the 4 defuzzification functions—COG,MOM,LOM and SOM—over the years covered by the study (2007,2015,2019). We will evaluate whether these functions are robust enough for our problem by comparing their train-validation and test set assessment measures. Additionally we will make a classifier comparison to determine the most appropriate one to predict a USP class. For both objectives, we will identify the optimal model for each classifier within the predefined hyperparameter combinations described in section 5.6, assessing their train-validation and test set performances.

6.4.1 Random Forests

6.4.1.1 North

Table 6.5 presents the hyperparameters of the optimal models, along with their cross-validation and test performance. The models were developed through a fine-tuning process in which multiple hyperparameters combinations were tested within a predefined hyperparameter space, described in Table 5.4.

Overall, the models exhibit notable similarities, with the exception of the criterion and min_samples_split parameters, which demonstrate a high degree of variability in values. Despite this, the fine-tuned models are relatively simple for a RF model, with the majority having the minimum number of available n_estimators and displaying minimal tree depth.

The simpler models were generated when using the LOM defuzzification function, with each model exhibiting lowest number of available estimators. Notwithstanding

	Criterion	max depth	min samples leaf	min samples split	n estimators	CV Balanced Accuracy	CV Precision	CV Recall	CV F1 score	Test Balanced Accuracy	Test Precision	Test Recall	Test F1 score	Average Test F1 score
2007-COG	Entropy	10	1	5	300	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	
2015-COG	Gini	10	1	2	300	0.93	0.94	0.93	0.93	0.94	0.94	0.95	0.94	**0.95** ± 0.01
2019-COG	Gini	15	1	3	700	0.91	0.92	0.91	0.91	0.95	0.95	0.95	0.94	
2007-LOM	Entropy	10	1	3	300	0.95	0.94	0.95	0.95	0.91	0.93	0.92	0.91	
2015-LOM	Entropy	15	1	6	300	0.95	0.96	0.95	0.95	0.96	0.96	0.96	0.96	0.93 ± 0.02
2019-LOM	Entropy	10	1	2	300	0.86	0.92	0.86	0.86	0.91	0.95	0.95	0.91	
2007-MOM	Gini	10	1	2	300	0.95	0.96	0.95	0.96	0.95	0.96	0.96	0.96	
2015-MOM	Gini	10	1	6	600	0.91	0.92	0.91	0.91	0.91	0.93	0.92	0.91	0.94 ± 0.02
2019-MOM	Gini	10	1	4	700	0.95	0.95	0.95	0.95	0.96	0.96	0.96	0.96	
2007-SOM	Gini	20	1	4	300	0.93	0.95	0.93	0.94	0.92	0.95	0.94	0.94	
2015-SOM	Entropy	10	1	3	300	0.95	0.96	0.95	0.95	0.86	0.91	0.9	0.87	0.90 ± 0.03
2019-SOM	Entropy	10	1	6	700	0.96	0.97	0.96	0.97	0.88	0.92	0.9	0.9	

Table 6.5: Cross-Validation and Test performance of the best models built-RF-NM geographic region

the exceptional cross-validation performances, the testing results did not diverge to a significant extent from those observed, thereby suggesting that the models are not exhibiting signs of overfitting. Nevertheless, some instances of overfitting were evident, particularly in the case of the datasets deffuzified by the SOM function, as observed in the 2015 and 2019 datasets. The datasets that yielded the most consistent model performance were those that employed the COG function, having the highest average F1 score along with the lowest standard deviation, highlighted in the table. The LOM and MOM functions also achieved good results with minimal deviations, with the SOM function presenting the worst overall results. Overall, the models presented a balanced tradeoff between precision and recall, evidenced by their high F1 score and similar precision and recall values.

6.4.1.2 South

A review of Table 6.6, reveals that the geographic region of SM results in the construction of more complex models. This is infered by an examination of the n_estimators parameter, which not only presents greater variability but also exhibits an increase when compared with the NM geographic region models. An analysis of the max_depth parameter also reveals that the models produced have deeper trees. The necessity for this elevated model complexity may be attributed to the SM geographic region datasets themselves, in which encompass more classes than the NM geographic region datasets, as observed in Table 6.4.

	Criterion	max depth	min samples leaf	min samples split	n estimators	CV Balanced Accuracy	CV Precision	CV Recall	CV F1 score	Test Balanced Accuracy	Test Precision	Test Recall	Test F1 score	Average Test F1 score
2007-COG	Gini	15	1	4	600	0.97	0.97	0.97	0.97	0.90	0.93	0.92	0.90	
2015-COG	Entropy	15	1	7	1000	0.88	0.9	0.88	0.88	0.84	0.88	0.86	0.84	**0.88** ± 0.03
2019-COG	Gini	10	1	3	600	0.88	0.89	0.92	0.88	0.89	0.91	0.90	0.90	
2007-LOM	Gini	15	1	2	800	0.87	0.91	0.97	0.88	800	0.90	0.89	0.87	
2015-LOM	Gini	15	1	3	800	0.89	0.92	0.89	0.89	0.78	0.83	0.82	0.78	0.82 ± 0.04
2019-LOM	Entropy	10	1	7	800	0.87	0.91	0.87	0.88	0.82	0.87	0.86	0.82	
2007-MOM	Entropy	10	1	4	400	0.97	0.97	0.97	0.97	0,97	0.86	0.86	0.84	
2015-MOM	Entropy	15	1	4	800	0.85	0.87	0.85	0.85	0.84	0.86	0.85	0.84	**0.88** ± 0.05
2019-MOM	Entropy	10	1	4	800	0.87	0.90	0.87	0.88	0.95	0.96	0.96	0.95	
2007-SOM	Entropy	10	1	7	800	0.89	0.93	0.89	0.90	0.89	0.93	0.92	0.90	
2015-SOM	Gini	15	1	5	800	0.89	0.91	0.89	0.89	0.85	0.90	0.88	0.87	0.85 ± 0.05
2019-SOM	Entropy	10	1	3	800	0.83	0.89	0.83	0.85	0.78	0.83	0.8	0.79	

Table 6.6: Cross-Validation and Test performance of the best models built-RF-SM geographic region

An examination of the CV-F1 score column reveals that the models have achieved commendable results, though they are slightly inferior to those of the geographic region

of NM models. In this instance, however, the elevated cross-validation performances translated in a greater incidence of overfitting, with the most pronounced example being the 2007-MOM dataset which, yielded a CV-F1 score of 0.97 but only achieved an F1 score of 0.84 upon testing. These overfitting cases are evident in all of the four deffuzification datasets. The COG and MOM presented equal average test F1-scores, followed by the SOM function with the LOM function presenting the less optimal results. Even so, the best overall results were obtained by the COG function, where, despite overfitting in 2007 it still achieved good "Test F1 score". Despite achieving a commendable balance between precision and recall, all models exhibit a tendency to prioritize precision over recall. This suggests that the number of false positive classifications should be minimal.

6.4.2 Ordinal Forests

6.4.2.1 North

Table 6.7 reveals a markedly elevated degree of model complexity when compared to the RF models. This elevated complexity is derived from the actual values evaluated in the hyperparameter space delineated in Table 5.5. A comparison between the ntreefinal and n_estimators parameters, which specify the number of trees each forest should comprise, reveals a substantital discrepancy in model complexity. The minimum value of the former is 2500, while the maximum value of the latter is 1000.

	ntreeperdiv	ntreefinal	npermtrial	nbest	CV Balanced Accuracy	CV Precision	CV Recall	CV F1 score	Test Balanced Accuracy	Test Precision	Test Recall	Test F1 score	Average Test F1 score
2007-COG	100	5000	700	10	0.98	0.98	0.97	0.97	0.97	0.95	0.95	0.95	
2015-COG	25	5000	700	5	0.97	0.96	0.95	0.95	0.97	0.96	0.95	0.96	0.96 ± 0.005
2019-COG	50	5000	600	10	0.95	0.93	0.93	0.93	0.97	0.97	0.96	0.96	
2007-LOM	50	2500	500	10	0.96	0.96	0.95	0.95	0.97	0.97	0.96	0.96	
2015-LOM	25	2500	500	5	0.96	0.95	0.94	0.94	0.95	0.94	0.93	0.93	0.90 ± 0.07
2019-LOM	25	5000	600	10	0.93	0.92	0.88	0.89	0.87	0.82	0.79	0.8	
2007-MOM	25	2500	500	5	0.97	0.97	0.96	0.96	1	1	1	1	
2015-MOM	25	2500	700	5	0.96	0.94	0.93	0.93	0.97	0.96	0.96	0.96	**0.98** ± 0.02
2019-MOM	25	2500	500	5	0.97	0.96	0.96	0.96	0.99	0.98	0.98	0.98	
2007-SOM	50	5000	500	5	0.94	0.95	0.92	0.93	1	1	1	1	
2015-SOM	50	2500	500	5	0.95	0.95	0.92	0.93	0.95	0.94	0.93	0.93	0.96 ± 0.03
2019-SOM	100	2500	700	5	0.97	0.97	0.96	0.96	0.97	0.96	0.96	0.96	

Table 6.7: Cross-Validation and Test performance of the best models built-OF-NM geographic region

The performance results of these models exhibit a similar pattern to that observed in the RF models of the same region, with highly favorable performance metrics and only one instance of overfitting, 2019-LOM. The models that demonstrated the most optimal performance were those that employed the MOM defuzzified datasets, with nearly maximum average test F1 scores with minimum standard deviation, highlighted in the table above. The models derived from the COG datasets exhibited a similar degree of consistency, with minimal discrepancies observed in their performance metrics. The SOM datasets yielded comparable performances, although, the considerable discrepancy between cross-validation to test perfromance in 2007 merits more scrutiny. The OF models demonstrated remarkable performance on both the cross-validation and test phases, with instances of perfect performances. Additionally, the models exhibited an excellent

precision-recall tradeoff, as evidenced by their high Test-F1 scores and the equilibrium observed in the comparison of the individual precision and recall of each model.

6.4.2.2 South

The models obtained demonstrate a stable hyperparameter,nbest, which serves an effective indicator of the actual hyperparameter space required for fine-tuning these models. The models exhibit a comparable level of complexity to those from the NM geographic region.

	ntreeperdiv	ntreefinal	npermtrial	nbest	CV Balanced Accuracy	CV Precision	CV Recall	CV F1 score	Test Balanced Accuracy	Test Precision	Test Recall	Test F1 score	Average Test F1 score
2007-COG	25	2500	600	5	0.96	0.95	0.95	0.95	0.97	0.94	0.95	0.94	
2015-COG	25	2500	500	5	0.94	0.92	0.90	0.90	0.91	0.86	0.84	0.83	0.87 ± 0.05
2019-COG	100	5000	700	5	0.95	0.94	0.92	0.93	0.9	0.89	0.84	0.85	
2007-LOM	100	2500	500	5	0.91	0.89	0.87	0.87	0.98	0.97	0.96	0.97	
2015-LOM	50	5000	500	5	0.93	0.93	0.88	0.89	0.96	0.96	0.92	0.94	**0.92** ± 0.05
2019-LOM	50	5000	500	5	0.93	0.92	0.88	0.89	0.91	0.85	0.85	0.85	
2007-MOM	25	2500	500	5	0.95	0.93	0.93	0.93	0.98	0.98	0.98	0.98	
2015-MOM	50	2500	600	10	0.93	0.89	0.88	0.88	0.89	0.83	0.82	0.82	0.91 ± 0.07
2019-MOM	100	5000	500	5	0.94	0.92	0.91	0.91	0.94	0.94	0.91	0.92	
2007-SOM	100	2500	700	5	0.94	0.93	0.91	0.91	0.94	0.93	0.9	0.91	
2015-SOM	100	5000	600	5	0.94	0.90	0.90	0.90	0.93	0.87	0.87	0.87	0.84 ± 0.07
2019-SOM	50	2500	700	5	0.89	0.90	0.83	0.84	0.84	0.78	0.73	0.74	

Table 6.8: Cross-Validation and Test performance of the best models built-OF-SM geographic region

As previously observed in the RF models in the SM geographic region, there are more evident cases of overfitting than in the geographic region regarding NM, with a notable performance decline in comparison to the NM geographic region models. These cases are more evidently present in models obtained in 2019-SOM and in the models built from the COG datasets (2015 and 2019), where a clear performance drop is observed from cross-validation to the testing phase. The LOM defuzzification function presented the highest average test F1 score with a value of 0.92, with the MOM following it with 0.91. These models exhibited a balanced precision-recall pattern, similar to that observed in the previous region and classifier. In this case, the focus was again on minimizing the number of false positives instead of false negatives.

6.4.3 KNN

To determine the most appropriate K, the process described in section 5.6.3 is applied. The learning curves resultant of this process are in section B.5 of appendix B. In the figures it is plotted the test missclassification error in red and in blue the Leave One Out Cross-Validation (LOOCV) missclassification error.

6.4.3.1 North

Table 6.9 presents the test performance of the best models constructed from the hyperparameter space defined. Overall, the K value does not exceed 15, with only one occurrence happening in 2019-MOM. When compared with the forest models, although generating simpler models, this algorithm does not reach the same level high performance as those models.

	K	Average LOOCV error	Average Preliminar Test error	Absolute Difference	Balanced Accuracy	Precision	Recall	F1 score	Average F1 score
2007-COG	13	0.133	0.13	0.003	0.87	0.83	0.82	0.82	
2015-COG	8	0.154	0.153	0.001	0.93	0.91	0.89	0.90	0.82 ± 0.07
2019-COG	13	0.204	0.195	0.009	0.83	0.83	0.74	0.74	
2007-LOM	8	0.114	0.04	0.074	0.91	0.89	0.88	0.88	
2015-LOM	12	0.173	0.173	0	0.87	0.81	0.80	0.80	0.83 ± 0.04
2019-LOM	5	0.177	0.163	0.014	0.87	0.79	0.82	0.81	
2007-MOM	5	0.13	0.112	0.018	0.92	0.89	0.88	0.88	
2015-MOM	5	0.172	0.171	0.001	0.89	0.84	0.84	0.84	0.81 ± 0.07
2019-MOM	19	0.194	0.196	0.002	0.81	0.78	0.71	0.72	
2007-SOM	5	0.121	0.112	0.009	0.91	0.92	0.87	0.89	
2015-SOM	9	0.146	0.154	0.008	0.91	0.90	0.86	0.87	**0.86** ± 0.03
2019-SOM	8	0.171	0.17	0.001	0.87	0.82	0.81	0.81	

Table 6.10: LOOCV and Test performance of the best models built-K-NN-NM geographic region

An initial examination of the absolute difference column reveals that the average test error and the average LOOCV error are not significantly disparate. However, there are instances where the test error is observed to be lower than the LOOCV error. Therefore, the use of the absolute difference between errors to determine the optimal K number of neighbors, is more appropriate than the difference alone. The models that demonstrated superior performance were those that underwent defuzzification via the SOM function, yielding the highest average test F1 score and also generating the simpler models overall. All of the other defuzzification functions presented similar average test F1 score metrics with small standard deviations. The SOM function has to be chosen over the other three, as its average test F1 score is relatively higher than them. The MOM and SOM generated predominantly simple models with low K values. However, the 2019-MOM model exhibited suboptimal data fit even with a high K value, and also presenting the poorest overall score of that year. The models follow the pattern of prioritizing precision over recall, a pattern observed in the other classifiers tested.

6.4.3.2 South

As evidenced by the data presented in Table 6.11, there is a discernible decline in performance, already observed in the forest models when regions are altered. The models constructed are overall more complex than the ones derived from the NM geographic region. The COG function produced the simpler models, although the MOM achieved the most favorable overall results.

The absolute distances between errors remain relatively consistent, with some instances exihibiting minimal variation. The overall performance decline in 2015 and 2019 is notable, primarily due to the fact that these years encompass 6 and 5 classes, respectively. In light of these observations, it is evident that none of the functions demonstrated consistent performance across all years. The performance observed in 2007 across all of functions can be attributed to the number of classes in that year, which was the same as in 2015 and 2019 for the geographic region of NM. This resulted in comparable performances across

	K	Average LOOCV error	Average Preliminar Test error	Absolute Difference	Balanced Accuracy	Precision	Recall	F1 measure	Average F1 score
2007-COG	5	0.17	0.171	0.001	0.90	0.84	0.84	0.84	
2015-COG	8	0.236	0.241	0.005	0.84	0.72	0.73	0.72	0.76 ± 0.06
2019-COG	7	0.167	0.168	0.001	0.84	0.72	0.73	0.72	
2007-LOM	13	0.172	0.172	0	0.91	0.88	0.86	0.86	
2015-LOM	8	0.259	0.259	0	0.78	0.61	0.62	0.61	0.72 ± 0.1
2019-LOM	5	0.213	0.214	0.001	0.81	0.70	0.69	0.69	
2007-MOM	15	0.154	0.154	0	0.96	0.93	0.93	0.93	
2015-MOM	9	0.248	0.241	0.007	0.86	0.77	0.77	0.77	**0.84** ± 0.07
2019-MOM	11	0.223	0.22	0.003	0.89	0.82	0.83	0.82	
2007-SOM	21	0.178	0.171	0.007	0.93	0.88	0.89	0.88	
2015-SOM	7	0.239	0.239	0.001	0.85	0.75	0.75	0.75	0.78 ± 0.08
2019-SOM	15	0.218	0.218	0	0.82	0.70	0.70	0.70	

Table 6.11: LOOCV and Test performance of the best models built-K-NN-SM geographic region

regions. The simpler models were obtained by defuzzifying the datasets using the COG function, whereas the MOM function generated the more complex models. In terms of performance, the MOM function demonstrated the most optimal results, exhibiting an equilibrium between precision and recall. While the models remain balanced with regard to precision and recall, the aforementioned tendency deviates from the norm in that the models now prioritize recall.

6.4.4 Summary

Table 6.12 presents the chosen defuzzification function for each classifier according to the optimal model obtained through the fine-tuning process described in section 5.6 whoose results were presented in the previous sections.

	Average Balanced Accuracy	Average F1 score	Defuzzification Function
RF-North	0.95 ± 0.02	0.95 ± 0.01	COG
OF-North	0.98 ± 0.02	0.98 ± 0.02	MOM
K-NN-North	0.89 ± 0.02	0.86 ± 0.03	SOM
RF-South	0.91 ± 0.05	0.88 ± 0.03	COG
OF-South	0.95 ± 0.04	0.92 ± 0.05	LOM
K-NN-South	0.90 ± 0.05	0.84 ± 0.07	MOM

Table 6.12: Optimal defuzzification function for each optimal model

The table above indicates that the COG and MOM defuzzification functions are suitable for this classification problem, as they were the best performing functions in at least one model for each region. The OF and RF classifiers proved to be the more adequate models for this task as they presented high balanced accuracy and F1 scores. The K-NN classifier had a decent accuracy, however, its prediction quality was not good, only achieving an average F1 score of 0.86 ± 0.03 for the NM geographic region and 0.84 ± 0.07 for the SM geographic region.

6.5 RF, OF, KNN Models comparison

We are going to compare the RF, OF,K-NN best constructed classification models, as it is important to compare them in order to understand which is best suited to the task at hand. We will focus the analysis on the optimal model of each classifier made for each year according to the region, regardless of the defuzzification function. This analysis will be conducted analyzing two different test set metrics: balanced accuracy and F1 score and will be visually aided with the obtained confusion matrices for each of the optimal models.

6.5.1 North

- In 2007, the OF presented a perfect performance by not missclassifying an instance. The RF algotithm also presented a really good performance as it presented a balanced accuracy of 0.97 and an F1 measure of 0.95. The worst performance in this year came from the K-NN classifier where its optimal model achieved a balanced accuracy metric value of 0.91 and an F1 score of 0.86, missclassifying mainly USP class 3 instances.

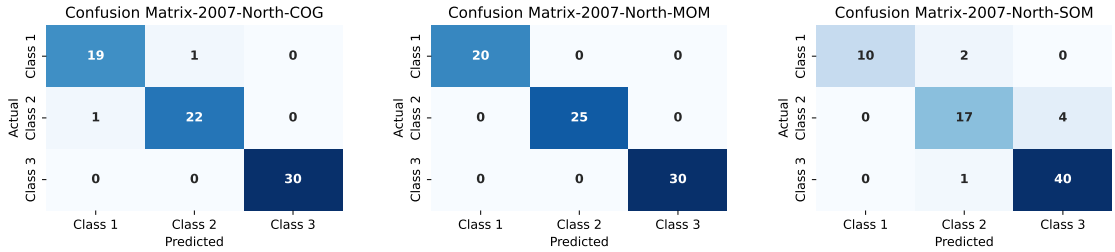


Figure 6.6: Confusion Matrices-2007-North (Random Forest-Ordinal Forest-K-NN)

- For 2015, the RF and OF had again really good performances, both with a F1 score metric value of 0.96, where OF lightly outperformed the RF model with a balanced accuracy value of 0.97 and the RF of 0.96. The K-NN classifier had a decent test set performance, where it achieved a balanced accuracy value of 0.93 and a its F1 score metric reached a value of 0.90.

- The OF forest model had near perfect performance, only misclassifying one instance, thus reaching a balanced accuracy value of 0.99 and a F1 score of 0.98. The RF model, missclassified 3 instances and achieved a metric value of 0.96 for balanced accuracy and F1 score. K-NN had bad performance USP class wise, missclassifying several instances from all of the classes, resulting in the lowest test performances of this geographic region for the optimal classifier models analyzed. K-NN achieved a balanced accuracy of 0.87 and a F1 score of 0.81. However, the K-NN's poor performance can be due to the nature of the dataset, which is highly unbalanced regarding the last USP class, as show in table 6.4.

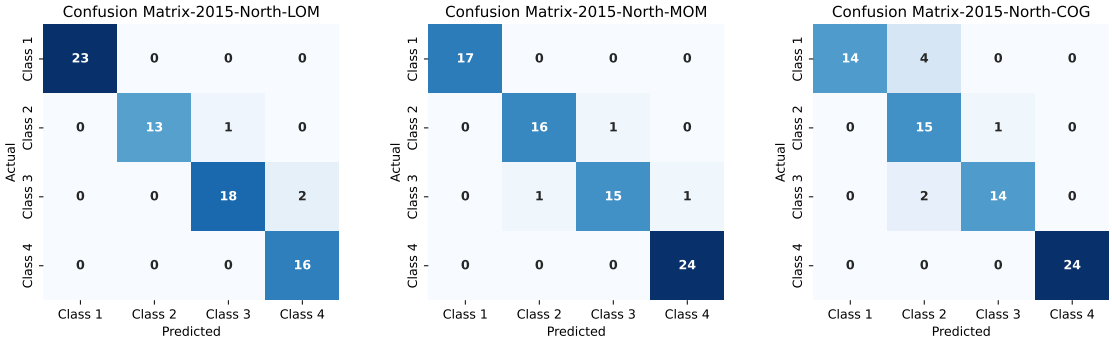


Figure 6.7: Confusion Matrices-2015-North (Random Forest-Ordinal Forest-K-NN)

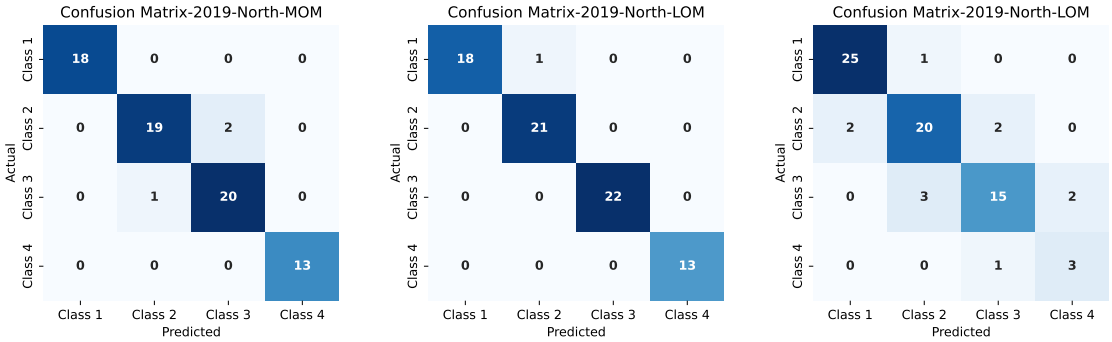


Figure 6.8: Confusion Matrices-2019-North (Random Forest-Ordinal Forest-K-NN)

6.5.2 South

- In 2007, the OF had an outstanding performance, only missclassifying one instance of USP class 3, which resulted in a value of 0.98 for both balanced accuracy and F1 score test set performance metrics. The K-NN classifier outperformed the RF model with a balanced accuracy of 0.96, while the RF model achieved 0.90. Regarding F1 scores of the K-NN and RF optimal models, the K-NN model also outperformed the RF with a value of 0.93 and the RF only 0.90. Overall, the K-NN and RF models had decent enough performances, however the OF classifier clearly outperformed them.

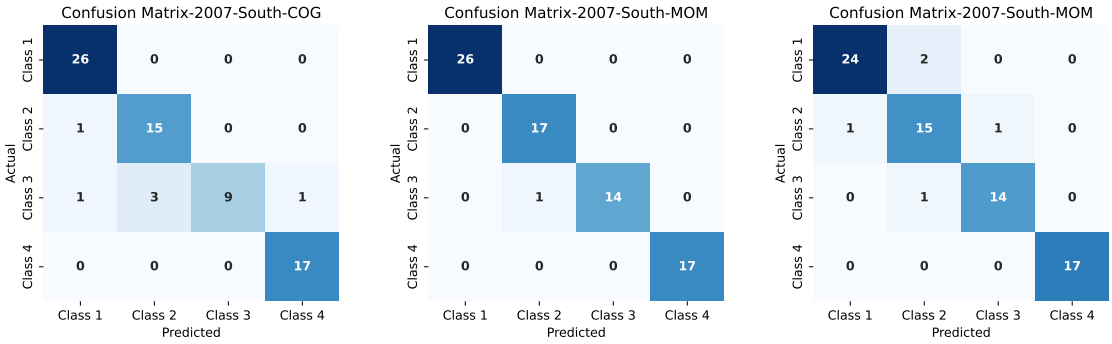


Figure 6.9: Confusion Matrices-2007-South (Random Forest-Ordinal Forest-K-NN)

- In 2015, the OF model outperformed the other two classifiers achieving a balanced accuracy of 0.96 and a F1 score of 0.94, showing great robustness regarding the increasing number of classes. The RF model missclassified some USP instances, specially the ones belonging to USP 3, resulting in a balaced accuracy value of 0.85 and a F1 score of 0.87. The K-NN was the model which performed the poorest, with many missclassification across all of the classes and with the lowest recorded F1 score metric, only reaching 0.77. However, this model achieved a decent balanced accuracy value of 0.86.

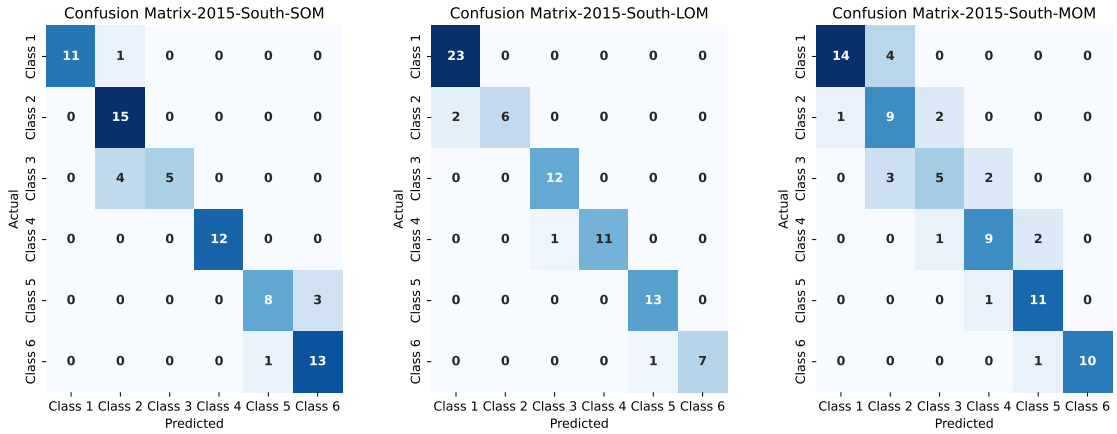


Figure 6.10: Confusion Matrices-2015-South (Random Forest-Ordinal Forest-KNN)

- In 2019, the same pattern as in 2015 is observed. The OF model adapted well to the number of classes with similar balanced accuracy and F1 score metrics, 0.94 and 0.92, respectively. Although the RF model missclassified less instances when compared to 2015 resulting in a a balanced accuracy value of 0.89 and a F1 socre metric value of 0.90, it still wasn't a performance good enough to beat the OF. The K-NN model did not perform greatly, missclassifying once again several instances with them belonging to every single class, achieving a low F1 score of 0.82 and decent balanced accuracy of 0.89.

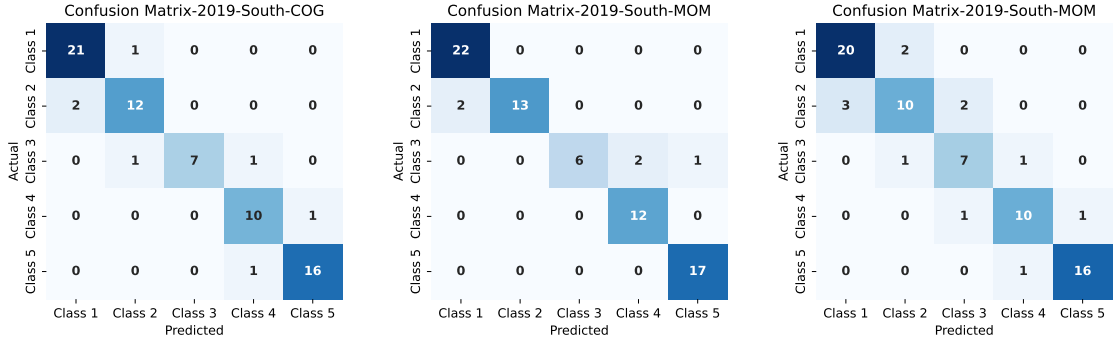


Figure 6.11: Confusion Matrices-2019-South (Random Forest-Ordinal Forest-KNN)

6.6 Summary

This section summarises the test set model performance based on the average performance metrics obtained for the optimal defuzzification functions. These results are summarized in Table 6.13.

	Average Balanced Accuracy	Average Precision	Average Recall	Average F1 score	Defuzzification Function
RF-North	0.95 ± 0.02	0.95 ± 0.02	0.96 ± 0.01	0.95 ± 0.01	COG
OF-North	0.98 ± 0.02	0.98 ± 0.02	0.98 ± 0.02	0.98 ± 0.02	MOM
K-NN-North	0.89 ± 0.02	0.88 ± 0.05	0.85 ± 0.03	0.86 ± 0.03	SOM
RF-South	0.91 ± 0.05	0.91 ± 0.02	0.89 ± 0.03	0.88 ± 0.03	COG
OF-South	0.95 ± 0.04	0.93 ± 0.07	0.91 ± 0.06	0.92 ± 0.05	LOM
K-NN-South	0.90 ± 0.05	0.84 ± 0.08	0.84 ± 0.03	0.84 ± 0.07	MOM

Table 6.13: Summary table of the test set performance

Given the performance results of the classifiers for the test sets and the comparison among the RF, OF, and K-NN models for each region-year the following aspects can be pointed out:

- The OF classifier outperformed the other two classifiers, with high average values across all of the test set performance metrics. However, when changing to the SM geographic region the classifier had a more unstable performance, which can be seen by the increase of the standard deviation values.
- The RF classifier presented also a good performance, specially with the COG defuzzification function, as for both geographic regions, the best models of this classifier were coupled with the former. This classifier also had a stable performance, characterized by the overall lower standard deviation values even when changing geographic regions.
- The K-NN classifier had the worst test set performance metrics in both geographic regions and had a not so stable test set performance shown by the higher standard deviation values in the SM region.
- A common trend seen in the test set classification performances is that as the number of USP increases the overall models' test set classification performance decreases, while also becoming more unstable. This is seen when changing geographic regions, in which the SM region datasets present a higher number of USP and by overall test set performance drop when comparing to the NM geographic region.
- Out of the 3 applied classifiers, the K-NN classifier, although being the fastest regarding fitting time presents the worst overall test set classification performances, degrading alot when changing to the SM geographic region. Figure 6.12 ilustrates the average fitting time for all of the optimal models determined for each classifier.

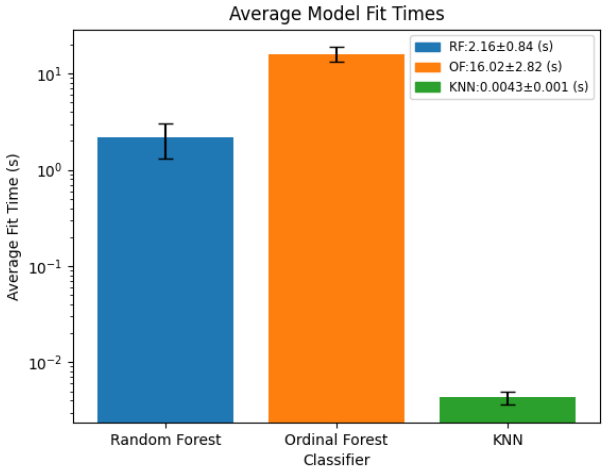


Figure 6.12: Average fitting times along with standard deviations for the optimal models for each classifier

As we can see in Figure 6.12, the OF surpasses greatly the RF and K-NN models. The K-NN model presents an extremely low fitting time due to its lazy learning nature.

7

Conclusion and future work

The work developed in this dissertation targeted the prediction of Upwelling Stability Period (USP) from wind data. To this intent wind and Sea Surface Temperature (SST) data were collected. The two data type collections suffered distinct processes to use them together First, the wind data was transformed into Wind Stress Anomaly (WSA) data, which was then applied a fuzzy membership function model the daily nature of it. The SST data was applied to the core-shell clustering framework to get ground truth regarding USP. With the ground truth determined, datasets were built and labeled according to the defuzzfication functions employed. Three classification models were utilized, namely Random Forest (RF), Ordinal Forest (OF) and K-Nearest Neighbors (K-NN).

It was observed that the two geographic regions presented different USP, with the North-Morocco (NM) geographic region having less and thus longer USP than the geographic region of South-Morocco (SM). Another characteristic observed when applying the four defuzzification functions, was that in the NM geographic region, the last USP usually has the higher USP class proportion, whereas in the SM geographic region, the first USP had the largest USP class proportion.

Regarding the models used, it was concluded that the forest based models performed well for this task, while the K-NN model presented some inconsistency, specially in the SM geographic region, independently of the defuzzification function used. Overall, all of the models of the geographic region of NM performed really well, with a performance drop when changing geographic region. The RF model performed outstandingly, however the OF classifier had the best performance overall, with some models of it performing perfectly when coupled with the right defuzzification function. However, it was observed that the OF takes a lot of computational time when compared to the K-NN and RF model, so it should be taken into consideration when choosing a specific classifier.

From the four defuzzification functions utilized, the ones which generate more central defuzzification values, the Center of Gravity (COG) and Middle of Maxima (MOM) presented the highest degree of consistency and robustness as functions of this type were the best ones in 4 of the 6 available region-year combinations. The Smallest of Maximum (SOM) and Largest of Maximum (LOM) functions tend to generate more unbalanced

datasets hence the lower performances of it.

Overall, the combination of using a RF model with a dataset defuzzified by the COG defuzzfication function seems to be more appropriate for predicting USP as this combination presents the best tradeoff between classification performance and computational fit time.

Regarding future work, the work made in this dissertation can be expanded the following way:

- Expand the study to other years as they are available to better understand the USP dynamics both in NM and SM geographic regions.
- Improvement of the wind's clustering-classification framework for application to new wind and SST data collections
- Explore different wind features such as the wind speed and wind stress itself.
- Test simpler hyperparameter combinations and models to avoid high computational fit times, specially in the latter.
- Experiment with other defuzzification functions to study the behavior of these in the datasets we have.
- Experiment with different ordinal data prediction models to assess if this the USP class should be treated as nominal or ordinal data.

Bibliography

[1] J. M. Lourenço. *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. 2021. URL: https://github.com/joaomlourenco/novathesis/raw/main/template.pdf (cit. on p. i).

[2] N. N. O. Service. *What is upwelling?* Figure showing the upwelling process. Accessed: 2024-09-26. 2024. URL: https://oceanservice.noaa.gov/facts/upwelling.html (cit. on p. 1).

[3] M. J. Artur Nowicki and L. Dzierzbicka-Głowacka. "Operational system for automatic coastal upwelling detection in the Baltic Sea based on the 3D CEMBS model". In: *Journal of Operational Oceanography* 12.2 (2019), pp. 104–115. DOI: 10.1080/1755876X.2019.1569748. eprint: https://doi.org/10.1080/1755876X.2019.1569748. URL: https://doi.org/10.1080/1755876X.2019.1569748 (cit. on pp. 1, 7).

[4] Z. E. Abidi and K. Minaoui. "An improved fusion method for detecting upwelling off the coast of northwest Africa from chlorophyll-a and sea surface temperature satellite images". In: *Remote Sensing Letters* 13.11 (2022), pp. 1110–1120. DOI: 10.1080/2150704X.2022.2123720. eprint: https://doi.org/10.1080/2150704X.2022.2123720. URL: https://doi.org/10.1080/2150704X.2022.2123720 (cit. on pp. 1, 8).

[5] S. Nascimento et al. "Piece-wise constant cluster modelling of dynamics of upwelling patterns". In: *Expert Systems* 40.10 (2023), e13446. DOI: https://doi.org/10.1111/exsy.13446. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/exsy.13446. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.13446 (cit. on pp. 1, 3, 26, 33, 51, 52, 82, 83, 89).

[6] A. G. Martins. "Unsupervised Spatio-Temporal Analysis of Coastal Upwelling from Sea Surface Temperature Images". Master in Computer Science. NOVA University Lisbon, 2022 (cit. on pp. 1, 22–27, 35, 52, 82).

[7] S. Nascimento, S. Mateen, and P. Relvas. "Sequential Self-tuning Clustering for Automatic Delimitation of Coastal Upwelling on SST Images". In: *Intelligent Data Engineering and Automated Learning – IDEAL 2020*. Ed. by C. Analide et al. Cham: Springer International Publishing, 2020, pp. 434–443. ISBN: 978-3-030-62365-4 (cit. on p. 2).

[8] D. Liu, J. Wang, and H. Wang. "Short-term wind speed forecasting based on spectral clustering and optimised echo state networks". In: *Renewable Energy* 78 (2015), pp. 599–608. ISSN: 0960-1481. DOI: https://doi.org/10.1016/j.renene.2015.01.022. URL: https://www.sciencedirect.com/science/article/pii/S0960148115000294 (cit. on p. 5).

[9] D. Zhang, K. Lee, and I. Lee. "Hierarchical trajectory clustering for spatio-temporal periodic pattern mining". In: *Expert Systems with Applications* 92 (2018), pp. 1–11. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2017.09.040. URL: https://www.sciencedirect.com/science/article/pii/S0957417417306450 (cit. on p. 5).

[10] B. Ramachandra, B. Dutton, and R. R. Vatsavai. "Anomalous Cluster Detection in Spatio-Temporal Meteorological Fields". In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 12.2 (2019), 193–218. DOI: 10.1002/sam.11398 (cit. on p. 5).

[11] F. Sambe and T. Suga. "Unsupervised clustering of Argo temperature and salinity profiles in the mid-latitude northwest Pacific Ocean and revealed influence of the Kuroshio Extension variability on the vertical structure distribution". In: *Journal of Geophysical Research: Oceans* 127 (2022). DOI: https://doi.org/10.1029/2021JC018138 (cit. on p. 5).

[12] F. Di Martino, W. Pedrycz, and S. Sessa. "Spatiotemporal extended fuzzy C-means clustering algorithm for hotspots detection and prediction". In: *Fuzzy Sets and Systems* 340 (2018). Theme: Clustering, pp. 109–126. ISSN: 0165-0114. DOI: https://doi.org/10.1016/j.fss.2017.11.011. URL: https://www.sciencedirect.com/science/article/pii/S0165011417304086 (cit. on p. 5).

[13] A. Konstantaras. "Deep Learning and Parallel Processing Spatio-Temporal Clustering Unveil New Ionian Distinct Seismic Zone". In: *Informatics* 7.4 (2020). ISSN: 2227-9709. DOI: 10.3390/informatics7040039. URL: https://www.mdpi.com/2227-9709/7/4/39 (cit. on p. 5).

[14] X. C. Chen et al. "Clustering Dynamic Spatio-Temporal Patterns in the Presence of Noise and Missing Data". In: *Proceedings of the 24th International Conference on Artificial Intelligence*. IJCAI'15. Buenos Aires, Argentina: AAAI Press, 2015, 2575–2581. ISBN: 9781577357384 (cit. on p. 6).

[15] J. P. Siemer et al. "Recent Trends in SST, Chl-a, Productivity and Wind Stress in Upwelling and Open Ocean Areas in the Upper Eastern North Atlantic Subtropical Gyre". In: *Journal of Geophysical Research: Oceans* 126.8 (2021). e2021JC017268 2021JC017268, e2021JC017268. DOI: https://doi.org/10.1029/2021JC017268. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2021JC017268. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021JC017268 (cit. on p. 6).

[16] R. F. Sánchez, P. Relvas, and M. Delgado. "Coupled ocean wind and sea surface temperature patterns off the western Iberian Peninsula". In: *Journal of Marine Systems* 68.1 (2007), pp. 103–127. ISSN: 0924-7963. DOI: https://doi.org/10.1016/j.jmarsys.2006.11.003. URL: https://www.sciencedirect.com/science/article/pii/S0924796306003472 (cit. on p. 6).

[17] S. Ferreira et al. "New Insights about Upwelling Trends off the Portuguese Coast: An ERA5 Dataset Analysis". In: *Journal of Marine Science and Engineering* 10.12 (2022). ISSN: 2077-1312. DOI: 10.3390/jmse10121849. URL: https://www.mdpi.com/2077-1312/10/12/1849 (cit. on p. 7).

[18] A. Lehmann, K. Myrberg, and K. Höflich. "A statistical approach to coastal upwelling in the Baltic Sea based on the analysis of satellite data for 1990–2009". In: *Oceanologia* 54.3 (2012), pp. 369–393. ISSN: 0078-3234. DOI: https://doi.org/10.5697/oc.54-3.369. URL: https://www.sciencedirect.com/science/article/pii/S0078323412500199 (cit. on pp. 7, 15).

[19] S. Zhang et al. "Mapping coastal upwelling in the Baltic Sea from 2002 to 2020 using remote sensing data". In: *International Journal of Applied Earth Observation and Geoinformation* 114 (2022), p. 103061. ISSN: 1569-8432. DOI: https://doi.org/10.1016/j.jag.2022.103061. URL: https://www.sciencedirect.com/science/article/pii/S1569843222002497 (cit. on p. 7).

[20] B. Mourre et al. "Intense wind-driven coastal upwelling in the Balearic Islands in response to Storm Blas (November 2021)". In: *State Planet* 1 (2023), p. 15. DOI: 10.5194/sp-1-osr7-15-2023 (cit. on p. 7).

[21] T. Georg, M. C. Neves, and P. Relvas. "The signature of NAO and EA climate patterns on the vertical structure of the Canary Current upwelling system". In: *Ocean Science* 19.2 (2023), pp. 351–361. DOI: 10.5194/os-19-351-2023. URL: https://os.copernicus.org/articles/19/351/2023/ (cit. on p. 8).

[22] A. El Aouni et al. "Physical and Biological Satellite Observations of the Northwest African Upwelling: Spatial Extent and Dynamics". In: *IEEE Transactions on Geoscience and Remote Sensing* 58.2 (2020), pp. 1409–1421. DOI: 10.1109/TGRS.2019.2946300 (cit. on p. 8).

[23] H. Belmajdoub et al. "A New Upwelling Index for the Moroccan Atlantic Coast for the Period between 1982ndash;2021". In: *Remote Sensing* 15.14 (2023). ISSN: 2072-4292. DOI: 10.3390/rs15143459. URL: https://www.mdpi.com/2072-4292/15/14/3459 (cit. on p. 8).

[24] J. D. Ramanantsoa et al. "Coastal upwelling south of Madagascar: Temporal and spatial variability". In: *Journal of Marine Systems* 178 (2018), pp. 29–37. ISSN: 0924-7963. DOI: https://doi.org/10.1016/j.jmarsys.2017.10.005. URL: https://www.sciencedirect.com/science/article/pii/S092479631730249X (cit. on p. 8).

[25] F. Ghorbani Afzal, M. Hasanlou, and S. Rajabi-Kiasari. "Monitoring and estimating coastal upwelling using Sentinel-3 satellite imagery (case study: The Caspian Sea)". In: *Continental Shelf Research* 261 (2023), p. 105010. ISSN: 0278-4343. DOI: https://doi.org/10.1016/j.csr.2023.105010. URL: https://www.sciencedirect.com/science/article/pii/S0278434323000870 (cit. on p. 9).

[26] F. Fallah and D. Mansoury. "Coastal upwelling by wind-driven forcing in the Caspian Sea: A numerical analysis". In: *Oceanologia* 64.2 (2022), pp. 363–375. ISSN: 0078-3234. DOI: https://doi.org/10.1016/j.oceano.2022.01.003. URL: https://www.sciencedirect.com/science/article/pii/S0078323422000148 (cit. on pp. 9, 15).

[27] C. Jayaram and P. D. Kumar. "Spatio-temporal variability of upwelling along the southwest coast of India based on satellite observations". In: *Continental Shelf Research* 156 (2018), pp. 33–42. ISSN: 0278-4343. DOI: https://doi.org/10.1016/j.csr.2018.02.003. URL: https://www.sciencedirect.com/science/article/pii/S0278434317301346 (cit. on p. 9).

[28] P. H. Kok et al. "Spatiotemporal trends in the southwest monsoon wind-driven upwelling in the southwestern part of the South China Sea". In: *PloS one* 12.2 (2017), e0171979 (cit. on p. 9).

[29] J. Shi et al. "Spatial and Temporal Variability of Upwelling in the West-Central South China Sea and Its Relationship with the Wind Field". In: *Applied Sciences* 13.9 (2023). ISSN: 2076-3417. DOI: 10.3390/app13095383. URL: https://www.mdpi.com/2076-3417/13/9/5383 (cit. on p. 9).

[30] Z. Huang, J. Hu, and W. Shi. "Mapping the Coastal Upwelling East of Taiwan Using Geostationary Satellite Data". In: *Remote Sensing* 13.2 (2021). ISSN: 2072-4292. DOI: 10.3390/rs13020170. URL: https://www.mdpi.com/2072-4292/13/2/170 (cit. on p. 10).

[31] V. Oerder et al. "Coastal Upwelling Front Detection off Central Chile (36.5–37°S) and Spatio-Temporal Variability of Frontal Characteristics". In: *Remote Sensing* 10.5 (2018). ISSN: 2072-4292. DOI: 10.3390/rs10050690. URL: https://www.mdpi.com/2072-4292/10/5/690 (cit. on p. 10).

[32] S. Yari, V. Mohrholz, and M. H. Bordbar. "Wind variability across the North Humboldt Upwelling System". In: *Frontiers in Marine Science* 10 (2023). ISSN: 2296-7745. DOI: 10.3389/fmars.2023.1087980. URL: https://www.frontiersin.org/articles/10.3389/fmars.2023.1087980 (cit. on p. 10).

[33] R. K. Walter et al. "Coastal upwelling seasonality and variability of temperature and chlorophyll in a small coastal embayment". In: *Continental Shelf Research* 154 (2018), pp. 9–18. ISSN: 0278-4343. DOI: https://doi.org/10.1016/j.csr.2018.01.002. URL: https://www.sciencedirect.com/science/article/pii/S0278434317302558 (cit. on p. 10).

[34] Z. Huang and X. H. Wang. "Mapping the spatial and temporal variability of the upwelling systems of the Australian south-eastern coast using 14-year of MODIS data". In: *Remote Sensing of Environment* 227 (2019), pp. 90–109. ISSN: 0034-4257. DOI: https://doi.org/10.1016/j.rse.2019.04.002. URL: https://www.sciencedirect.com/science/article/pii/S0034425719301397 (cit. on p. 10).

[35] H. Peng et al. "MuSTC: A Multi-Stage Spationdash;Temporal Clustering Method for Uncovering the Regionality of Global SST". In: *Atmosphere* 14.9 (2023). ISSN: 2073-4433. DOI: 10.3390/atmos14091358. URL: https://www.mdpi.com/2073-4433/14/9/1358 (cit. on p. 10).

[36] A. Azhar and H. Hashim. "A Review of Wind Clustering Methods Based on the Wind Speed and Trend in Malaysia". In: *Energies* 16.8 (2023). ISSN: 1996-1073. DOI: 10.3390/en16083388. URL: https://www.mdpi.com/1996-1073/16/8/3388 (cit. on p. 11).

[37] C. Y. Janse van Vuuren and H. J. Vermeulen. "Clustering of wind resource data for the South African renewable energy development zones". In: *Journal of Energy in Southern Africa* 30.2 (2019), 126–143. DOI: 10.17159/2413-3051/2019/v30i2a6316. URL: https://energyjournal.africa/article/view/6316 (cit. on p. 11).

[38] M. Yesilbudak. "Clustering analysis of multidimensional wind speed data using k-means approach". In: (2016), pp. 961–965. DOI: 10.1109/ICRERA.2016.7884477 (cit. on p. 12).

[39] A. Clifton and J. K. Lundquist. "Data Clustering Reveals Climate Impacts on Local Wind Phenomena". In: *Journal of Applied Meteorology and Climatology* 51.8 (2012), pp. 1547–1557. DOI: https://doi.org/10.1175/JAMC-D-11-0227.1. URL: https://journals.ametsoc.org/view/journals/apme/51/8/jamc-d-11-0227.1.xml (cit. on p. 12).

[40] A. Di Bernardino et al. "Classification of synoptic and local-scale wind patterns using k-means clustering in a Tyrrhenian coastal area (Italy)". In: *Meteorology and Atmospheric Physics* 134.30 (2022). DOI: 10.1007/s00703-022-00871-z. URL: https://doi.org/10.1007/s00703-022-00871-z (cit. on p. 12).

[41] M. C. Bueso et al. "Characterization of Vertical Wind Speed Profiles Based on Ward's Agglomerative Clustering Algorithm". In: *Journal of Modern Power Systems and Clean Energy* 11.5 (2023), pp. 1437–1449. DOI: 10.35833/MPCE.2022.000703 (cit. on p. 12).

[42] V. Güldal and H. Tongal. "Cluster analysis in search of wind impacts on evaporation". In: *Applied Ecology and Environmental Research* 6 (2008-12). DOI: 10.15666/aeer/0604_069076 (cit. on p. 12).

[43] I. Colak et al. "A data mining approach: Analyzing wind speed and insolation period data in Turkey for installations of wind and solar power plants". In: *Energy Conversion and Management* 65 (2013). Global Conference on Renewable energy and Energy Efficiency for Desert Regions 2011 "GCREEDER 2011", pp. 185–197. ISSN: 0196-8904. DOI: https://doi.org/10.1016/j.enconman.2012.07.011. URL: https://www.sciencedirect.com/science/article/pii/S019689041200297X (cit. on p. 12).

[44] G. Ratto, R. Maronna, and G. Berri. "Analysis of Wind Roses Using Hierarchical Cluster and Multidimensional Scaling Analysis at La Plata, Argentina". In: *Boundary-Layer Meteorology* 137.3 (2010), 477–492 (cit. on p. 12).

[45] M. R. Kousari, H. Ahani, and H. Hakimelahi. "An investigation of near surface wind speed trends in arid and semiarid regions of Iran". In: *Theoretical and Applied Climatology* 114.1-2 (2013), 153–168 (cit. on p. 12).

[46] A. Dokuz et al. "Year-Ahead Wind Speed Forecasting using a Clustering-Statistical Hybrid Method". In: (2018-09) (cit. on p. 13).

[47] V. Kushwah, R. Wadhvani, and A. K. Kushwah. "Trend-based time series data clustering for wind speed forecasting". In: *Wind Engineering* 45.4 (2021), pp. 992–1001. DOI: 10.1177/0309524X20941180. eprint: https://doi.org/10.1177/0309524X20941180. URL: https://doi.org/10.1177/0309524X20941180 (cit. on p. 13).

[48] J. M. Angosto et al. "Wind classification through cluster analysis for the development of predictive statistical models on atmospheric pollution". In: (2002). URL: https://api.semanticscholar.org/CorpusID:40274469 (cit. on p. 13).

[49] J. Liu et al. "A Process-Oriented Spatiotemporal Clustering Method for Complex Trajectories of Dynamic Geographic Phenomena". In: *IEEE Access* 7 (2019), pp. 155951–155964. DOI: 10.1109/ACCESS.2019.2949049 (cit. on p. 13).

[50] S. Beaver and A. Palazoglu. "Cluster Analysis of Hourly Wind Measurements to Reveal Synoptic Regimes Affecting Air Quality". In: *Journal of Applied Meteorology and Climatology* 45.12 (2006), pp. 1710 –1726. DOI: https://doi.org/10.1175/JAM2437.1. URL: https://journals.ametsoc.org/view/journals/apme/45/12/jam2437.1.xml (cit. on p. 13).

[51] E. AZIZI et al. "Wind Speed Clustering Using Linkage-Ward Method: A Case Study of Khaaf, Iran". In: *Gazi University Journal of Science* 32.3 (2019), 945–954. DOI: 10.35378/gujs.459840 (cit. on p. 13).

[52] B. Zhu et al. "A prediction model for wind farm power generation based on fuzzy modeling". In: *Procedia Environmental Sciences* 12 (2012). 2011 International Conference of Environmental Science and Engineering, pp. 122–129. ISSN: 1878-0296. DOI: https://doi.org/10.1016/j.proenv.2012.01.256. URL: https://www.sciencedirect.com/science/article/pii/S1878029612002575 (cit. on p. 14).

[53] N. G. Society. *What is the Coriolis effect?* Accessed: 2023-11-12. 2023. URL: https://education.nationalgeographic.org/resource/coriolis-effect/ (cit. on p. 15).

[54] *What is the Coriolis Effect?* Accessed on 2023-11-12. URL: https://scijinks.gov/coriolis/ (cit. on pp. 15, 16).

[55] Gaines, Steve and Airame, Satie. *Upwelling and the Creation of Rich Marine Habitats*. Accessed on 2023-11-12. 2017. URL: https://oceanexplorer.noaa.gov/explorations/02quest/background/upwelling/upwelling.html (cit. on p. 16).

[56] M. O. 101. *9.5: Ekman Spiral and Ekman Transport*. Accessed: 2024-02-06. 2024. URL: https://libretexts.org/@go/page/123456 (cit. on p. 18).

[57] Accessed: 2023-12-26. 2021. URL: https://www.offshoreengineering.com/oceanography/ekman-current-upwelling-downwelling/ (cit. on p. 18).

[58] P. T. Strub and C. James. "Evaluation of Nearshore QuikSCAT 4.1 and ERA-5 Wind Stress and Wind Stress Curl Fields over Eastern Boundary Currents". In: *Remote Sensing* 14.9 (2022). ISSN: 2072-4292. DOI: 10.3390/rs14092251. URL: https://www.mdpi.com/2072-4292/14/9/2251 (cit. on p. 19).

[59] EUMETSAT. *ASCAT - EUMETSAT*. Accessed on 03 Nov 2023. 2023. URL: https://www.eumetsat.int/ascat (cit. on p. 20).

[60] O. S. W. Team. *Advanced Scatterometer Data Products*. Accessed: 2023-11-03. 2023. URL: https://www.star.nesdis.noaa.gov/socd/oswt/ascat/ (cit. on p. 20).

[61] C. C. C. Service. *Climate reanalyses*. Accessed: 2023-11-03. 2023. URL: https://climate.copernicus.eu/climate-reanalysis (cit. on p. 20).

[62] C. C. C. Service. *ERA5 hourly data on single levels from 1940 to present*. Accessed on: 2023-12-05. Year. URL: https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview (cit. on p. 20).

[63] L. Ricciardulli and N. C. for Atmospheric Research Staff. *CCMP: Cross-Calibrated Multi-Platform wind vector analysis*. Climate Data Guide. Retrieved from https://climatedataguide.uca data/ccmp-cross-calibrated-multi-platform-wind-vector-analysis on 2023-11-28. 2022. DOI: 10.1002/2013eo130001 (cit. on p. 21).

[64] C. Mears et al. *Cross-Calibrated Multi-Platform (CCMP) Ocean Vector Wind Analysis*. https://www.remss.com/measurements/ccmp/. Accessed: 2023-12-05. 2022. URL: https://www.remss.com/measurements/ccmp/ (cit. on p. 21).

[65] S. Nascimento et al. "Core–shell clustering approach for detection and analysis of coastal upwelling". In: *Computers Geosciences* 179 (2023), p. 105421. ISSN: 0098-3004. DOI: https://doi.org/10.1016/j.cageo.2023.105421. URL: https://www.sciencedirect.com/science/article/pii/S0098300423001255 (cit. on pp. 22, 26, 82, 83, 89).

[66] S. Nascimento et al. "Novel Cluster Modeling for the Spatiotemporal Analysis of Coastal Upwelling". In: *Proceedings of the Conference on Spatiotemporal Data Analysis*. NOVA University, Lisboa, Portugal. 2024 (cit. on pp. 24, 26).

[67] J. P. B. Vargues. "Characterizing Time Ranges of Coastal Upwelling via Unsupervised Clustering". MA thesis. NOVA University Lisbon, 2024 (cit. on p. 26).

[68] B. Mirkin. "Clustering for Data Mining: A Data Recovery Approach". In: (2005-04), p. 266. DOI: 10.1201/9781420034912 (cit. on pp. 26, 27).

[69] S. Nascimento and N. Madaleno. "Unsupervised Initialization of Archetypal Analysis and Proportional Membership Fuzzy Clustering". In: *Intelligent Data Engineering and Automated Learning – IDEAL 2019*. Ed. by H. Yin et al. Cham: Springer International Publishing, 2019, pp. 12–20. ISBN: 978-3-030-33617-2 (cit. on pp. 27, 86, 90).

[70] S. Nascimento et al. "Automated computational delimitation of SST upwelling areas using fuzzy clustering". In: *Computers Geosciences* 43 (2012), pp. 207–216. ISSN: 0098-3004. DOI: https://doi.org/10.1016/j.cageo.2011.10.025. URL: https://www.sciencedirect.com/science/article/pii/S0098300411003608 (cit. on pp. 27, 86).

[71] L. Breiman. "Random forests". In: *Machine Learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324 (cit. on p. 27).

[72] A. Ziegler and I. R. König. "Mining data with random forests: current options for real-world applications". In: *WIREs Data Mining and Knowledge Discovery* 4.1 (2014), pp. 55–63. DOI: 10.1002/widm.1114 (cit. on pp. 27, 28).

[73] S. Christiansen. "Ischemic stroke thrombus characterization through quantitative magnetic resonance imaging". PhD thesis. 2021-04. DOI: 10.13140/RG.2.2.12667.64806 (cit. on p. 28).

[74] P. Probst, M. N. Wright, and A.-L. Boulesteix. "Hyperparameters and tuning strategies for random forest". In: *WIREs Data Mining and Knowledge Discovery* 9.3 (2019), e1301. DOI: 10.1002/widm.1301 (cit. on pp. 28, 48).

[75] M. Fernández-Delgado et al. "Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?" In: *Journal of Machine Learning Research* 15.90 (2014), pp. 3133–3181. URL: http://jmlr.org/papers/v15/delgado14a.html (cit. on p. 28).

[76] S. S et al. "Credit Risk Customers Categorization with Random Forest Classifier using Various Searching Techniques". In: *2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT)*. 2023, pp. 1–6. DOI: 10.1109/EASCT59475.2023.10392797 (cit. on p. 29).

[77] Y. Qi. "Random Forest for Bioinformatics". In: *Machine Learning in Bioinformatics*. Princeton, NJ: NEC Labs America, 2024, pp. 1–16 (cit. on p. 29).

[78] A Ajesh, J. Nair, and P. S. Jijin. "A random forest approach for rating-based recommender system". In: *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. 2016, pp. 1293–1297. DOI: 10.1109/ICACCI.2016.7732225 (cit. on p. 29).

[79] R. Hornung. "Ordinal Forests". In: *Journal of Classification* 37 (2019), pp. 4–17. DOI: 10.1007/s00357-018-9302-x (cit. on pp. 29, 30).

[80] R. Hornung. *Ordinal Forests: Prediction and Variable Ranking with Ordinal Target Variables*. R package version 2.4-3. 2022. URL: https://CRAN.R-project.org/package=ordinalForest (cit. on p. 29).

[81] S. Janitza, G. Tutz, and A.-L. Boulesteix. "Random forest for ordinal responses: Prediction and variable selection". In: *Computational Statistics Data Analysis* 96 (2016), pp. 57–73. ISSN: 0167-9473. DOI: https://doi.org/10.1016/j.csda.2015.10.005. URL: https://www.sciencedirect.com/science/article/pii/S0167947315002601 (cit. on p. 30).

[82] G. Tutz. "Ordinal Trees and Random Forests: Score-Free Recursive Partitioning and Improved Ensembles". In: *Journal of Classification* 39 (2022), pp. 241–263. DOI: 10.1007/s00357-021-09406-4 (cit. on p. 30).

[83] P. Cunningham and S. J. Delany. "k-Nearest Neighbour Classifiers - A Tutorial". In: *ACM Computing Surveys* 54.6 (2021), pp. 1–25. DOI: 10.1145/3459665 (cit. on pp. 30, 31, 50).

[84] E. Arslan. *Evren Arslan on Medium*. https://arslanev.medium.com/. Accessed: 2024-09-27. 2020 (cit. on p. 30).

[85] "The Earth Mover's Distance as a Metric for Image Retrieval". In: () (cit. on p. 31).

[86] Y. Li et al. "Predicting the Number of Nearest Neighbor for kNN Classifier". In: *IAENG International Journal of Computer Science* 46.4 (2019), Advance online publication: 20 November 2019. URL: http://www.iaeng.org/IJCS/issues_v46/issue_4/IJCS_46_4_16.pdf (cit. on p. 31).

[87] X. Zhang and Q. Song. "Predicting the number of nearest neighbors for the k-NN classification algorithm". In: *Intelligent Data Analysis* 18.3 (2014), pp. 449–464. DOI: 10.3233/IDA-140650 (cit. on p. 31).

[88] S. Yang et al. "Progressive neighbors pursuit for radar images classification". In: *Applied Soft Computing* 109 (2021), p. 107194. ISSN: 1568-4946. DOI: https://doi.org/10.1016/j.asoc.2021.107194. URL: https://www.sciencedirect.com/science/article/pii/S1568494621001174 (cit. on p. 31).

[89] IBM. *What is the K-nearest neighbors algorithm?* Accessed: 31 August 2024. 2021. URL: https://www.ibm.com/topics/knn (cit. on p. 31).

[90] A. Adeniyi, Z. Wai, and Y. Yongquan. "Automated Web Usage Data Mining and Recommendation System using K-Nearest Neighbor (KNN) Classification Method". In: *Applied Computing and Informatics* 36 (2014-10). DOI: 10.1016/j.aci.2014.10.001 (cit. on p. 31).

[91] M. A. Mukid et al. "Credit scoring analysis using weighted k nearest neighbor". In: *Journal of Physics: Conference Series*. Vol. 1025. IOP Publishing, 2018, p. 012114. DOI: 10.1088/1742-6596/1025/1/012114 (cit. on p. 31).

[92] Y. Wang et al. "Improved Handwritten Digit Recognition using Quantum K-Nearest Neighbor Algorithm". In: *International Journal of Theoretical Physics* 58 (2019), pp. 2331–2340. DOI: 10.1007/s10773-019-04124-5 (cit. on p. 31).

[93] H. Hersbach et al. *ERA5 hourly data on single levels from 1940 to present*. (Accessed on 06-02-2024). 2023. DOI: 10.24381/cds.adbb2d47 (cit. on pp. 33, 51).

[94] J. R. Holton and G. J. Hakim. *An Introduction to Dynamic Meteorology*. 5th. Elsevier, 2013. ISBN: 978-0-12-384866-6 (cit. on p. 33).

[95] Z. Okba, O. Cherkaoui Dekkaki, and H. Elouizgani. "UPWELLING ANALYSIS OF MOROCCAN ATLANTIC COAST BASED ON SATELLITE DATA." In: *Journal of Survey in Fisheries Sciences* 10 (2023-08). DOI: 10.53555/sfs.v10i2.1134 (cit. on p. 34).

[96] MetPy Developers. *Calculate the U, V wind vector components from the speed and direction*. Accessed: 2024-01-28. 2024. URL: https://unidata.github.io/MetPy/latest/api/generated/metpy.calc.wind_components.html (cit. on p. 34).

[97] Google. *Google Earth*. Accessed: 2024-01-28. URL: https://www.google.com/earth/ (cit. on p. 34).

[98] C. R. Harris et al. "Array programming with NumPy". In: *Nature* 585.7825 (2020-09), pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2 (cit. on p. 34).

[99] P. Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2 (cit. on p. 34).

[100] J. D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55 (cit. on p. 39).

[101] L. Liu and R. Vuillemot. "Categorizing Quantities using an Interactive Fuzzy Membership Function". In: *The 12th International Conference on Information Visualisation Theory and Applications*. Available online: https://hal.science/hal-03081984. Ecole Centrale de Lyon, Univ. Lyon, SICAL, LIRIS CNRS, France. 2021. DOI: 10.5220/0010270801950202 (cit. on p. 40).

[102] Q. Hamarsheh. *Different Types of Membership Functions in Fuzzy Logic Systems*. Online; accessed 26-June-2024. 2024 (cit. on p. 41).

[103] K. S. Gilda and S. L. Satarkar. "Analytical overview of defuzzification methods". In: *International Journal of Advance Research, Ideas and Innovations in Technology* 6.2 (2020), pp. 359–365. URL: https://www.researchgate.net/publication/356343424 (cit. on p. 43).

[104] A. Bhatt et al. "Explainable Artificial Intelligence in Retinal Imaging for the detection of Systemic Diseases". In: (2022-12). DOI: 10.48550/arXiv.2212.07058 (cit. on p. 46).

[105] T.-T. Wong. "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation". In: *Pattern Recognition* 48.9 (2015), pp. 2839–2846. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2015.03.009. URL: https://www.sciencedirect.com/science/article/pii/S0031320315000989 (cit. on pp. 47, 49).

[106] DataScience-ProF. "Unleashing the Power of GridSearchCV: A Comprehensive Guide to Hyperparameter Tuning". In: *Medium* (2023). URL: https://medium.com/@TheDataScience-ProF/unleashing-the-power-of-gridsearchcv-a-comprehensive-guide-to-hyperparameter-tuning-93054706d700 (cit. on p. 47).

[107] M. Olugbenga. "Balanced Accuracy: When Should You Use It?" In: *Neptune Blog* (2023). Accessed: 2024-09-25. URL: https://neptune.ai/blog/balanced-accuracy-when-should-you-use-it (cit. on p. 47).

[108] A. Unknown. "Understanding Precision and Recall in Machine Learning". In: (2024). Accessed: 2024-09-25. URL: https://www.analyticsvidhya.com/blog/2020/10/confusion-matrix-is-no-more-a-confusion/ (cit. on p. 47).

[109] P. Probst and A.-L. Boulesteix. "To tune or not to tune the number of trees in random forest?" In: *Journal of Machine Learning Research* 18 (2017-05). DOI: 10.48550/arXiv.1705.05654 (cit. on p. 48).

[110] S. Saxena. "Understanding and Tuning Random Forest Hyperparameters". In: (2024). Accessed: 2024-09-14. URL: https://www.example.com/random-forest-hyperparameters (cit. on p. 48).

[111] R. Hornung. *ordinalForest: Ordinal Forests: Prediction and Variable Ranking with Ordinal Target Variables*. R package version 2.4-3. 2022. URL: https://CRAN.R-project.org/package=ordinalForest (cit. on p. 49).

[112] S. Nascimento, B. Mirkin, and F. Moura-Pires. "Modeling proportional membership in fuzzy clustering". In: *IEEE Transactions on Fuzzy Systems* 11.2 (2003), pp. 173–186. DOI: 10.1109/TFUZZ.2003.809889 (cit. on p. 85).

[113] S. Almeida. *Fuzzy Clustering via Proportional Membership Model*. English. 1st. Vol. 119. Frontiers of Artificial Intelligence and Applications. Netherlands: IOS Press, 2005. ISBN: 978-1-58603-489-4 (cit. on p. 85).

[114] B. Mirkin and S. Nascimento. "Additive spectral method for fuzzy cluster analysis of similarity data including community structure and affinity matrices". In: *Information Sciences* 183.1 (2012), pp. 16–34. ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2011.09.009. URL: https://www.sciencedirect.com/science/article/pii/S0020025511004592 (cit. on pp. 87–89).

[115] gufotta. "Rayleigh-Ritz theorem". In: (2013-03). Accessed on 11 January 2024. URL: https://planetmath.org/rayleighritztheorem (cit. on p. 88).

[116] S. Nascimento and B. Mirkin. "Ideal type model and an associated method for relational fuzzy clustering". In: *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. 2017, pp. 1–6. DOI: 10.1109/FUZZ-IEEE.2017.8015473 (cit. on p. 89).

[117] T. Danka. *How the dot product measures similarity*. Accessed: 2024-01-11. 2024. URL: https://tivadardanka.com/blog/how-the-dot-product-measures-similarity (cit. on p. 89).

[118] B. O'Connor. *Cosine Similarity, Pearson Correlation, and OLS Coefficients*. Web Page. Accessed: 2024-01-11. 2012-03. URL: https://brenocon.com/blog/2012/03/cosine-similarity-pearson-correlation-and-ols-coefficients/ (cit. on p. 89).

[119] M. Meilă. "Comparing clusterings—an information based distance". In: *Journal of Multivariate Analysis* 98.5 (2007), pp. 873–895. ISSN: 0047-259X. DOI: https://doi.org/10.1016/j.jmva.2006.11.013. URL: https://www.sciencedirect.com/science/article/pii/S0047259X06002016 (cit. on p. 90).

[120] S. Kulczyński. *Die pflanzenassoziationen der pieninen*. Vol. 3. Imprimerie de l'Université, 1928 (cit. on p. 90).

[121] F. R. Zakani et al. "Kulczynski similarity index for objective evaluation of mesh segmentation algorithms". In: *2016 5th International Conference on Multimedia Computing and Systems (ICMCS)*. 2016, pp. 12–17. DOI: 10.1109/ICMCS.2016.7905611 (cit. on p. 90).

A

Initial proposed framework

A.1 Introduction

The primary aim of this dissertation is to investigate the correlation between wind patterns and automatically segmented coastal upwelling regions derived from SST maps, obtained from MODIS-Aqua satellite imagery.

Recently, it was developed a spatiotemporal clustering framework and its software tool, the core-shell clustering framework [5, 6, 65] for the automatic segmentation of coastal upwelling regions and its spatiotemporal analysis from SST maps derived from the MODIS-Aqua satellite images. The framework was successfully applied to 16 years of SST data of the Canary current upwelling system (CCUS) (Portuguese coast, North and South Morocco).

The core-shell clustering framework unsupervisedly derives stable time ranges to define upwelling periods of (relative) stability along the upwelling seasons.

Taking as an example the region of South Morroco (20ºN-28ºN; 8ºW-13ºW), the upwelling season period of 1st of April to the 31st of October 2007, corresponding to 23 weekly SST grid maps or SST instants. The core-shell clustering framework unsupervisedly found four time ranges, where each time range is a set of consecutive SST instants. In this case, the derived ranges were the following: 1-9, 10-13, 14-19, 20-23 (designation *i-j* denotes a time range starting at SST instant *i* and ending at SST instant *j*).

Figure A.1 (top row) shows, for the second time range (span), its SST instant maps after the pre-processing stage, followed (second row) by the obtained core-shell segmentations. The core region, highlighted in yellow, defines the "core" of upwelling, which is almost constant along the SST grids of the time range, while the shells, corresponding to the evolving upwelling region, vary.

We are interested in studying the variations and trends of winds, as they are the main driver of coastal upwelling formation and its spatiotemporal dynamics.

Figure 1.2 shows, for each time range (top to bottom), the corresponding core-shell cluster followed by the sequence of wind instant maps.

Interestingly, we can observe a wind intensity pattern in each time range as well as

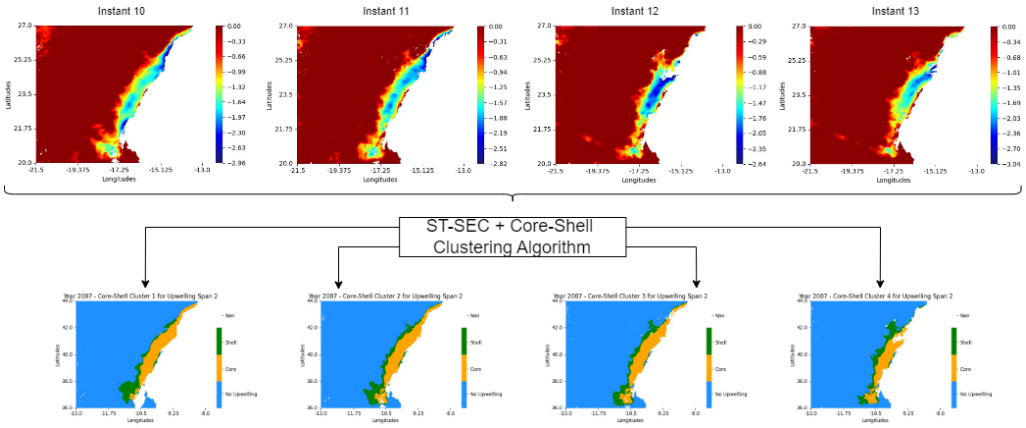


Figure A.1: Core-shell segmentations for the second time range of 2007

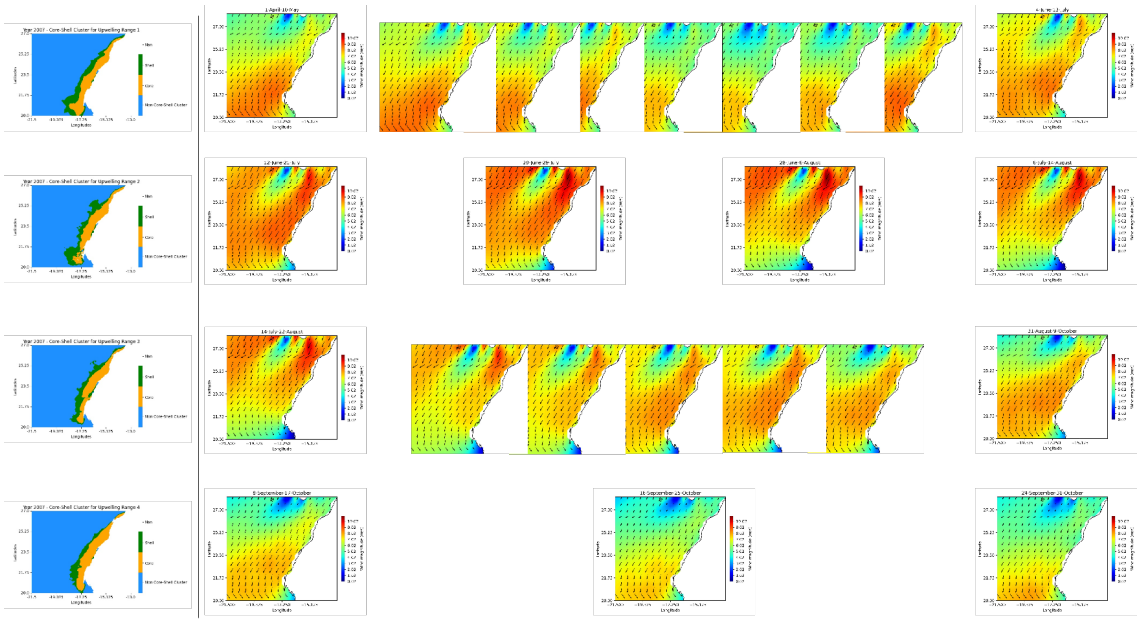


Figure A.2: Wind map sequences for the year 2007

wind intensity transitions between the four core-shell time ranges.

We will use fuzzy clustering techniques to segment wind maps. The relationship between the resulting wind clusters and coastal upwelling regions will be investigated. Features extracted from these clusters will be used to construct time series, allowing the study of wind variations and trends over an extended period. This analysis will be contrasted with the long-term spatiotemporal study of coastal upwelling, as detailed in [5, 65].

A.2 Fuzzy Clustering with proportional membership

A.2.1 Mathematical model

The mathematical model behind Fuzzy clustering with proportional membership (FCPM) comes from the idea of Data-Driven Cluster Modeling. Here, the structure underlies the data in the form of a simple statistical equation:

$$observeddata = modeldata + noise. \tag{A.1}$$

As the model is not prespecified, but rather derived from the data, '$noise$' can be considered as the set of differences between '$observeddata$' and '$modeldata$', where the differences should be minimized by fitting the data model.

In this clustering model, it is assumed that the observed entities share part of the prototypes. In this way, an entity's membership in a cluster expresses the proportion of the cluster's prototype reflected in the entity.

Thus, equation A.1 is adapted to fit a *Generic proportional Membership Model* formally described by:

$$y_{kh} = u_{ik}v_{ih} + e_{ikh}, \tag{A.2}$$

where y_{kh} is the entity, u_{ik} is the membership value, v_{ih} the prototype, and e_{ikh} the residual values, small as possible. $k = \{1, \ldots, n\}$, $h = \{1, \ldots, p\}$, and $i = \{1, \ldots, c\}$. Where n is the number of entities, p is the number of features, and c is the number of clusters.

To get the smallest possible residual values, the *Generic Square-Error criterion* (A.2) can be defined as the minimization of all the residual values:

$$E_0(U, V) = \sum_{i=1}^{c}\sum_{k=1}^{n}\sum_{h=1}^{p}(y_{kh} - u_{ik}v_{ih})^2, \tag{A.3}$$

concerning the fuzzy constraints:

$$0 \leq u_{ik} \leq 1, \text{for all } i = 1, \ldots, c, k = 1, \ldots, n, \tag{A.4}$$

$$\sum_{i=1}^{c} u_{ik} = 1, \text{for all } k = 1, \ldots, n. \tag{A.5}$$

The problem with criterion (A.3) is that by intuition, we might only consider meaningful proportions the ones that have a high enough membership value, and thus a lower residual

value. To smooth the influence of high e_{ikh}, we weight the residuals by a power of m ($m = 1, 2$) leading to the final equation to be minimized:

$$E_m(U, V) = \sum_{i=1}^{c} \sum_{k=1}^{n} \sum_{h=1}^{p} u_{ik}^{m} (y_{kh} - u_{ik} v_{ih})^2. \tag{A.6}$$

Regarding the fuzzy constraints defined by (A.4) and (A.5)

The generic clustering criterion defined by A.6 leads to a special case when $m = 0$, equation A.2. There are other variants of this algorithm, however, we will focus on the FCPM-0 and FCPM-2 versions ($m = 0$ and $m = 2$). According to the work of Nascimento et al. [112], an initial idea we have is that FCPM-0 appears to perform better in the lower-dimensional cases. In contrast, FCPM-2 is more effective in the higher-dimensional case. This suggests that, according to the typological model, these variants can serve as reliable indicators of the natural number of clusters in the data. It has also been concluded that among all of the variants of the FCPM-m algorithm, the fastest one is when $m = 0$. This is mainly due to the lesser time spent in minor iterations during gradient projection and the faster convergence to an optimum of the minor iteration cycle.

A.2.2 Algorithm

The main steps of the algorithm are the following [112]:

1. Choose the adequate number of clusters, c, the parameter m, stopping threshold ϵ, and the maximum number of iterations N

2. Initialize the cluster centroids (prototypes) matrix, V, the membership matrix U, and the iteration counter, t

3. Repeat until convergence or the maximum number of iterations is reached

    a) Given V, update the matrix U by solving a constrained optimization problem using the gradient projection method

    b) Given U, update V by solving a system of linear equations for each prototype

    c) Increment t by 1 and check the stopping condition based on the difference between the current V matrix and the old V matrix. If the difference is less than ϵ stop the algorithm and return U and V as the final solution

A detailed description of the FCPM algorithm may be consulted in [113].

A.2.3 FCPM properties

The model assumes the existence of some prototypes that serve as "ideal" patterns as data entities. To relate the prototypes to the observations, we assume that the entities share parts of the prototypes, such that an entity may bear 70% of a prototype V_1 and 30% of prototype V_2, which simultaneously expresses the entity's membership to the respective clusters. The underlying structure of this model can be described by a fuzzy K-partition defined in such a way that the membership of an entity to a cluster expresses the proportion of the cluster's prototype present in the entity.

The FCPM-m leads to distinct fuzzy clustering c-partitions depending on the fuzziness parameter m: cluster structures with central prototypes (FCPM-0, FCPM-1), closely matching the popular Fuzzy C-means (FCM), as well as cluster structures with extremal prototypes (FCPM-2), close to the concept of ideal, or archetypal type [69].

Although the FCPM algorithm requires solving a more complex optimization problem to meet fuzziness constraints, its major advantage is the capacity to reveal extreme and ideal prototypes, making it more suitable to model typological structures.

This is interesting to explore clustering wind from wind maps since wind clusters represented by central/extreme prototypes should provide meaningful wind patterns for our proposed study. Overall, this algorithm is suitable for this study because the usual conditions for triggering upwelling are usually extreme compared to the typical weather, and upwelling regions are not uniformly distributed. Using the concept of *proportional membership*, we can better understand how strongly an area is influenced by upwelling. The ideal prototypes can represent prototypical winds as trigger conditions for coastal upwelling.

A.2.4 FCPM Segmentation and Visualization of wind maps

To segment and visualize the wind instant maps via FCPM we will apply a similar approach as the one in the FUZZYUPWELL system [70]. Each $L_1 \times L_2$ wind map is converted into a set $W = w_i$ $(i = 1 \ldots N)$, $N = L_1 \times L_2$ wind feature vectors. Each feature vector, w_{i1}, is the average/maximum wind speed value. The FCPM fuzzy c-partitions result in c fuzzy wind clusters. To fine-tune the number of clusters in FCPM partitions we will adopt the experimental protocol proposed in [69].

Each fuzzy c-partition is visualized as a wind map following the approach in [70]. Clustered wind points are displayed in the corresponding spatial location of the wind map and colored according to the color label (prespecified) of its cluster and its membership value.

A.3 Fuzzy Additive Spectral Clustering

A.3.1 Mathematical model

The mathematical model of the Fuzzy Additive Spectral Clustering (FADDIS) assumes the observed similarity matrix B is the sum of K fuzzy clusters A_K and a residual similarity E, denoted by the following equation:

$$B = A_1 + A_2 + \cdots + A_K + E. \tag{A.7}$$

Where E is to be minimized over unknown clusters.

Each fuzzy cluster A is defined by a membership vector and an intensity value μ such that:

$$A = \mu^2 u.u^T, \tag{A.8}$$

where $a_{ij} = (\mu u_i)(\mu u_j)$ and $\mu^2 u_i u_j$ is the contribution of the cluster to similarity a_{ij} between i and j.

This method does not require a defined number of K of clusters, instead, it extracts them one by one with the following criterion:

$$\min_{u,\zeta} \sum_{i,j \in I} (b_{ij} - \zeta u_i u_j)^2. \tag{A.9}$$

Concerning unknown weight $\zeta > 0$ ($\zeta = \mu^2$) and fuzzy membership vector $u = (u_i)$, given similarity value b_{ij}.

The data scatter measures how much the similarity values vary in the data matrix. It is defined as the sum of the squares of the similarity values [114]:

$$S = \sum b_{ij}^2. \tag{A.10}$$

We can then find ζ by minimizing A.9 for an arbitrary u or maximize $G(u)$ which is the contribution of the cluster to the data scatter:

$$\max_{|||u||\neq 0} \frac{u^T B u}{u^T u}. \tag{A.11}$$

Here, the maximum value is the maximum value of matrix B, reached at the corresponding eigenvector, by the Rayleigh-Ritz theorem [115].

So here enters the Spectral clustering approach, from the B matrix, we will get the maximum eigenvalue, λ, and the corresponding eigenvector, normalizing the latter one. By doing this, the eigenvector becomes a unitary vector simplifying computations and improving their stability, avoiding the curse of dimensionality.

This will extract each cluster sequentially, starting with the one with the highest contribution, and subtracting it from the observed similarity matrix, B, obtaining a new similarity matrix:

$$B_{K+1} = B_K - \mu_K^2 u_k . u_k^T. \tag{A.12}$$

The extraction will stop when either the eigenvalue for the next cluster is negative or its contribution is below a defined threshold.

A.3.2 Algorithm

The main steps to the algorithm are the following [114]:

1. Compute the similarity matrix and choose the threshold of the contribution of an individual cluster, $\epsilon > 0$

2. Initialize by setting $k = 1$ and $S = \sum_{i,j} b_{ij}^2$

3. Find all of the set of positive eigenvalues $\Lambda = \{\lambda\}$ and corresponding normed eigenvectors $Z = \{z_\lambda\}$ for matrix B

4. If Λ is empty stop the computation and output the clusters found

5. Take the eigenvectors z and $-z$ corresponding to the maximum eigenvalue, λ, and compute the fuzzified projections

6. Take the eigenvector that maximizes the contribution, $G(u)$, as u_k along with the corresponding $\mu = u_k^{'} B u_k$ and contribution $G(u_k)$

7. If $G(u_k) < \epsilon$ stop the computation, with k being k the number of clusters found. Otherwise, add 1 to k, and set $B = B$ - $\mu^2 u_k u_k^{'}$ and go to step 2

8. Output K, the number of clusters, the cluster membership matrix U, the cluster intensity values, and the cluster's contributions

A.3.3 FADDIS properties

FADDIS [114, 116] is a fuzzy clustering algorithm that iteratively extracts clusters one by one having a stopping condition derived from the clustering criterion, that allows to control the number of clusters to retrieve from data. FADDIS was applied to the analysis of research tendencies in Data Science with a proper similarity measure proposed by the authors [116]. The intensities of clusters are related to the clusters' contribution to the total data scatter, being useful for interpreting cluster results. The ability to handle different types of similarity data is also useful, allowing for the analysis of the phenomenon.

A.3.4 FADDIS Segmentation and Visualization of wind maps

The approach used here differs from FCPM, as both input and output are distinct. The input of FADDIS algorithm is a squared similarity matrix measuring the proximity between pairs of entities to be clustered.

In our work, we will take each $L_1 \times L_2$ wind map and build a $N \times N$ similarity matrix ($N = L_1 \times L_2$) using the inner product of two wind points intensities, taking as similarity measure the inner product [117].

Geometrically interpreting it, given two vectors $\mathbf{x}$ and $\mathbf{y}$, we can decompose $\mathbf{x}$ in two parts, x_0 and x_1, where one will be orthogonal to $\mathbf{y}$ and the other parallel to $\mathbf{y}$, defined like this:

$$x_1 = cy, (c \in \mathbb{R}),$$
$$\langle x_0, y \rangle = 0.$$

As mathematically derived in [117], applying the inner product properties to this set of equations, we can see that if $\mathbf{x}$ and $\mathbf{y}$ are both unitary vectors, their inner product yields the magnitude of the orthogonal projection, which measures the degree of similarity between them. Several other variants of the inner product exist [118], each with their characteristics regarding output range and geometric interpretation, but this similarity measure suffices since the vectors are expected to be normalized before computations. The use of the inner product a similarity measure was proposed in an extension of FADDIS in [116].

We will apply the FADDIS algorithm having as input inner-product wind similarity matrices and will fine-tune its stop condition to model the number of clusters to provide the best wind map segmentations.

A.4 Validation of Wind Map Segmentations

The SST upwelling regions segmentations obtained from the analysis of almost 20 years of SST data of the CCUS (including the Portuguese coast, North and South Morocco) had been validated by expert oceanographers as well as with popular validity indices [5, 65]. Therefore, they constitute reliable ground truth to be used in our study.

We propose to apply a collection of set validity indices (ARI [119], NMI [119], Kulczynski similarity index [120, 121]) to assess the quality of wind partitions obtained with our clustering algorithms as well as corresponding fuzzy validity indices as applied in [69].

B

Experimental study appendix

B.1 Appendix structure

This appendix presents the results obtained in the phases of the experimental study. The results are presented chronologically regarding the proposed experimental methodology.

B.2 North Morroco coastline angles

Starting Latitude	Ending Latitude	Coastline Angle
30°N	30.25°N	140°
30.5°N	31.25°N	180°
31.5°N	32.25°N	209°
32.5°N	33.25°N	221º
33.5°N	34°N	240°
34.25°N	35°N	202°

Table B.1: Coastline angles for the coast of North Morroco

B.3 Sample dataset for the North Morroco region-Year 2007

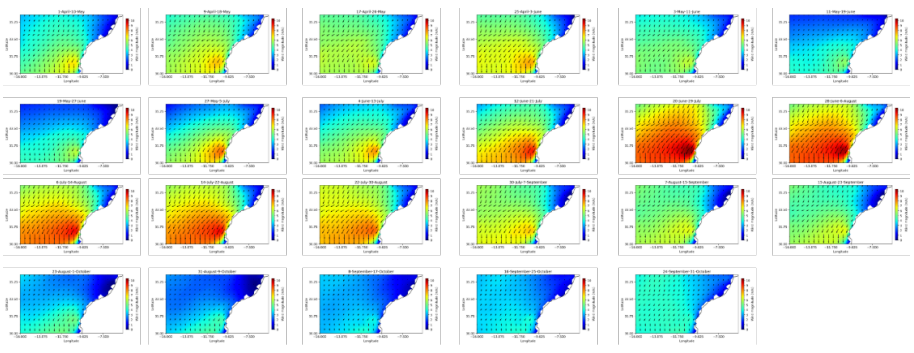


Figure B.1: Full length of the instants derived from applying the pipeline to North Morroco. Year 2007, from April to October

B.4 Average clustered wind stress anomaly evolution

In this section are presented the cluster average WSA evolution plots for the years 2007 and 2015 for the NM and SM geographic regions.

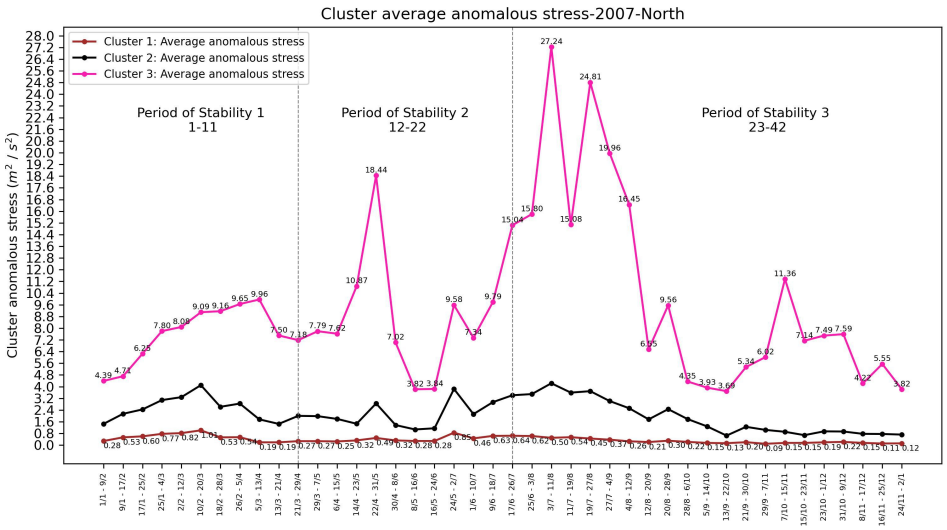


Figure B.2: Evolution of the average cluster WSA in 2007-NM geographic region

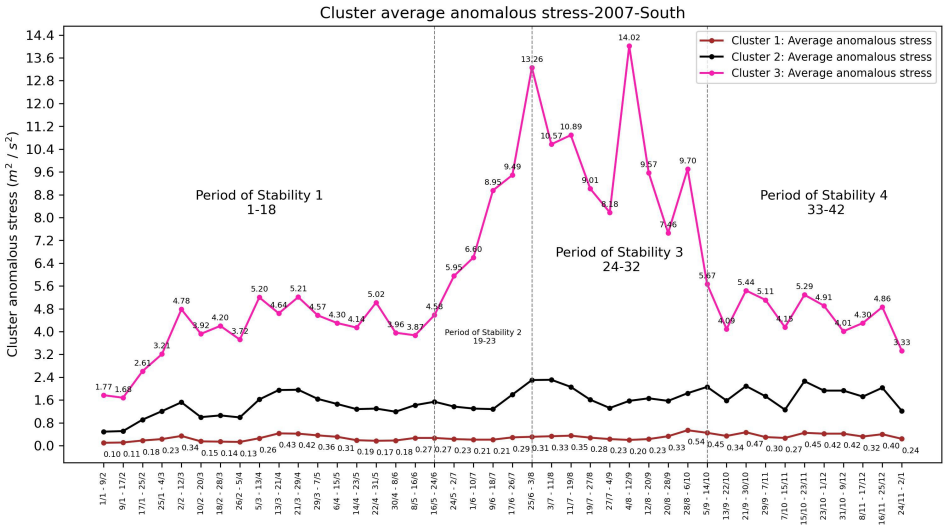


Figure B.3: Evolution of the average cluster WSA in 2007-SM geographic region

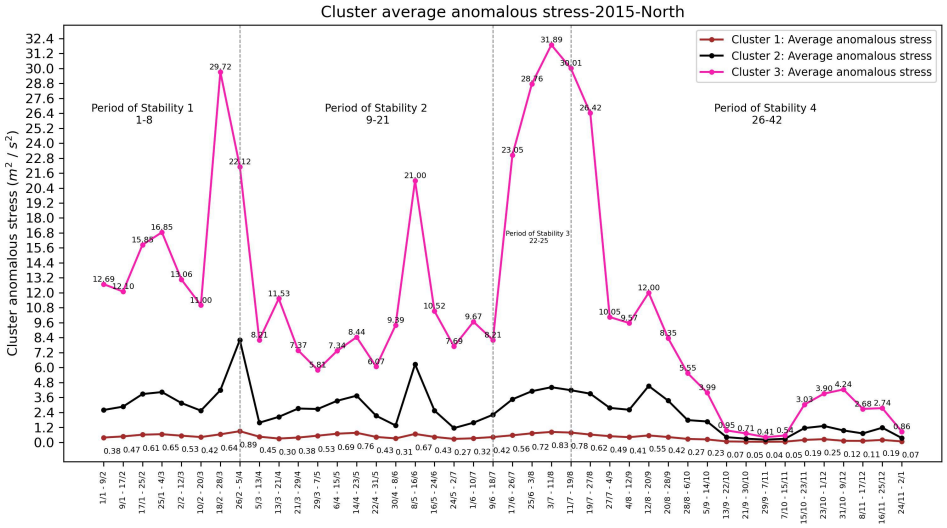


Figure B.4: Evolution of the average cluster WSA in 2015-NM geographic region

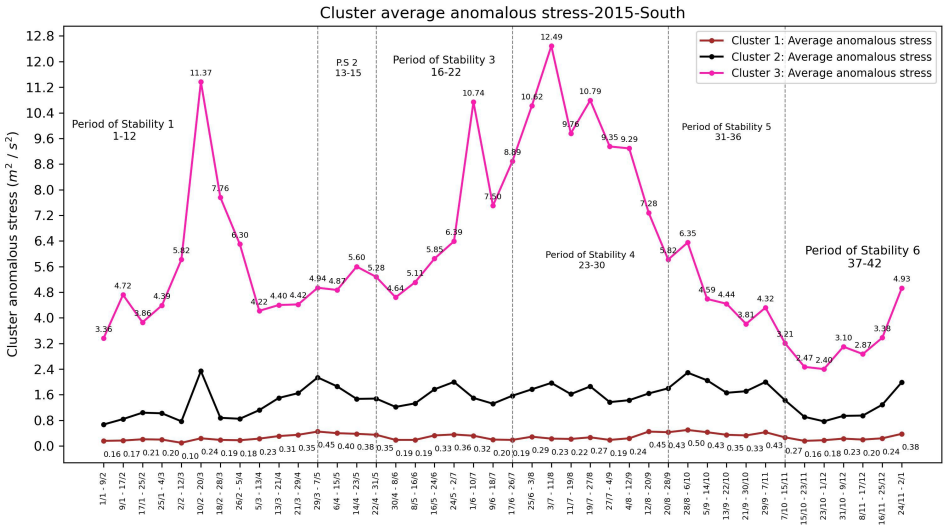


Figure B.5: Evolution of the average cluster WSA in 2015-SM geographic region

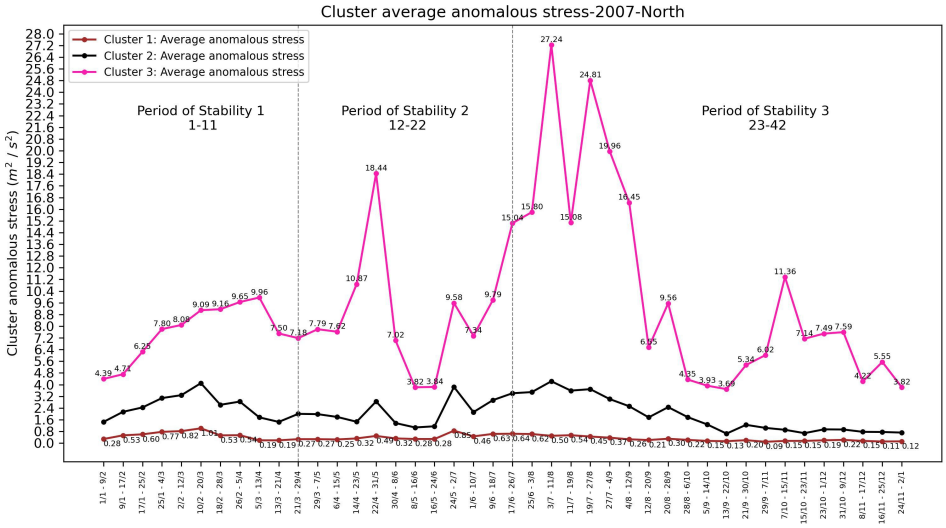


Figure B.6: Evolution of the average cluster WSA in 2007-NM geographic region

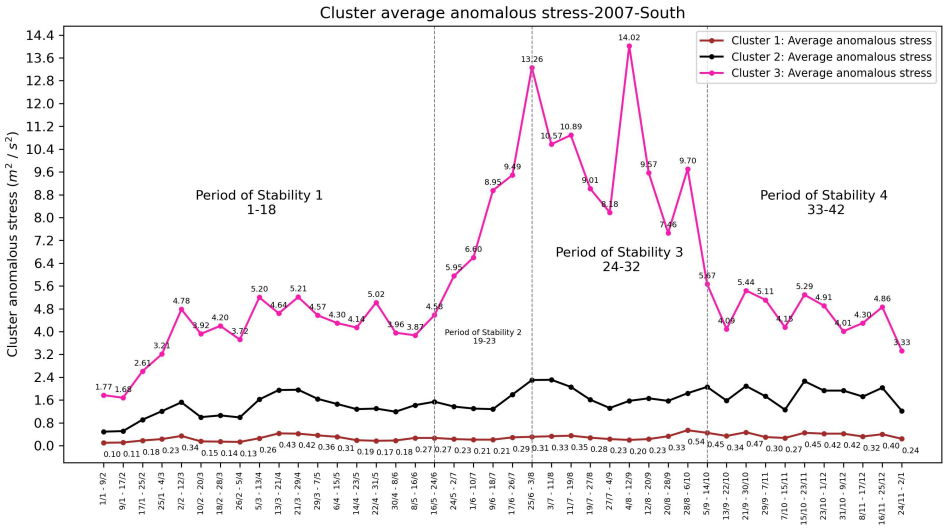


Figure B.7: Evolution of the average cluster WSA in 2007-SM geographic region

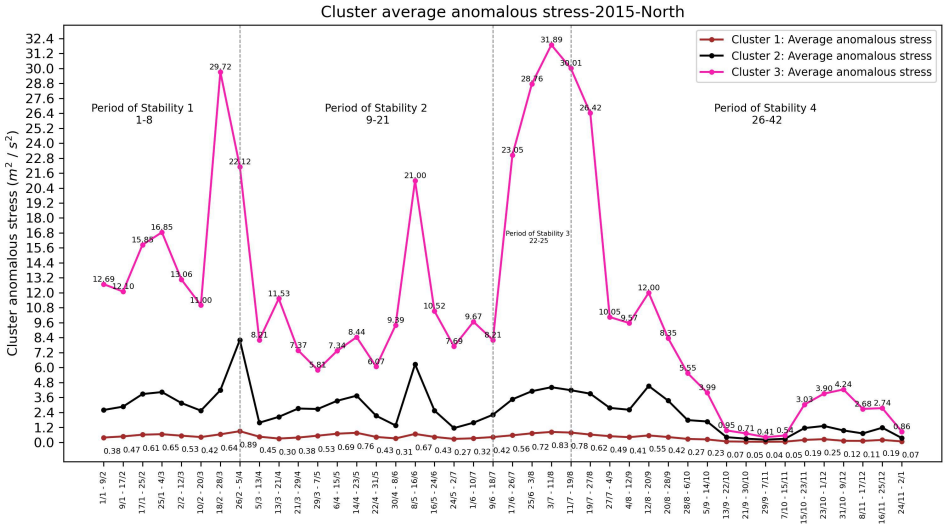


Figure B.8: Evolution of the average cluster WSA in 2015-NM geographic region

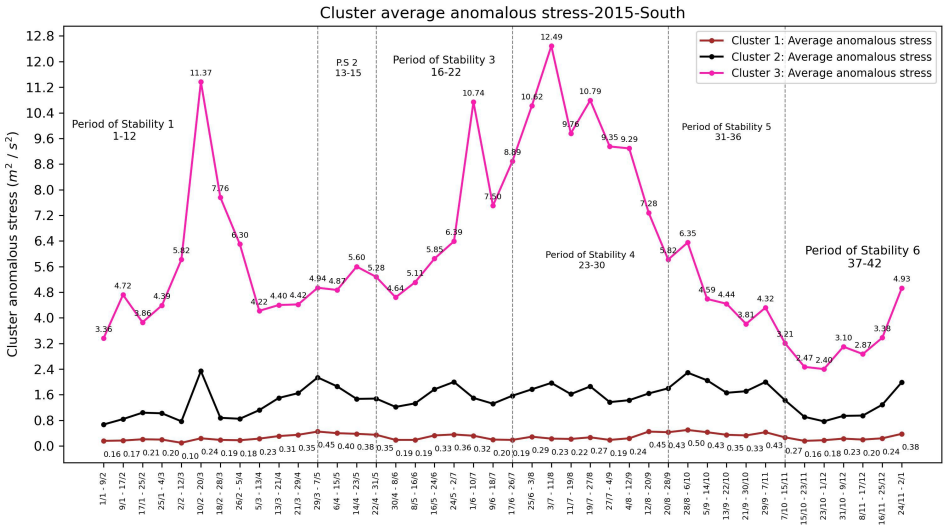
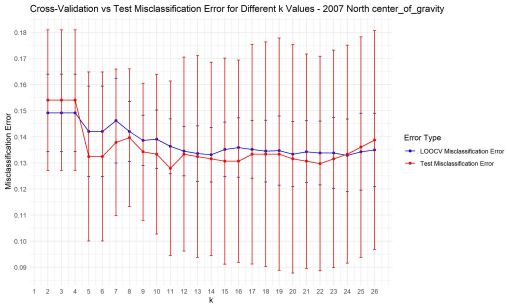


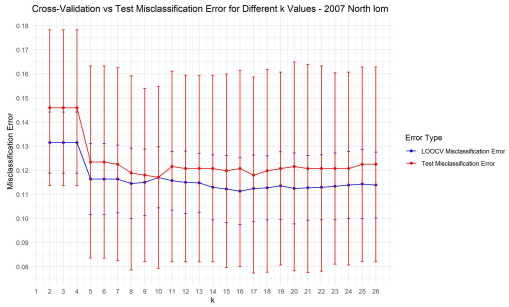
Figure B.9: Evolution of the average cluster WSA in 2015-SM geographic region

B.5 Learning curves for KNN models

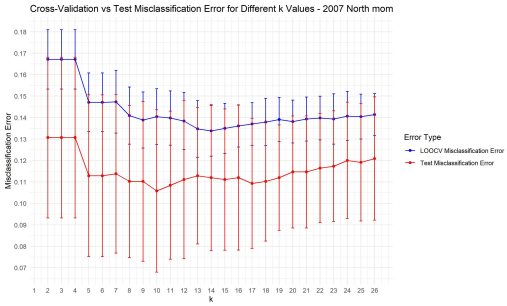
This section presents the learning curves for each K-NN model constructed.



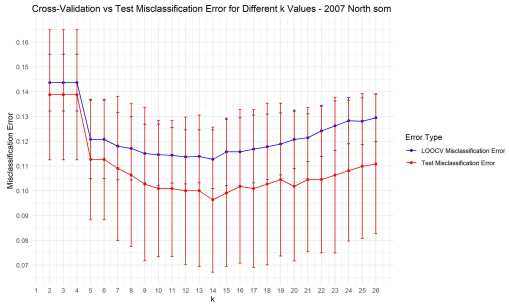
(a) 2007-NM-COG



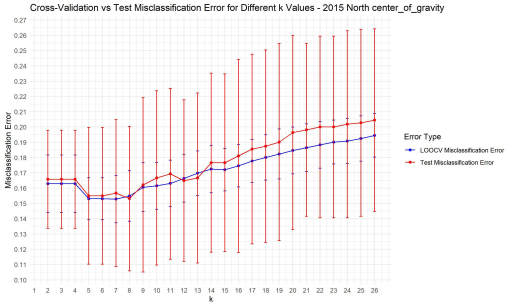
(b) 2007-NM-LOM



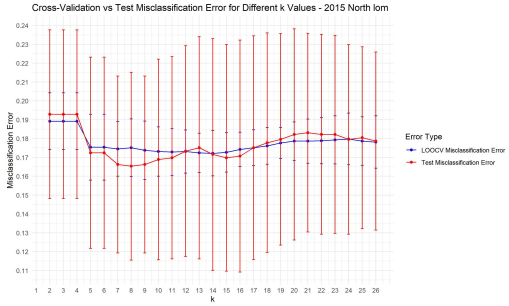
(c) 2007-NM-MOM



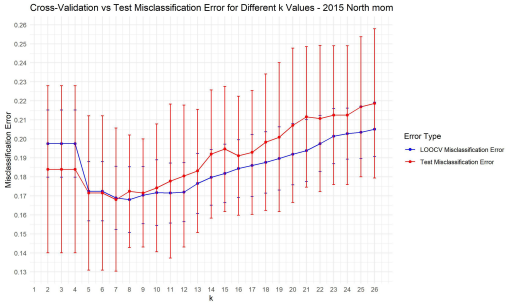
(d) 2007-NM-SOM



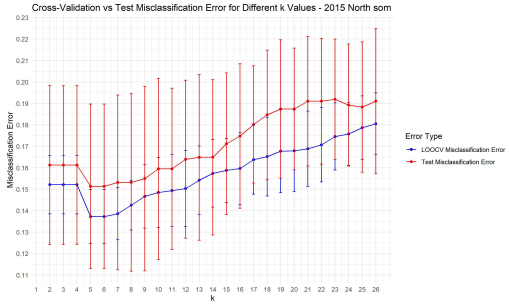
(a) 2015-NM-COG



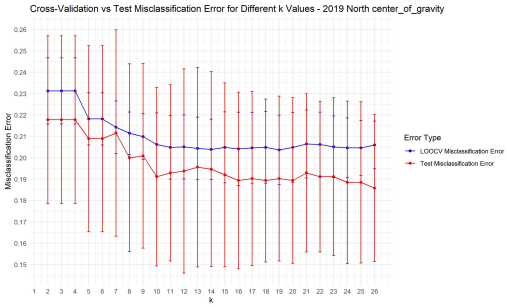
(b) 2015-NM-LOM



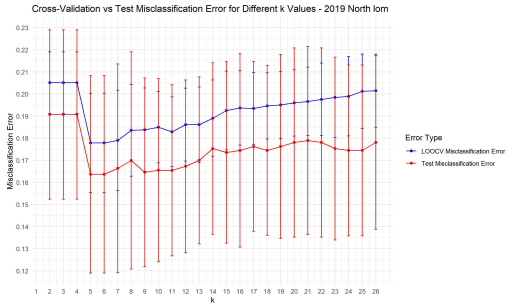
(c) 2015-NM-MOM



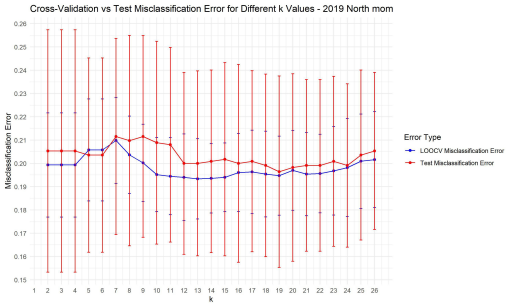
(d) 2015-NM-SOM



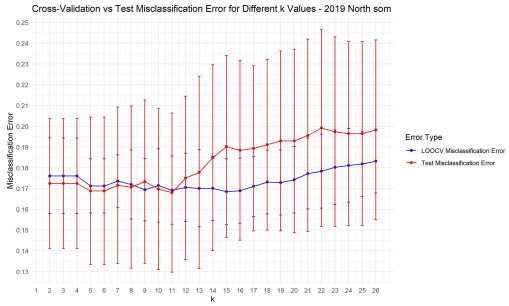
(a) 2019-NM-COG



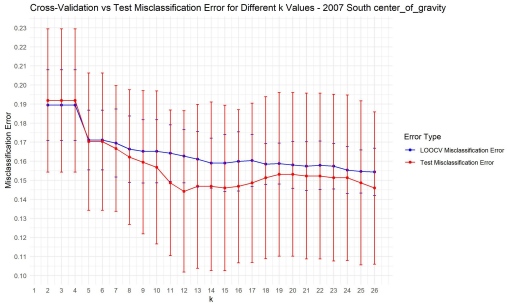
(b) 2019-NM-LOM



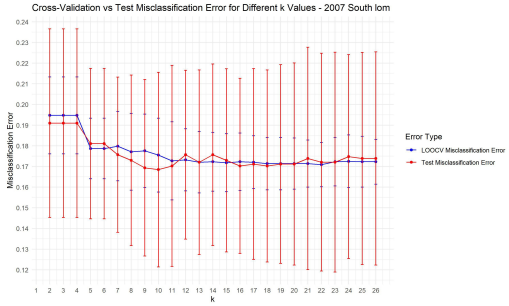
(c) 2019-NM-MOM



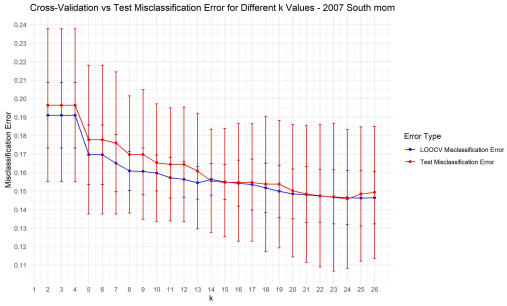
(d) 2019-NM-SOM



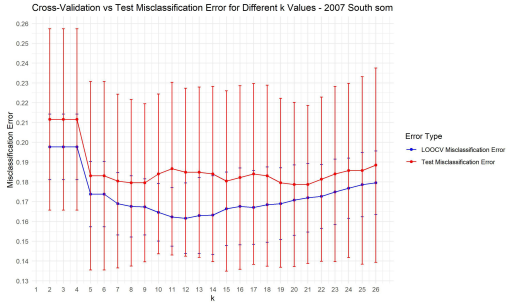
(a) 2007-SM-COG



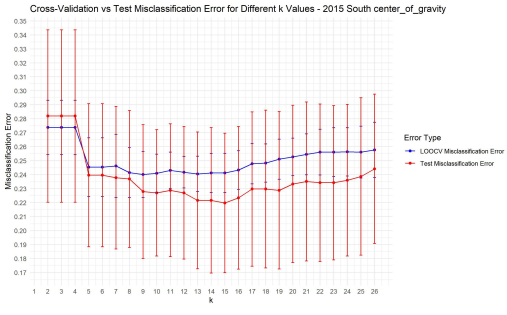
(b) 2007-SM-LOM



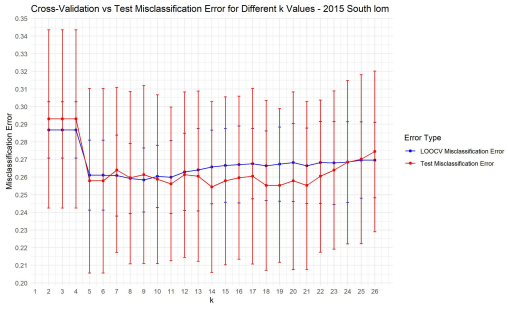
(c) 2007-SM-MOM



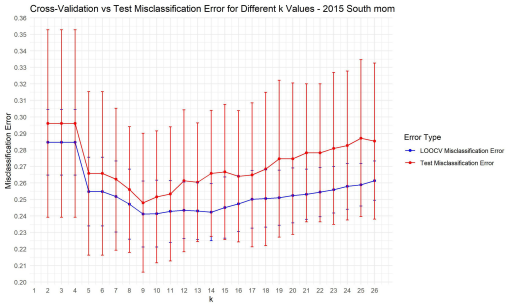
(d) 2007-SM-SOM



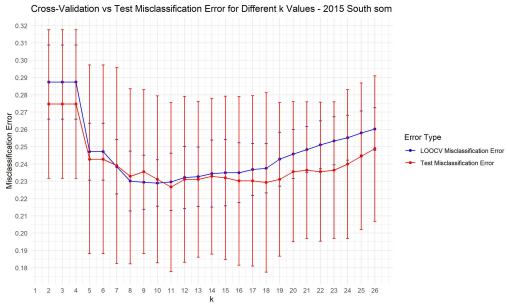
(a) 2015-SM-COG



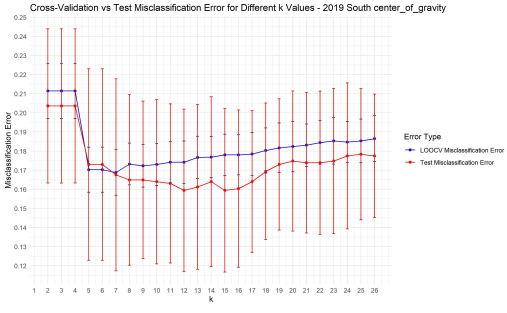
(b) 2015-SM-LOM



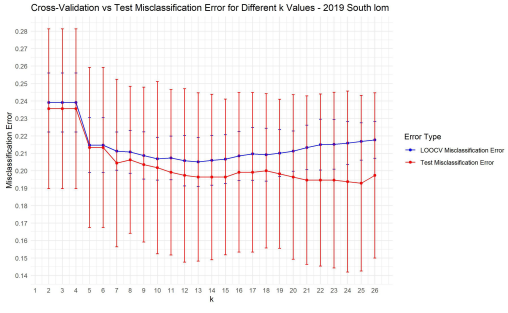
(c) 2015-SM-MOM



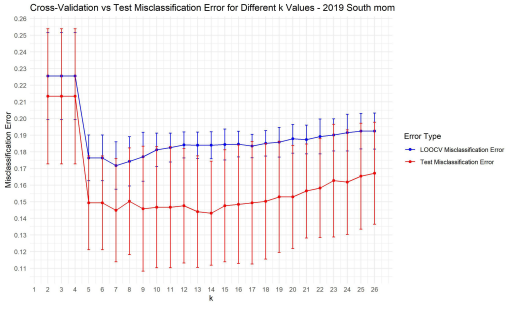
(d) 2015-SM-SOM



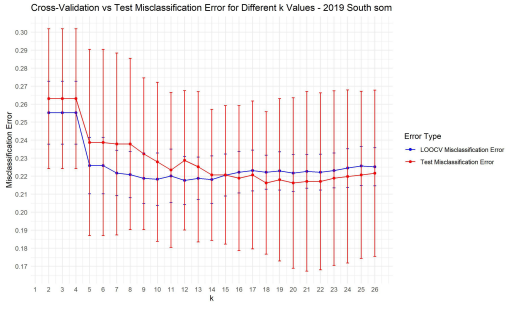
(a) 2019-SM-COG



(b) 2019-SM-LOM



(c) 2019-SM-MOM



(d) 2019-SM-SOM