Università degli Studi di Ferrara

DOTTORATO DI RICERCA IN

" SCIENZE DELLA TERRA E DEL MARE"

Settore Scientifico

Disciplinare GEO02

CICLO XXXV

COORDINATORE

Prof. Paolo Ciavola

PREDICTION OF POTENTIAL FISHING AREAS AROUND THE GALAPAGOS ISLANDS FOR A SUSTAINABLE EXPLOITATION OF THE FISHERY RESOURCES

Dottorando

CPFG-EMT Chávez Fernando

Tutore

Prof. Roberta Ivaldi

Prof. Massimo Coltorti

Prof. Guido Sciavicco (Co-tutor)

Anni 2019/2022

DEDICATION

I dedicate this thesis to my family, who gave me the support and the understanding to successfully complete this stage of my professional life. The time spent in the development of this thesis has forced me to postpone spending time with them, so part of the effort made is thanks to the immense love of my wife and my daughters.

Likewise, I would like to dedicate this thesis to my tutors Massimo Coltorti, Roberta Ivaldi, and Guido Sciavicco, thanks to whom I have managed to present in this document a proposal that I hope will be very useful for the Ecuadorian Navy and especially for the Ecuadorian State.

ACKNOWLEDGEMENTS

Firstly, I would like to thank God for the life he has given me and because without him, nothing would be possible.

To my parents who take care of me from heaven thanks to their example and dedication, I have become a good man with principles and values that have accompanied me throughout my life.

To my wife and daughters, thanks to their company and understanding, I am completing a very important stage in my professional life.

To my siblings and aunts, who have trusted and supported me throughout my life.

And finally, to my tutors, since without their teachings and support, it would not have been possible to successfully complete this research, which will surely be of great value for Ecuador and the protection of the fishery resources of our country.

TABLE OF CONTENTS

TABLE CAPTIONS

FIGURE CAPTIONS

ABSTRACT

The increase of the world population and the lack of protein worldwide have led to an overexploitation of fishery resources in recent years, seriously affecting the world's marine ecosystems, especially for those countries whose economy is based on the exploitation of this type of resources, as the case of Ecuador. Year after year, due to the wealth of nutrients and marine resources around the Galapagos Islands, the country has been affected by overfishing both inside and outside the exclusive economic zone, in fact the particular morphology and geodynamics of seabed, the interface processes and water masses circulation are the main factors and constrains to the marine resources and the presence of marine protected area in the Galapagos Island. In this context, the main objective of this study is to provide the Ecuadorian Navy with a tool that can give timely information to determine the potential fishing areas around the Galapagos Islands in order to properly plan naval operations that combat illegal, unreported and unregulated fishing. To this end, two types of experiment have been carried out: for the first experiment, chlorophyll has been taken as an objective variable in order to identify the places where there is a higher level of biomass, previously predicted using machine learning techniques, such as artificial neural networks, and optimized multivariate regression using Fourier series; for the second experiment, the statistical analysis of the fishing effort has been used to determine the ranges in which the ocean variables must be so that there is a greater probability of fishing. With this information, a computer tool has been created, which proposes daily potential fishing areas, under the aforementioned perspectives, as well as presents information on the historical behaviour of the national and international fishing fleet around the Galapagos Islands and the daily conditions of the oceanographic variables considered in this study.

KEYWORDS: Illegal fishing, chlorophyll, potential fishing areas, Galápagos.

ABSTRACT

L'aumento della popolazione mondiale e la carenza di proteine a livello mondiale hanno portato negli ultimi anni a un sovrasfruttamento delle risorse ittiche, colpendo gravemente gli ecosistemi marini mondiali, soprattutto per quei paesi la cui economia si basa sullo sfruttamento di questo tipo di risorse, come il caso dell'Ecuador. Anno dopo anno, a causa della ricchezza di nutrienti e risorse marine intorno alle Isole Galapagos, il paese è stato colpito dalla pesca eccessiva sia all'interno che all'esterno della zona economica esclusiva, infatti la particolare morfologia e geodinamica dei fondali, i processi di interfaccia e le masse d'acqua la circolazione sono i principali fattori e vincoli alle risorse marine e alla presenza di un'area marina protetta nell'isola delle Galapagos. In questo contesto, l'obiettivo principale di questo studio è fornire alla Marina ecuadoriana uno strumento in grado di fornire informazioni tempestive per determinare le potenziali aree di pesca intorno alle Isole Galapagos al fine di pianificare adeguatamente le operazioni navali che combattono la pesca illegale, non dichiarata e non regolamentata. A tal fine sono stati condotti due tipi di esperimenti: per il primo esperimento, la clorofilla è stata assunta come variabile oggettiva al fine di identificare i luoghi in cui è presente un livello più elevato di biomassa, precedentemente previsto utilizzando tecniche di apprendimento automatico, come reti neurali artificiali e regressione multivariata ottimizzata utilizzando le serie di Fourier; per il secondo esperimento, l'analisi statistica dello sforzo di pesca è stata utilizzata per determinare gli intervalli in cui devono trovarsi le variabili oceaniche affinché vi sia una maggiore probabilità di pesca. Con queste informazioni è stato creato uno strumento informatico che propone giornalmente potenziali zone di pesca, secondo le suddette prospettive, oltre a presentare informazioni sul comportamento storico della flotta peschereccia nazionale e internazionale intorno alle Isole

Galapagos e le condizioni quotidiane del mare oceanografico variabili considerate in questo studio.

PAROLE CHIAVE: Pesca illegale, clorofilla, potenziali zone di pesca, Galápagos.

RESUMEN

El incremento de la población mundial y la falta de proteína en el mundo ha provocado que en los últimos años, exista una sobreexplotación de los recursos pesquero, afectando gravemente a los ecosistemas marinos del mundo, especialmente para aquellos países que su economía se centra en la explotación de este tipo de recursos, como es el caso del Ecuador, ya que alrededor de las Islas Galápagos por la riqueza de nutrientes y recursos marinos, año tras año ha sido afectado por la sobre explotación pesquera tanto dentro, como fuera de la zona económica exclusiva, de hecho, la morfología y geodinámica particulares del lecho marino, los procesos de interfaz y la circulación de las masas de agua son los principales factores y limitantes de los recursos marinos y la presencia de áreas marinas protegidas en las Islas Galápagos. En este contexto este estudio, tiene como objetivo fundamental, proporcionar a la Armada del Ecuador, una herramienta que le proporcione información oportuna para determinar las zonas potenciales de pesca alrededor de las Islas Galápagos, a fin de planificar adecuadamente las operaciones navales que combaten la pesca ilegal, no declarada y no reglamentada. Para el efecto, se ha realizado dos tipos de experimento, el primero que tomado como variable objetivo a la clorofila para identificar los lugares donde existen mayor nivel de biomasa, la misma que ha sido predecida empleando en primer lugar técnicas de machine learning como redes neuronales artificiales y la regresión multivariable optimizada mediante series de Fourier; y la segunda a través del análisis estadístico del esfuerzo pesquero para determinar los rangos en los que deben estar las variables del océano para que exista mayor probabilidad de pesca. Con esta información, se ha preparado una herramienta informática que propone diariamente las zonas potenciales de pesca, bajo las perspectivas antes indicadas, así como también presenta información del comportamiento histórica de la flota pesquera nacional e internacional alrededor de las islas Galápagos y las condiciones diarias de las variables oceanográficas consideradas dentro de este este estudio.

Palabras Claves: Pesca Ilegal, clorofila, áreas potenciales de pesca, Galápagos.

INTRODUCTION

The substantial increase in the world population and the demand for fish protein is causing pressure on the oceans that is depleting their resources (United Nations Educational, 2021), so that the countries with the largest populations and some developed countries are unable to meet their fishing quotas with catches in their own seas (Wilmoth, 2019), being forced to cross borders and search for fish on the high seas or in jurisdictional waters of other countries, which is causing environmental damage, loss of sustainability, and economic losses to the affected countries (Agnew et al., 2013).

The commercial fishing exploitation in the Galapagos Islands began in the mid-1900s, estimating a catch from 1950 to 2010 of 797,000 tons, of which 80% represented by the tuna industrial catch. It is also alarming that shark finning practices have increased since the 1980s (Schiller et al, 2015), which are also carried out within the Galapagos Marine Reserve, as like the capture of the Chinese fishing vessel Fu Yuan Yu Leng 999 with nearly 300 tons (Guardian, 2020). This event highlights the serious problem of illegal, unreported and unregulated fishing around the Galapagos marine protected areas (Valery, 2017). To further emphasize the need to contribute to the protection of the marine species of the islands, the international community is on alert for news presented during July 2020, which reported that more than 260 ships are currently under international waters on the outskirts of the Exclusive Economic Zone (EEZ) and warned that the aggressiveness of this vast fleet is putting at risk the delicate marine ecosystem and the natural balance of species in Galapagos (Guardian, 2020).

Additionally, this kind of illegal, unreported and unregulated fishing is considered a stress factor for the ocean, due to the over-fishing that occurs, which is influencing in the ecosystem, by changing the abundance of fish and the performance of other organisms. This leads to changes in predator-prey dynamics and competition between species and intra-species. This factor has a negative impact on marine life, which directly affects sustainable development, being necessary to formulate policies that allow managing the multiple stress factors that the ocean has nowadays (IOC-UNESCO, 2022). Even the "Artisanal Fishing Community of Galapagos" is concerned about the protection of fishery resources, as many of them agree that the country should take action to try to curb illegal fishing; however, despite being aware of the impact that large-scale fishing has on the marine ecosystem, they do not agree with the imposition of fishing quotas or the implementation of strict regulatory rules to control fishing (Castro, 2005). For this purpose, Art. 61 of the United Nations Convention on the Law of the Sea (UNCLOS) establishes that the coastal State has the right to set fish catch quotas in its exclusive economic zone, as it is its duty, together with the regional fisheries management organizations, to take measures that seek to maintain or restore fish stocks for the conservation of the marine ecosystem (Harrison & Morgera, 2012).

Despite the vastness of the ocean, fish stocks are declining. This resource is being overexploited and Ecuador is no exception, so it is necessary not only to monitor and control marine ecosystems, but also to manage them in a comprehensive manner, giving them urgent attention (Detsis et al., 2012). For this reason, Ecuador has implemented a series of fishing management practices. One of the most important is the participation in several regional fishing organizations, such as the Commission on Inland Fisheries and Aquaculture for Latin America and the Caribbean (COPESCAALC), the Inter-American Tropical Tuna Commission (IATTC), the International Whaling Commission (IWC), the Latin American Organization for Fisheries Development (OLDEPESCA), the Permanent Commission for the South Pacific (CPPS) and the South Pacific Regional Fisheries Management Organization (SPRFMO) (FAO, 2011).

However, Ecuador has still relevant problems to solve, such as strengthening its legal framework, improving monitoring, control and surveillance, increasing research and providing technological training to artisanal fishermen, among others. This study aims to contribute to the control and monitoring of fishery resources and suggest actions to the Ecuadorian Navy for surveillance planning.

This thesis consists of six chapters: the first one explains the identified problem and the importance of its research, as well as the general and specific objectives that are sought to be achieved with this study.

In chapter two, the frame of reference is defined, the background of the problem is analyzed, a review of the state of the art is made, the theoretical and conceptual foundations are established and finally the hypothesis to be demonstrated is put forward.

In chapter three, a model for the prediction of chlorophyll around the Galapagos Islands is presented, with which it is intended to identify the areas that would have the highest concentration of biomass so that the Ecuadorian Navy can concentrate its efforts to monitor and control fishing activity in the identified areas. To make the prediction, machine learning tools such as the multivariate linear regression and artificial neural networks have been used. The models are trained with data recorded by the Copernicus space program. The model is configured using a dataset of 1 month of the year 2018 and the test is carried out with a dataset of 10 days.

In chapter four, an optimization of the multivariate linear regression model that was developed in the previous chapter is proposed, using the discrete Fourier transform theory. The lag that exists between the physical and biogeochemical variables with respect to chlorophyll is identified to restructure the 6-month dataset of the year 2021 used to configure the machine learning model, in order to predict chlorophyll up to 15 days later. Subsequently, its accuracy is evaluated with a dataset of 6 months of the same year. To this end, the data recorded by the Copernicus program (https://www.copernicus.eu) is also used.

In chapter five, another alternative is presented in order to determine the potential daily fishing zones, for which the fishing effort that has existed around the Galapagos Islands from 2017 to 2020 is analyzed, through the identification of the range in which the oceanographic conditions must be for the presence of fishing vessels. Statistical tools such as kernel density estimation have been used in order to estimate the heat maps that allow to identify the gravity centers of the different potential fishing areas. The datasets used for this experiment come from the Copernicus program and the Global Fishing Watch organization (https://www.globalfishingwatch.org).

In chapter 6, the webpage that has been developed to show the results of the experiments is presented, that will be used by the Ecuadorian Navy. The scripts developed in Python throughout the development of the research have been stored in a virtual server so that they are executed automatically, allowing to provide updated information on the prediction of possible fishing areas, as well as the behaviour of oceanographic variables in the study area. Finally, the conclusions of this research are presented.

CHAPTER I

THE PROBLEM

1.1. Problem Statement

The world population is growing rapidly, from 2.6 billion to 7.7 billion people since the Second World War, with China (1,449 million) and India (1,390 million) being the countries with the greatest demographic expansion, representing 19% and 18% of the world population, respectively (United Nations, 2019). Therefore, the sea resources are fundamental for the provision of the protein necessary for the nutrition of human beings, being the exclusive right of the States to exploit for their benefit the resources available in their exclusive economic zones (EEZ). Outside these zones, the existing resources are freely available to the countries, here the over exploitation of fishery resources begins, especially to satisfy the request from the Asian countries.

Ecuador has a jurisdictional maritime space of 1092140.25 km^2, where it exercises sovereignty, control and exploitation of marine resources in a sustainable and rational manner (SENPLADES, 2017). This maritime territory is strongly influenced by a series of marine currents, which positively affect its biodiversity as they distribute nutrients throughout the geographical space. The main currents the northward nutrient-rich and cold Peruvian Coastal Current (PCC, known as Humboldt Current), the southward warm Ecuador-Peru Coastal Current (EPCC, known as El Niño Current) and the Cronwell countercurrent in the Galapagos Archipelago, which distributes nutrients around the islands.

Then also these currents (Figure 1) impact on the distribution and abundance of fishing resources, being the Humboldt Current the most important since it carries a large amount of nutrients, plankton and krill.

The abundance of nutrients attracts a lot of marine resources that are born and grown in equatorial waters, so Ecuador has the privilege of taking advantage of a large amount of resources, being fishing one of its main commercial activities. A large number of inhabitants of the Ecuadorian coast depend on this activity, which generates great economic benefits for the country and represents the third most important item of the Gross Domestic Product (GDP), which contributes USD $4,813 million dollars, equivalent to 22% of GDP. Table 1 shows in detail the economic impact of the fishing industry. It is important to mention that this activity includes the commercialization of both primary and industrialized products.

The Ecuadorian fishing industry is the second largest and most modern in the whole American continent. The main exported products are: tuna, dorado, billfish, swordfish, wahoo, horse mackerel, sardine, mackerel and corvina, among others (SENPLADES, 2017).

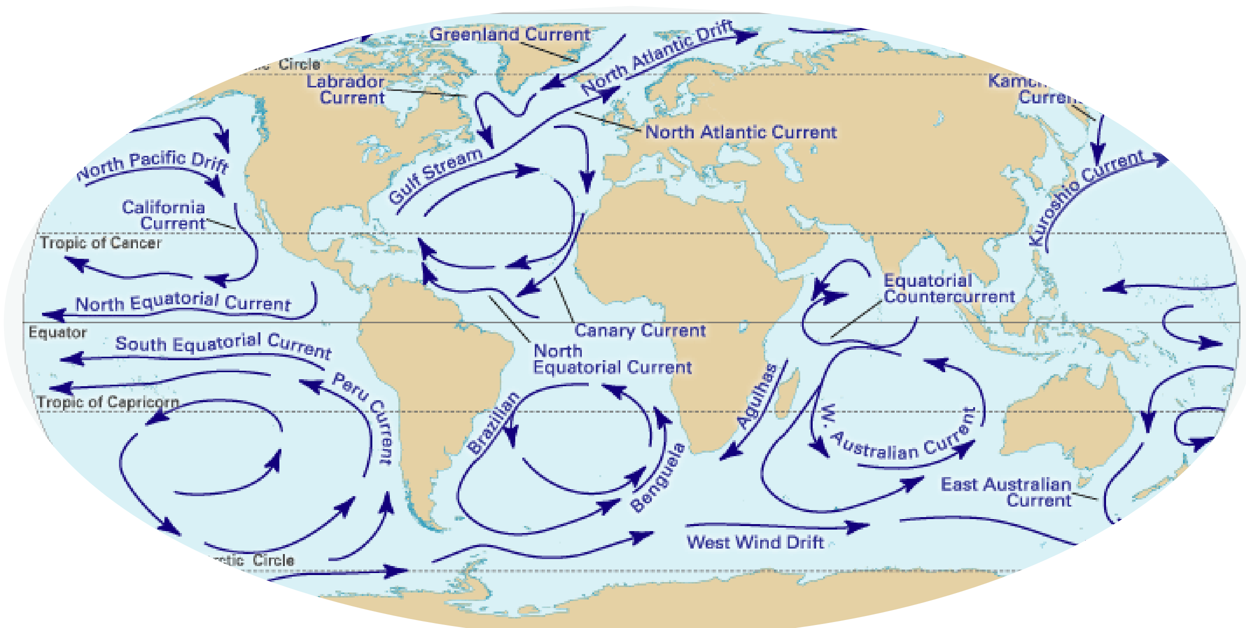


Figure 1. Global marine currents (Encyclopædia, 2007)

Table 1: National exports *(2019) - Marine products. Source: Central Bank of Ecuador*

(1) Thousands of USD FOB (2) Includes whole tuna, tuna steaks and other steaks.

TOTAL PRIMARY MARINE PRODUCTS (2)	TOTAL INDUSTRIALIZED MARINE PRODUCTS	TOTAL MARINE PRODUCTS EXPORTS	TOTAL NATIONAL EXPORTS	% PARTICIPATION MARINE PRODUCTS EXPORTS
307,960	4,505,909	4,813,869	22,329,379	22%

The sea richness is however being threatened by illegal activities in and outside of jurisdictional maritime areas. Illegal fishing is mainly related to the capture of fish or shellfish that contravenes national and international laws and regulations on fishing in jurisdictional and non-jurisdictional waters (COGMAR, 2013).

The different activities related to illegal fishing can be grouped into the following identified categories, according to Sumaila et al. (2006):

▪ Illegal fishing, related to vessels of countries that belong to a fishing organization, but that operate in violation of its rules, in jurisdictional waters of a country without permission, or on the high seas without displaying a flag or other distinctions.

▪ Unreported catches, which corresponds to fishing that is not reported to the authorities of the State to which this vessel belongs. Misreporting and unreported catches are included in this category.

▪ Unregulated fishing, commonly carried out by vessels flying the flag of countries that are not parties or participants in a fishing organization, and therefore are not required to comply with its rules.

Illegal, unregulated and unreported (IUU) fishing occurs not only on the high seas, but also within EEZs.

1.2. Identification of the root cause of the problem

Root causes of illegal fishing include moral, social and economic factors (Sumaila et al, 2006), since the fishermen who decide to participate in IUU acts make a decision taking into account the following:

a. Economic benefits, or other benefits that could be obtained by participating in illegal IUU activities.
b. Probability of detection of the illegal activities by the authorities.
c. Penalties that the fisherman could face if he is detected while committing acts of IUU fishing.
d. Social and moral values of the fishermen who participate in IUU fishing.

Analyzing the drivers of IUU fishing, it can be stated that the greater the economic return from legal fishing, the lower the tendency to participate in IUU fishing; however, when the fisherman takes a long time to find fish, he needs to spend more time on the high seas, involving higher operating costs (fuel, labor, licenses, etc.), which added to the low probability of detection of schools, converts acts related to IUU fishing into a profitable option (Castro, 2005).

Likewise, when the States do not have the technology and the appropriate means to carry out the control of the jurisdictional maritime zones, there is a greater probability that the fisherman will turn to illegal fishing. It is worth mentioning that the efficiency of the detection systems encourages fishermen to commit IUU infractions or not, therefore it is beneficial for the authorities to have information that allows them to predict the areas where fishing activity could occur, in order to identify possible cases of IUU fishing.

The severity of the laws applied when a fisherman is caught committing infractions related to IUU fishing is also an important driver that encourages this type of crime. When the sanction is very high, economically or legally, there will be less incentive to commit these crimes. In this sense, it is important that there are agreements between countries so that crimes are punished with the same severity, aiming to prevent them worldwide.

Moral and social conditions of the fishermen can also play an important role in the decision to participate or not in IUU fishing, since the fisherman who has no scruples can engage in IUU fishing activities regardless of the personal, family or social effects that this activity may cause. Some fishermen only want to profit economically without caring about the consequences. Also, the operation of criminal organizations is very common, which can force fishermen to commit this type of crime. In figure 2, a cause-effect diagram can be observed for better understanding.

However, it is not only domestic fishermen who are involved in illegal fishing, but also other countries that have depleted their fishery resources and are looking to other horizons to meet their food needs. China currently has the largest fishing fleet in the world and some of its fishing vessels and thousands of people working on them and in related industries are registered in China's Third China Sea Force, called China's People's Armed Forces Maritime Militia (PAFMM) (Kennedy & Erickson, 2017). The PAFMM is an armed organization whose members are seafarers, working in the civilian economy and trained to defend and promote China's maritime territorial claims and support the Chinese Navy in times of war.

The Regional Fisheries Management Organizations (RFMOs) are responsible for managing high seas fisheries, especially migratory and straddling fish stocks, which seek to control high seas fisheries. States that share common interests often organize themselves to act jointly in actions related to stock control, catches, assessments, decisions, and monitoring; however, the implementation of corrective actions in exclusive economic zones is the responsibility of the flag State (Metuzals et al., 2010).

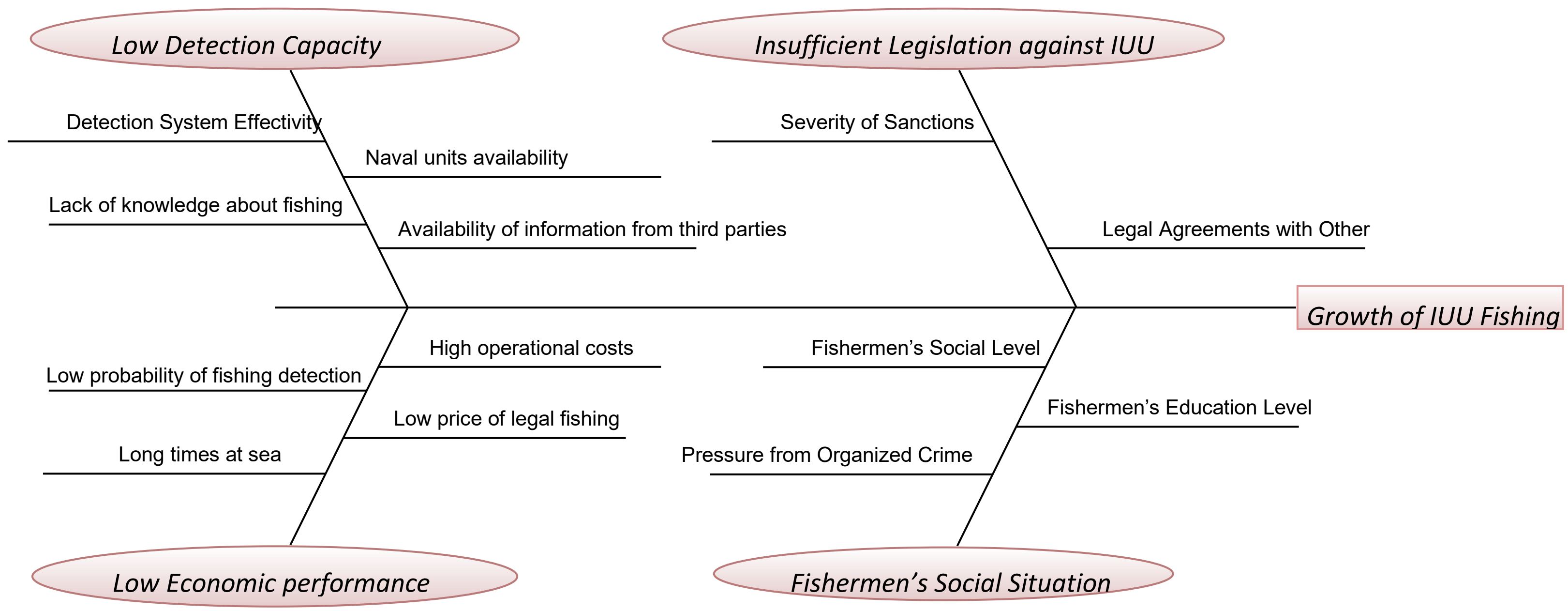


Figure 2. Cause-effect diagram of the problematic situation.

1.3. Formulation of the problem to be solved

Early detection of fishing vessels crossing the ocean is very difficult if these vessels disconnect their VMS system, or if the country concerned does not have the support of satellite surveillance or the permanent support of aeromaritime scanning, so that vessels are detected late, when they have reached the fishing point. The Galapagos Islands, due to their geographic location and the influence of currents and countercurrents, provide the appropriate conditions for a high presence of fish, which attracts both national and foreign fishermen, involving risk of illegal, unreported and unregulated fishing.

For this reason, it is necessary to determine if it is feasible to predict the potential fishing zones in the Galapagos Islands, in order to provide information to the Ecuadorian Navy to contribute to the surveillance, protection and sustainable exploitation of fishery resources.

1.4. Justification and significance

1. IUU fishing has a serious impact on the sustainability of fisheries (Sumaila et al., 2006), as inaccurate catch data do not allow for accurate assessments that contribute to sustainable catch estimates; however, according to Detsis et al. (2012), IUU catches are on the order of 10 - 20% of commercial catches, depending on the species and location.
2. Illegal fishing is a serious global problem that not only destroys marine ecosystems but will also bring about the future collapse of the fishing sector, threaten the population's food supply and cause poverty in the society that lives of seafood products and carries out legal fishing. Unfortunately, illegal fishing involves from artisanal fishermen to large fleets with connections to criminal organizations, since IUU fishing is an attractive option to obtain more profits with low operating costs (Metuzals et al., 2010).
3. Figure 3, developed by (Detsis et al., 2012) shows how illegal fishing operates: a real owner of the vessel, a declared owner and a State in which the vessel is formally registered. Subsequently, all activities related to the logistical support of illegal fishing vessels, transfer of fishing and finally marketing of the illegal product are shown.

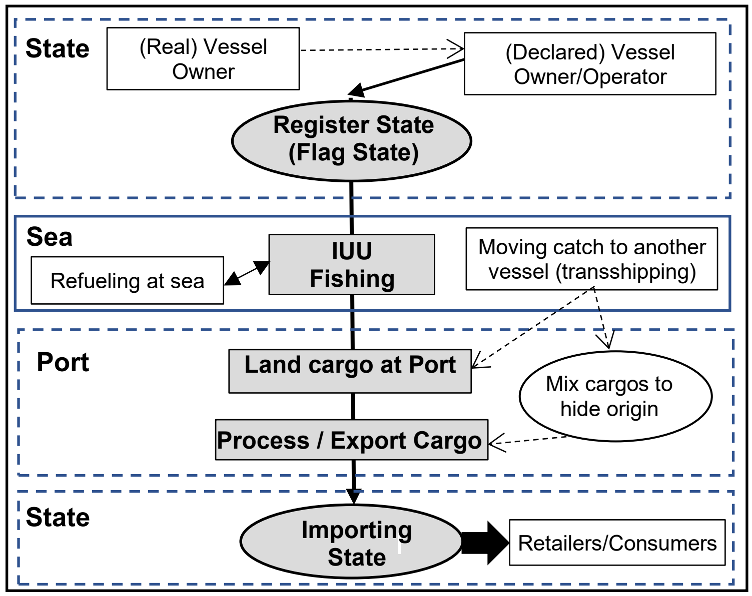


Figure 3. IUU Fishing Supply Chain (Detsis et al., 2012)

1.5. Objectives of the Research

1.5.1. General Objective

Determine potential fishing zones around the Galapagos Islands to provide information that will contribute to the planning of naval operations to combat illegal, unreported and unregulated fishing.

1.5.2. Specific Objectives

a. Identify potential fishing zones by predicting chlorophyll level in the sea and using machine learning tools to relate the physical and biogeochemical variables, considering the particular bottom physiography and geological setting of the ocean currents.

b. Analyze the movement of national and international fishing vessels in order to determine the ranges of ocean physical and biogeochemical variables that favor the presence of fishing vessels to predict potential fishing zones by using statistical tools.

c. Develop a computer tool that allows to automatically predict potential fishing areas to contribute with the planning of naval operations that combat illegal, unreported and unregulated fishing.

1.6. Contribution of the degree project

The main aim of this degree project is to provide the Ecuadorian Navy with a tool to contribute to the planning of naval operations to combat illegal, unreported and unregulated fishing. For this purpose, artificial intelligence and statistical tools are used to identify the areas where the most favorable conditions for the existence of fish will exist, so that control actions can be carried out, and to avoid the depredation of the seas in the Insular Exclusive Economic Zone.

CHAPTER II

REFERENCE FRAMEWORK

2.1. State of the art

2.1.1. Fishing Monitoring and Control technologies

Currently, there are several systems that allow States to monitor and control the areas under their jurisdiction. These systems are known for monitoring, control and surveillance systems, which mainly consider four types according Detsis et al. (2012):

a. Vessel Monitoring Systems (VMS)

b. Automatic Identification Systems (AIS)

c. Long Range Identification and Tracking System (LRIT).

d. Synthetic Aperture Radar (SAR) (Satellite system).

These systems have independent communications protocols but are complementary to each other to achieve optimal maritime management within a country's Exclusive Economic Zone.

Briefly, the VMS system securely transmits the location of the vessel and other sensitive information such as catch weight and species to the organization supervising the fishing operation. The AIS system, on the other hand, assists in the management of maritime traffic; the information is normally transmitted to all vessels and stations that can receive the messages. The LRIT system provides long-range monitoring of vessels around the world. And the synthetic aperture radar installed on the satellites allows the position of ships at sea to be identified at all times, even when other systems are out of service. This radar can also be installed on an air-sea reconnaissance aircraft, but due to its autonomy and the altitude at which it operates, it does not allow permanent control of the area of responsibility.

The use of current navigation and fish detection equipment, such as GPS, echo sounders, multi-beam sonar, among others, have increased the capacity to catch fish, allow fishing vessels to safely go farther from the ports of origin and have on board the capacity to process and preserve on board the fish caught for long times. Similarly, satellite communications and other control systems such as VMS, AIS and Global Maritime Distress Safety System (GMDSS) have also evolved and improved safety on board, reducing the probability of casualties at sea and improving the management of fishery resources; however, there has also been an increase in fish catches and the use of destructive fishing techniques, which has also increased damage to marine ecosystems (Garcia & De Leiva Moreno, 2003)

2.1.2. Copernicus Marine Environment Monitoring Service – CMEMS

According to its High-Level Service Evolution Strategy (Von Schuckmann et al., 2016), CMEMS is one of the 6 components of the European Union's Earth Observation Program, called Copernicus. This program, initiated in 2014, has planned to put into orbit until 2030 approximately 20 satellites, but it also includes data from a lot of ground, maritime and airborne sensors. CMEMS's main objective is to provide free of charge updated data of the physical, biogeochemical, and dynamic state of the world's oceans and European regional marine areas. This information is used in various applications like maritime safety, coastal marine environmental monitoring, marine resource monitoring and climate forecasting and control, thus providing the European Union with a powerful tool to monitor and assess ocean variability and change in near real time. CMEMS also provides short-term forecasts of physical and biogeochemical parameters. Some of the applications of the information provided by CMEMS include climate monitoring.

In order to identify the variables to be studied and considered within the Copernicus database, Global Climate Observing System (GCOS) and Global Ocean Observing System (GOOS) programs developed the concepts of essential climate variables (ECVs) and essential ocean variables (EOVs), these physical, chemical, and biological variables are necessary to characterize the land, oceans and climate (Lindstrom et al., 2012).

The use of ocean-color remote sensing provides the ability to have temporal and spatial information on important physical, geological, chemical and biological parameters of the ocean (Chang, 2015). Watercolor and transparency are oceanographic parameters that are closely related to physical and biogeochemical processes in the ocean, and whose variations are indicators of changes in the optically active components of the water (Wang et al., 2019).

Oceanographic data are essential for the monitoring and management of marine living resources because of the stric relationship between the physiology of marine organisms and their environment. However, although oceanographic data are invaluable for understanding the processes that govern the dynamics and behaviour of marine resources, they are not being used for fishery resource management (Payne & Lehodey, 2019).

According to Hollandet al. (2019), the satellite data made available by Copernicus Marine Service can be used by decisions makers in those regions where there is a lack of in-situ data, especially for permanent monitoring of physical and biological variables. Lack of data leads to limited capacity in the decision-making process, especially for marine ecosystem protection.

Similarly, climate change over the last decade is affecting the distribution and abundance of small fish stocks, influencing the distribution of commercial fish and altering fisheries management policies for fisheries sustainability (Brunel et al., 2018).

2.1.3. Machine learning in oceanography

Machine learning (ML) is a subfield of artificial intelligence (AI) that enables us to take decisions based on analysis of data. In this way, AI allows us to reduce the demands on human specialists in oceanographic research. The principal uses of machine learning in oceanography include wave modeling, species identification, habitat modeling and distribution, management of marine resources, coastal water monitoring, and prediction of ocean weather and climate (Ahmad, 2019).

The ocean is large, dynamic, and complex. The data structure of the ocean is becoming more complicated and vaster; typical data analysis techniques are time-consuming and expensive, and in certain circumstances, conventional analysis is impracticable. ML approaches are powerful, efficient, and quite precise, depending by the quality of input data.

The two fundamental subfields of machine learning are supervised learning (SL) and unsupervised learning (UL). SL primarily attempts to model the connection between the inputs and their following outputs using the training data so that we may anticipate the output based on its prior understanding of the input-output relationship. One of the most important applications of supervised learning is generating forecasting based on observed behaviours or features from previously recorded data (the data history). Supervised learning allows you the discovery of patterns in historical data. There are two principal types of SL: Classification and regression. A classification system predicts a category, while a regression predicts a number (Learned-Miller, 2014; Nasteski, 2017).

Finding correlations between variables is the aim of the linear regression algorithm. It shows the modeling connection between a continuous scalar dependent variable Y, and one or more independent variables, often known as explanatory variables, marked by the letter X. The objective of regression analysis is to forecast a continuous target variable.

The artificial network-neural network (ANN) or neural network is a machine learning technique based on the concept of imitating the human brain. Each neuron is a cell that performs a basic job, such as responding to an incoming signal, and contains three essential components: node characteristics, network topology, and learning rules. Character node determines the number of inputs and outputs, the weights associated with each input and output, and the activation function. The network topology defines how nodes are structured and interconnected. The initialization and adjustment of weights are determined by the learning rules. Node characteristics, such as the number of inputs and outputs, the weights associated with each input and output, and the activation function, impact how signals are processed by the node. The network topology defines how nodes are structured and interconnected. The initialization and adjustment of weights are determined by the learning rules. The artificial neural network can do complicated tasks, as voice and picture recognition, with speed and precision. The ANN is capable of modeling complicated nonlinear interactions in contrast to typical regression techniques. In addition, it has fault-tolerance and is scalable with parallel processing (Zou, 2008).

2.2. Theoretical foundations

2.2.1. Concept of Marine Ecosystems

In 1935, the botanist Arthur Transley first used the term ecosystem to identify the integration of different living beings (biotic factors) that share a physical environment. Years later, in 1942, Raymond Lindeman carried out a study on the flow of energy in an aquatic ecosystem, both of its biotic and abiotic components (Reyes, 2007). This concept has evolved as the one developed by Schowalter (2016) who indicates that ecosystems are complex systems, represented by a set of species and abiotic resources, which do not have easily identified boundaries and whose vital processes are interrelated. The ecosystem is therefore considered the fundamental unit of an ecological organization.

Within an ecosystem, a variety of biotic factors are related, such as microorganisms, plants, animals, as well as a series of abiotic resources (light, water, solar radiation, atmosphere, temperature, humidity, minerals, etc.) found in the environment. This interrelation between biotic and abiotic factors forms trophic or food chains. It is important to mention that the development of microorganisms and other living beings depends on the environment in which the ecosystem develops.

Making an analogy of the general concept of ecosystem, we can define that the marine ecosystem is the existing interrelationship between marine animals and plants with the marine environment (sea or ocean and bottom morphology). Then because the oceans correspond to approximately 70% of the earth's surface; the marine ecosystems are one of the largest and most complex ecosystems on the planet, whose balance is fundamental for human existence.

2.2.2. Marine ecosystem functioning

Ecosystem models are being used more frequently to create global-scale methods for estimating the distribution of marine biomass and the potential catch of fisheries. This helps us in understanding the effects of fishing and the potential effects of climate change under various hypothetical scenarios. Ecosystem models offer a tool to disentangle the trophic dynamics of marine ecosystems and to assess the impact of anthropogenic activities, facilitating analyses of food webs. An ecosystem's food web, which is made up of species at various trophic levels, plays a significant role in how well it works. Using food webs, researchers can calculate the biomass at each trophic level in a specific ecosystem and infer how well that ecosystem is functioning overall while taking into account trophic flows (Pontavice, 2019).

In figure 4, there is an example of a model of marine ecosystem functioning, a PISCES-v2 (Pelagic Interactions Scheme for Carbon and Ecosystem Studies volume 2) model, which simulates marine primary production with explicit phytoplankton and zooplankton components. This model describes the biogeochemical cycle of carbon and the main nutrients (nitrate, ammonium, phosphate, silicic acid and iron). PISCES is a biogeochemical model with five nutrients (NO_3^-, NH_4^+, PO_4, Si and Fe), two groups of phytoplankton (diatoms and nanophytoplankton), two groups of zooplankton (micro and mesozooplankton) and two non-living compartments (particulate and organic dissolved) (Aumont et al., 2015)

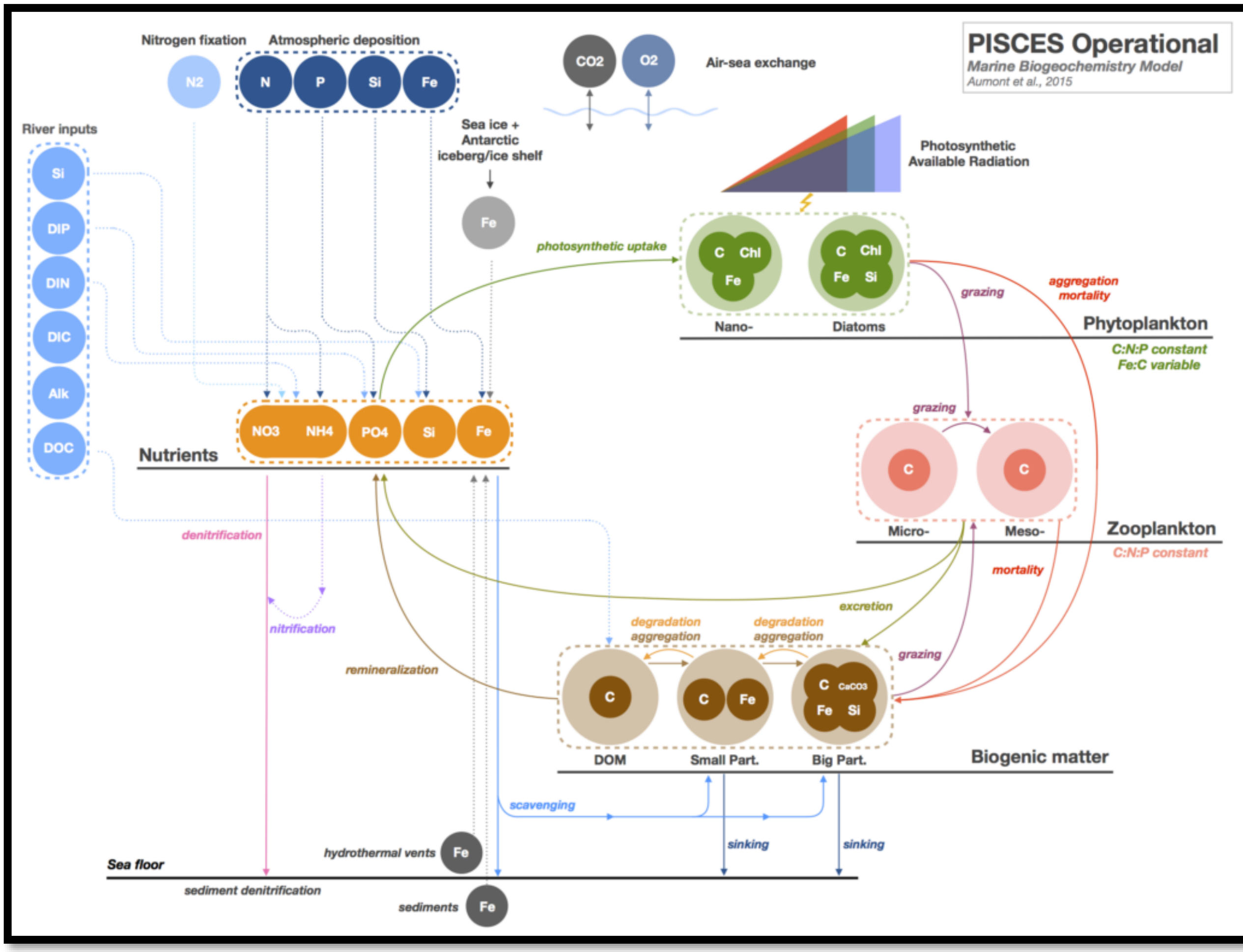


Figure 4. Pisces Model (Aumont et al., 2015)

2.2.3. Marine food webs

The food webs are part of biotic factors building a chain though which energy is transferred from the small organism to the final predator. There are many trophic chains within an ecosystem, so the trophic chains between species are interconnected. The flow of energy in food webs within an ecosystem provides the necessary interactions to maintain the trophic levels of the food chain. In a more detailed way, we can say that food webs are a series of biotic factors that are interrelated by their feeding cycles, so these webs are the way by which energy is transferred from smaller organisms to larger organisms, i.e. from plants to herbivores, then to carnivores and finally to top predators, highlighting the trophic level of each species and the interactions between producers, consumers and decomposers (Trites, 2003), being the energy transfer between each trophic level, about 10% (Schulz et al., 2004).

This same concept is applied to marine ecosystems as seen in figure 5, where a basic marine feeding network is shown. As it can be seen, a lot of organisms are grouped together in the food web and play an important part to make a pyramid of different trophic levels. Pyramids have different levels: 1st level is made up of producers, 2nd level is made up of herbivores and so on. Trophic levels have different numbers of organisms but less biomass overall at trophic level. Energy level is transferred from an organism to a next level organism as from producer to consumer at trophic level and it is supported by millions of primary producers. Consumers can be herbivores or carnivores. Moreover, herbivores consumers in the marine oceanic ecosystem are zooplankton. The Phytoplankton is consumed by organisms at the 2nd level of the pyramid. Therefore, most of the top marine predators live in the sea; humans are another top-level consumer in the marine food web.

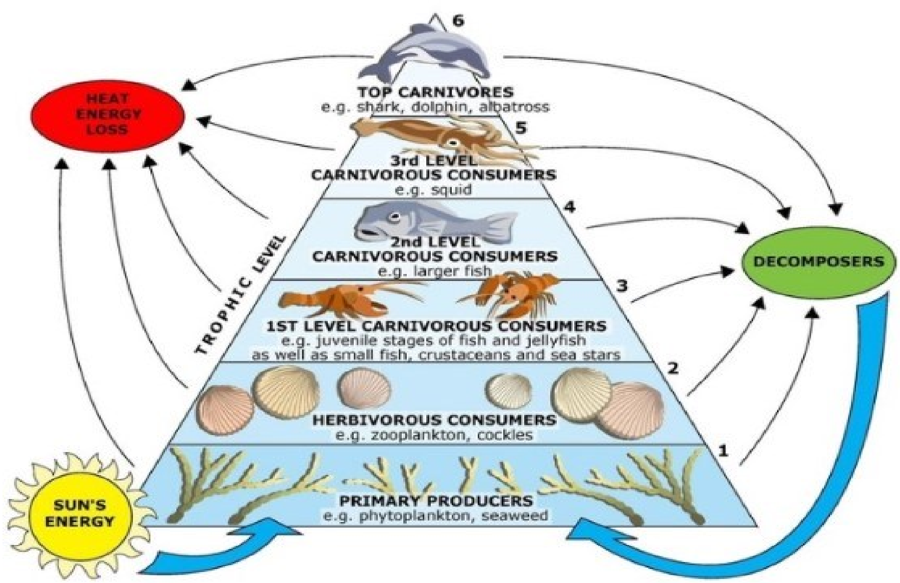


Figure 5. Marine Food Pyramid (Waikato, 2009)

Normally, marine food chains have phytoplankton as their main producer, which is composed of microscopic algae. Despite their small size, they are the base of the food chain and are simply known as producers. Primary consumers feed on it, and indirectly consumers at higher levels also feed on phytoplankton. Food web removes or reduces the effect of species or organisms found in marine food chains. The level of biodiversity varies from high to low in marine food webs. The removal of plant species will negatively target organisms and species on the complete food web. The loss of one plant may have little or no effect on herbivores or carnivores.

Some species in a food web of ocean are called as keystone species. This keystone species is very important and root for a major change in food web ecosystem. If one keystone species is removed from the food web ecosystem, the whole equilibrium become disturbed. Hence, the bridge between producers, consumers and decomposers causes the structure of food web to collapse. For example, sea star is a keystone species in the New Zealand that controls the number of the species it feeds on like mussels. If sea star would vanish from the system, the number of mussels would enhance causing a larger number of mussels throughout the food web (Alsterberg et al., 2013)

Food webs are often useful tools for a conceptual understanding of ecosystems, identifying their deficiencies and understanding the importance of each species, identifying species that are at risk and which processes are critical for the conservation of the ecosystem (Trites, 2003), and sometimes the interconnection between species may not occur simultaneously or may change according to the seasons or years.

According to Christensen (1996), in unexploited areas, marine ecosystems can be expected to be in equilibrium, often with relative abundances of predatory fishes. Initially, fisheries will catch for the larger and more expensive species and later, when the larger species become scarce, fishing will move towards the smaller species and, consequently the ecosystem will be damaged, and a fishing crisis will occur.

Importantly, phytoplankton is generally capable of producing its own energy through the process of photosynthesis. This energy is transmitted to the next level, composed of zooplankton and other small fish and finally, transferred to a further level when these small fish are ingested by larger fish. Once the predators die, the energy returns to the ecosystem when their body is a food source for detritivores, which are organisms that feed on decaying organic matter. Depending on the amount of energy that a lower level can transmit to a higher level, organisms of these levels must consume a greater number of trophic organisms of a lower level. Because of this, it will always be necessary to have a greater number of lower level organisms than higher ones to maintain a balanced ecosystem.

It should also be noted that larger fish (predators) prefer smaller prey, approximately a quarter to a third of their own size, as they are limited by the size of their jaws. However, trophic webs are generated by the feeding behaviour of fish, since they have multiple predators, multiple prey and multiple competitors (Curyet al., 2003), since in the early stages of their life they can feed on the larvae of their own predators or even of their own species.

2.2.4. Essential oceanic variables - EOVs

EOVs are physical, geological, chemical or biological variables that contribute directly to the characterization of the oceans. The ocean plays a fundamental role in all Earth systems, such as its capacity to absorb and emit heat; it influences weather and climate regulation, its capacity to recycle nutrients and produce biomass, it is a source of biodiversity for the planet, as well as a generator of economic wealth for nations that depend on the sea (Lindstrom et al., 2012). Many of the EOC variables are considered within the EOVs. The Global Ocean Observing System (GOOS) is a program implemented by the Intergovernmental Oceanographic Commission (IOC) of UNESCO, which uses the “Framework for Ocean Observing” to monitor the oceans and to implement an integrated and sustained observing system. The variables considered by GOOS, as essential oceanic variables, are those listed in table 3.

2.2.4.1. Surface sea temperature

Sea surface temperature (SST) is a fundamental ocean variable to determine the heat exchange between the ocean and the atmosphere. Temperature has always been considered as one of the basic parameters for research and prediction of climate variability and change, since SST can provide information about the heat balance on Earth. The information provided by satellites have a typical accuracy of 0.5° when compared to other instruments installed in situ (Von Schuckmann et al., 2016). CMEMS also provides data on the temperature below the water surface.

Table 2: Essential Oceanic Variables (GOOS, 2012)

Variables		
Physics	**Biogeochemistry**	**Biology and Ecosystems**
1. Sea state 2. Ocean surface stress 3. Sea ice 4. Sea surface height 5. Sea surface temperature 6. Subsurface temperature 7. Surface currents 8. Subsurface currents 9. Sea surface salinity 10. Subsurface salinity 11. Ocean surface heat flux 12. Ocean colour 13. Ocean Sound	1. Oxygen 2. Nutrients 3. Inorganic carbon 4. Transient tracers 5. Particulate matter 6. Nitrous oxide 7. Stable carbon isotopes 8. Dissolved organic carbon	1. Phytoplankton biomass and diversity 2. Zooplankton biomass and diversity 3. Fish abundance and distribution 4. Marine turtles, birds, mammals abundance and distribution 5. Hard coral cover and composition 6. Seagrass cover and composition 7. Macroalgal canopy cover and composition 8. Mangrove cover and composition 9. Microbe biomass and diversity 10. Invertebrate abundance.

It is worth mentioning that the spatial distribution of fish is being affected by the increase in ocean temperature, which is also altering the dynamics of food webs in terms of the predator-prey relationship, whereby temperature conditions are factors that control the distribution of a species and the abundance of prey (MacKenzie et al., 2014).

2.2.4.2. Plankton

The term plankton describes organisms that move near or on the surface of the water and do not have the ability to swim against currents, tides or winds (Santhanam & Begum, 2019). Plankton constitutes the main biomass within the aquatic food web and provides food to different marine animals, from microorganisms to whales. Therefore, the plankton is the most significant part of the oceanic ecosystem because of diverse functions in the oceanic food web. A recent study shows that plankton is more critical to protect marine ecosystems during trophic levels consumption by organisms (Lumini & Nanni, 2019).

Plankton is of paramount ecological importance, forming pelagic food webs and constituting the largest source of protein in the oceans for most fishes. Being a primary producer, within the food chain, their absence or reduction often affects fisheries. In many countries, it has been observed that the decline in fisheries has been attributed to the reduction in plankton population (Santhanam & Begum, 2019). The plankton is made up of phytoplankton (microscopic plants or plant plankton) and zooplankton (animal plankton) that are small animals present near the surface of aquatic environments. These primary producers feed everything, from zooplankton to multi-ton whales, small fish, invertebrates and other animals.

Phytoplankton is composed by unicellular organisms, great majority photosynthetic and move with the waves and perform oxygenic photosynthesis. It is found in the top waters of all aquatic ecosystems. It acquires all the energy from light and resolved inorganic solutes as they are photolithotrophs and its growth rate depends on source of energy supplied in photolithotrophic. Phytoplankton is also mixotrophs and can get their energy and carbon from inorganic compounds, light and inorganic carbon such as sapromixotrophs and phagomixotrophs,

Phytoplankton is usually algae and cyanobacteria that float freely in a body of water and depend on the water column for their resources. When phytoplankton uses energy and carbon from other existing organic or inorganic compounds, they are referred to as mixotrophs. Some mixotrophic algae can grow even in the absence of light (Raven & Maberly, 2009). Phytoplankton photosynthesis also contributes almost 50% of the oxygen in the Earth's atmosphere and consumes a large amount of carbon (Von Schuckmann et al., 2019).

The size of a phytoplankton cell is approximately between 0.1 μm3 and 108 μm3. The size is affected by diverse prospects like physiology, ecology and population, communities. The structure and function of phytoplankton in a water body are most important to explain the marine biogeochemical cycles within an ecosystem (Marañón et al., 2015). Phytoplankton species are found widely in the ocean, and they can also make colonies in both aquatic and freshwater oceans, lakes, lagoons, and marine areas. Biomass changes are based on temporal and spatial scales (Naselli-Flores et al., 2003).

Phytoplankton is classified into five groups according to their cell structure and cell wall arrangement. These groups are diatoms, dinoflagellates, blue-green algae, green algae and coccolithophores. It is worth mentioning that the decrease in phytoplankton abundance has a negative impact on the abundance of zooplankton, which decreases the abundance of pelagic fish and leads to the decrease of top predators, such as large pelagic fish, demersal fish, seabirds, and mammals (Cury et al., 2003). The marine ecosystem also responds drastically to environmental events, such as the phenomena of El Niño, affecting the structure of plankton and the spatial distribution of fish and invertebrates, as well as intensive fishing, which alters the functioning of marine ecosystems.

Zooplankton, another fundamental component of the marine ecosystem, is composed by organisms which float or move with water currents and has a significant biogeochemical property as compared to phytoplankton. Some organisms in zooplankton consume organic particles at night in the surface and mixed it with carbon dioxide during daytime. Some zooplanktons are herbivores, and some are carnivores, and they involve invertebrates, water fleas, and rotifers found in lakes or streams having slow motion. Basically, zooplankton is classified according to habitat, depth distribution, size and life cycle, and the zooplankton imparts an important part in the oceanic nitrogen cycle and consume organic nitrogen to control on primary production and growth of phytoplankton by nitrogen excretion. Dissolved organic nitrogen (DON) is released by preparing food or excretion done by zooplankton.

They also help heterotrophic bacterial growth for the principal pathway from phytoplankton towards bacteria. Zooplankton species transfer particulate as well as dissolved organic nitrogen through production of shrinking fecal pellets. Diet is vertically moved towards migrating zooplanktons that covers the vertical export of nitrogen by during the production process mostly at nighttime (Wang et al., 2020).

2.2.4.3. Chlorophyll and phytoplankton

Phytoplankton is microscopic, single-celled, floating marine organisms capable of photosynthesis, absorbing carbon dioxide dissolved in water in the presence of sunlight to produce organic material. All higher pelagic organisms, including fish, depend on phytoplankton for nutrition as phytoplankton are the main producers in the sea. Phytoplankton is present everywhere in the sunlit layers of the ocean in varying concentrations (Von Schuckmann et al., 2019). The basis of the aquatic food web and the primary producers is phytoplankton. Phytoplankton is recognized as an (ECV) in the Global Climate Observing System implementation plan (Zemp, et al., 2022), with chlorophyll being a measure of its concentration.

Chlorophyll-a (Chl-a) is an indicator of the biological type of marine ecosystems as it is closely related to almost all photosynthetic organisms, and its concentration is commonly used to estimate phytoplankton biomass and water quality (Maslukah, et al., 2022). The concentration of Chl-a is one of the main products of ocean-color remote sensing. To date, several algorithms have been developed to determine the Chl-a concentration based on the ocean color spectrum (González et al., 2011). Marine life is directly dependent on phytoplankton and Chl-a its main photosynthetic pigment, has been used as an indicator to monitor the level of phytoplankton which, as previously explained, is the basis of the marine food web.

According to Holland et al. (2019), ocean chlorophyll can be used to assess its health, productivity and marine life that depend on phytoplankton. Low chlorophyll concentrations and the associated decrease in primary production can negatively affect fish and marine life, which is important for food security and the economy of countries that depend on fisheries. Chlorophyll is an excellent indicator that helps us determine whether the phytoplankton community grew or declined.

To validate the chlorophyll concentration obtained by satellite, two in situ methods are used, the first is by the fluorometric or spectrophotometric and the other by high performance liquid chromatography (Valente et al., 2016).

Satellites provide global coverage of ocean surface color, including the concentration of Chl-a (a proxy for phytoplankton biomass). Ocean color radiance (OCR) is the wavelength-dependent solar energy captured by an optical sensor looking at the sea surface. This glow from the water contains information about the optical components of the seawater, in particular the pigments (mainly Chl-a) contained in the phytoplankton and helps to calculate the different ocean biogeochemical variables and other parameters related to marine ecosystems. Other important OCR products are particulate organic carbon (POC), suspended sediments and others. OCR data products are the only measurements related to biological and biogeochemical processes in the ocean that can be obtained on a permanent basis and are used to assess the health and productivity of ocean ecosystems and the role of the oceans in the global carbon cycle, to manage living marine resources and to quantify the impacts of climate variability and change (Secretariat, 2010).

2.5.1. United Nations Convention on the Law of the Sea (UNCLOS)

The United Nations Convention on the Law of the Sea was adopted in 1982 (United Nations, 2006). It establishes guidelines for all uses of the oceans' resources, thorough framework of law and order throughout the world's oceans and seas. It also regulates the peaceful uses of seas and oceans. It is the first convention accepted worldwide that clearly establishes the scope of national jurisdiction over the maritime areas. It delimits and proposes criteria for the delimitation of the maritime boundary between States.

According to Article 56 of UNCLOS, the passage of foreign vessels through jurisdictional areas can be restricted. The legal tools to do so are similar to those used to restrict freedom of navigation in an area of high ecological sensitivity such as the Galapagos Islands, establishing environmental protection mechanisms. These mechanisms have legal support and international support from the Convention of the Law of the Sea, through the application of Paragraph 1, Article 73, which empowers the coastal State to search and detain fishing vessels that violate these laws. Further, in the case of actions of pollution of the marine environment, the Convention provides for the search of the vessel polluter, according to Paragraphs 3 and 6 of Article 220, Part XII, (United Nations, 2006).

CHAPTER III

TEMPORAL ASPECTS OF CHLOROPHYLL-a PRESENCE PREDICTION AROUND GALAPAGOS ISLANDS

3.1. Abstract

Chl-a is a specific form of chlorophyll used in oxygenic photosynthesis which has been linked to nutrient presence in sea waters (Dutta et al., 2016; Zainuddin, 2011) and being able to correctly determine its concentrations may turn out to be a key step in helping to prevent and controlling illegal fishing activities in certain areas. In this work, we consider open access data taken from the Copernicus program (currently used in the European Union for Earth observation and monitoring) that includes several physical and biogeochemical ocean data of the area surrounding Galapagos Islands (Ecuador). We use such data in an attempt to build a reliable spatial temporal model that can be used to forecast the presence of chlorophyll-a, using a novel technique called spatial-temporal regression. Our initial results can be used to design a more complex, reliable, and implementable prediction model for real-time forecasting of chlorophyll-a presence.

3.2. Framework

As the quantity, the complexity, and the availability of oceanographic data grows, machine learning-based approaches to their analysis are becoming ever more common (Ahmad, 2019). Typical applications range from climate prediction, habitat modelling, and climate change analysis to species distribution and identification, resource management, and environmental protection (see Thessen (2016) for a recent review).

Examples of concrete applications include species identification (Guisande, 2010), automatic detection and classification of ocean pollution, oil spills, alga bloom, plastic pollution (Ahmad, 2019, Del Frate et al., 2000), as well as several fishing control-related applications (Gladju et al., 2022). Fishing control, and connected activities, in particular are of special interest in this work.

The surveillance of illegal fishing activities and the detection of abnormal shing vessel behaviours are critical issues for the management of marine resources. Machine learning techniques were employed for fishing gear recognition starting from vessel monitor system (VMS) data to detect abnormal VMS patterns of shing vessels in Indonesia (Marzuki et al., 2017). Moreover, while the coastal fisheries in national waters are closely monitored, at least by some countries, in high seas, there is a lot of uncertainty. For the automatic control of fishing activities in high seas it is necessary to understand the general behaviour of fishing fleets, to enforce fisheries management and conservation measures worldwide. Satellite-based automatic information systems (S-AISs) are now commonly installed on the vessels with the function to control ships' positions and have been proposed as a tool for monitoring the movements of fishing fleets in near real-time.

Using these data, models have been developed to detect potential fishing activity from trawlers, longliners and purse seiners (De Souza et al., 2016). However, illegal shing control is related with ocean resource and habitat management because it affects the conservation of fishery resources, and taking the correct decisions is often a hard problem, due to the non-availability of specific data. According to Thessen (2016), machine learning techniques have demonstrated the potential to eliminate data gaps, predict future events, and increase the accuracy of the results. The satellite remote sensing for marine applications started in the early 1960s, with the first pictures of the Earth. Currently, satellites are being used in the indirect detection of fishes, via measuring sea surface temperature, which is the most used environmental parameter in investigations concerning the relationship between environment and fish abundance (Santos, 2000). Nevertheless, other oceanographic variables exist that can be used to increase the accuracy of the prediction.

The lack of data that include such oceanographic variables taken directly from permanent in-situ observations have contributed to satellite remote sensing playing a pivotal role, considering that satellites offer the opportunity to measure and monitor multivariate oceanographic variables at the same time (Monolisha et al., 2017). This approach has been applied on large spatial scales with high temporal resolutions in coastal waters, but while oceanic color satellites suffer of serious limitations, such as the low spatial resolution of sensor systems, using machine learning techniques (artificial neural networks and support vector machines) allowed to develop chlorophyll-a models (Kwon et al., 2018).

Sea surface temperature and chlorophyll images are considered fundamental for the identification of fishing zones, which in turn is essential for illegal fishing detection. Features such as eddies, gyres, meanders, and upwelling that are indicative of fish abundance areas, can be derived from satellite information and are in relation to the complexity of the bottom morphology and the meteoclimate conditions. One of the elements that indicate the quality of the water is the concentration of chlorophyll-a, whose presence is highly correlated with the phytoplankton biomass (Desortová, 1981).

Phytoplankton is the base of the food chain in the marine ecosystems, and it is the main responsible for the primary production. A very recent multivariate statistical approach to predict chlorophyll-a levels in coastal marine ecosystems, taking into account 20 variables from 64 observations is reported in Franklin et al. (2020). The main differences between this approach and ours are that the former uses costal data extracted from sensor, instead of high sea data from satellite, and it does not include a spatio-temporal study of the cause-effect relationships.

The currents around Galápagos Islands carry waters plenty of nutrients from the sea bottom to the surface and has extremely high productivity areas with diverse marine organisms, attracting various species towards the exclusive economic zone of the Galápagos and its surroundings. For this reason, this zone receives pressure from industrial fishing, principally from Asia, as shown in figure 6. Very often, the activities around the maritime limits become in illegal fishery, which is very difficult to control due to the immense Galapagos' maritime territory.

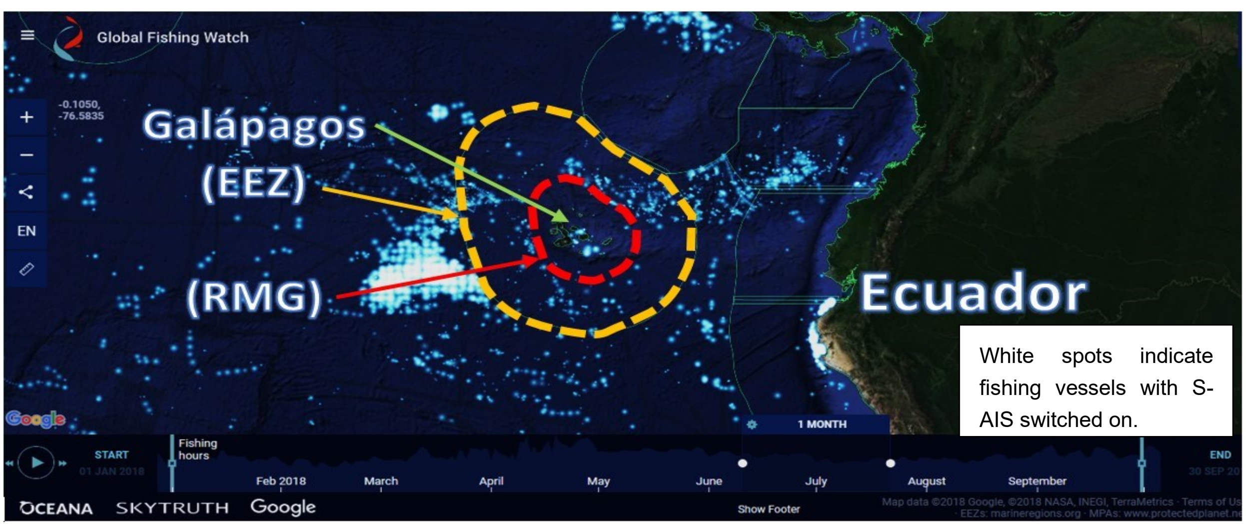


Figure 6. An aerial view of fishing ships around Galapagos exclusive economic zone. Snapshot taken in August, 2019. Source: http://www.global shingwatch.org.

In Annex A there is a comparison of the data provided by the Copernicus program versus the data recorded in situ.

3.3. Chlorophyll prediction

This research has been based on the marine food pyramid, in which chlorophyll is the photosynthetic pigment that allows estimating the concentration of phytoplankton and therefore, the biological activity. It is a basic indicator of primary production. Under this concept, a methodology based on machine learning techniques was proposed to predict chlorophyll levels especially around the Galapagos Islands.

The prediction of chlorophyll will be made using spatiotemporal data analysis techniques, on datasets extracted from the Copernicus program (COPERNICUS, 2019). In particular, the two techniques will be considered, the first one corresponding to multivariate linear regression and the second one to artificial neural networks.

3.3.1. Multivariate linear regression

The multivariate linear regression, also known as multivariate variable regression, it is used when we want to predict the value of a variable (independent variable) based on the value of two or more variables (dependent variables). In our case the dependent variable will be the chlorophyll and the independent variables will be others physical and biogeochemical variables. There are studies, such as the one carried out by Sunyé and Servain (1998), in which they establish that multivariate regression analyses showed considerable correlation between oceanographic data and level of catches (Sunye & Servain, 1998). The mathematical analysis is detailed below.

Given a data set A with *n* independent variables A1, . . ., An and one observed variable B, solving a linear regression problem consists of finding *n + 1* parameters (or coefficients) *c0, c1, . . . , cn* so that the equation:

$$B = c_0 + \sum_{i=1}^{n} (c_i \,.\, A_i + \epsilon) \qquad (1)$$

Where ϵ is a random value, is satisfied. Starting from a data set of observations:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & \dots & a_{2n} & b_2 \\ \dots & \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} & b_m \end{bmatrix} \quad (2)$$

The regression problem is usually solved by suitably estimating the coefficients *ci* so that, for each *1 ≤ j ≤ m*:

$$b_j \approx c_0 + \sum_{i=1}^{n} (c_i \,.\, a_{ij} + \epsilon) \quad (3)$$

There are several available, and well-known algorithms to solve such an inverse problem. The performance of such an estimation can be measured in several (standard) ways, such as correlation, covariance, mean absolute error, among others. When A is a multivariate time series, composed by *n* independent and one dependent time series, then data are temporally ordered and associated to a time-stamp:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} & b_1 & t_1 \\ a_{21} & a_{22} & \dots & a_{2n} & b_2 & t_2 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} & b_m & t_m \end{bmatrix} \quad (4)$$

Using linear regression to explain *B*, then, entails that finding optimal coefficients for:

$$B(t) = c_0 + \textstyle\sum_{i=1}^{n} (c_i \,.\, A_i(t) + \epsilon) \quad (5)$$

Because we aim to explain B at a certain point in time t using the values *A1(t), . . . , An(t).* Equations *(1)* and *(5)* model exactly the same problem, only in the latter the temporal component is made explicit.

Lag (or temporal) (linear) regression consists of solving a more general equation, formulated as:

$$B(t) = c_0 + \textstyle\sum_{i=1}^{n} \sum_{l=0}^{p_i} (c_{i,l} .\, A_i(t-l) + \epsilon) \quad (6)$$

In other words, we use the value of each independent variable A_i not only at time *t*, but also at time *t − 1, t − 2, . . . , t − pi* , to explain *B* at time *t*; each $A_i(t-l)$ is associated to a coefficient c*i,l,* which must be estimated, along with *each maximum lag pi*. There are available techniques, based on standard regression algorithms that allow one to solve the inverse problem associated to (6). The question to be addressed is how to estimate p, for each variable Ai. Moreover, our problem presents a further degree of complexity. Even assuming that we have solved (6), it would not be very useful for illegal fishing control, because it would be a prediction for the current day, while it may be necessary more than one day to reach the geographical points with higher chlorophyll-a predictions (which are candidates for illegal fishing areas). To solve this inconvenience, we use linear temporal forecasting regression with a temporal range of *k* units:

$$B(t+k) = c_0 + \sum_{i=1}^{n} \sum_{l=0}^{p_i} (c_{i,l}.\ A_i(t-l) + \epsilon) \tag{7}$$

In the case in which all p_{i}s are equal, which is the case in our experiments, we denote the unique maximum lag of the problem by using p.

3.3.2. Artificial Neural Network

An artificial neural network (ANN) works similarly to a human brain and consists in a network composed of several layers of neurons. The first layer is the input, after are the hidden layers and finally the output layer follow. Each layer is composed of one or several neurons. The ANN transmits data and information from their input layer to an output layer. In figure 7 you can see the structure of a neural network for chlorophyll prediction where the input variables are SPCO2 (Surface CO2), O2 (Dissolved Oxygen), NO3 (Nitrate), po4 (Phosphate), Si (Dissolved Silicate), pH, Fe (Dissolved iron), Uo (Northward sea current water velocity), Vo (Eastward sea current water velocity), mlotst (Mixed layer depth), So (Salinity), zos (Sea surface height), SST (Sea water Potential temperature – 0 m) and T-40m (Temperature – 40 m), and the output variable is Chl-a (Total Chlorophyll – a)

The number of hidden layers and neurons depends on the complexity of the model and the amount of training data. Each connection between neurons is associated with a numeric number called weight and unlike with multivariate linear regression, the neural network can capture linear and non-linear relationships between statistical inputs and output data models (Amini et al., 2021). Also, each neuron has associated an activation function. This function is responsible for introducing non-linearity in the relationship. The mathematical analysis is detailed below.

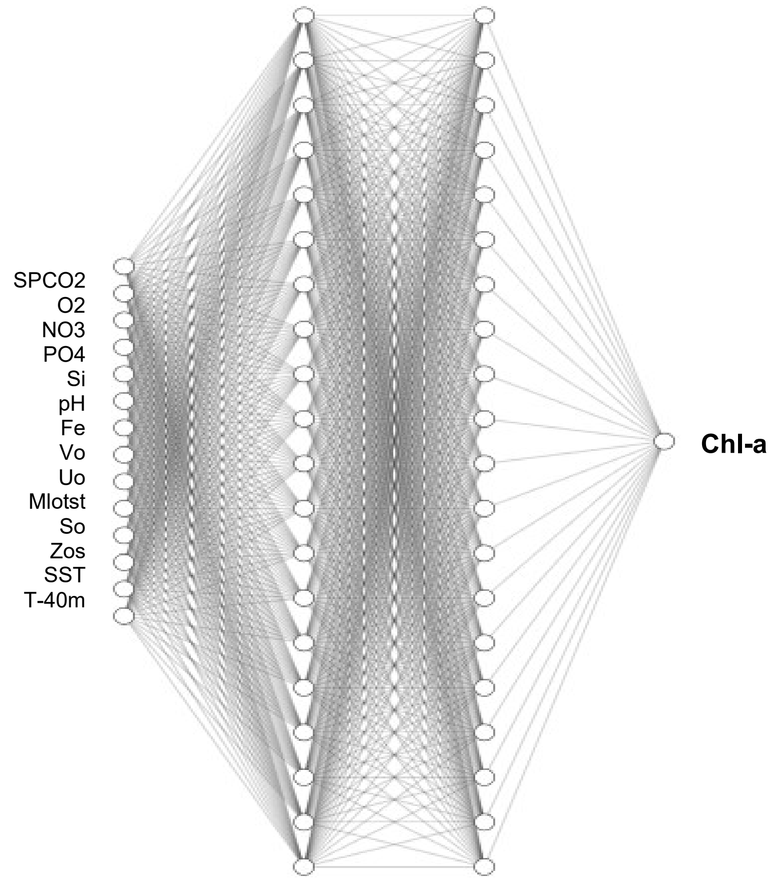


Figure 7. Structure of a neural network for chlorophyll prediction

The equation of the output function (*hi*) of neuron *i* in the hidden layer is:

$$h_{i=\sigma\left(\sum_{J=1}^{N} w_{ij} x_j + T_i^{hid}\right)} \quad (8)$$

Where σ is the activation function, N is the number of input neurons, V_{ij} are the weights, x_j inputs to the input neurons, and T_i^{hid} the threshold terms of the hidden neurons. The purpose of the activation function is introducing nonlinearity into the neural network. A common activation function is the sigmoid (or logistic) function (Wang, 2003). This function is used in our model.

$$\sigma(u) = \frac{1}{1+e^{(-u)}} \quad (9)$$

$$\sigma(z_m) = \frac{1}{1 + e^{(-zm)}} \tag{10}$$

In this case, the model will use a multilayer perceptron (MLP) trained with backpropagation method to predict the chlorophyll. The input vector would look like:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} \tag{11}$$

The weight matrix can be generalized as follows:

$$w = \begin{bmatrix} w11 & w12 & \dots & w1n \\ w21 & w22 & \dots & w2n \\ \dots. & \dots & \dots & \dots \\ wn1 & wn2 & \dots & wnn \end{bmatrix} \tag{12}$$

The output equals to the multiplication of the vector x and the matrix W. To perform the multiplication, it is necessary to transpose the matrix W.

$$z = W_T \times x = \begin{bmatrix} z_1 \\ z_2 \\ \dots. \\ z_n \end{bmatrix} \tag{13}$$

3.3.3. Models Evaluation

To validate a model, it is necessary to use a fit performance metrics. For this study we have analyzed the following statistical measures (Hallman, 2019):

a. Mean squared error.
b. Mean absolute error
c. R^2
d. k-fold Cross-Validation
e. Correlation

3.3.3.1 Mean squared error

The Mean Square Error (MSE) measures the average of the squares of the errors and shows whether the estimated values match the actual values, and is given by:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \tag{14}$$

The prediction is represented by $\hat{f}(x_i)$. High MSE value implies that the prediction not corresponding with the true value while allow value means that the prediction is near to the real value.

3.3.3.2 Mean Absolute error

Mean Absolute Error (MAE) calculates the mean of all recorded absolute errors, ignoring the sign of the error and is given by:

$$\frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{f}(x_i)\right| \tag{15}$$

This method is easy to interpret, and it is robust to outliers because there is no square operation involved.

3.3.3.3 R2

R2 also knowing as coefficient of determination is a method to measure the validity of a regression model. The output ranges from 0 to 1, where a value of 1 indicates that every point on the regression line fits the data perfectly. A value of 0.7 mean that 70% of the data points corresponding with the result of the regression line. This value can be increased if there are more independent variables. The following equation defines R2.

$$R^2 = \frac{\sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2} \tag{16}$$

3.3.3.4 K-fold Cross-Validation

K-fold Cross Validation is a method that describes how well a machine learning model performs on different data sets and randomly splits the data into k groups, hence its name K-fold. In our analysis it has been divided into 10 groups. Each unique group is divided into training and test data. To fit the model, we use the training data and evaluate it against the test dataset. After evaluating the k groups, the evaluation score of the model is obtained. It is important to note that each observation is used once in the entire training and testing procedure. The splitting of the dataset to use this method is shown graphically in the figure 8.

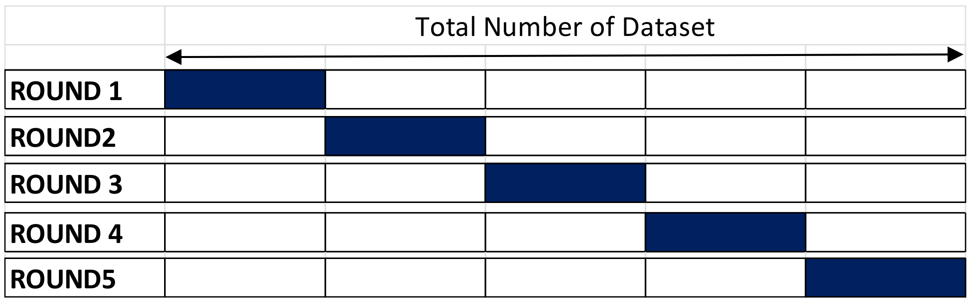


Figure 8. An illustration of 5-fold cross validation

3.3.3.5 Pearson's Correlation

Correlation is a statistical measure that expresses the linear relationship between two variables. It is a tool that helps in describing the strength of the relationships between variables, which can be positive or negative.

If the correlation is close to zero, the linear relationship is weak, the positive value indicates that both variables tend to increase o decrease together, and the negative value indicates the inverse trend between the behaviour of the variables.

The Pearson's correlation coefficient between two samples of size *n* with x_i and y_i (i = 1, 2, …,n) is defined as:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} \tag{17}$$

Where $\bar{x}$ is the average of the sample values x_i and $\bar{y}$ is the average of the sample values y_i

3.4. Data

The Copernicus encompasses three complete constellations, each one with two satellites plus an additional single satellite. This system provides 150 TB of open access data every day, including fundamental measurements or estimates of several physical, chemical, and biological oceanic variables. Ocean color information from the Sentinel satellites is employed for monitoring water quality through chlorophyll-a and phytoplankton analysis; other oceanic variables such as wave height, tide, sea currents, salinity, temperature, nutrient, and dissolved oxygen are used to run hydrodynamic models to forecast the evolution of ocean variables relevant for any kind of use, from aquaculture to oil spill combat. Moreover, temperature, salinity, mixed layer thickness, wind, sea currents, wave heights, mixed layer thickness, chlorophyll, phytoplankton, zooplankton, and nutrients are used to run models on oceanic conditions and fish's habitat spatial distribution (Prince Albert II of Monaco et al, 2019). Since our ultimate goal is to build a spatial-temporal model that explains, and therefore predicts, chlorophyll-a concentrations, phytoplankton related values have been excluded from this analysis, as they can be considered a duplication of Chl-a data.

The research area is located around the Galapagos Islands, namely between 6∘N and 10∘S and between 85∘ W and 116∘W. Our data come from the data base Global Analysis Forecast-PHY-CPL-001-012 (COPERNICUS, 2020), containing the values of physical variables, and the data base Global Analysis Forecast-BIO-001-028, containing biogeochemical variables (COPERNICUS, 2019). A summary of the considered variables can be found in table 3 and table 4.

Table 3: A description of the physical variables used in this experiment.

Variables	**Definition**	**Unit**
SST	Sea water Potential temperature – 0 m.	[°C]
T-40m	Temperature – 40 m.	[°C]
so	Salinity	[1e-3]
zos	Sea surface height	[m]
mlotst	Mixed layer depth	(m]
Uo	Northward sea current water velocity	[ms-1]
vo	Eastward sea current water velocity	[ms-1]

Table 4: A description of the biogeochemical variables used in this experiment.

Variables	Definition	Unit
Chl-a	Total Chlorophyll	[mg m-3]
	Mass concentration of chlorophyll in sea water – 40 meters.	
Fe	Dissolved Iron	[mmol m-3]
	Mole concentration of disolved iron in sea water	
NO3	Nitrate Mole concentration of nitrate in sea water	[mmol m-3]
Nppv	Total Primary Production of Phyto	[mg m-3 day-1]
	Net primary production of biomass expressed as carbon per unit volume in sea water	
O2	Dissolved Oxygen	[mmol m-3]
	Mole concentration of disolved molecular oxygen in sea water	
pH	pH	[1]
	Sea water ph reported on total scale	
Phyc	Total Phytoplankton	[mg m-3]
	Mole concentration of phytoplankton expressed as carbon in sea water	
po4	Phosphate	[mmol m-3]
	Mole concentration of phosphate in sea water	
Si	Dissolved Silicate	[mmol m-3]
	Mole concentration of silicate in sea water	
SPCO2	Surface CO2	[Pa]
	Surface partial pressure of carbon dioxide in sea water	

Our data set was constructed with daily mean values of these variables during January, 2018 ($A_{training}$) and during February, 2018 (A_{test}), with a spatial resolution of $\frac{1}{4}$ nautical mile in every direction, for a total of 8125 geographically distinct points per day. Our data contained no missing values. Basic statistical values of all 14 physical and biogeochemical variables can be found in table 5, and their one-to-one correlation can be found in table 6. As it can be appreciated, chlorophyll-a presents a high positive correlation with nutrients (NO3, PO4, Si, Fe), and some negative correlation with oxygen, pH, temperature, and mixed layer depth. Other parameters like surface CO2, salinity, sea surface height, and sea current water velocity seem to have a low correlation.

Table 5: Basic statistical values of physical and biogeochemical variables recorded during January, 2018.

	Mean	Median	Maximum	Minimum	Variance	Skewness	Kurtosis	Std Dev.
SPCO2	47.48	47.79	62.06	32.77	9.70	-0.30	0.80	3.12
O2	177.70	188.66	230.27	57.95	1630.82	-0.68	-0.64	40.38
NO3	9.96	8.19	26.88	0.20	32.98	0.63	-0.64	5.74
PO4	1.15	1.09	2.10	0.43	0.08	0.53	-0.08	0.28
Si	7.38	6.77	21.65	2.63	9.79	0.82	-0.01	3.13
pH	7.93	7.95	8.03	7.66	0.00	-1.41	2.01	0.06
Fe	9.8E-05	7.0E-05	6.4E-04	5.2E-06	7.7E-09	1.5E+00	2.2E+00	8.8E-05
Chl-a	0.45	0.32	2.03	0.13	0.09	1.69	2.47	0.31
Vo	-0.05	-0.03	1.11	-1.42	0.05	-0.51	1.85	0.23
Uo	-0.03	-0.03	1.26	-1.41	0.07	-0.06	1.45	0.27
Mlotst	13.30	9.00	87.30	2.90	103.16	1.83	3.77	10.16
So	34.88	34.96	36.68	32.90	0.19	-0.89	1.21	0.44
Zos	0.21	0.20	0.46	0.04	0.00	0.61	-0.01	0.06
ST	24.35	24.36	29.10	18.35	1.89	-0.02	-0.11	1.37
DT	21.19	21.99	28.18	11.41	9.87	-0.51	-0.82	3.14

Table 6: Correlation matrix for our variables recorded during January, 2018.

	Chl-a	**SPCO2**	**O2**	**NO3**	**PO4**	**Si**	**pH**	**Fe**	
	1	-0.15	-0.87	0.88	0.85	0.83	-0.81	0.83	**Chl-a**
Chl-a	1	1	0.22	-0.05	0.08	-0.01	0.07	-0.04	**SPCO2**
ST	0.22	1	1	-0.94	-0.89	-0.90	0.87	-0.84	**O2**
DT	-0.48	0.33	1	1	0.97	0.87	-0.88	0.87	**NO3**
Vo	-0.06	0.07	0.06	1	1	0.85	-0.90	0.83	**PO4**
Uo	-0.02	0.12	-0.04	-0.05	1	1	-0.75	0.88	**Si**
Mlotst	-0.42	0.07	0.49	0.14	0.04	1	1	-0.68	**pH**
So	-0.31	-0.43	-0.19	0.14	0.03	0.39	1	1	**Fe**
Zos	-0.11	0.52	0.58	-0.01	-0.08	0.11	-0.49	1	
	Chl-a	**ST**	**DT**	**Vo**	**Uo**	**Mlotst**	**So**	**Zos**	

3.5. Experiments and Results

Our strategy consists of the following three steps:

i Estimate the maximum lag *p* using a selection of 1000 randomly chosen geographical points from the data set $A_{training}$, so that the average correlation for temporal forecasting regression with *k* units, with $1 \leq k \leq 10$ is sufficiently high.

ii Evaluate the decision by applying temporal forecasting regression with *k* units, with $1 \leq k \leq 10$, with maximum lag *p* using the data set A*training* with all available points recorded during January 2018.

iii Validate the obtained model by testing it with $1 \leq k \leq 10$ on the data set A_{test} with all available points recorded during February 2018.

3.5.1. Experiment No. 1: Estimation of the maximum lag *p* using a selection of 1000 randomly chosen geographical points from the data set Atraining

In our first experiment, 1000 random samples were selected from the $A_{training}$ to fit a multivariate regression model to estimate the levels of Clorophyll-a, and, in particular, to evaluate the effect on the predicting capabilities of our model in adding longer histories to each of our variables. Thus, we instantiate with p from 1 to 10, and with k from 1 to 10. Our goal is to establish how the 10-fold cross-validation values for each value of k increases as p increases. This experiment will be performed for both multivariate Linear Regressions.

As we have seen, the resulting graph has the classical elbow-shaped aspect, allowing one to estimate at which point adding new lag variables is no longer worth the expected improvement. Also, we can see how our results are consistent for different values of k, suggesting that our approach is stable. In table 7 and figure 9 the results of the cross validation - MLR are shown.

Table 7: Cross Validation - MLR January 2018 - Experiment I

Day	CV Day 1	CV Day 2	CV Day 3	CV Day 4	CV Day 5	Mean - CV
1	0.86589262	0.82750112	0.80639193	0.7918245	0.77844698	0.81401143
2	0.87359864	0.83542564	0.81559288	0.80235774	0.78773991	0.82294296
3	0.8792239	0.84181512	0.82168224	0.80552125	0.78942531	0.82753356
4	0.88313508	0.84742742	0.82650792	0.80916332	0.79115306	0.83147736
5	0.88563424	0.85111674	0.83030623	0.81115197	0.7911967	0.83388118
6	0.88948017	0.85594259	0.83277022	0.8119144	0.79351738	0.83672495
7	0.89143346	0.85632327	0.83238277	0.81284185	0.79175	0.83694627
8	0.89181251	0.85752268	0.83591975	0.81333979	0.79508053	0.83873505
9	0.89351424	0.86118484	0.83686863	0.81707149	0.79880351	0.84148854
10	0.89575193	0.86222813	0.84031263	0.82083976	0.8036838	0.84456325

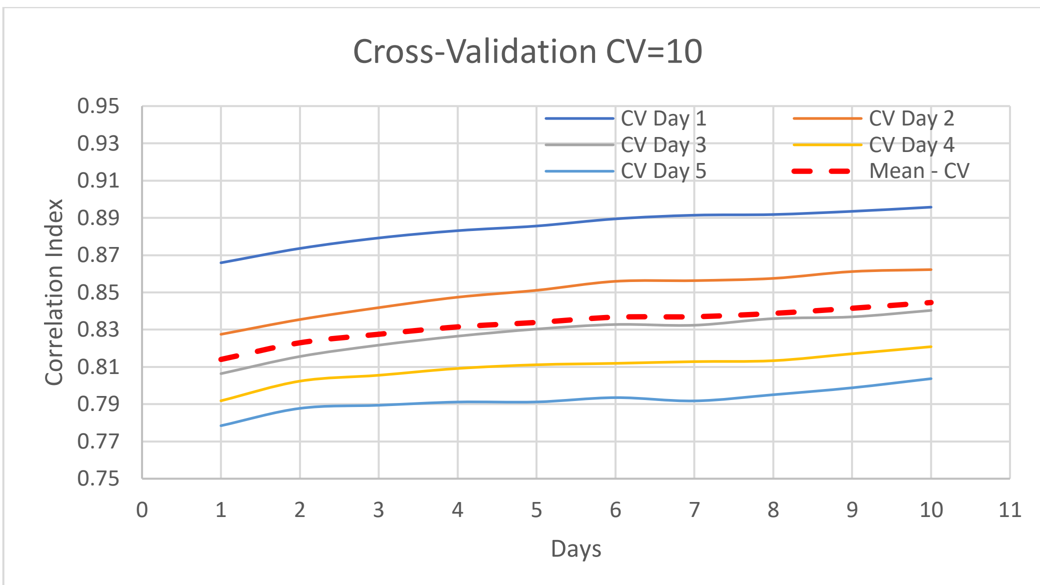


Figure 9. Cross Validation using MLR – January 2018 - Experiment I

3.5.2. Experiment No. 2: Temporal forecasting regression using data set $A_{training}$

In this second experiment, we took a dataset from January 2018, to perform a broader analysis of our models, analyzing model performance indicators such as: Mean absolute error, R^2 and K-fold Cross-Validation. We took as a maximum number of lags = 10 (p=10), as well as a maximum prediction of 10 days (k=10). The dataset used is divided into Atraining and Atest, the former composed of 80% of the data and the latter with the remaining 20%, selected randomly to give greater validity to the results. The results of the aforementioned indicators are presented below.

3.5.2.1. Mean absolute error

Multivariate linear regression

Table 7 and figure 9 indicate the results for this indicator when evaluating the multivariate Linear Regression model. It can be identified that in all cases the error is less than 11%, highlighting that when we want to evaluate after 10 days of prediction, the range of this error is between 6% and 10%.

Table 8: Mean absolute error - MLR January, 2018 - Experiment II

	Mae D=1	Mae D=2	Mae D=3	Mae D=4	Mae D=5	Mae D=6	Mae D=7	Mae D=8	Mae D=9	Mae D=10
LAG 1	0.07	0.08	0.08	0.09	0.09	0.09	0.1	0.1	0.1	0.11
LAG 2	0.07	0.08	0.08	0.08	0.09	0.09	0.1	0.1	0.1	0.11
LAG 3	0.06	0.07	0.08	0.08	0.09	0.09	0.1	0.1	0.1	0.11
LAG 4	0.06	0.07	0.08	0.08	0.09	0.09	0.1	0.1	0.1	0.1
LAG 5	0.06	0.07	0.08	0.08	0.09	0.09	0.1	0.1	0.1	0.1
LAG 6	0.06	0.07	0.08	0.08	0.09	0.09	0.09	0.1	0.1	0.1
LAG 7	0.06	0.07	0.08	0.08	0.09	0.09	0.09	0.1	0.1	0.1
LAG 8	0.06	0.07	0.08	0.08	0.09	0.09	0.09	0.1	0.1	0.1
LAG 9	0.06	0.07	0.08	0.08	0.09	0.09	0.09	0.09	0.1	0.1
LAG 10	0.06	0.07	0.08	0.08	0.09	0.09	0.09	0.09	0.09	0.1

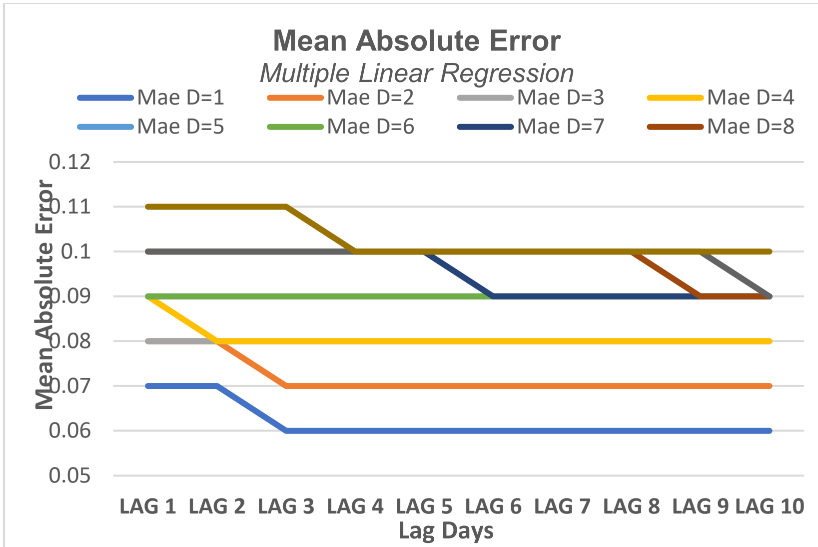


Figure 10. Mean absolute error using MLR – January 2018 - Experiment II

Artificial neural network regression

Table 9 and figure 11 show the results for this indicator when evaluating the artificial neural network regression model. It can be identified that in all cases the error is less than 9%, highlighting that, when we want to evaluate after 10 days of prediction, the range of this error is between 6% and 9%. The results obtained do not show linearity, since the neural networks are adapted and configured taking into consideration the data under analysis.

Table 9: Mean absolute error - ANN January, 2018. - Experiment II

	Mae D=1	Mae D=2	Mae D=3	Mae D=4	Mae D=5	Mae D=6	Mae D=7	Mae D=8	Mae D=9	Mae D=10
LAG 1	0.06	0.07	0.09	0.07	0.08	0.08	0.09	0.09	0.08	0.09
LAG 2	0.06	0.07	0.07	0.07	0.07	0.08	0.08	0.08	0.08	0.08
LAG 3	0.07	0.06	0.07	0.07	0.07	0.07	0.08	0.08	0.08	0.08
LAG 4	0.06	0.07	0.07	0.07	0.07	0.08	0.08	0.08	0.08	0.07
LAG 5	0.06	0.07	0.07	0.07	0.08	0.08	0.08	0.08	0.08	0.08
LAG 6	0.08	0.07	0.08	0.07	0.07	0.07	0.08	0.08	0.08	0.08
LAG 7	0.06	0.07	0.07	0.07	0.07	0.07	0.07	0.08	0.08	0.09
LAG 8	0.07	0.07	0.07	0.07	0.08	0.08	0.08	0.08	0.08	0.08
LAG 9	0.06	0.06	0.07	0.07	0.08	0.09	0.08	0.07	0.08	0.08
LAG 10	0.06	0.06	0.06	0.08	0.07	0.08	0.07	0.08	0.07	0.07

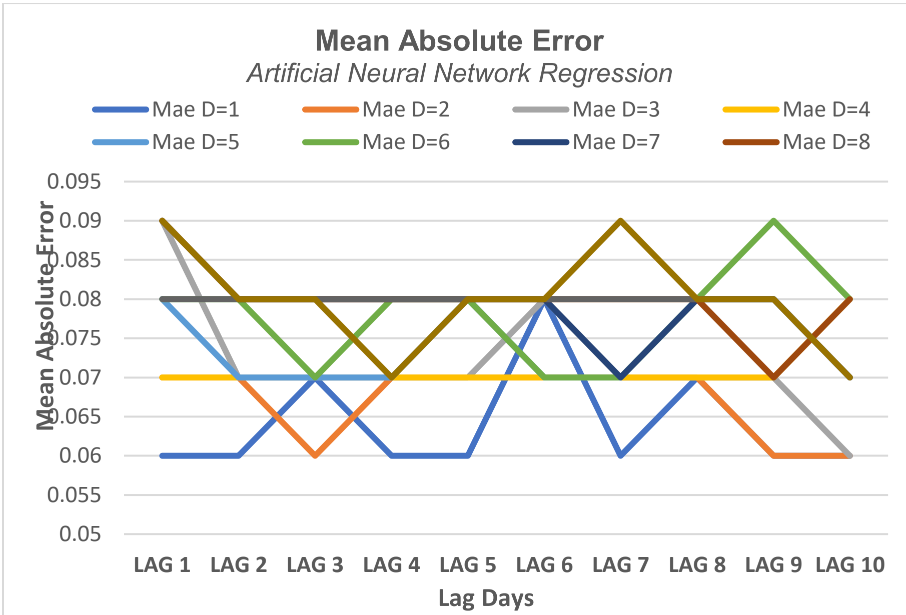


Figure 11. Mean absolute error using ANN – January 2018 - Experiment II

3.5.2.2. R^2

Multivariate linear regression

It should be remembered that this indicator is useful for measuring the validity of a regression model. Table 10 and figure 12 indicate the results for this indicator when the multivariate linear regression model is evaluated. It can be verified that when only 1 lag is used, this model goes from an effectiveness of 86% for a 1 day prediction and 58% for a 10 days of prediction, highlighting then the need to verify how many lags are necessary to add to improve the level of prediction. In this case, we see that, to improve the 1-day prediction model to 90%, 8 lags will be sufficient, while to improve the 10-day prediction to 70% it is necessary to use 10 lags.

Table 10: R^2- MLR January, 2018 - Experiment II

	R^2 D=1	R^2 D=2	R^2 D=3	R^2 D=4	R^2 D=5	R^2 D=6	R^2 D=7	R^2 D=8	R^2 D=9	R^2 D=10
LAG 1	0.86	0.81	0.78	0.76	0.73	0.71	0.67	0.65	0.62	0.58
LAG 2	0.87	0.82	0.8	0.77	0.75	0.72	0.68	0.65	0.63	0.6
LAG 3	0.88	0.84	0.81	0.77	0.75	0.72	0.69	0.67	0.63	0.62
LAG 4	0.89	0.84	0.81	0.78	0.75	0.73	0.69	0.67	0.65	0.62
LAG 5	0.89	0.84	0.82	0.79	0.76	0.74	0.7	0.68	0.65	0.63
LAG 6	0.89	0.85	0.82	0.79	0.76	0.74	0.7	0.68	0.67	0.64
LAG 7	0.89	0.85	0.83	0.79	0.77	0.74	0.72	0.7	0.66	0.65
LAG 8	0.9	0.86	0.83	0.8	0.77	0.75	0.73	0.71	0.69	0.66
LAG 9	0.9	0.86	0.83	0.81	0.78	0.76	0.74	0.73	0.7	0.69
LAG 10	0.9	0.86	0.84	0.81	0.79	0.77	0.76	0.73	0.73	0.7

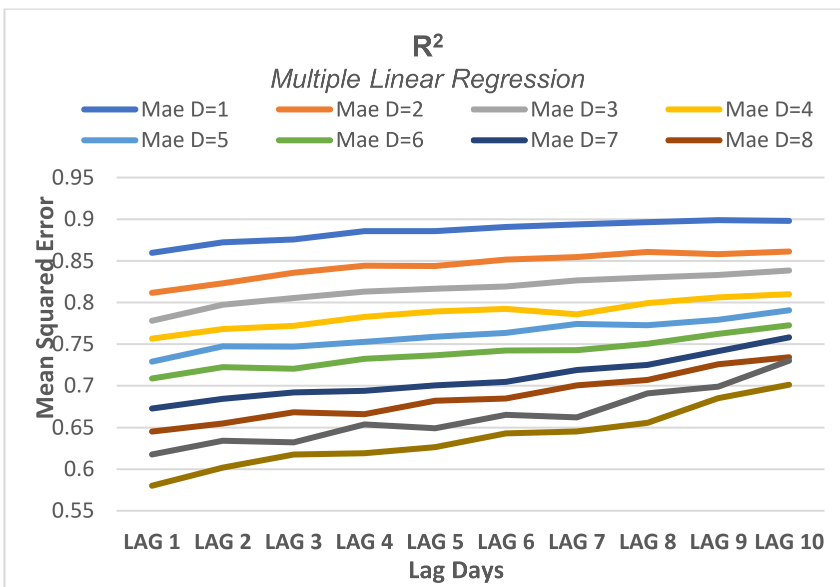


Figure 12. R^2 using MLR – January 2018 - Experiment II

Artificial neural network regression

Table 11 and figure 13 indicate the results for this indicator when the artificial neural network regression model is evaluated, identifying again the non-linearity of its result and verifying that when only 1 lag is used, this model goes from an effectiveness of 88% for a 1 day prediction to 73% for a 10-day prediction. To improve the 1-day prediction model to 90%, 7 lags will be sufficient, while to improve the 10-day prediction to 84%, it will be necessary to use up to 10 lags.

Table 11: R^2- ANN January, 2018 – Experiment II

	R^2 D=1	R^2 D=2	R^2 D=3	R^2 D=4	R^2 D=5	R^2 D=6	R^2 D=7	R^2 D=8	R^2 D=9	R^2 D=10
LAG 1	0.88	0.85	0.79	0.82	0.81	0.79	0.78	0.78	0.72	0.73
LAG 2	0.9	0.83	0.81	0.83	0.84	0.83	0.76	0.78	0.78	0.76
LAG 3	0.88	0.88	0.86	0.85	0.83	0.84	0.79	0.8	0.8	0.79
LAG 4	0.9	0.87	0.86	0.82	0.83	0.81	0.81	0.81	0.78	0.83
LAG 5	0.89	0.85	0.86	0.85	0.78	0.81	0.82	0.83	0.78	0.8
LAG 6	0.86	0.86	0.79	0.83	0.84	0.85	0.8	0.78	0.81	0.75
LAG 7	0.9	0.84	0.84	0.85	0.85	0.85	0.83	0.78	0.78	0.79
LAG 8	0.88	0.88	0.88	0.84	0.84	0.84	0.84	0.78	0.79	0.8
LAG 9	0.9	0.88	0.86	0.86	0.83	0.81	0.76	0.82	0.8	0.8
LAG 10	0.91	0.87	0.87	0.84	0.85	0.84	0.79	0.84	0.84	0.84

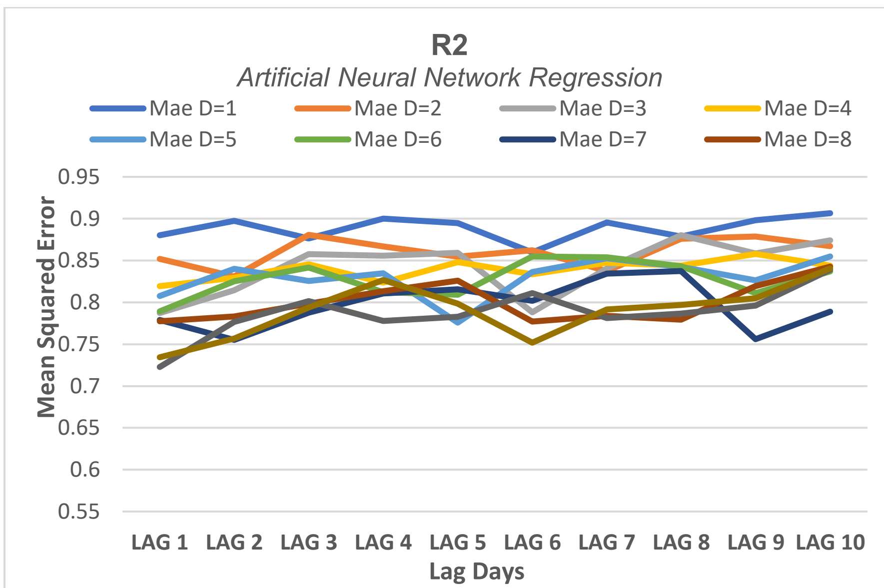


Figure 13. R^2 using ANN – January 2018 – Experiment II

3.5.2.3. K-fold Cross-Validation

Multivariate linear regression

This cross-validation method was set up to be performed with 10 k-fold and repeated for chlorophyll prediction from 1 day to 10 days and considering from 1 to 10 historical lags of each variable. With our training data, recorded in January 2018, our multivariate linear regression model shows interesting cross-validation values ranging from 0.74 (10 days ahead prediction) to 0.87 (1 day ahead prediction). Table 12 and figure 14 indicate the results for this indicator.

Table 12: k-fold Cross Validation - MLR January 2018 – Experiment II

	CV D=1	CV D=2	CV D=3	CV D=4	CV D=5	CV D=6	CV D=7	CV D=8	CV D=9	CV D=10
LAG 1	0.87	0.83	0.81	0.8	0.78	0.77	0.75	0.73	0.72	0.7
LAG 2	0.88	0.84	0.82	0.8	0.79	0.77	0.76	0.74	0.72	0.7
LAG 3	0.88	0.85	0.83	0.81	0.79	0.77	0.75	0.73	0.72	0.7
LAG 4	0.89	0.85	0.83	0.81	0.79	0.77	0.76	0.74	0.72	0.7
LAG 5	0.89	0.86	0.83	0.81	0.79	0.77	0.75	0.73	0.72	0.7
LAG 6	0.89	0.86	0.84	0.81	0.79	0.77	0.75	0.74	0.72	0.71
LAG 7	0.9	0.86	0.84	0.81	0.79	0.77	0.76	0.74	0.72	0.71
LAG 8	0.9	0.86	0.84	0.81	0.79	0.78	0.76	0.75	0.73	0.72
LAG 9	0.9	0.86	0.84	0.81	0.8	0.78	0.77	0.75	0.74	0.73
LAG 10	0.9	0.87	0.84	0.82	0.8	0.79	0.78	0.76	0.75	0.74

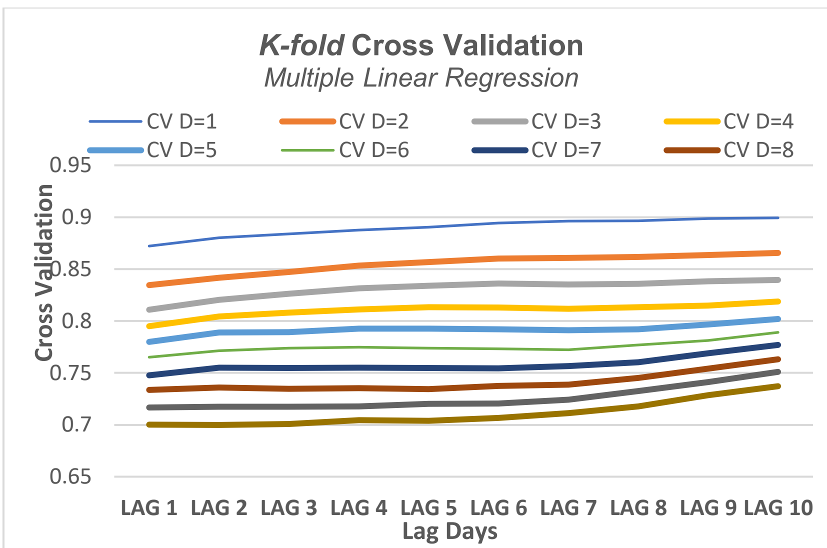


Figure 14. k-fold Cross Validation using MLR – January 2018 – Experiment II

Artificial neural network regression

The behaviour of this indicator is very similar to the previous case; however, the linearity of its prediction cannot be visualized due to the characteristics of a neural network, so the range of values of the indicator must be analyzed. It can be observed that, in some cases, it exceeds the unit, which could mean a possible network overfitting, making it difficult to establish the prediction limit parameters. It can therefore be estimated that the efficiency of the method can range from 0.85 for the 10-day prediction to 0.91 for the 1-day prediction. The table of results and the respective graph are presented in table 13 and figure 15.

Table 13: k-fold Cross Validation - ANN January, 2018 – Experiment II

	CV D=1	CV D=2	CV D=3	CV D=4	CV D=5	CV D=6	CV D=7	CV D=8	CV D=9	CV D=10
LAG 1	0.91	0.85	0.89	0.98	1.05	1.05	1.17	1.02	1.13	1.21
LAG 2	0.87	0.92	1.05	0.87	1.12	0.88	1.08	1.1	1.19	1.43
LAG 3	0.87	0.87	1.02	0.83	1.02	0.88	1.1	0.86	1.05	1.28
LAG 4	0.84	0.85	0.99	0.9	0.98	0.87	0.93	0.94	0.96	1.09
LAG 5	0.91	0.87	0.84	0.85	1.04	0.91	0.87	0.89	0.96	0.96
LAG 6	1.13	0.83	0.85	0.87	0.98	0.91	0.86	0.95	0.96	1.06
LAG 7	0.81	0.93	0.88	0.79	1.06	0.83	1.05	1.05	0.98	0.89
LAG 8	0.86	0.82	0.86	0.83	0.84	0.8	0.92	1.02	1.06	0.91
LAG 9	0.72	0.75	0.85	0.85	0.91	0.87	0.9	0.97	0.96	0.92
LAG 10	0.81	0.84	0.78	0.91	0.93	0.9	0.96	0.9	0.81	0.85

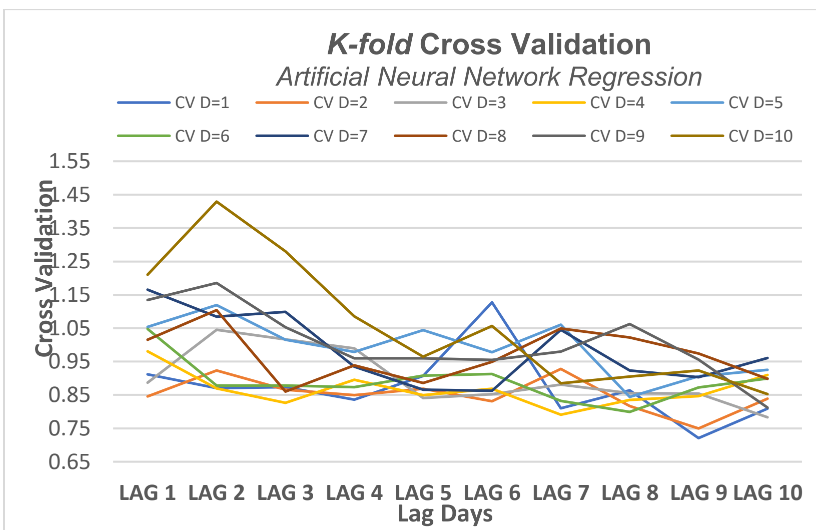


Figure 15. k-fold Cross Validation using ANN – January 2018 – Experiment II

3.5.3. Experiment No. 3: Validation the model, using data set A test

This experiment is designed to evaluate the efficiency of the method, applying two machine learning models to the February 2018 data test. Following our strategy, we applied a 10-days lag transformation to the data set $A_{training}$, and evaluated the correlation, the root squared error and mean absolute error of multivariate regression in 10-folds cross-validation mode.

3.5.3.1. Mean absolute error

Multivariate linear regression

Table 14 and figure 16 indicate the results for this indicator when evaluating the multivariate linear regression model. It can be identified that in all cases the error is less than 8%, highlighting that when we want to evaluate after 10 days of prediction, the range of this error is between 7% and 8%.

Table 14: Mean absolute error - MLR - Experiment III

	Mae D=1	Mae D=2	Mae D=3	Mae D=4	Mae D=5	Mae D=6	Mae D=7	Mae D=8	Mae D=9	Mae D=10
LAG 1	0.04	0.04	0.05	0.05	0.05	0.06	0.06	0.06	0.07	0.07
LAG 2	0.04	0.05	0.05	0.06	0.06	0.06	0.06	0.07	0.07	0.07
LAG 3	0.04	0.05	0.06	0.06	0.06	0.06	0.07	0.07	0.07	0.07
LAG 4	0.05	0.05	0.06	0.06	0.06	0.06	0.06	0.07	0.07	0.07
LAG 5	0.05	0.05	0.06	0.06	0.06	0.06	0.06	0.06	0.07	0.07
LAG 6	0.05	0.05	0.06	0.06	0.06	0.06	0.06	0.06	0.07	0.07
LAG 7	0.04	0.05	0.05	0.05	0.06	0.06	0.06	0.06	0.07	0.07
LAG 8	0.04	0.05	0.05	0.06	0.06	0.06	0.06	0.06	0.07	0.07
LAG 9	0.04	0.05	0.05	0.06	0.06	0.06	0.07	0.06	0.07	0.07
LAG 10	0.04	0.05	0.06	0.06	0.07	0.07	0.06	0.07	0.07	0.08

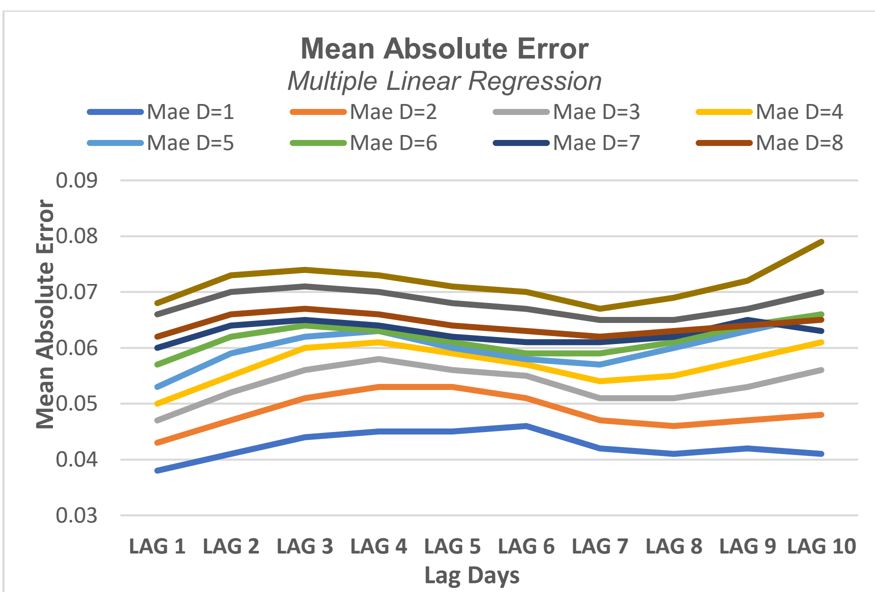


Figure 16. Mean absolute error using ANN - Experiment III

Artificial neural network regression

Table 15 and figure 17 show the results for this indicator when evaluating the artificial neural network regression model. It can be identified that in all cases the error is less than 8%, highlighting that when we want to evaluate after 10 days of prediction, the range of this error is between 8% and 9%. The results obtained do not show linearity, since the neural networks are adapted and configured taking into consideration the data under analysis.

Table 15: Mean absolute error - ANN January 2018. - Experiment III

	Mae D=1	Mae D=2	Mae D=3	Mae D=4	Mae D=5	Mae D=6	Mae D=7	Mae D=8	Mae D=9	Mae D=10
LAG 1	0.03	0.05	0.07	0.05	0.06	0.06	0.07	0.08	0.07	0.08
LAG 2	0.04	0.05	0.06	0.06	0.06	0.07	0.07	0.07	0.08	0.08
LAG 3	0.06	0.05	0.06	0.06	0.06	0.07	0.07	0.07	0.07	0.07
LAG 4	0.05	0.05	0.06	0.06	0.06	0.07	0.07	0.07	0.07	0.08
LAG 5	0.05	0.06	0.06	0.07	0.06	0.07	0.07	0.07	0.07	0.08
LAG 6	0.07	0.05	0.05	0.06	0.06	0.07	0.06	0.07	0.08	0.08
LAG 7	0.05	0.05	0.05	0.06	0.06	0.07	0.07	0.07	0.08	0.09
LAG 8	0.05	0.05	0.06	0.06	0.06	0.07	0.07	0.07	0.07	0.08
LAG 9	0.04	0.05	0.05	0.06	0.06	0.08	0.06	0.07	0.07	0.08
LAG 10	0.05	0.05	0.05	0.07	0.06	0.08	0.06	0.07	0.07	0.08

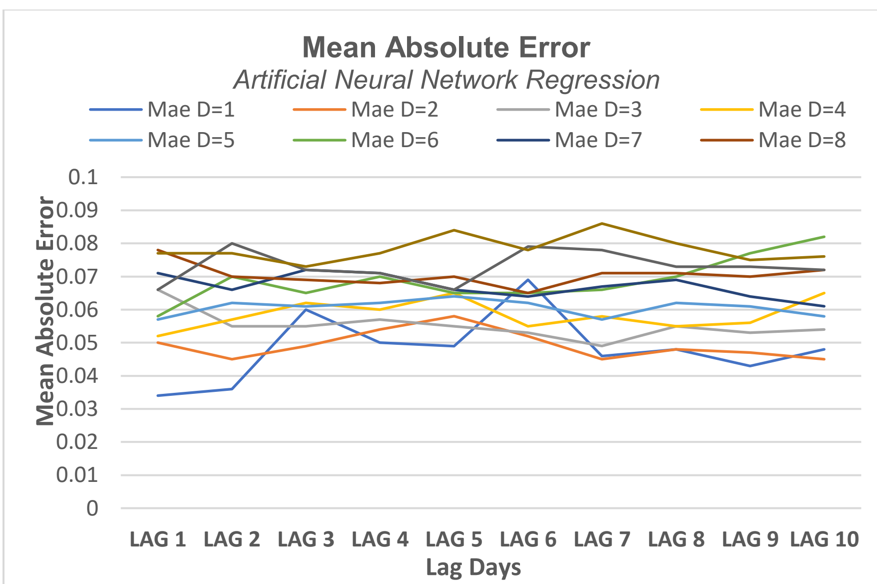


Figure 17. Mean absolute error using ANN - Experiment III

3.5.3.2. R2

Multivariate linear regression

The table 16 and figure 18 indicate the results for this indicator when the multivariate linear regression model is evaluated. It can be verified that when only 1 lag is used, this model goes from an effectiveness of 96% for a 1-day prediction and 85% for 10 days of prediction, while the effectiveness low to 94% to 77% when use 10 lags. In this case it is necessary to verify how many lags are necessary to add to improve the level of prediction. In this case, we see that to improve the 1-day prediction model to 96%, 1 lag will be sufficient, while to improve the 10-day prediction to 85% it is necessary to use 7 lags.

Table 16: R^2- MLR January, 2018 - Experiment III

	R^2 D=1	R^2 D=2	R^2 D=3	R^2 D=4	R^2 D=5	R^2 D=6	R^2 D=7	R^2 D=8	R^2 D=9	R^2 D=10
LAG 1	0.96	0.94	0.93	0.92	0.91	0.89	0.88	0.87	0.86	0.85
LAG 2	0.95	0.93	0.92	0.91	0.89	0.88	0.87	0.86	0.85	0.83
LAG 3	0.95	0.93	0.91	0.89	0.88	0.88	0.87	0.86	0.84	0.83
LAG 4	0.94	0.92	0.9	0.89	0.88	0.87	0.87	0.86	0.84	0.83
LAG 5	0.93	0.91	0.9	0.89	0.88	0.88	0.87	0.86	0.85	0.84
LAG 6	0.93	0.92	0.91	0.9	0.89	0.89	0.88	0.87	0.86	0.85
LAG 7	0.94	0.92	0.91	0.9	0.89	0.88	0.87	0.86	0.85	0.85
LAG 8	0.94	0.92	0.91	0.89	0.87	0.86	0.86	0.85	0.84	0.83
LAG 9	0.93	0.91	0.9	0.88	0.86	0.84	0.84	0.84	0.83	0.8
LAG 10	0.94	0.91	0.89	0.86	0.84	0.83	0.84	0.83	0.8	0.77

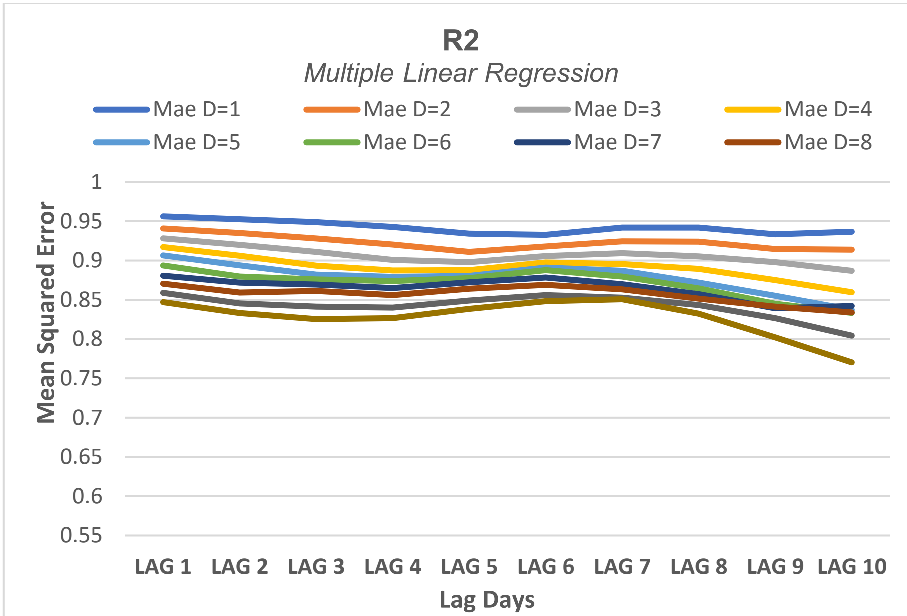


Figure 18. R^2 using MLR - Experiment III

Artificial neural network regression

Table 17 and figure 19 indicate the results for this indicator when the artificial neural network regression model is evaluated, identifying again the non-linearity of its result and verifying that when only 1 lag is used, this model goes from an effectiveness of 0.96% for a 1 day prediction to 0.81% for a 10-day prediction. To improve the 1-day prediction model to 96%, 2 lags will be sufficient, while to improve the 10-day prediction to 82%, it will be necessary to use up to 3 lags.

Table 17: R^2- ANN January, 2018 – Experiment III

	R^2 D=1	R^2 D=2	R^2 D=3	R^2 D=4	R^2 D=5	R^2 D=6	R^2 D=7	R^2 D=8	R^2 D=9	R^2 D=10
LAG 1	0.96	0.92	0.88	0.9	0.89	0.88	0.84	0.82	0.83	0.81
LAG 2	0.96	0.92	0.89	0.88	0.87	0.84	0.84	0.82	0.8	0.8
LAG 3	0.91	0.92	0.9	0.88	0.86	0.86	0.83	0.84	0.83	0.82
LAG 4	0.93	0.9	0.88	0.85	0.85	0.84	0.83	0.83	0.8	0.8
LAG 5	0.92	0.87	0.88	0.83	0.83	0.82	0.84	0.81	0.84	0.71
LAG 6	0.88	0.89	0.87	0.88	0.86	0.82	0.82	0.84	0.76	0.77
LAG 7	0.93	0.9	0.9	0.87	0.85	0.84	0.83	0.81	0.81	0.78
LAG 8	0.91	0.9	0.87	0.86	0.85	0.81	0.83	0.81	0.79	0.75
LAG 9	0.9	0.88	0.85	0.86	0.84	0.79	0.8	0.78	0.77	0.77
LAG 10	0.9	0.9	0.86	0.83	0.84	0.78	0.82	0.79	0.79	0.75

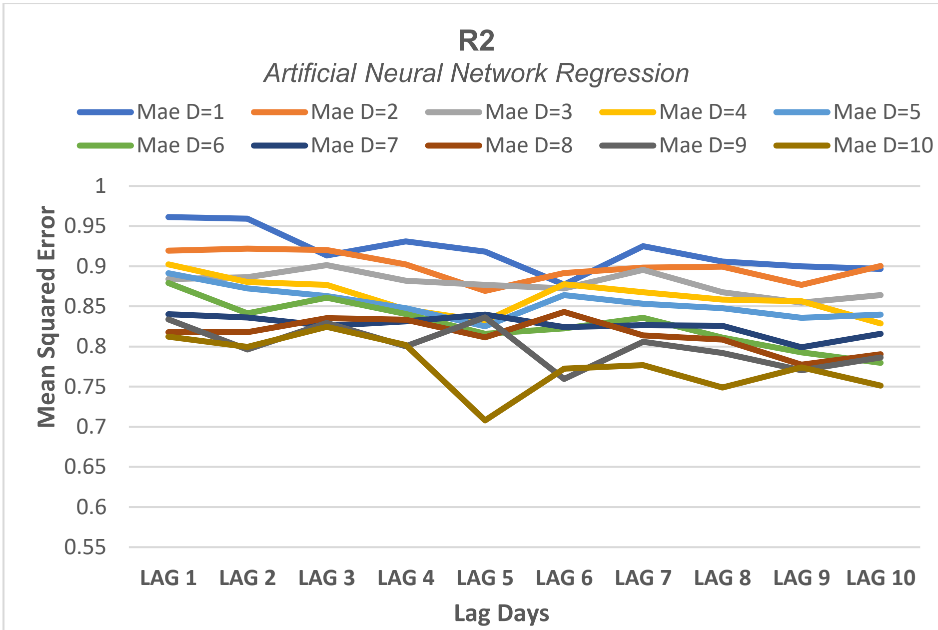


Figure 19. R^2 using ANN – January 2018 – Experiment III

3.5.3.3. Correlation

Multivariate linear regression

In order to establish the validity of our approach, the computed model was tested on our data set Atest, with values recorded during February 2018. As it turned out, the correlation values are consistent, ranging from 0.91 to 0.98. The results are shown in table 18 and figure 20.

Table 18: Correlation - MLR January, 2018 – Experiment III

	CV D=1	CV D=2	CV D=3	CV D=4	CV D=5	CV D=6	CV D=7	CV D=8	CV D=9	CV D=10
LAG 1	0.98	0.98	0.97	0.97	0.96	0.96	0.95	0.94	0.94	0.93
LAG 2	0.98	0.97	0.97	0.96	0.96	0.95	0.95	0.94	0.94	0.93
LAG 3	0.98	0.97	0.96	0.96	0.95	0.95	0.95	0.94	0.94	0.93
LAG 4	0.98	0.97	0.96	0.95	0.95	0.95	0.94	0.94	0.93	0.93
LAG 5	0.97	0.96	0.95	0.95	0.95	0.95	0.94	0.94	0.93	0.93
LAG 6	0.97	0.96	0.96	0.95	0.95	0.95	0.95	0.94	0.93	0.93
LAG 7	0.97	0.97	0.96	0.95	0.95	0.94	0.94	0.94	0.93	0.93
LAG 8	0.97	0.96	0.96	0.95	0.94	0.94	0.93	0.93	0.93	0.92
LAG 9	0.97	0.96	0.95	0.94	0.93	0.92	0.92	0.92	0.92	0.91
LAG 10	0.97	0.96	0.95	0.93	0.92	0.92	0.93	0.92	0.91	0.89

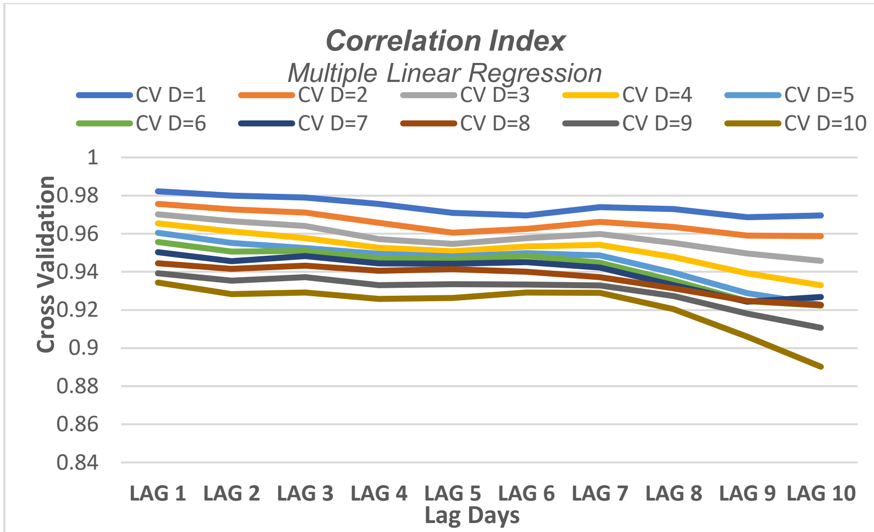


Figure 20. Correlation using MLR – January 2018 - Experiment III

Artificial neural network regression

In the same way, was evaluated the correlation index to ANN regression was evaluated, using the same data set Atest, with values recorded during February 2018. As it turned out, the correlation values are consistent, ranging from 0.88 to 0.98. The results are shown in table 19 and figure 21.

Table 19: Correlation - ANN January 2018 – Experiment III

	CV D=1	CV D=2	CV D=3	CV D=4	CV D=5	CV D=6	CV D=7	CV D=8	CV D=9	CV D=10
LAG 1	0.98	0.97	0.96	0.96	0.94	0.94	0.94	0.91	0.91	0.91
LAG 2	0.98	0.96	0.95	0.94	0.94	0.94	0.92	0.9	0.89	0.88
LAG 3	0.97	0.96	0.95	0.94	0.94	0.94	0.92	0.92	0.9	0.9
LAG 4	0.97	0.96	0.94	0.94	0.94	0.92	0.92	0.91	0.9	0.88
LAG 5	0.97	0.95	0.93	0.94	0.91	0.92	0.91	0.9	0.9	0.86
LAG 6	0.96	0.96	0.95	0.94	0.93	0.93	0.92	0.9	0.89	0.88
LAG 7	0.97	0.95	0.95	0.94	0.94	0.92	0.92	0.92	0.9	0.89
LAG 8	0.97	0.95	0.94	0.93	0.93	0.91	0.9	0.89	0.89	0.89
LAG 9	0.96	0.94	0.92	0.92	0.9	0.91	0.9	0.91	0.88	0.88
LAG 10	0.95	0.94	0.92	0.93	0.91	0.9	0.88	0.89	0.89	0.88

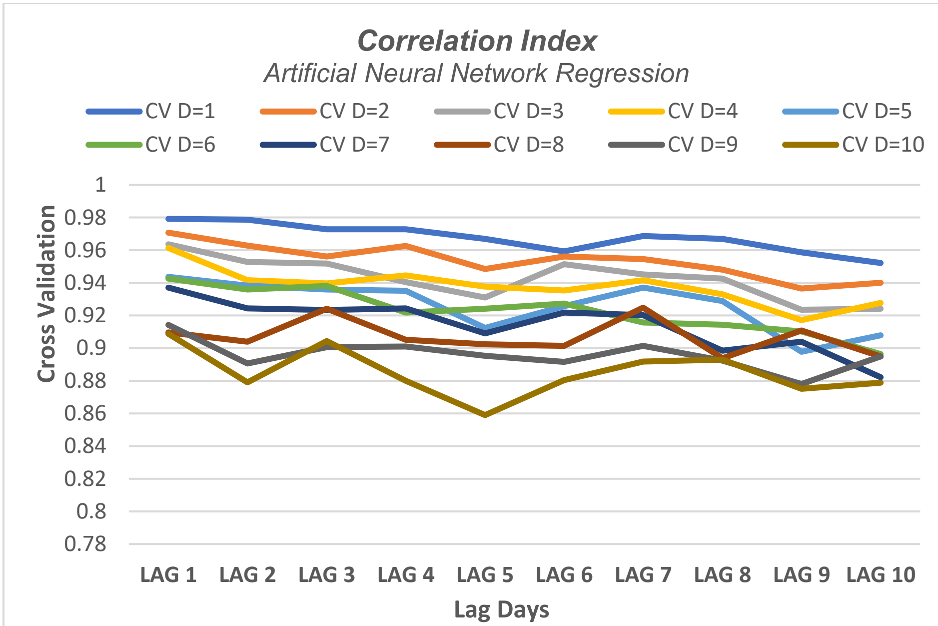


Figure 21. Correlation using ANN – January 2018 - Experiment III

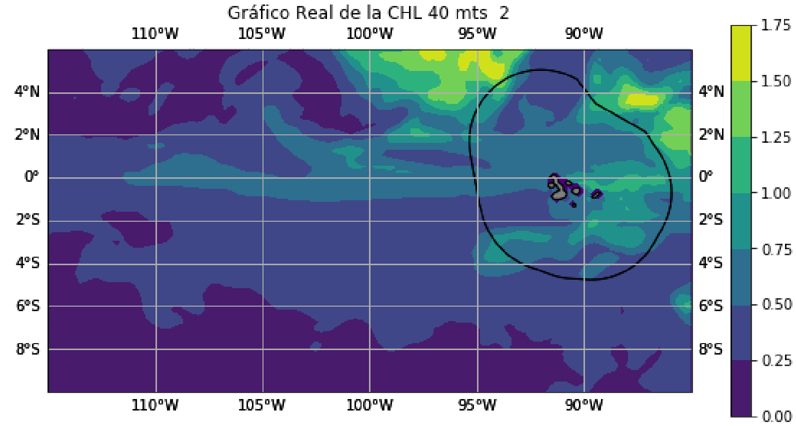
3.5.3.4. Visualization maps applying regression methods

A summary of the graphs obtained will be presented where the similarity in the prediction can be visualized graphically, both using the multivariate linear regression and artificial neural network regression methods. It is resume why mentioning that only the graphs for 1-day prediction and 1 lag, as well as for 10-day prediction with 10 lags, have been taken for the respective comparison.

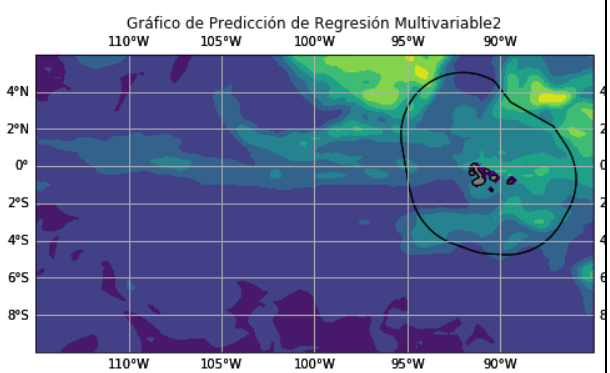
Forecast for 01 day with 01 lag

The results of the prediction using both methods are presented and can be compared with the actual chlorophyll graph, identifying that the prediction of values higher than 0.5 mg are predicted with good accuracy, while low chlorophyll values are not as accurate (Figure 22).

Real chlorophyll-a chart



Chlorophyll-a forecast chart using method Multivariate Linear Regression



Chlorophyll-a forecast chart using method Artificial Neural Network Regression

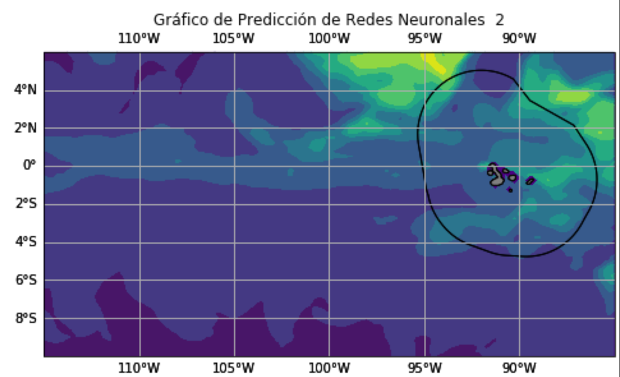


Figure 22. Chlorophyll graph - Forecast for 01 day with 01 lag (Real vs MRL – ANN)

The correlation graph (Figure 23) shows that using the multivariate linear regression method, the data are more clustered and practically not overestimated when chlorophyll is greater than 1. On the other hand, with the Artificial Neural Network method, the correlation is more dispersed and in many cases the prediction is overestimated when chlorophyll is greater than 1.

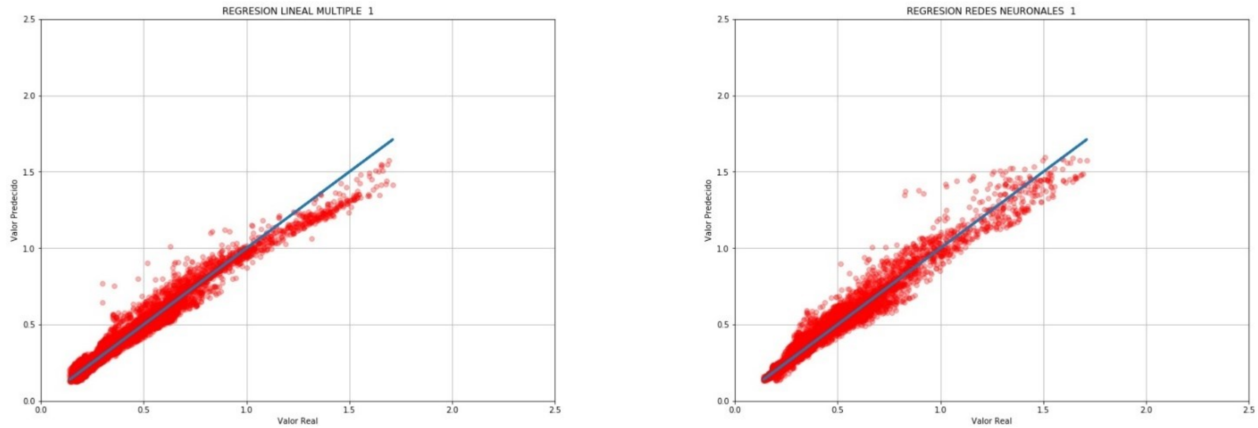
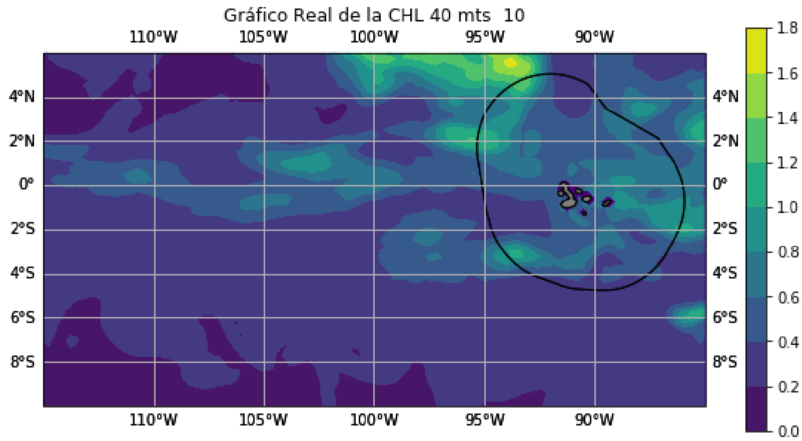


Figure 23. Correlation graph (MRL vs ANN)

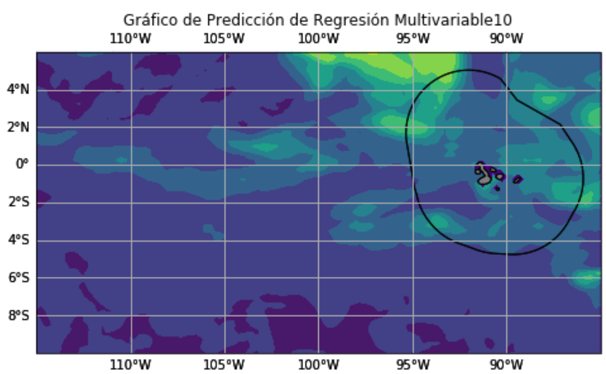
Forecast for 01 day with 10 lags

For the single day prediction using 10 lags, it is evident that the graph based on the multivariate linear regression method is closer to the real graph, especially when Chl-a is higher than 0.5. On the other hand, the other method overestimates high values of chlorophyll, which is not convenient for the prediction of potential fishing zones since it can divert control and surveillance towards unsuitable areas. Similarly, low chlorophyll values are better predicted with the second method (Figure 24).

Real chlorophyll-a chart



Chlorophyll-a forecast chart using method Multivariate Linear Regression



Chlorophyll-a forecast chart using method Artificial Neural Network Regression

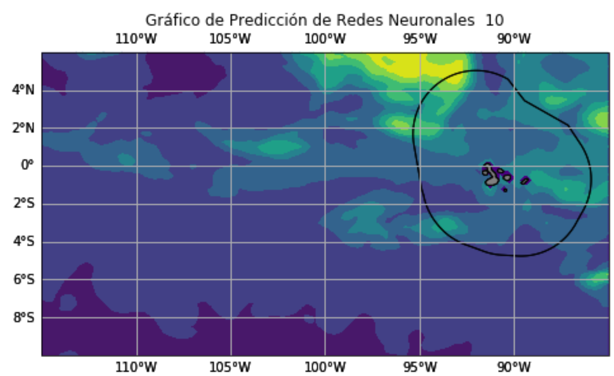


Figure 24. *Chlorophyll graph - Forecast for 01 day with 10 lags (Real vs MRL – ANN)*

The same form in the previous case, the correlation graph shows that using the multivariate linear regression method, the data are more clustered and practically not overestimated when chlorophyll is greater than 1. The correlation calculated, using the Artificial Neural Network method, the correlation is more dispersed and in many cases the prediction is overestimated when chlorophyll is greater than 1 (Figure 25).

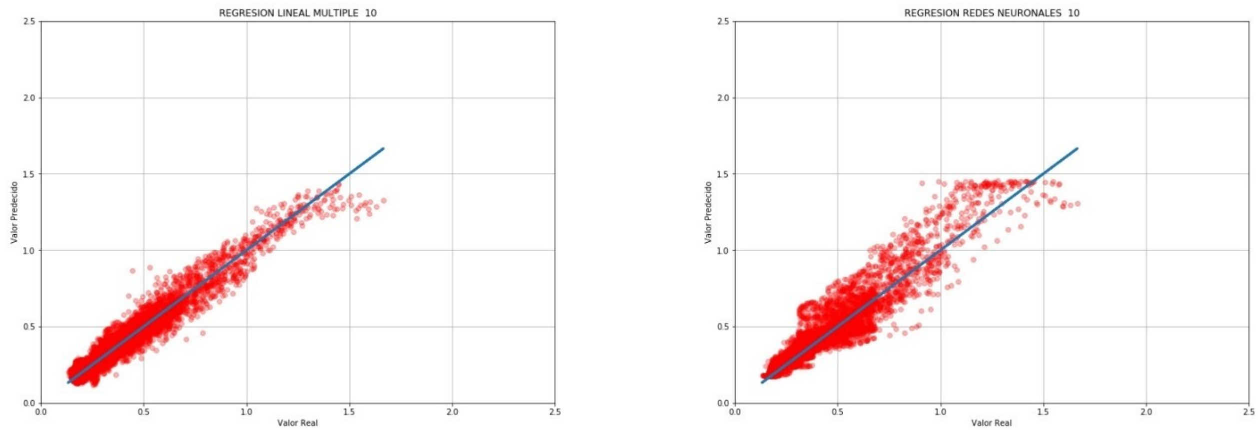
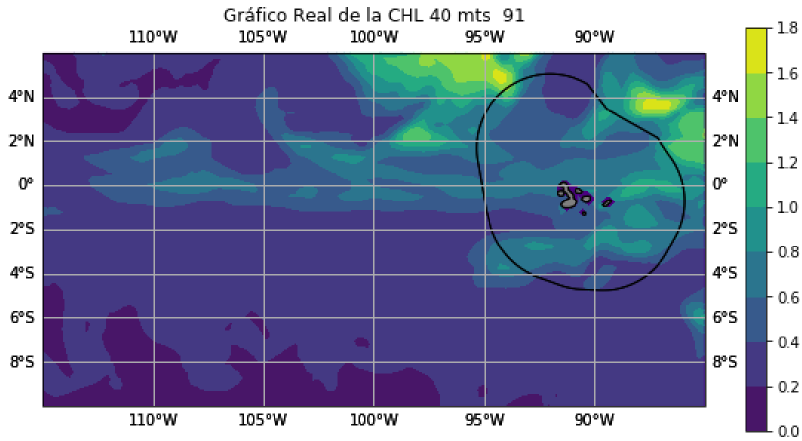


Figure 25. Correlation graph (MRL vs ANN) - Forecast for 01 day with 10 lags

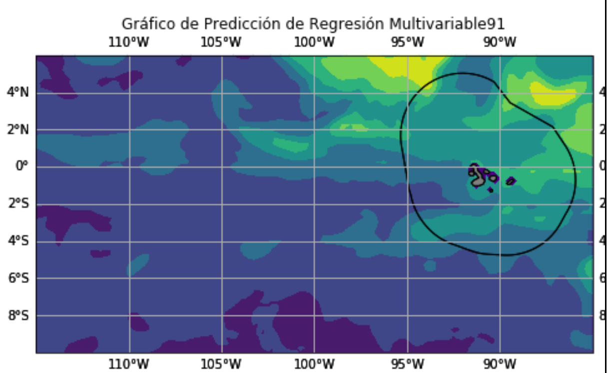
Forecast for 10th day and 01 lag

For the prediction of chlorophyll on the 10th day using 1 lag, it can be seen that both methods present a significant distortion for both high and low chlorophyll values, while for intermediate values, the prediction appears to be correct (Figure 26).

Real chlorophyll-a chart



Chlorophyll-a forecast chart using method Multivariate Linear Regression



Chlorophyll-a forecast chart using method Artificial Neural Network Regression

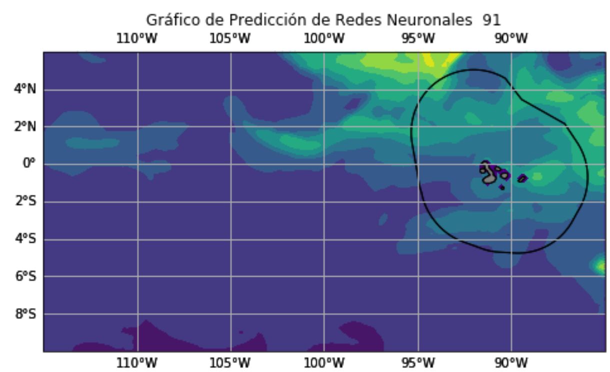


Figure 26. Chlorophyll graph - Forecast for 10 days with 01 lag (Real vs MRL – ANN)

Regarding the correlation graph, the correlation resulting from applying the multivariate linear regression method shows less dispersed data compared to the other method, which again overestimates the high values of chlorophyll, confusing the identification of potential fishing areas (Figure 27).

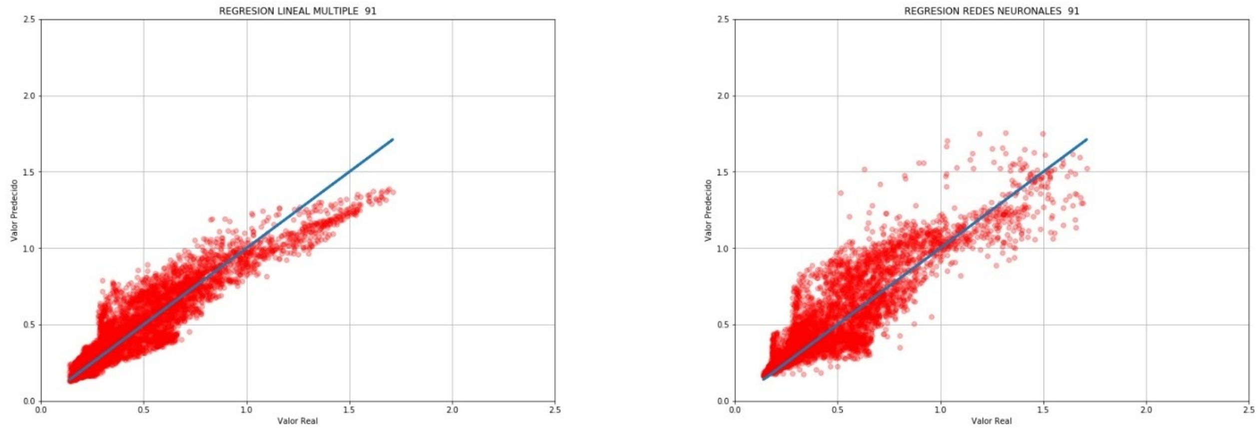
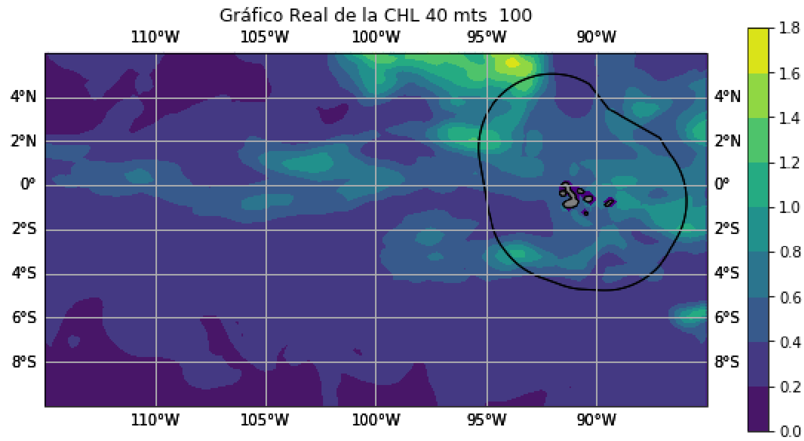


Figure 27**.** Correlation graph (MRL vs ANN) - Forecast for 10 day with 01 lags

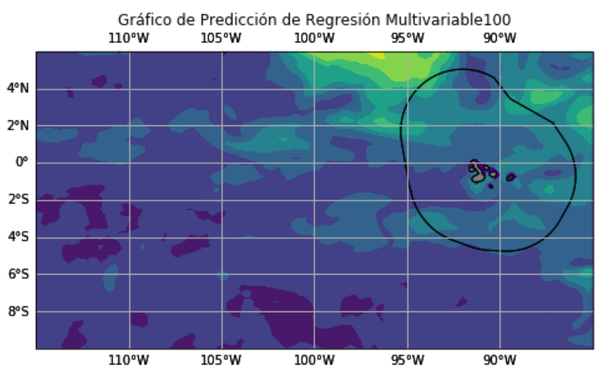
Forecast for 10 days and 10 lag

For the chlorophyll prediction of the 10th day, using 10 lags, it can be seen that the distortion in the chlorophyll prediction graphs decreases using the two methods, compared to the real chlorophyll map. However, the multivariate linear regression method presents better predictions in the medium and high chlorophyll values, while the artificial neural network method is more robust in predicting the low chlorophyll values (Figure 28).

Real chlorophyll-a chart



Chlorophyll-a forecast chart using method Multivariate Linear Regression



Chlorophyll-a forecast chart using method Artificial Neural Network Regression

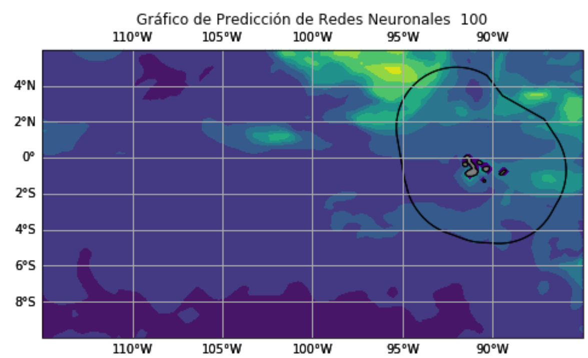


Figure 28. Chlorophyll graph - Forecast for 10 days with 10 lags (Real vs MRL – ANN)

Regarding the correlation graph (Figure 29), it can be seen again that the results applying the multivariate linear regression method show less dispersed data compared to the other method, which again overestimates the high chlorophyll values. This explains why this method was chosen to continue with the prediction optimization process.

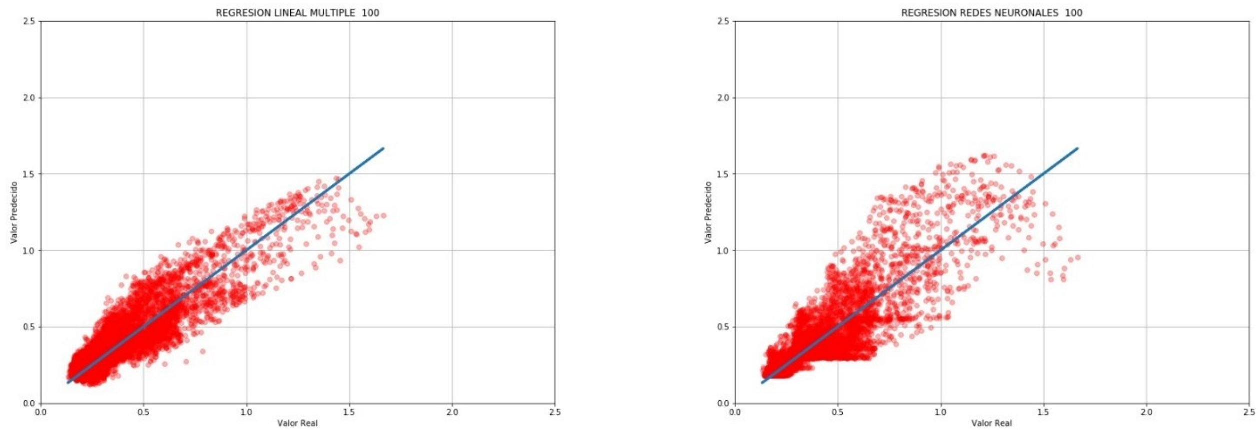


***Figure 29*.** Correlation graph (MRL vs ANN) - Forecast for 10 days with 10 lags

3.6. Conclusions

In the first experiment, as explained, we evaluated the optimal value of p. Using a random choice of geographical points, we are sure that the value of p is not biased by local conditions and, using only a small subset of the available points, that our methodology is, in fact, applicable. As we seen, the resulting graph shows the classical elbow-shaped aspect, allowing to estimate when adding new variables, the expected improvement is not worth. Also, we can see how our results are consistent for different values of k, suggesting that our approach is stable. In the second experiment, we set $p = 10$. This means that for one specific prediction we considered 10 days of history per variable. With our test data A_{test}, recorder in February 2018, our multivariate linear regression model shows interesting cross-validations values ranging from 0.91 (10 days ahead prediction) to 0.98 (1 day ahead prediction) in a multivariate linear regression method, and very similar in an artificial neural network method. From figure 22 to 29, we show how the error in the prediction is geographically distributed on a particular, random, day in February 2018. As seen, the 'bulk' of the concentration of chlorophyll-a is correctly predicted even in the 10-days ahead model; the longer the forecast, however, the more difficult is to predict the correct values in low-level concentration areas.

Finally, in this experiment we have implemented a first prediction model for chlorophyll-a concentration in the waters surrounding Galapagos Islands, Ecuador. The main motivation behind this study is the design of an implementable, cost-effective system that allows naval forces to guess possible illegal fishing areas with enough time to intervene, and enough accuracy to minimize false alarms. Using open access data, we tacked this problem as a spatial-temporal multivariate linear regression problem, and we focused to begin with, on the temporal component. While the results seem encouraging, the role of the spatial and the temporal component of this problem are yet to be identified clearly. The ultimate goal of this project is, indeed, to build a long-term explanation/prediction model, so that naval operations can be planned in a proper way.

CHAPTER IV

FOURIER-BASED OPTIMIZATION FOR MULTIVARIATE SPATIAL-TEMPORAL REGRESSION IN CHLOROPHYLL-a PRESENCE PREDICTION AROUND GALAPAGOS ISLANDS.

4.1 Abstract

Chlorophyll-a (Chl-a) is an indicator of phytoplankton biomass, which can be used to predict the presence of fish in the ocean. By predicting the Chl-a with sufficient time, this data can be used to better plan naval operations to combat illegal, unreported and unregulated fishing by increasing surveillance of the identified areas. Here, a new technique is proposed, based on the application of the discrete Fourier transform theory to optimize multivariate spatial-temporal regression model, which considers physical and biogeochemical ocean variables to predict the presence of chlorophyll-a around Galápagos Islands. This work considers open access data taken from the Copernicus program, of the European Union.

4.2 Framework

The constant formation process of the islands, coupled with the influence of the main ocean currents, positively affects biodiversity, producing and distributing nutrients, plankton and krill throughout this geographical area. In addition to the prevailing currents (Guarderas Sevilla, 2019), another aspect to consider is the behaviour of the study area's climate, a region with only two seasons, the warm from December to May and the cold one, from June to November, (SENPLADES, 2017).

The prediction model for the concentration of chlorophyll-a developed in Chávez et al., (2020), had the main motivation to design a model based on multivariate linear regression, which allows forecasting the possible areas of illegal fishing.

In oceanography, spectral analysis of temporal series can be done by making use of the Fourier series method, since it can research the interrelation between observed dynamic processes. Any periodic function following certain requirements, mainly convergence, can be represented as a series of complex exponential functions, allowing the transformation of a time signal on the frequency domain, with the objective of analyzing its frequency content in terms of amplitude and phase, because the Fourier coefficients of the transformed function represent the contribution of each sine and cosine function at a given frequency (Rodriguez, 2002).

The concept of frequency spectrum was introduced in the field of analysis of oceanographic temporal series between late 1940s and early 1950s. It was first used in the study of oceanic wind waves around 1950. During the last five decades, its use has been quickly developed and spread, despite the fact that it has shown some drawbacks for the analysis of temporal series observed in nature, as most applications use real values, while the Fourier transform uses complex arithmetic, transforming a sequence of real data of the time domain in a sequence of complex numbers in the domain of frequency (Rodriguez, 2002).

The discrete Fourier transform (DFT) is the discrete version of the Fourier transform (FT). This technique allows doing the spectral analysis of signals and can be used in various areas like science and engineering. A data series can be considered as a linear combination of frequency components, where the time domain can be represented as the sum of the frequency components, and so all waveforms can be reconstructed with an infinite number of sinusoids of various frequencies. In practice, only a finite number of sinusoids is used to reconstruct the wave (Sundararajan, 2001).

In Fourier decomposition, any N points of signal can be decomposed into N+2 signals, half of them in sine waves and the others in cosine waves where the lowest frequency corresponds to the DC signal. It is possible to create any waveform from superimposed sinusoids. Fourier decomposition provides a direct analysis of the waveform composition, including information about the frequency, phase and amplitude of its components (Smith, 2003). The variables related with the ocean constantly oscillate their values and can be considered as signals that can be decomposed into its components, in order to analyze their behaviour depending on the seasonality where they are found.

It is often difficult to analyze biological signals due to their non-linear and non-stationary characteristics. To solve this problem, time-frequency decomposition methods are used to analyze the very slight changes that exist between these signals and that may be related with an external phenomenon. If a signal is uniformly sampled and its characteristics change slowly with time, it can be considered that the seasonality is maintained during this time interval, and then the Fourier transform can be applied to this part of the signal (Wang & Veluvolu, 2017).

4.3 Experiments

In this work, we consider open access data taken from the Copernicus space program to predict chlorophyll-a presence surrounding Galapagos Islands. The main objective is to optimize the prediction of chlorophyll levels, that is conducted by using the multivariate spatial-temporal regression technique, for which the discrete Fourier transform is applied in order to first determine the behaviour of the physical, geological, chemical and biological variables (independent variables), in relation to the dependent variable (Chl-a), to subsequently optimize the model by restructuring the dataset, taking into account the lag (delay) between the variables and the Chl-a. Also, we analyzed the bathymetry around the Galápagos Islands because the bathymetry is an important factor that constrains the dynamics of the water masses, circulation and marine processes.

In order to compare the variables, a normalization of the data series is performed, to subsequently apply a decomposition of the data series applying the one-dimensional discrete Fourier transform. Next, the data series is recomposed again by applying the inverse discrete Fourier transform in n-point using the first 15 main frequencies. In this way, the noise that can occur in the data series is eliminated, and the measurement of the lag between variables is possible. Once the lag is determined, the dataset is finally reconstructed, considering the lag between different variables. Although the method is used to determine the variables which are ahead and behind in relation to the chlorophyll, in the prediction model, only those variables which are ahead of chlorophyll are taken into account, as their variation affects the Chl-a level. With the reconstructed dataset, the multivariate spatial-temporal regression is applied, predicting the chlorophyll, and improving on average 6% of the correlation between the calculated value and the real value, up to 18 days of prediction. Finally, up to 10-days lag of each variable are considered, and the dataset is restructured again to create the regression model, obtaining an improvement of up to 10% in the Chl-a prediction; nevertheless, this method is also optimal up to 18 days of prediction. The analysis is conducted in the 2090 geographical points that compose the dataset, and for a period of 1 year. Half of the dataset is used to train the model and, the other half, to test the model.

A prediction model can help to plan better and quicker maritime surveillance routes with enough anticipation, in order to ensure an early vessel detection. The presence of chlorophyll-a in combination with relevant physical, chemical, and biological variables at a certain geographical point can be thought as a multivariate spatial-temporal series, where the Chl-a plays the role of dependent variable, which is interesting to predict and used as an indicator of phytoplankton biomass (Barocio et al, 2007). Likewise, multivariate spatial-temporal regression can be used to estimate not only the functional model, but also the temporal component for each predictor. Multivariate spatial-temporal regression is simply a multivariate regression in which the spatial-temporal component is explicitly considered via suitable data transformations. In its simplest form, it consists of adding suitable lagged data to the original ones so that the temporal history of an element plays a role in the regression. In Chávez et al. (2020) it is proven that lagged data improves the chlorophyll-a prediction, but some limitations could be appreciated, as the data used for testing consists of one month and there is no detailed survey and bottom mapping study the entire ZEE of Galapagos inside and outside, only data from the General Bathymetric Chart of the Oceans (GEBCO; www.gebco.net) and low resolution bathymetric acquisitions are available to describe the study area. In this work, the use of oceanographic data of Copernicus program from 15-December-2020 to 16-December-2021 makes it possible to consider the seasonality, which also influences the chlorophyll-a behaviour. In addition, a bathymetric study was included in order to select the best study area.

4.4 The bottom morphology and the marine ecosystem in Galápagos Islands

It is known that the principal drivers of change in marine ecosystems are the state and changes of the abiotic component of the ecosystem, like as natural meteorological conditions, hydrology and hydrochemistry, the state of the coasts, and the geomorphological characteristics of the seafloor that determine the functioning features of the ecosystem in the area under study (Medinets et al., 2020).

The Galápagos Islands are of volcanic origin and stand on a raised submarine platform whose development is connected to a hotspot that, in addition to forming the archipelago, contributes magma to the mountain ridge, located between the Cocos and Nazca plate. This ride is called the Galapagos Spreading Center (GSC). Numerous seamounts, some oriented linearly, are located in the northwest, near Wolf and Darwin Islands. These seamounts seem to be of special significance because they exist at various depths as well as provide hard surfaces for the production of coral and sponge in deep water. They can also be the locations of hydrothermal mineralization and concentrations of commercially significant resources, as well as the source of local upwellings of nutrients that bolster fish populations. Typically, seamounts are conical in shape, with relatively flat summits and tiny summit craters, varying substrates and slopes, and diverse species at varying depths. There are also seamounts in deeper water, with base depths of around 2,000 meters and peaks of approximately 1,000 meters (Carey et al., 2016)

On the website of GEBCO it is possible to find information about seamounts, ridges, hills, banks and other underseas features existent in Galápagos Islands. Figure 30 shows the principal features around Galápagos.

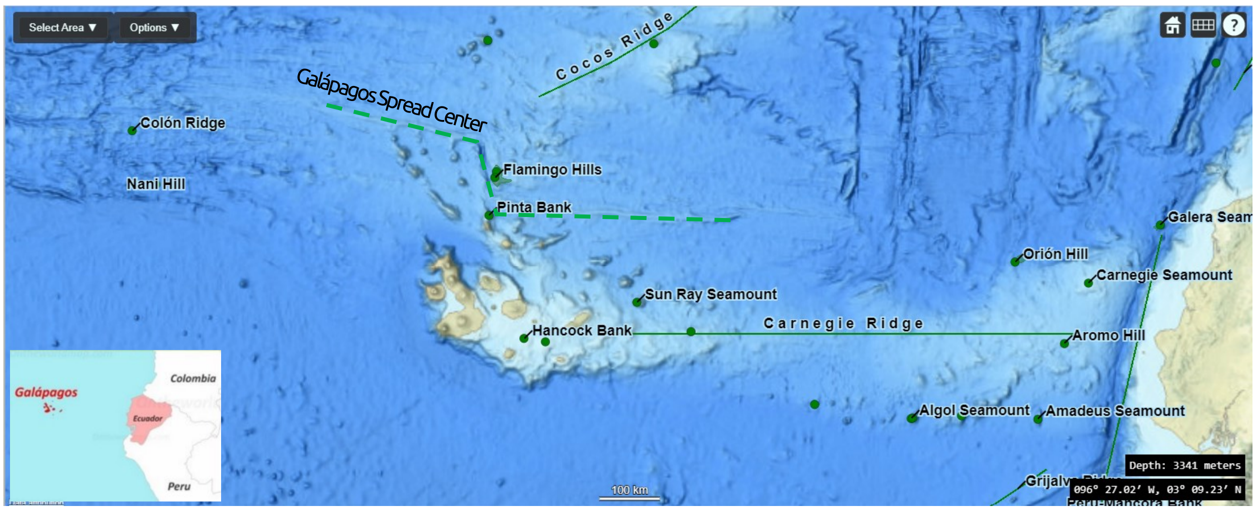


Figure 30. Galápagos undersea features. (GEBCO, 2020)

For this reason, the bathymetric and geomorphological information is a crucial element for coastal and marine modeling in the Galapagos Islands, which have unique geographical and geologic configurations in the tropical Pacific and consist of several islands and rocky outcroppings. The seabed in Galápagos is notoriously challenging because its nonstationary and complicated formations, which include several undersea features such as small channels with variable orientations and deep canyons among sections of gentle slopes.

Between San Cristobal and Fernandina, the amount of submerged platform is roughly 200 meters below sea level. The sea bottom of the surrounding Pacific Ocean is deeper than 3,000 meters. This shows that unique mechanisms are involved in the formation of the Galápagos Islands. In this case the plate tectonics is an important factor in the development of the Galápagos Islands. The plate tectonics hypothesis proposes that the earth's crust is composed of wide and small plates that move in relation to one another. The crustal motions assume many shapes, and their collisions are largely responsible for the wide range of landscapes observed today (Wooster, 1966).

In Galápagos the increased of phytoplankton and zooplankton production has been linked with the upwelling, the same is mainly generated by several variables like the local topography, the local ocean surface and deep circulation and the local meteorology (winds) that move surface waters far the coast calling deep waters, rich of nutrients, to replace the superficial ones. This phenomenon is responsible for the abundance of fish in the Galapagos Islands. The Galápagos phytoplankton stock is maintained by upwelling and the systematic association between elevated surface chlorophyll-a concentration (an indicator of phytoplankton levels) and decreased sea surface temperature (SST) measurements around the archipelago (an indicator of cool, nutrient-rich deep waters reaching the surface (Forryan & Naveira, 2021)).

4.5 Study area

The Galápagos Islands, located 965 km westward of mainland Ecuador in the eastern Pacific Ocean, have had a continuous evolving process since Cretaceous, for more than 20 million years ago (m.a.), and are the result of the ongoing interaction between the Galapagos hotspot located on the Isabela Island and the Cocos Nazca Plates (CN) (Jean-Yves, 2009). This interaction has allowed the formation of the islands' platform composed of 13 large islands, 9 small islands, and 107 islets, as well as the Carnegie and Coco's submarine volcanic ridges (Dunn et al., 2005).

The hot spot of the islands' platform is under Isabel Island, considered the most active of the archipelago (Jean-Yves et al., 2009), having 190 km from this point to the CN ridge. The expansion of this ridge ranges from 90° to 98° West, and the strong influence of the hot spot on the western side has caused the formation of underwater ridges in a depth range of -3500 to -1000 m. It is noteworthy that the evolving process of the Galapagos Islands and submarine ridges is permanent, since the Cocos Plate moves 8 cm/year NE, while the Nazca plate moves 5.8 cm/year E (Trenkamp et al., 2002). In the figure 31, the submarines ridges of Galápagos are shown.

Based on Galápagos' ocean geomorphology, we have had to divide the study area into 12 sub-areas related to different depths in order to increase model's performance (Figure 32). For this purpose, we have performed a statistical analysis of both the bathymetry of the area and the variation of chlorophyll, to select the area that will be the basis of this study. It should be mentioned that, according to the study carried out by Renteria et al (2019), national and international fishing effort is mainly concentrated in the west of the Galapagos Islands, including inside the Marine Reserve and outside the Insular Exclusive Economic Zone.

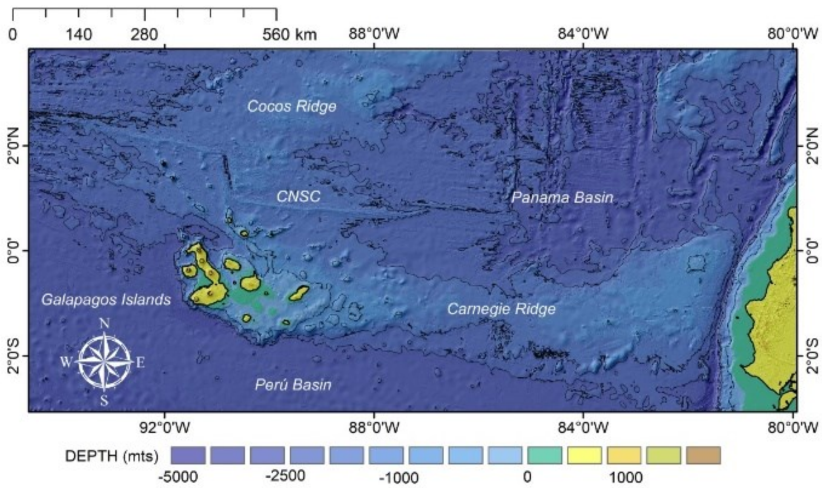


Figure 31. Galápagos Ridges

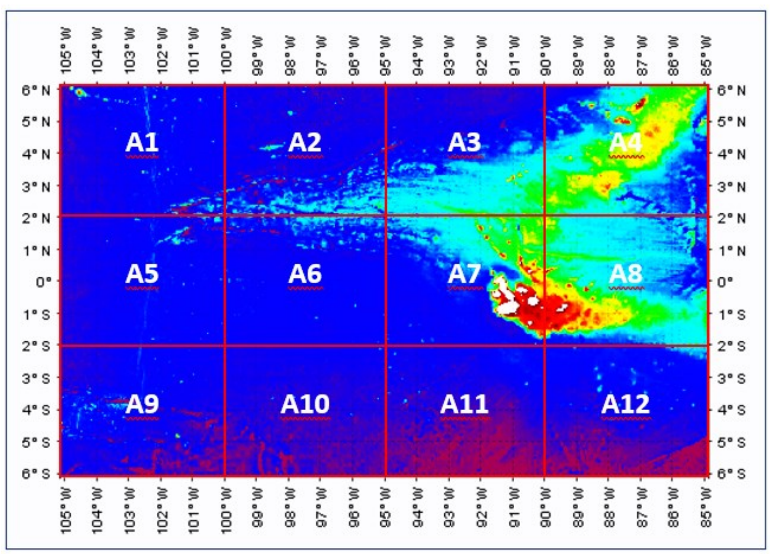


Figure 32. Study area division

The bathymetry study has been used to optimize the chlorophyll prediction model, since it has allowed us to subdivide the study area in such a way that the data considered in the model are homogeneous to increase the accuracy of the correct chlorophyll prediction. The depth has been evaluated in the development of the model as a “static variable”, which intervenes indirectly in the prediction model.

Finally, the A6 area was selected because it is located to the west of Galápagos Islands and includes the insular exclusive economic zone (IEEZ), as well as after having analyzing the behaviour of the bathymetry in that area. Variance, and standard deviation of this variable, is low compared with the results obtained in other subareas (table 20). The basic statistical values of the variable of interest in this study (Chlorophyll-a) were also analyzed and area A6 has a wide range, which will help to generalize the prediction model, its variance and standard deviation are also relatively low (Table 21). It should be mentioned that in area A6, constant overfishing activities have been identified.

Table 20: Basic statistical values of bathymetry.

AREA	Max. m	Min. m	Mean m	Range m	Var.	Std. Dev.	% CV
A1	-2875	-4222	-3482	1347	52130	228	0.17
A2	-2703	-4214	-3626	1511	31488	177	0.12
A3	-2703	-4221	-3677	1518	71010	266	0.18
A4	-2191	-4152	-3513	1961	123058	351	0.18
A5	-2154	-4048	-3353	1894	35509	188	0.10
A6	-2349	-3737	-3273	1388	36317	191	0.14
A7	703	-3712	-2601	4415	944219	972	0.22
A8	-52	-3452	-2322	3400	450137	671	0.20
A9	-2151	-4679	-3365	2528	48572	220	0.09
A10	-2333	-3953	-3322	1620	108416	329	0.20
A11	-844	-4018	-2995	3174	278410	528	0.17
A12	-2151	-4679	-3365	2528	48572	220	0.09

Table 21: Basic statistical values of chlorophyll-a during year, 2021.

AREA	Max. mg/m3	Min. mg/m3	Mean mg/m3	Range mg/m3	Var.	Std. Dev.	% CV
A1	2.127	0.135	0.385	1.992	0.070	0.264	0.13
A2	2.146	0.140	0.409	2.006	0.130	0.361	0.18
A3	1.940	0.142	0.387	1.798	0.052	0.228	0.13
A4	1.899	0.165	0.487	1.735	0.061	0.246	0.14
A5	1.572	0.165	0.426	1.406	0.026	0.162	0.12
A6	2.175	0.169	0.486	2.006	0.052	0.228	0.11
A7	1.645	0.111	0.513	1.534	0.038	0.194	0.13
A8	1.976	0.152	0.663	1.825	0.090	0.300	0.16
A9	1.589	0.135	0.271	1.454	0.018	0.134	0.09
A10	2.146	0.140	0.409	2.006	0.130	0.361	0.18
A11	2.146	0.142	0.391	2.004	0.061	0.246	0.12
A12	1.899	0.165	0.487	1.735	0.061	0.246	0.14

4.6 Data, method and results

This study was conducted in the archipelagic waters of Galápagos Islands, in the subarea located between 2° N and 2° S and between 95° W and 100° W.

4.6.1 Data

The used datasets from the Copernicus database are physical variables from the "Global Analysis Forecast-PHY-001-024-TDS¨ (COPERNICUS, 2020), and biogeochemical variables from the "Global Analysis Forecast-BIO-001-028-TDS¨ (COPERNICUS, 2019).

The variables from the downloaded datasets, considered in this study, are described in table 22, and structured with daily mean values. The Training dataset is composed by values recorded from December-2020 to May-2021 (Atraining) while the Test dataset considers the values from Jun-2021 to November-2021 (Atest), with a spatial resolution of 0.1 degrees in every direction, for 2091 geographically distinct points per day, in total 7259952 measurements. Our datasets did not contain missing values. The correlation one-to-one of 16 physical and biogeochemical variables can be found in table. 23.

Table 22: Description of the physical and the biogeochemical variables

	Variable	Description	Unit
Biogeochemical variables	Chl-a	Total chlorophyll-a	mg/m3
	Fe	Dissolved iron	mmol/m3
	NO3	Nitrate	mmol/m3
	O2	Dissolved oxygen	mmol/m3
	pH	pH	
	PO4	Phosphate	mmol/m3
	Si	Dissolved silicate	mmol/m3
	SPCO2	Surface CO2	Pa
	Nppv	Net primary production	mmol/m3
	Phyc	Phytoplankton concentration in carbon	mmol/m3
Physical variables	ST	Sea water -40m temperature	ºC
	So	Salinity	1/e3
	Zos	Sea surface height	m
	Mlots	Mixed layer depth	m
	Uo	Northward sea current water vel.	m/s
	Vo	Eastward sea current water velocity	m/s

Table 23: Correlation matrix for our variable recorded during Dic-2020 to Nov-2021

	Chl-a	**Chl-a**	**CO2**	**ST**	**So**	**Mlots**	**Zos**	**Vo**	**Uo**	
Chl-a	1	1	0.68	-0.81	0.75	-0.66	-0.69	0.02	0.67	**Chl-a**
Nppv	0.79	1	1	-0.62	0.81	-0.55	-0.61	0.11	0.63	**CO2**
phyc	0.87	0.75	1	1	-0.88	0.78	0.89	-0.06	-0.71	**ST**
Fe	0.91	0.84	0.78	1	1	-0.76	-0.81	0.08	0.69	**So**
Si	0.88	0.83	0.78	0.97	1	1	0.76	0.01	-0.56	**Mlots**
NO3	0.88	0.75	0.71	0.94	0.92	1	1	-0.03	-0.63	**Zos**
PO4	0.85	0.76	0.78	0.89	0.95	0.90	1	1	0.05	**Vo**
O2	-0.77	-0.71	-0.55	-0.89	-0.85	-0.92	-0.75	1	1	**Uo**
pH	-0.80	-0.73	-0.72	-0.83	-0.92	-0.83	-0.93	0.74	1	
	Chl-a	**Nppv**	**Phyc**	**Fe**	**Si**	**NO3**	**PO4**	**O2**	**pH**	

4.6.2 Chlorophyll-a prediction (30 days)

First, the chlorophyll-a prediction is conducted using the multivariate spatial-temporal regression method, in order to determine the efficacy of this method without applying any kind of optimization. Two datasets are configured; the first one will be used for model training and the second one will be used for its evaluation. Each dataset contains data of 180 days, evaluated from each of the 2091 geographical points and the 16 analyzed variables. The correlation between the real value and the predicted values are evaluated, and this will be our reference value to evaluate the method optimization. In the figure 33, the results from the preliminary evaluation can be observed. The prediction without optimization of the chlorophyll substance decreases from 0.85 to 0.39, this prediction is made during the following 30 days, on each day considered in the training dataset.

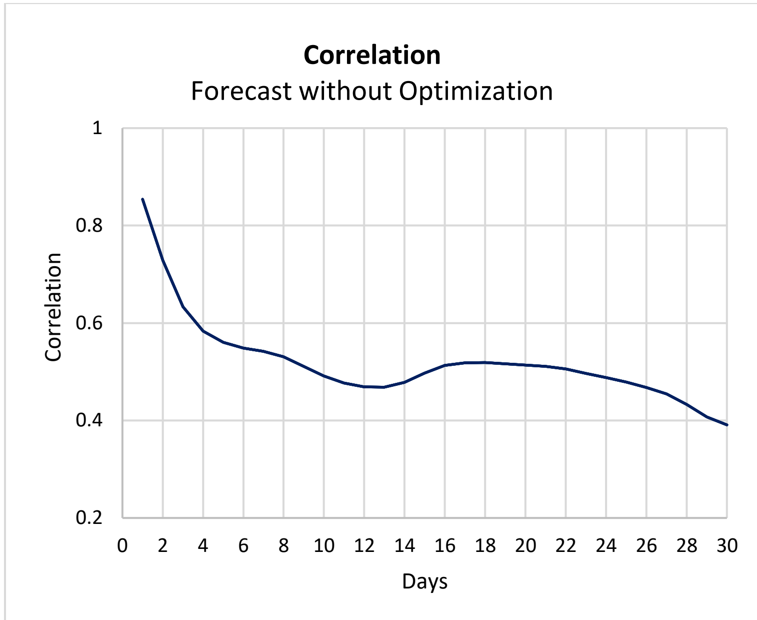


Figure 33. Correlation of chlorophyll-a prediction without optimization

4.6.3 Optimization model and results

DFT and IDF. A periodic time domain sequence $x(n)$, with period N, can be expressed as summation of a set of N harmonically related complex sinusoids with coefficients $X(k)$ as

$$x(n) = \sum_{K=0}^{N-1} X(k) e^{j\frac{-2\pi}{N}nk}, n = 0,1, \dots, N-1 \tag{18}$$

Applying Euler's formula, we can express this equation in terms of sines and cosines:

$$x(n) = \sum_{K=0}^{N-1} x_n . \left[cos\left(\frac{2\pi}{N}kn\right) - i. sin\left(\frac{2\pi}{N}kn\right)\right] \tag{19}$$

Discrete Fourier transforms (DFT) are extremely useful because they indicate the periodicities in the data of a signal, as well as the contribution of each of its periodic components, and can be used on samples containing certain numbers of points. In our case, the analyzed data consist on daily samples of physical, chemical and biological values, represented as a data series, to which we can apply DFT, determining its components.

After applying DFT, it has been chosen to reconstruct the data series using exclusively its first 15 components of the data series, this way it is possible to remove noise and understand the behaviour of chlorophyll in relation with other variables. For this effect, the Inverse Discrete Fourier Transform (IDFT) is used, serving as passband filter, where the first few components are selected to reconstruct the signal in N-dimensional complex vectors.

$$x(n) = \frac{1}{N}\sum_{K=0}^{N-1} X(k)e^{j\frac{2\pi}{N}nk}, n = 0,1,\ldots,N-1 \tag{20}$$

In the figure 34 to figure 48, after applying the DFT and then the IDFT, it can be visualized that reconstructing a signal with the first 15 components is enough to understand the trend that chlorophyll may have. One point has been randomly selected for this demonstration.

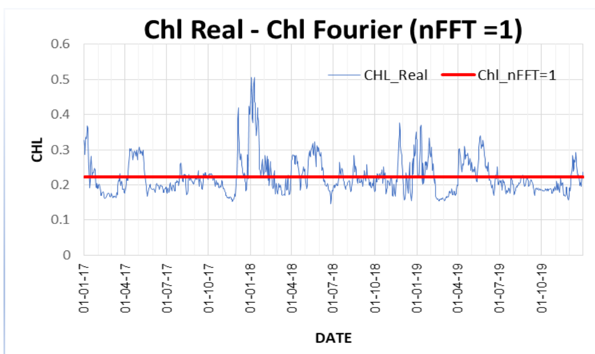


Figure 34. *Chl-a real vs Chl-a Fourier (nFFT=1)*

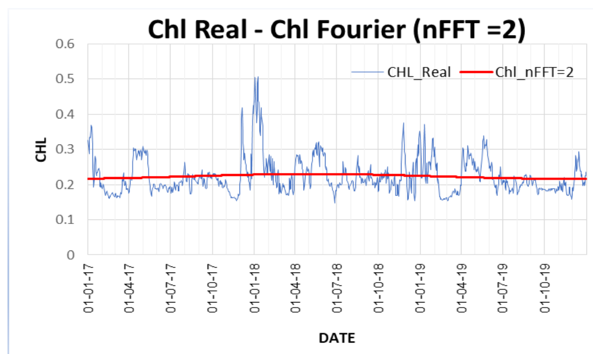


Figure 35. *Chl-a real vs Chl-a Fourier (nFFT=2)*

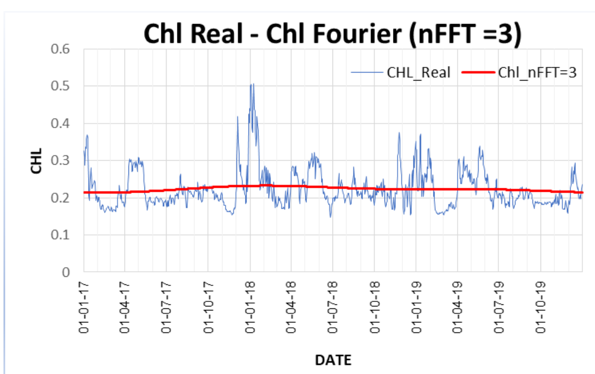


Figure 36. *Chl-a real vs Chl-a Fourier (nFFT=3)*

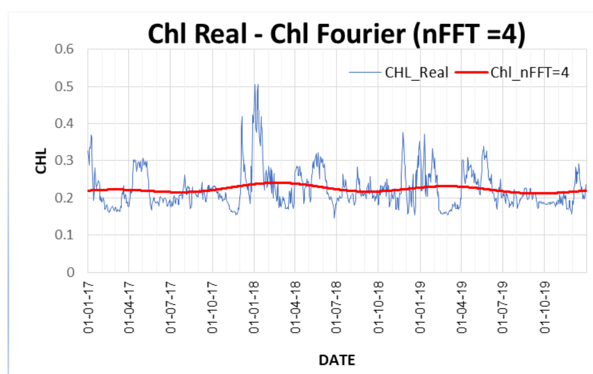


Figure 37. *Chl-a real vs Chl-a Fourier (nFFT=4)*

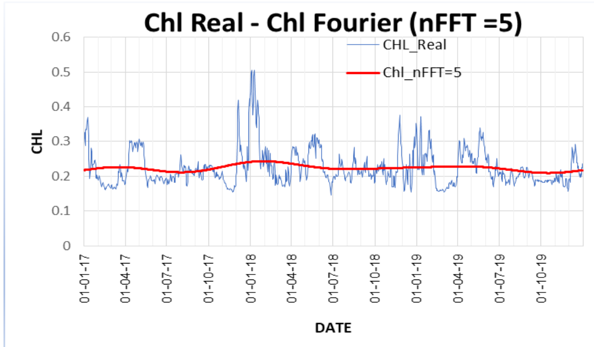


Figure 38. *Chl-a real vs Chl-a Fourier (nFFT=5)*

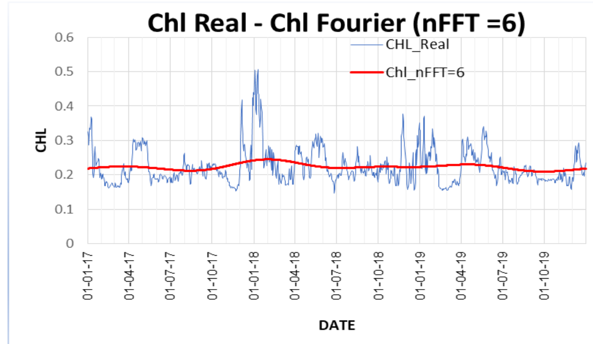


Figure 39. *Chl-a real vs Chl-a Fourier (nFFT=6)*

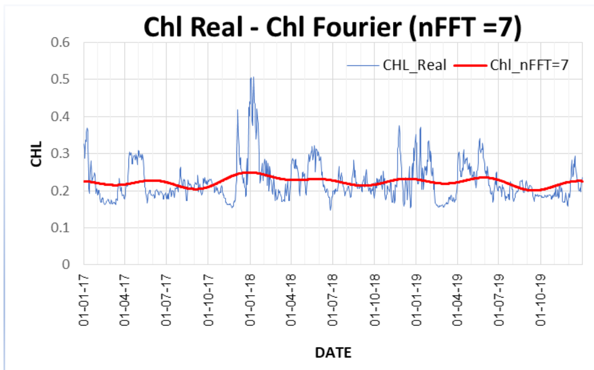


Figure 40. *Chl-a real vs Chl-a Fourier (nFFT=7)*

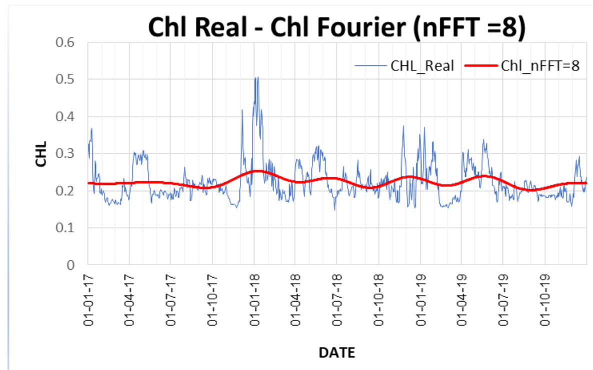


Figure 41. *Chl-a real vs Chl-a Fourier (nFFT=8)*

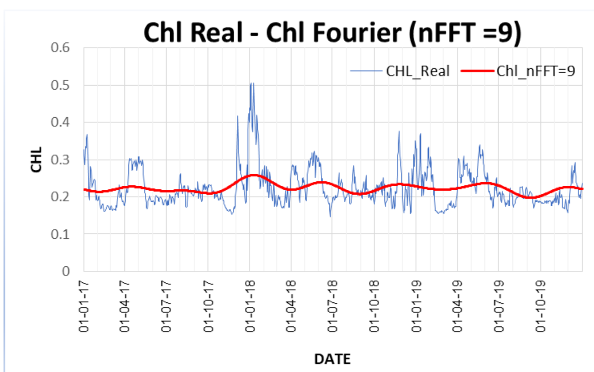


Figure 42. *Chl-a real vs Chl-a Fourier (nFFT=9)*

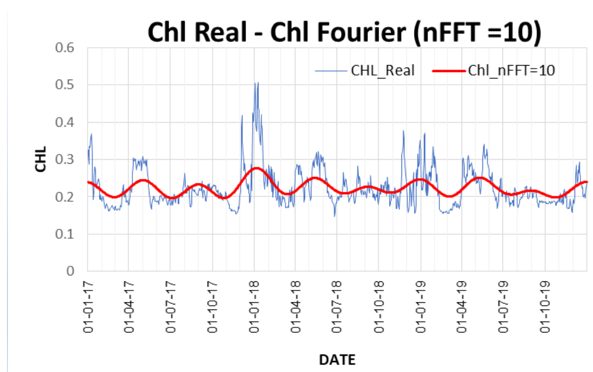


Figure 43. *Chl-a real vs Chl-a Fourier (nFFT=10)*

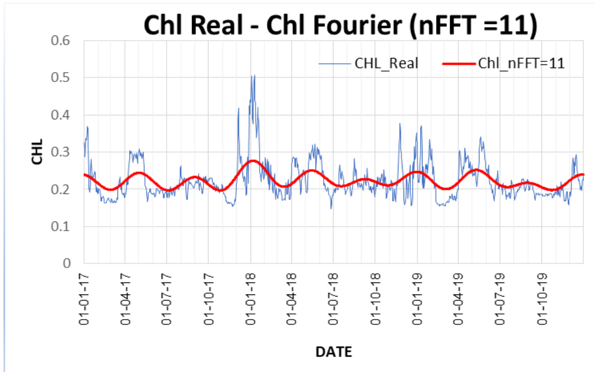


__Figure 44.__ Chl-a real vs Chl-a Fourier (nFFT=11)

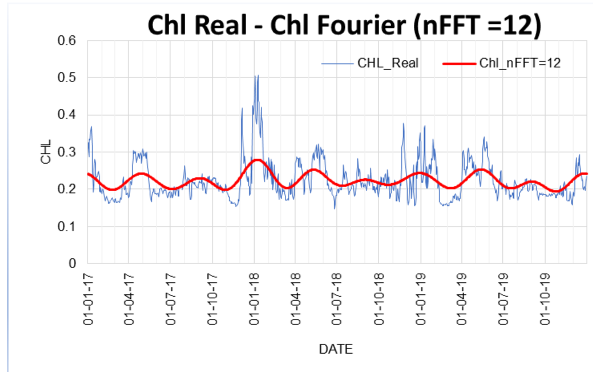


__Figure 45.__ Chl-a real vs Chl-a Fourier (nFFT=12)

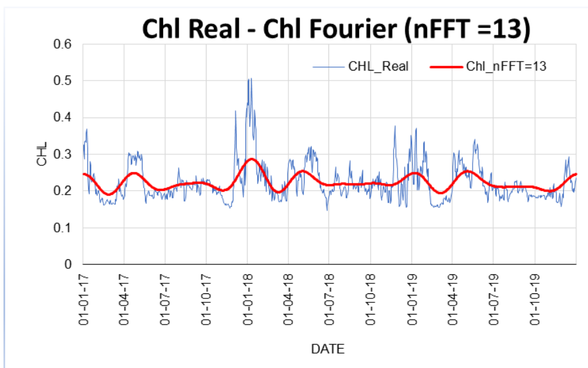


__Figure 46.__ Chl-a real vs Chl-a Fourier (nFFT=13)

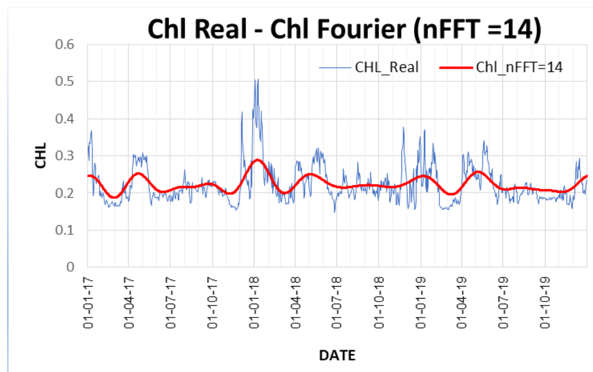


__Figure 47.__ Chl-a real vs Chl-a Fourier (nFFT=14)

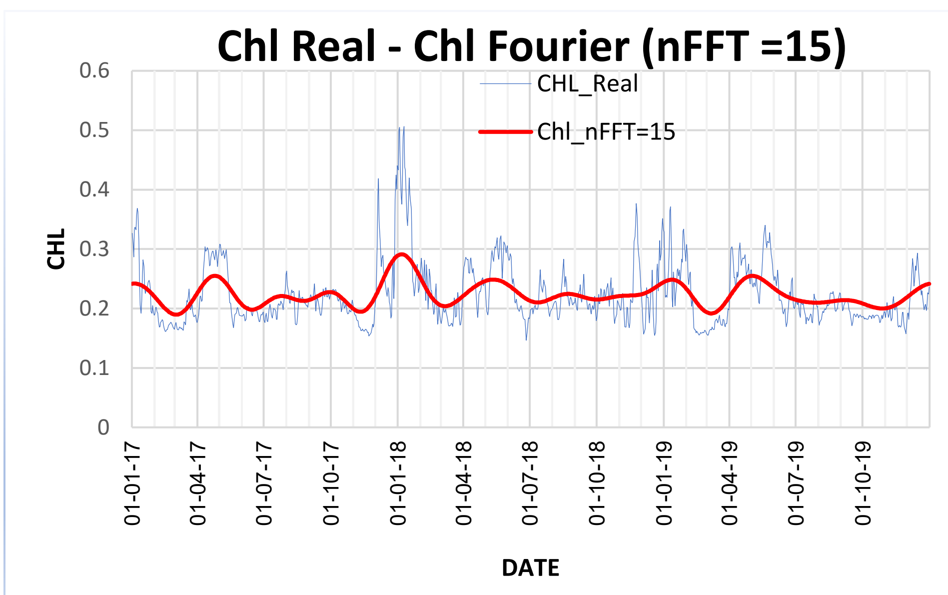


__Figure 48__. Chl-a real vs Chl-a Fourier (nFFT=15)

As it can be observed in figure 48, the Chl-a, reconstructed with 15 components, it is sufficient to shows the tendency of the variables observed in the analysis. This transformation must be performed for every variable and for every one of the 2091 geographical points in the dataset, per day. To perform these transformations, we used Python.

Afterwards, a comparison between the different variables was performed, relative to the Chl-a, to determine the phase lag between these waves. Since the analyzed waves use different units and magnitudes, it was first necessary to normalize the datasets.

The phase angles are calculated from a one cycle discrete Fourier transform (DFT), based on phase to phase between data series Chl-a and each variable in all 2091 points. For this calculation, specialized Python libraries of signal processing were used. This process was used to increase the correlation in the dataset to design the multivariate spatial-temporal regression model. It is estimated that if the waves are in phase, the Chlorophyll prediction must improve, being this the hypothesis of our study. It should be mentioned that it was necessary to identify the variables that have a proportional behaviour to chlorophyll and those that behave inversely. Points have been randomly selected. Their demonstration is presented for physical variables from figure 49 to 54 and for biogeochemical variable from figure 55 to 58.

Physical Variables

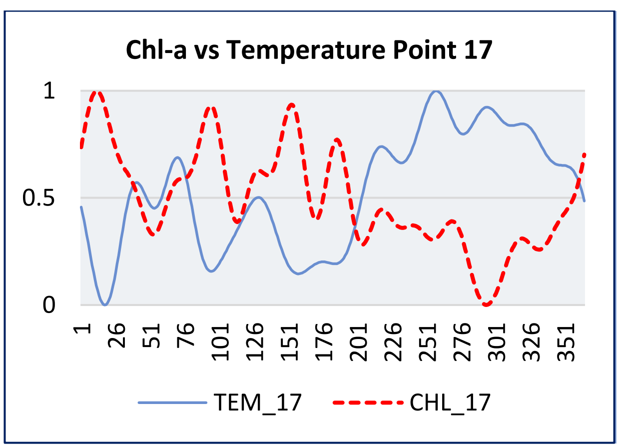


Figure 49. *Chl-a vs Temperature (nFFT=15)*

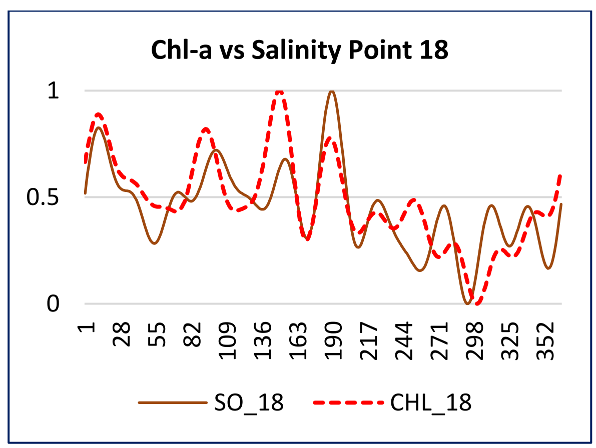


Figure 50. *Chl-a vs Salinity (nFFT=15)*

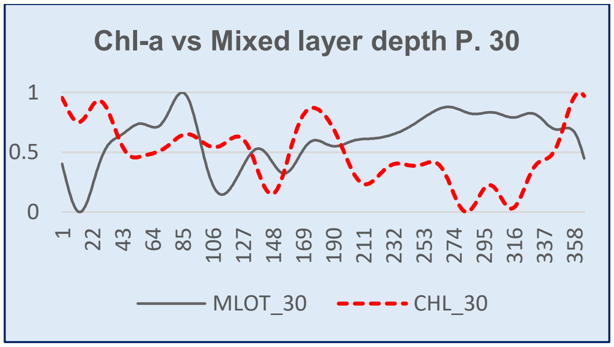


Figure 51. *Chl-a vs MLOTS (nFFT=15)*

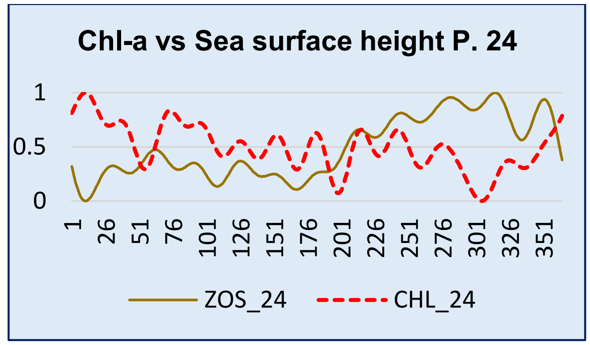


Figure 52. *Chl-a vs ZOS (nFFT=15)*

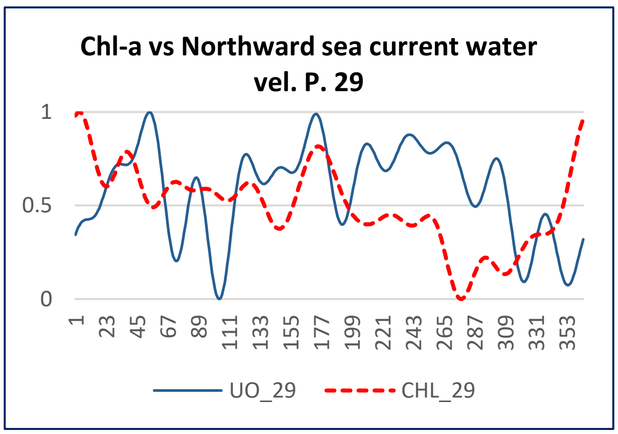


Figure 53. *Chl-a vs Uo (nFFT=15)*

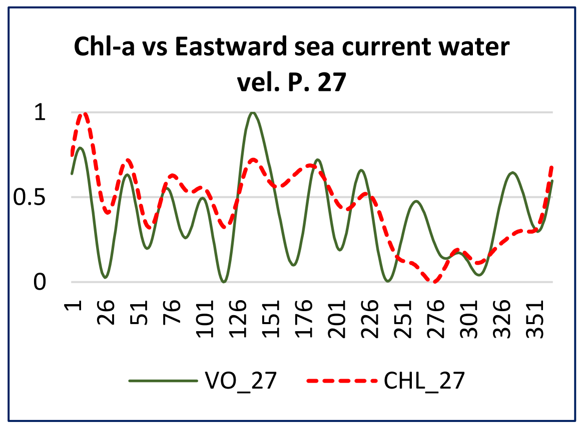


Figure 54. *Chl-a vs Vo (nFFT=15)*

Biogeochemical variables

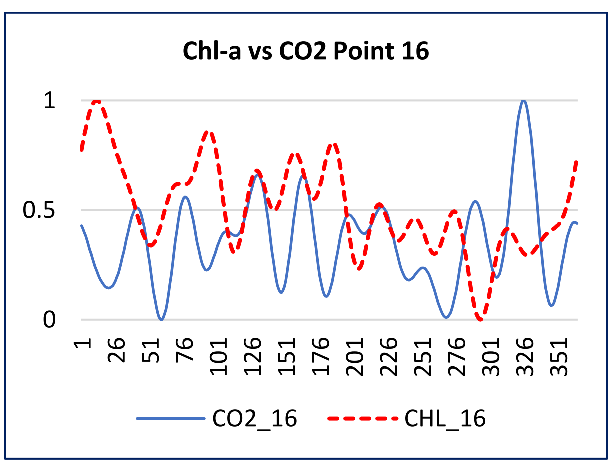


Figure 55. *Chl-a vs CO2 (nFFT=15)*

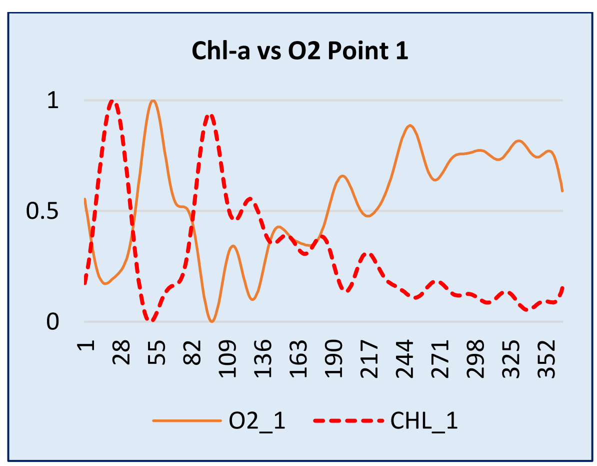


Figure 56. *Chl-a vs O2 (nFFT=15)*

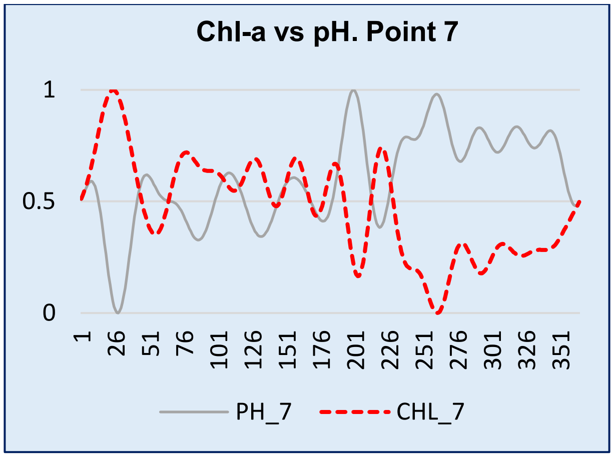


Figure 57. Chl-a vs pH (nFFT=15)

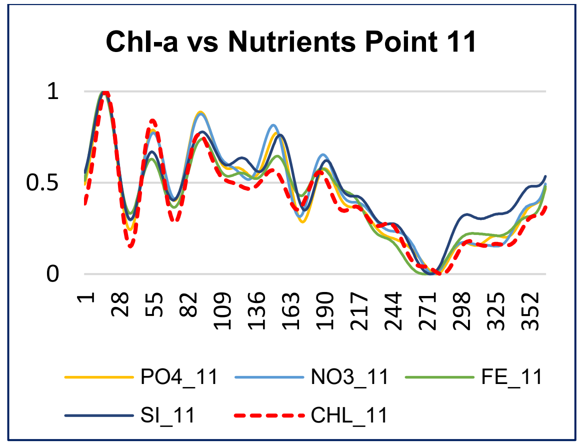


Figure 58. Chl-a vs Nutrients (nFFT=15)

Once the variables that are inversely proportional to chlorophyll are detected, they are inverted to determine the gap between them in order to reorganize the dataset for the application of the multivariate spatial-temporal regression method. The figures 59 to 68 show the results of the inversion of the variables that are inverse to chlorophyll, in the random points analyzed above.

Physical Variables

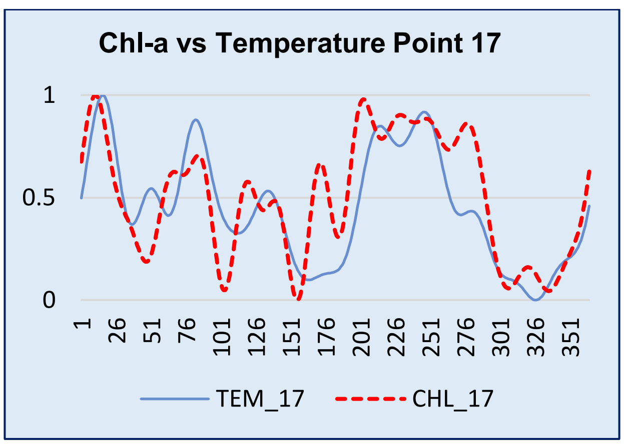


Figure 59. Chl-a vs Temperature (nFFT=15)

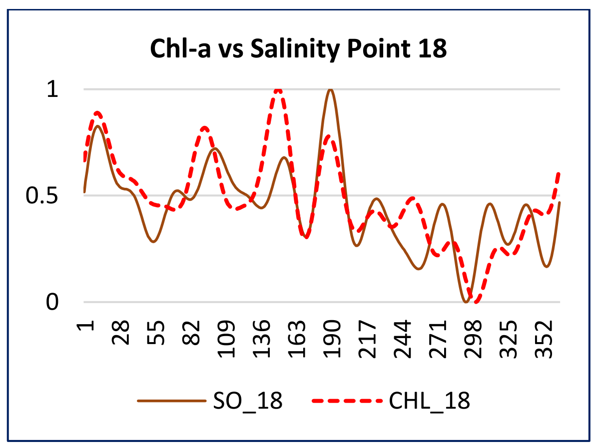


Figure 60. Chl-a vs Salinity (nFFT=15)

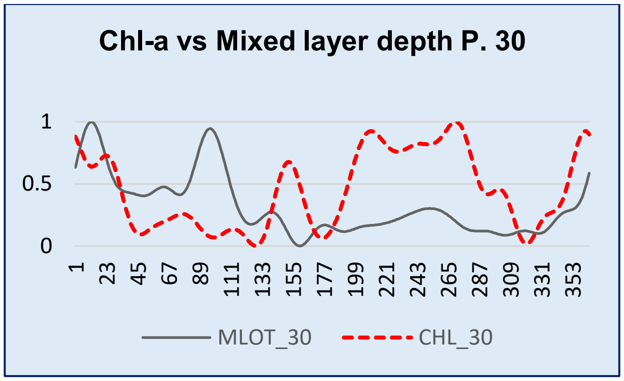


Figure 61. *Chl-a vs MLOT (nFFT=15)*

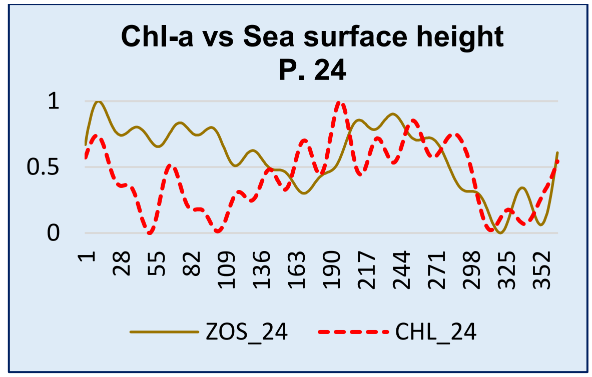


Figure 62. *Chl-a vs ZOS (nFFT=15)*

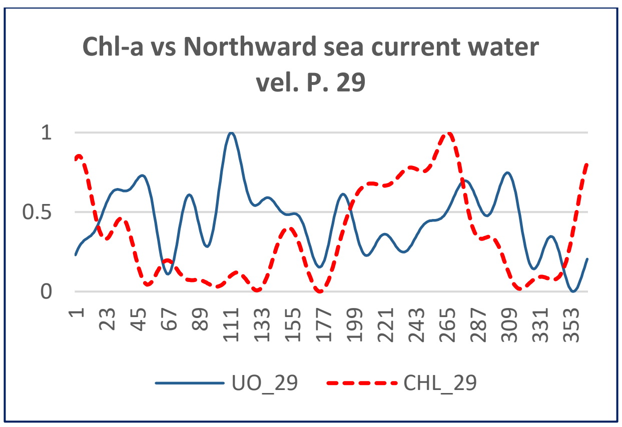


Figure 63. *Chl-a vs Uo (nFFT=15)*

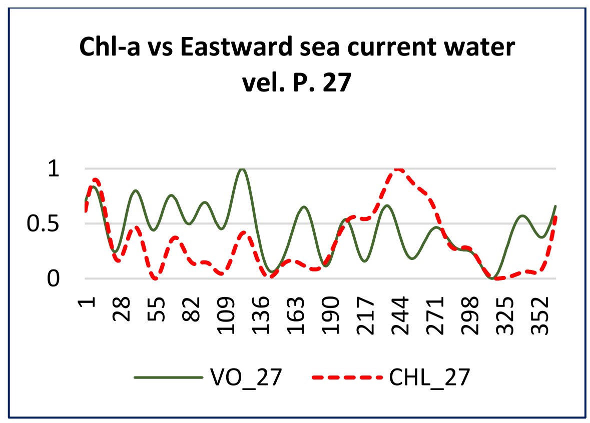


Figure 64. *Chl-a vs Vo (nFFT=15)*

Biogeochemical variables

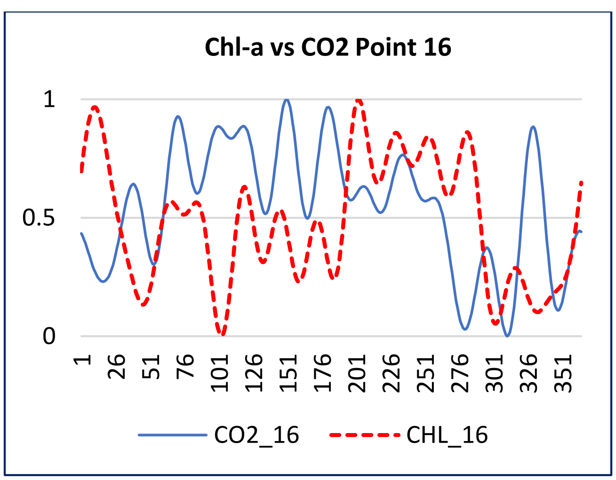


Figure 65. *Chl-a vs CO2 (nFFT=15)*

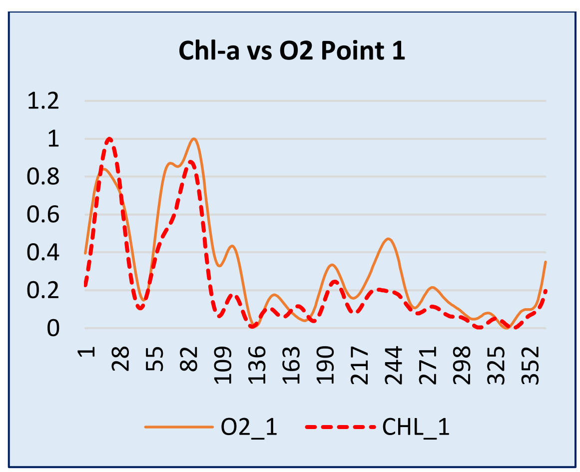


Figure 66. *Chl-a vs O2 (nFFT=15)*

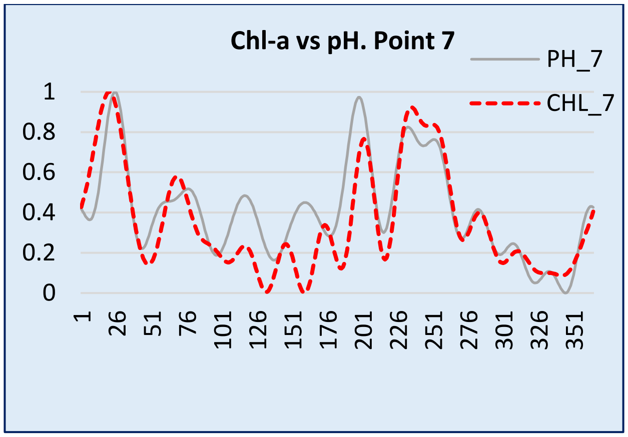


Figure 67. Chl-a vs pH (nFFT=15)

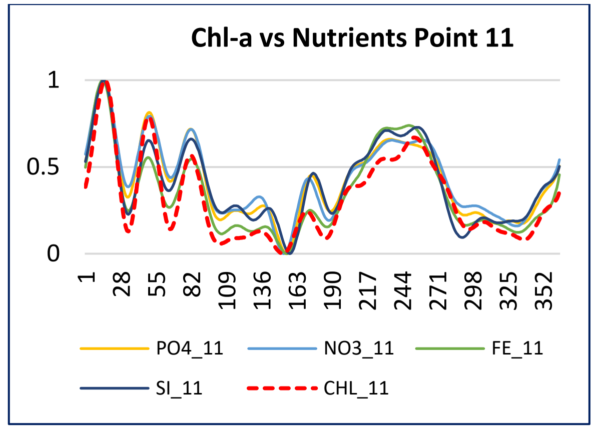


Figure 68. Chl-a vs Nutrients (nFFT=15)

In the figure 69, we can see the example of the gap between Chlorophyl-a and CO2 in a random point of the dataset. The calculation of these gaps of each variables in all geographycal point was made with algorithms developed in Python.

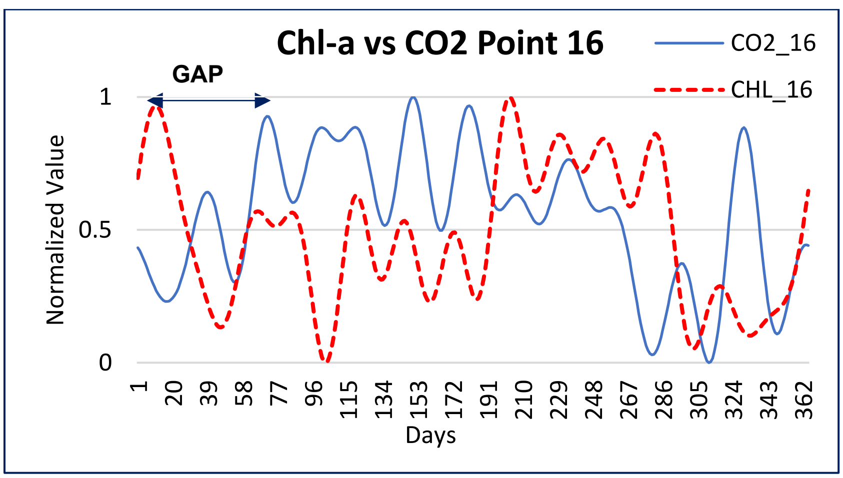


Figure 69 Gap example between Chl-a and CO2 (nFFT=15)

Once the phase lag is determined for each of the variables, the dataset is once again restructured, taking into consideration the variables that are ahead of phase with respect to Chl-a, given that these are the ones that influence its variation. For the variables that are delayed with respect to Chl-a, their last known value was kept. Also we tested without considering the delayed variables, but the performance of the model did not improve.

After the dataset restructuration, the prediction model was executed, and its result can be observed in figure 70.

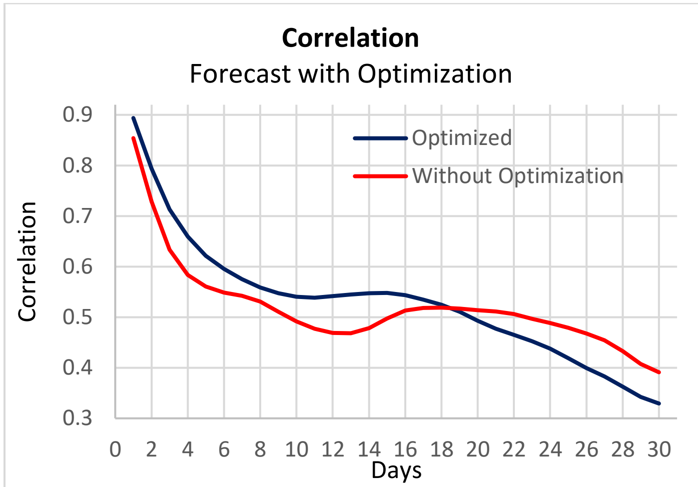


Figure 70. *Correlation of chlorophyll-a prediction with optimization in a month*

On this graph, when considering the phase lag of the variables with respect to Chl-a, the Chlorophyll prediction improved on average 6% up to the 18th day, after which the optimization model turns out to not be very efficient.

4.6.4 Chlorophyll-a prediction (30 days) with lags

The *multivariate spatial-temporal regression mode* was performed, including 10-days lag transformation to the data set Atraining, determining that the model becomes even more optimized, then obtaining an improvement of up to 10% in the Chl-a prediction on days 10 and 11; nonetheless, the effectiveness does not improve after the 18th day, remaining at 52%. From this day, this optimization model is not suitable. In figure 71, the model optimization is presented, considering the number of phase delays with better performance for each prediction day.

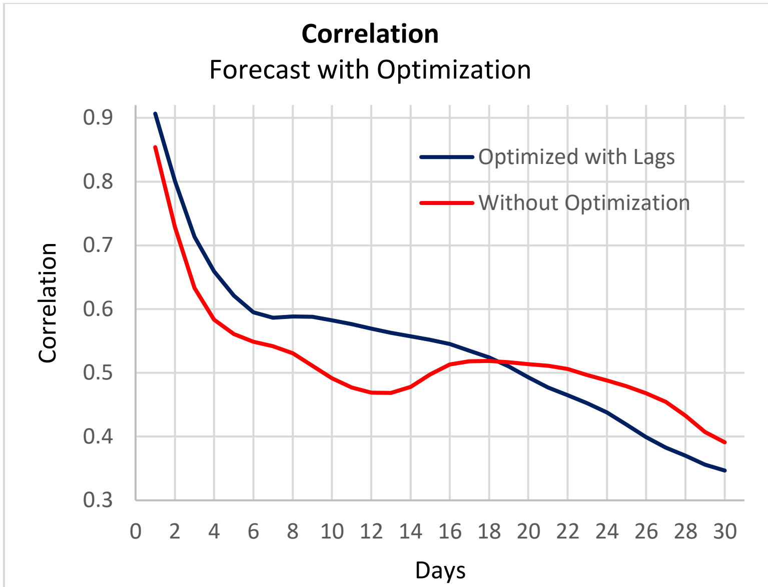


Figure 71. *Correlation of chlorophyll-a prediction with optimization including lags*

4.7 Conclusion

A model that optimizes the prediction of the concentration of Chlorophyll around the Galapagos Islands was designed, making use of the discrete Fourier transform theory that allows, on one hand, to understand the relationship between physical and biogeochemical variables of the ocean in relation to the Chl-a. On the other hand, it permitted to restructure the dataset, considering the displacement of the variables that are ahead of the Chlorophyll, in order to optimize the prediction using the multivariate spatial-temporal regression model. This prediction makes it possible to identify the most productive areas in the study zone, to suggest the naval forces to intensify maritime surveillance at these points. The best correlation is given in a time gap of 3 days (>0,7), while it remains above 0,5 up to 18 days. It is necessary to mention that the Ecuadorian Navy takes 0.6 days (14 hours) to send a Coast Guard from San Cristobal Island to the furthest point in Galapagos (200 nautical miles), and 3 days of navigation from Manta (Mainland Ecuador) to the furthest point (930 nautical miles).

In this experiment the bathymetry was considered as variable constant but in the future there will be the need to consider the integration between the climate variables (Meteo Marine variables) with Chlorophyll or other biogeochemical variables and the processes in the water masses caused by the presence of particular water depth and undersea features.

CHAPTER V

IDENTIFICATION OF POTENTIAL FISHING AREAS THROUGH DAILY ANALYSIS OF PHYSICAL AND BIOGEOCHEMICAL VARIABLES OF THE OCEAN

5.1 Abstract

Illegal and unregulated fishing is a grave issue that is altering the global marine ecosystem; Ecuador, given its geographical situation and the nutrients transported by the currents circulating through its waters, also suffers from over-fishing, especially around the Galapagos islands. For this reason, given the large marine territory under its possession, great national efforts are required to exercise control and sovereignty over its national waters. Since it does not depend solely on chlorophyll levels to predict the presence of fish, an analysis of fishing effort was carried out to identify the ranges of physical and biogeochemical variables in the places where fishing activity was recorded.

This research presents a methodology through which the Ecuador Navy can daily identify prime fishing areas, prioritizing its guarding, considering fishing events from previous years as well as physical, chemical and biological values of geographical areas where illegal fishing activities took place. To this end, we have used datasets belonging to the EU Copernicus Programme and the Global Fishing Watch organization. We have also made use of statistical tools such as kernel density estimation, to obtain the heat maps that allow to identify the gravity centres of the different potential fishing areas.

In this chapter firs, we analysed the fishing effort around the jurisdictional maritime areas of Ecuador, taking as a reference the fishing activities in the year 2020. Each of the geographic points where fishing activities took place were identified and recorded, and then the physical and biogeochemical conditions of the ocean where fishing took place were identified. This information was analysed on a daily basis, in order to determine the pattern of behaviour of these variables on a monthly basis, which serves as a reference to look for similar conditions depending on the day being analysed. Using the kernel density estimation (KDE) tool, it was estimated the zones in which fishing activity were concentrated. This analysis is performed by grouping the oceanic variables into physical, chemical, biological and nutrients.

5.2 Framework

As indicated by Witt and Godley (2007), in order to conserve marine ecosystems it is necessary to understand in a holistic manner the behaviour of fishing fleets, to have sustainable fisheries. This is a need to develop spatial models that show their distribution, for which satellite monitoring systems are the fundamental tool to determine their behaviour patterns (Witt & Godley, 2007).

As fish stocks are depleted, fisheries management faces control challenges, especially due to the increase in illegal, unregulated and unreported fishing. This situation is further exacerbated for countries with island territories, forcing them to have the necessary resources to monitor and guard their territorial sea (Cimino et al., 2019). Similarly, oceanographic conditions are a predominant predictor during seasonal variation in fishing effort, where productivity, migration pattern and fish distribution are influenced by environmental changes in temperature, chlorophyll-a concentration, currents, salinity and wind (Syah et al., 2020).

Currently, satellite remote sensing (SRS) is a tool used for environmental monitoring of the marine ecosystem and for fisheries management. Satellite images of oceanographic variables, such as chlorophyll concentration (Chl-a) and sea surface temperature (SST), are a source of information to explain the environment, presence and abundance of fish. Currently, the data provided by the SRS are fundamental to understand the habitats of marine species and their relationship with oceanographic conditions (Chassot et al., 2011).

For this reason, remote sensing methods can also play an important role in guiding fleets to optimal fishing grounds, which will lead to more efficient fishing effort, with a corresponding increase in economic benefits. However, this optimization will also affect the marine ecosystem, since it will cause an increase in pressure on marine resources, which will lead fishing authorities to also seek ways to regulate the exploitation of resources, and it is important for States to carry out fisheries management by analysing the relationship between the biotic and abiotic factors of the sea that govern the species distribution in the marine ecosystem (Santos, 2000). It should be mentioned that traditional methods that study the distribution of species do not have sufficient information compared to methods related to remote sensing, to efficiently manage fishery resources.

It is worth mentioning that Resolution 61/105 on 8 December 2006, adopted by the United Nations Convention General Assembly, about sustainable fisheries, highlights the importance of the long-term conservation, management and sustainable use of marine ecosystems and the obligations of States to cooperate to this end and to implement conservation and management measures related to overfishing, bycatch, pollution and protection of the marine environment. The resolution urges countries to record data on catches and fishing related activities related specially to straddling fish and highly migratory fish found inside and outside the exclusive economic zones of each country, as well as to commit to combating illegal, unreported and unregulated fishing and to adopt all necessary measures (United Nations, 2006) .

The kernel density estimation (KDE) is a non-parametric neighbour-based smoothing function that has been used in ecology to identify hot spots in vulnerable marine ecosystems (VME), i.e. areas of relatively high biomass/abundance. The KDE technique counts the number of samples in a given area and creates a field from these objects, whereas other techniques such as interpolation estimated data values based on other data and requires these to be continuous and normally distributed (Kenchington et al., 2014). KDEs are more efficient estimators than histograms and do not depend on interval width and grid origin and deliver more uniform results, allowing to identify characteristics such as skewness and outliers (Salgado Ugarte et al., 2000). Similarly, the satellite monitoring devices provides important information that allows determining the distribution of marine fishing activities to evaluate fishing pressure, which through the use of techniques such as KDE allows quickly mapping the hot spots where there is greater activity (Chen et al., 2022).

5.3 Data and study area

5.3.1 Data

The data selected for this experiment comes from 2 datasets; the first one is provided by the Global Fishing Watch organization (www.globalfishingwatch.org), which provides information on the transit of fishing vessels and location of fishing events. The records provided from the year 2020 have been used and its dataset is organized as follows in table 24:

Table 24: Description of the fishing vessel positions used in this experiment

Variable	Description	Unit
Date	Event date	YYYY-MM-DD
cell_ll_lat	Latitude of the lower left corner of the grid cell	decimal degrees
cell_ll_lon	Longitude of the lower left corner of the grid cell	decimal degrees
mmsi	Maritime Mobile Service Identity, the identifier for AIS	Real number
Vessel_hours	Hours that the vessel was present in this grid cell on this day	Hour
fishing_hours	Hours that the vessel was fishing in this grid cell on this day	Hour

Global Fishing Watch, has developed an algorithm based on changes in speed of a vessel and its movements to determine whether a vessel is fishing; In fact, its database records the transit and fishing positions of the vessels in each sector analyzed.

The second dataset comes from the database of the EU Copernicus Program. It provides the data of physical and biogeochemical oceanic variables for the year 2020 mentioned in the previous Chapter. In table 25, the variables used for this experiment are described.

Table 25: Description of the physical and the biogeochemical variables used in this experiment.

	Variable	**Description**	**Unit**
Biogeochemical variables	Chl-a	Total chlorophyll-a	mg/m3
	Fe	Dissolved iron	mmol/m3
	NO3	Nitrate	mmol/m3
	O2	Dissolved oxygen	mmol/m3
	pH	pH	
	PO4	Phosphate	mmol/m3
	Si	Dissolved silicate	mmol/m3
	SPCO2	Surface CO2	Pa
	Nppv	Net primary production	mmol/m3
	Phyc	Phytoplankton concentration in carbon	mmol/m3
Physical variables	ST	Sea water -40m temperature	ºC
	So	Salinity	1/e3
	Zos	Sea surface height	m
	Mlots	Mixed layer depth	m

During 2020, according to the database provided by Global Fishing Watch, 281492 fishing vessel positions were recorded in the study area, of which 24029 positions were categorized as fishing activities. To propose this methodology, we took into consideration only the fishing activity that took place during the month of June 2020, which corresponds to 6293 events.

Another additional consideration was taken, which is related to the level of chlorophyll, as this indicator determines the presence of biomass in the ocean. During the month of June 2020, 2489 fishing events were identified with a Chl-a level higher than 0.5 mg/m3; according to Filatov et al. (2007), levels lower than 0.5 mg/m^3 are considered low; for this reason, only fishing data above this value will be analyzed, to further guarantee the certainty of the prediction. It is worth mentioning that chlorophyll-a ([Chl-a]) is used as an indicator of phytoplankton biomass (Barocio et al., 2007), which is the most elementary level in the ocean food chain.

5.3.2 Study area

Our research area is located at continental shelf and around the islands of the exclusive economic zone of Ecuador, whose limits fall between 6ºN and 6ºS and between 75º W and 100ºW (Figure 72).

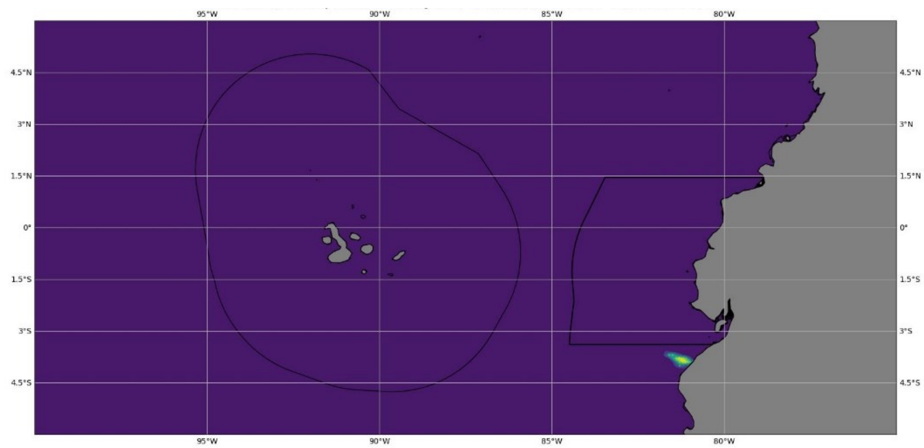


Figure 72. *Study area division*

5.4 Methodology and experiments.

5.4.1 Kernel density estimation

Heat maps allow to obtain a spatial interpretation of the data. These maps are intended to indicate the geographic clustering of a phenomenon, thus showing the event concentration at specific points and identifying possible trends. There are several probability density estimation methods to identify hot spots, being the kernel density estimation method, also known as the Parzen-Rosenblatt window method, as one of the most popular. This method has been used in other experiments for the visualization of biological and statistical analysis (Anderson, 2009).

Kernel density estimation, known as KDE, is a non-parametric technique to estimate the probability density function that, unlike the traditional method known as histogram, which shape is discontinuous and flat in each bin, KDE is smooth and represents a continuous variable (Zhang et al., 2018). The kernel estimates of f(x) and is defined as:

$$f(x) = \frac{1}{nh}\sum K\left(\frac{x - Xi}{h}\right) \tag{21}$$

Where h is the bandwidth, called smoothing parameter, and K is the kernel, a non-negative function, that must satisfy the following:

$$K(x) \geq 0 \;\; for\; all\; x \in R \tag{22}$$

$$\int_R K(x)dx = 1 \tag{23}$$

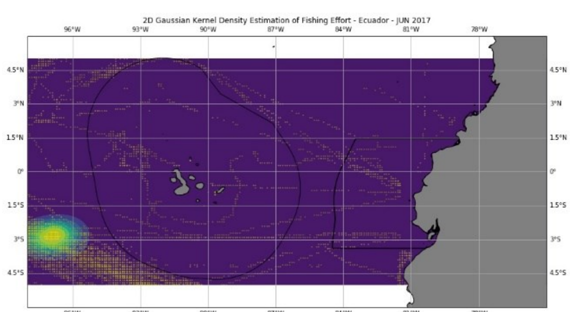
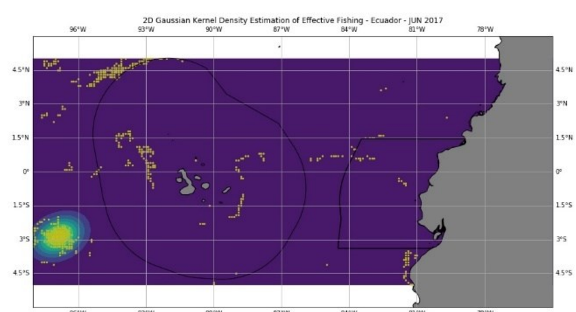
There are some kernel functions which are commonly used, such as Epanechnikov, biweight, triangular, Gaussian, rectangular and others. In this case, the Gaussian function has been used since its estimate is more symmetric and smoother. It is represented by the following expression:

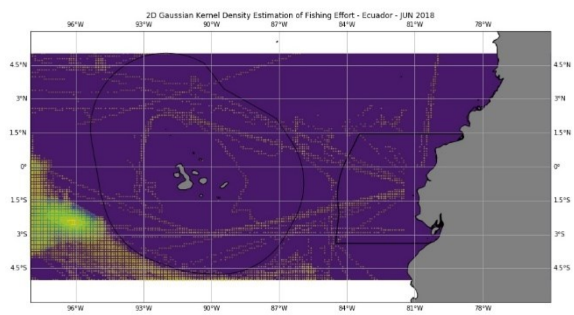
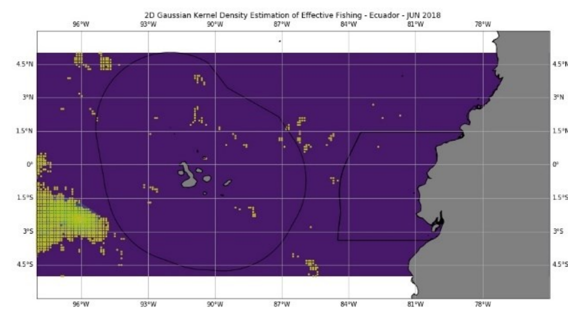
$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{24}$$

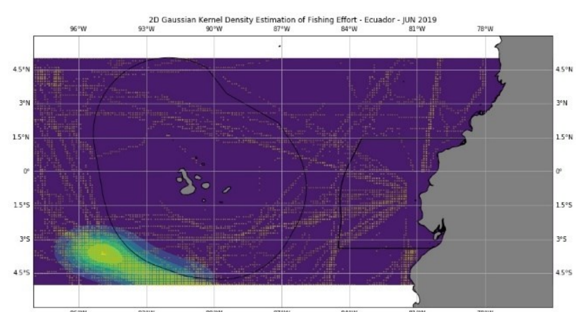
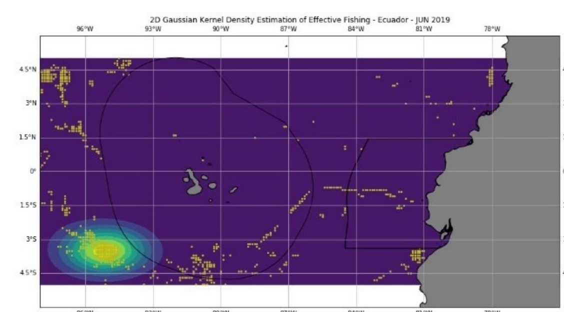
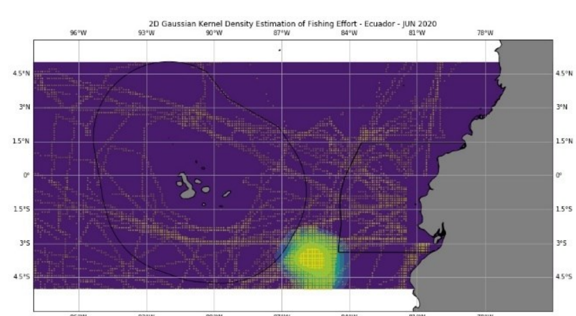
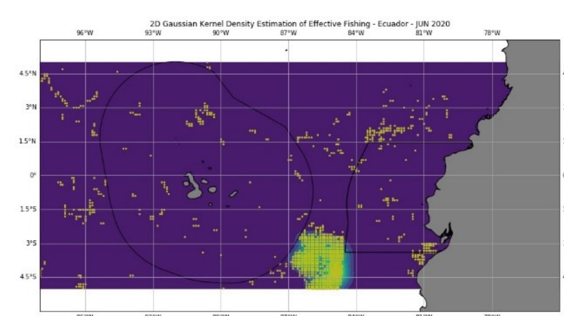
5.4.2 Experiments

Firstly, a spatial and temporal analysis of the distribution of fishing events from 2017 to 2020 was carried out, whereby it was determined that these events do not present the behaviour of a temporal pattern which is repeated in all months of every year, in the same sector. This indicates that the identification of potential fishing areas is related to the physical and biogeochemical conditions of the ocean, which must be in the appropriate ranges for the presence of fish. See figures 72 to 80.

In the graphs that follow, the orange dots represent vessels that have been detected in latitude and longitude, in fishing effort, or effective fishing activities. The kernel probability density is showed on the graph with a yellow colour. The place with the highest concentration of activity in the cluster will have the most intense colour.

Fishing Effort	**Effective Fishing**
2017	**2017**
 ***Figure 73.*** *Ecuador Fishing Effort - 2017*	 ***Figure 74.*** *Ecuador Effective Fishing – 2017*

2018	2018
 ***Figure 75.*** *Ecuador Fishing Effort - 2018*	 ***Figure 76.*** *Ecuador Effective Fishing - 2018*

Fishing Effort	Effective Fishing
2019	**2019**
 ***Figure 77.*** *Ecuador Fishing Effort - 2019*	 ***Figure 78.*** *Ecuador Effective Fishing - 2019*
2020	**2020**
 ***Figure 79.*** *Ecuador Fishing Effort - 2020*	 ***Figure 80.*** *Ecuador Effective Fishing - 2020*

In the figure 81 to figure 84, an example of the fishing effort during the month of June of the years indicated above is shown, by making use of heat maps using the KDE technique.

June 2017

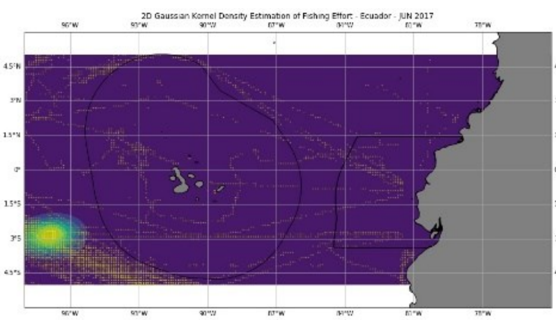


Figure 81. *Ecuador Fishing Effort – Jun-2017*

June 2018

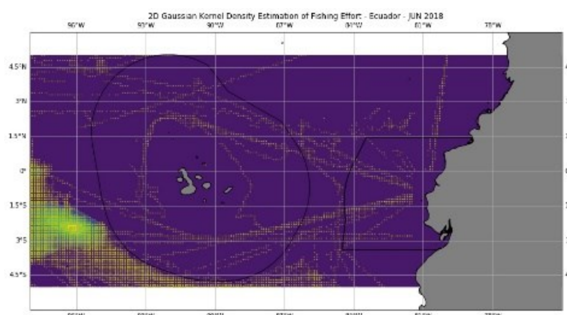


Figure 82. *Ecuador Fishing Effort – Jun-2018*

June 2019

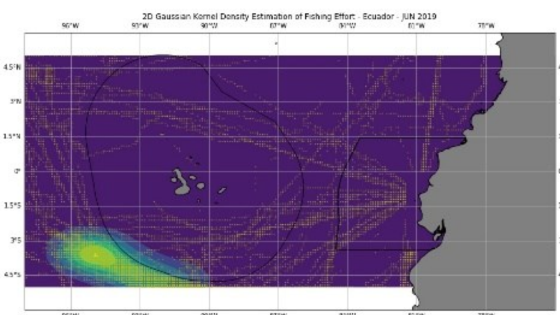


Figure 83. *Ecuador Fishing Effort – Jun-2019*

June 2020

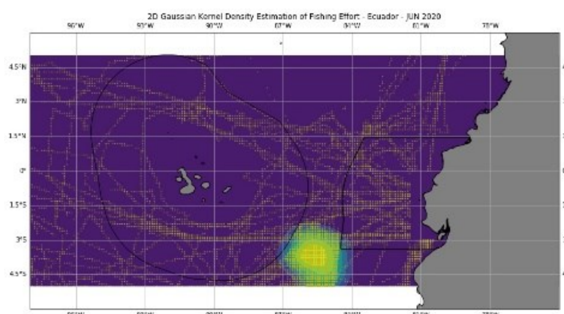


Figure 84. *Ecuador Fishing Effort – Jun-2020*

In this experiment, the effective fishing events of June 2020 will be analyzed, identifying the latitude and longitude of each one (Figure 85). Subsequently, the physical and biogeochemical conditions of the ocean at these points are recorded in a database in order to identify if these conditions are repeated on another day but in this case June 2022. We select June 9, 2022 as the specific day of the test.

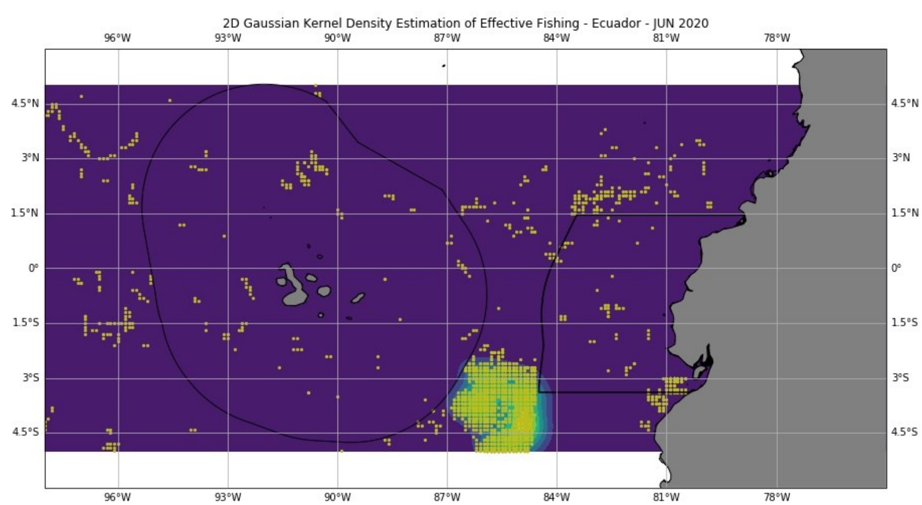


Figure 85. *Fishing events occurred during June – 2020*

This data is processed in Python and by making use of a probability density analysis tool, such as histograms. The range of every oceanographic variable where there was a greater number of fishing events was identified considering the restriction that chlorophyll must be in a range greater than 0.5 mg/m3. Figure 86 shows the process of extracting data from the database of the physical and biogeochemical variables of the ocean, which are related to the fishing events of the month under analysis.

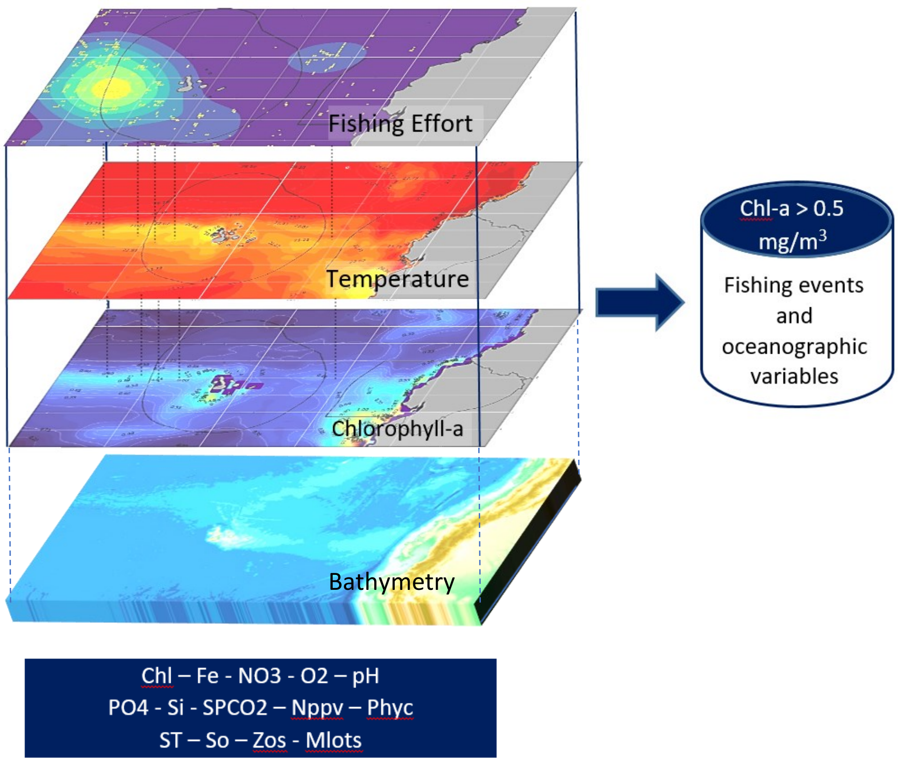


Figure 86. *Relation between Fishing Effort and Oceanographic Conditions*

Taking into consideration only the data analyzed for the year 2020 and using the algorithms developed in Python, we could determine the ranges of each variable by months, which influenced the existence of fishing activity. From table 26 to 29, the range of values of greater fishing activity by variables type and by are presented.

Table 26: Range of values of greater fishing activity for each month during 2020 Physics Variables

	PHYSICS VARIABLES							
	thetao	**thetao**	**so**	**so**	**zos**	**zos**	**mlotst**	**mlotst**
	low	**high**	**low**	**high**	**low**	**high**	**low**	**high**
ENE	22.03	23.98	33.84	34.45	0.19	0.23	10.53	11.79
FEB	21.61	23.46	33.93	34.54	0.19	0.24	10.53	10.95
MAR	19.21	21.96	34.74	35.15	0.16	0.22	10.53	10.91
ABR	24.57	26.18	34.72	34.84	0.17	0.19	10.53	11.94
MAY	19.26	21.41	34.76	35.03	0.15	0.18	10.53	11.80
JUN	19.89	21.80	34.93	35.29	0.14	0.16	10.53	13.17
JUL	20.55	22.83	34.65	35.20	0.13	0.17	10.53	14.75
AGO	20.49	22.74	34.69	35.15	0.10	0.13	10.53	14.82
SEP	17.44	19.90	34.73	35.18	0.10	0.15	22.05	27.81
OCT	19.40	21.68	33.76	34.18	0.15	0.18	10.53	12.93
NOV	17.85	19.62	34.31	34.67	0.10	0.13	10.53	13.39
DIC	16.22	18.34	34.88	35.29	0.11	0.14	10.53	12.30
PROM.	**19.87**	**21.99**	**34.49**	**34.91**	**0.14**	**0.18**	**11.49**	**13.88**

Table 27: Range of values of greater fishing activity for each month during 2020 Biological Variables

	BIOLOGICAL VARIABLES					
	Chl-a	**Chl-a**	**phyc**	**phyc**	**nppv**	**nppv**
	low	**high**	**low**	**high**	**low**	**high**
ENE	0.50	1.17	1.82	4.07	11.78	28.97
FEB	0.50	1.22	1.52	4.03	13.19	30.36
MAR	0.50	0.91	1.19	2.81	12.37	20.50
ABR	0.50	0.65	2.00	2.52	16.14	20.22
MAY	0.50	1.28	1.72	3.72	7.74	16.54
JUN	0.50	0.98	3.09	4.35	14.73	27.71
JUL	0.50	0.97	3.07	4.73	16.56	27.98
AGO	0.50	1.11	3.21	5.03	20.09	35.04
SEP	0.50	1.01	3.00	4.24	23.81	31.88
OCT	0.50	1.13	3.05	4.68	19.11	30.06
NOV	0.50	0.91	2.81	4.21	15.97	24.84
DIC	0.50	1.10	1.45	3.04	7.38	18.52
PROM.	**0.50**	**1.04**	**2.33**	**3.95**	**14.91**	**26.05**

Table 28: Range of values of greater fishing activity for each month during 2020 Chemical Variables

	CHEMICAL					
	ph	ph	o2	o2	spco2	spco2
	low	high	low	high	low	high
ENE	8.01	8.05	200.54	214.41	37.21	40.80
FEB	7.96	8.00	187.48	202.93	38.52	41.98
MAR	7.91	7.96	161.06	191.23	50.63	55.24
ABR	7.88	7.91	183.61	204.63	56.21	58.34
MAY	7.83	7.87	117.20	141.38	38.63	43.00
JUN	7.90	7.95	189.64	216.31	52.72	57.58
JUL	7.88	7.94	194.69	224.61	56.64	61.81
AGO	7.86	7.91	193.50	221.81	52.11	58.02
SEP	7.88	7.92	202.86	229.53	57.40	62.89
OCT	7.88	7.92	126.24	151.56	36.02	40.35
NOV	7.92	7.97	147.88	174.03	45.55	48.91
DIC	7.90	7.95	142.21	169.84	35.29	38.65
PROM.	**7.90**	**7.95**	**170.58**	**195.19**	**46.41**	**50.63**

Table 29: Range of values of greater fishing activity for each month during 2020 Nutrients Variables

	NUTRIENTS							
	fe	fe	no3	no3	po4	po4	si	si
	low	high	low	high	low	high	low	high
ENE	1.05E-04	4.24E-04	3.07	5.73	0.95	1.16	1.17	2.92
FEB	9.77E-05	5.10E-04	5.49	8.21	0.84	1.05	2.86	4.98
MAR	9.33E-05	4.00E-04	9.49	14.16	0.88	1.20	7.09	10.21
ABR	9.42E-05	1.73E-04	6.02	8.92	0.94	1.08	5.27	7.77
MAY	4.96E-04	8.77E-04	15.61	18.98	1.47	1.71	9.93	12.57
JUN	7.99E-05	4.22E-04	10.13	13.97	1.07	1.33	4.50	7.48
JUL	6.85E-05	4.37E-04	9.59	14.19	1.20	1.52	4.74	8.09
AGO	7.84E-05	4.42E-04	8.68	12.91	1.16	1.45	4.41	7.21
SEP	5.14E-05	4.41E-04	8.16	11.94	1.11	1.38	4.01	6.99
OCT	7.05E-05	4.48E-04	11.14	14.71	1.06	1.32	8.75	11.32
NOV	6.76E-05	4.65E-04	11.79	15.52	1.15	1.41	9.05	11.68
DIC	7.53E-05	3.86E-04	9.71	14.11	1.26	1.55	8.20	11.62
PROM.	**1.15E-04**	**4.52E-04**	**9.07**	**12.78**	**1.09**	**1.35**	**5.83**	**8.57**

Since the day under analysis is on June, the range of each physical and biogeochemical variable was determined during that month in the year 2020, in the places where there was greater fishing activity and where the chlorophyll was greater than 0.5 mg/m3. Results are shown in table 30.

Table 30: Range of values of greater fishing activity – June 2020

PHYSICS VARIABLES							
ST ºC		**So 1/e3**		**Zos m**		**Mlotst m**	
low	**high**	**low**	**high**	**low**	**high**	**low**	**high**
19.89	21.80	34.93	35.29	0.14	0.16	10.53	13.17
NUTRIENTS							
Fe mmol/m3		**NO3 mmol/m3**		**PO4 mmol/m3**		**Si mmol/m3**	
low	**high**	**low**	**high**	**low**	**high**	**low**	**high**
7.99E-05	4.22E-04	10.13	13.97	1.07	1.33	4.50	7.48
BIOLOGICAL VARIABLES							
Chl-a mg/m3		**phyc mmol/m3**		**nppv mmol/m3**			
low	**high**	**low**	**high**	**low**	**high**		
0.50	0.98	3.09	4.35	14.73	27.71		
CHEMICAL							
Ph		**O2 mmol/m3**		**SPCO2 Pa**			
low	**high**	**low**	**high**	**low**	**high**		
7.83	7.87	117.20	141.38	38.63	43.00		

Once the range of values of each oceanographic variable was determined, the physical and biogeochemical conditions of the ocean in the study area on the analysed day (June 9, 2022) were downloaded from the Copernicus database, and by making use of a software developed in Python, a match of values for each of the ranges established for the variables considered in this study is searched.

In order to increase the certainty in the prediction and facilitate the analysis, the variables were grouped depending on their type, that is, physical, chemical, biological and nutrient. In the geographical location where there was a greater match with the values established for the range of each variable, we considered that this location met the oceanographic conditions for repeated fishing events.

Finally, to identify a possible cluster or hot spot where a fishing event can be repeated, the kernel density function was used, and thus the area where the Navy should intensify patrolling to contribute to the protection of the marine ecosystem was determined. These charts will also be useful for planning the approach route to the hot spot.

5.5 Discussion of results.

With the methodology used, it was developed a tool that helps us to understand the physical and biogeochemical conditions that fishermen seek to carry out fishing tasks. Once the ranges in which each of the variables are known, it would be possible to search for similar conditions that are repeated on the chosen day.

The analysis was carried out by type of variables, mentioned above. Algorithms developed in Python were applied to search on the selected day and determine if those conditions were repeated. Two types of graphs were obtained: the first one, corresponding to the graphic representation of the match of the ranges of values in the sea and the second one, which was linked to the hot spot that identified each group of variables.

The analysis was performed for June 9, 2022, and for physical variables, it was possible to identify that in the northern region of the Galapagos Islands and Continental Ecuador, no conditions that were mainly related to temperature, salinity, sea surface height and mixed layer depth were presented, so the analysis had to be carried out in the south-west area of Galapagos, as well as in the area between the continental economic exclusive zone and insular. In figure 87, the main areas where similar conditions existed as in 2020 are shown.

Likewise, this figure helps the Navy to determine the routes to follow in order to reach the hot spot, to further increase the probability of detecting illegal vessels. This approach must be carried out through a route that crosses the areas with the highest intensity on the map and taking into account the range of their radars.

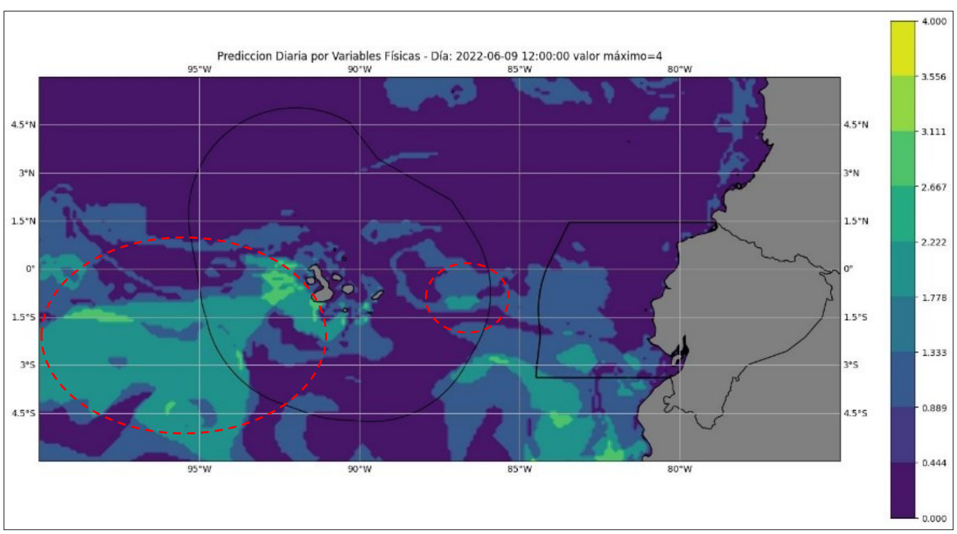


Figure 87. *Prediction of Fishing Areas – Type: Physical variables - June 9th, 2022*

The hot spot was identified using the kernel distribution density technique, the same technique applied to the data analyzed in order to determine the maps which identify the areas where similar fishing conditions existed. In this case, the area that should be given the highest attention was within the exclusive economic zone, to the south-west of the Galapagos Islands. Likewise, outside the EEZ, favourable conditions were also presented, which could attract the international fishing fleet. Figure 88 shows graphically what was described above.

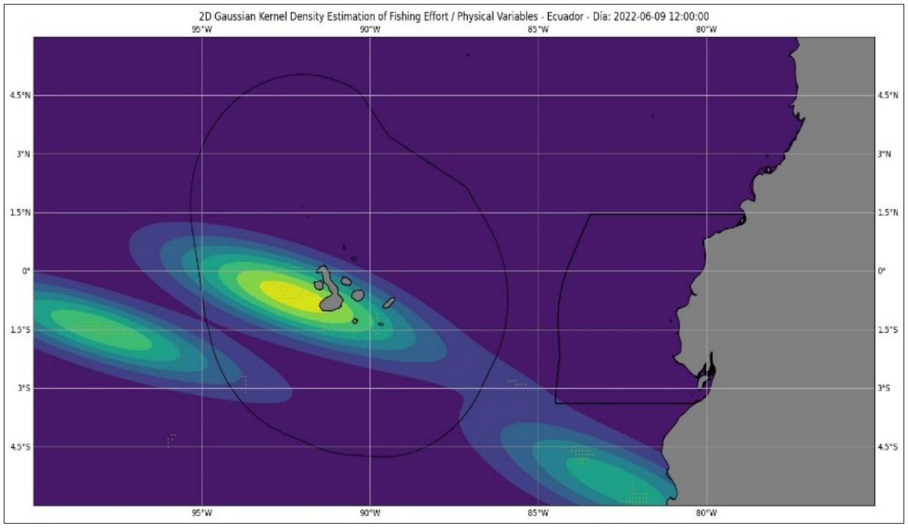


Figure 88. *Fishing hot spot – Type: Physical variables - June 9th, 2022*

A similar analysis was carried out for the other groups of variables, in the case of biological variables (Chlorophyll-a, net primary production and phytoplankton). In figure 89, we can observe that there are also favourable conditions in the south-west of the Galapagos Islands and in the south-west of Continental Ecuador, extending to the north of Perú.

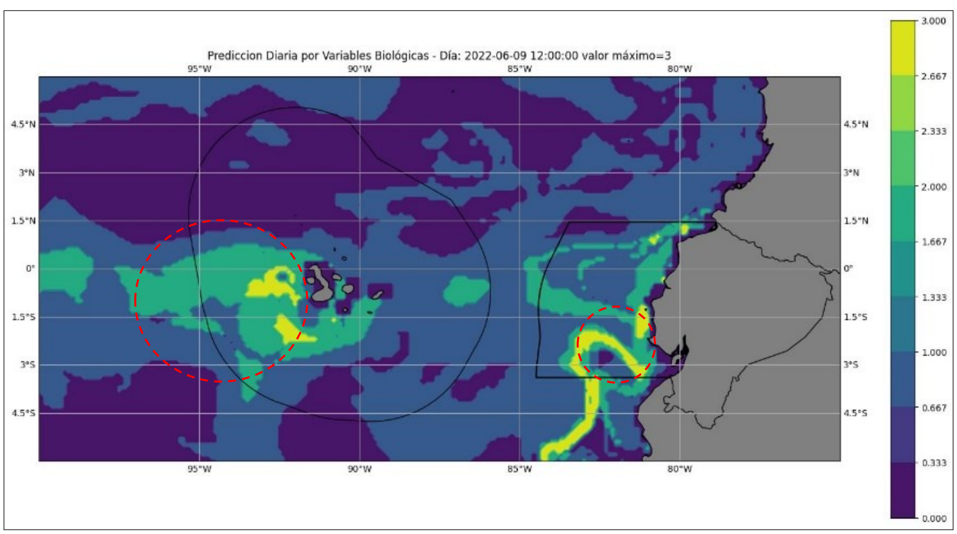


Figure 89. *Prediction of Fishing Areas – Type: Biological variables - June 9th, 2022*

The representation of the hot spot of fishing location considering the biological variables in the insular region, is very close to the location determined with the analysis of the physical variables; however, there would also be another potential area inside the continental EEZ. Figure 90 shows what was detailed above.

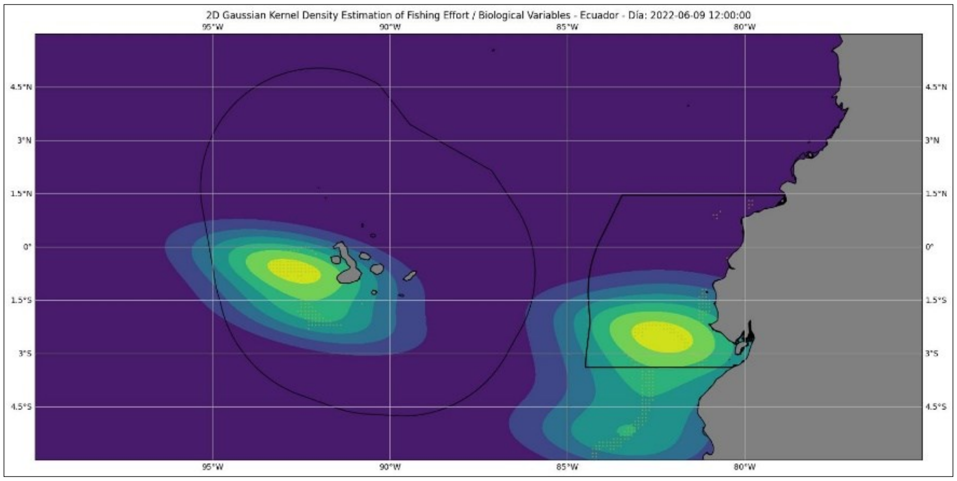


Figure 90. *Fishing hot spot – Type: Biological variables - June 9th, 2022*

From the analysis of the chemical parameters, it is evident in figure 91 that, in the northern area of Continental and Insular Ecuador, there are no conditions similar to those occurred in June 2020, while in the south-west sector, both inside and outside the exclusive economic zone, as well as between both exclusive zones, appropriate chemical conditions would also exist.

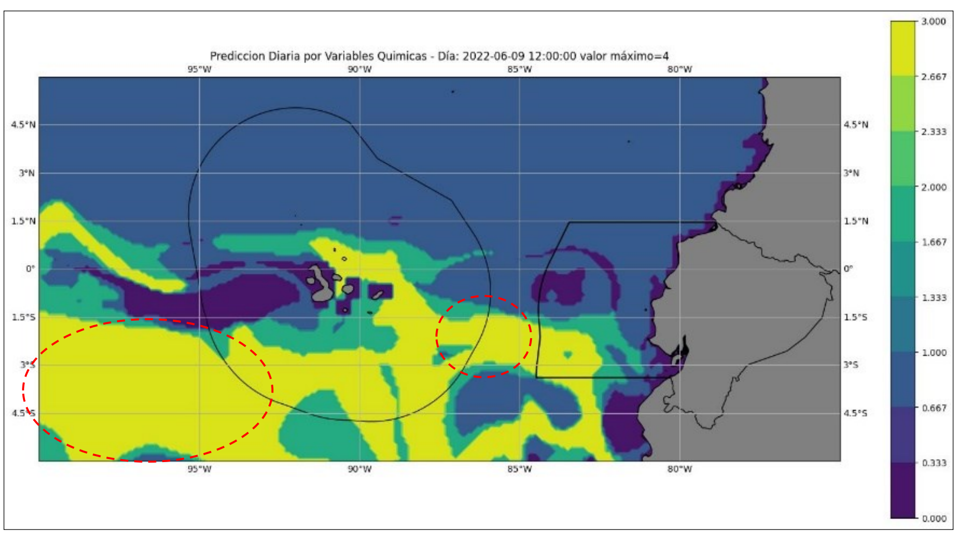


Figure 91. *Prediction of Fishing Areas – Type: Biological variables - June 9th, 2022*

When the hot spot is established, in reference to the chemical variables, it is again identified in figure 92 that similar conditions are found in the south west of the Galapagos Islands;

however, it can also be observed that between the exclusive territorial and insular zones, there is a hot spot with less intensity linked to this type of variable.

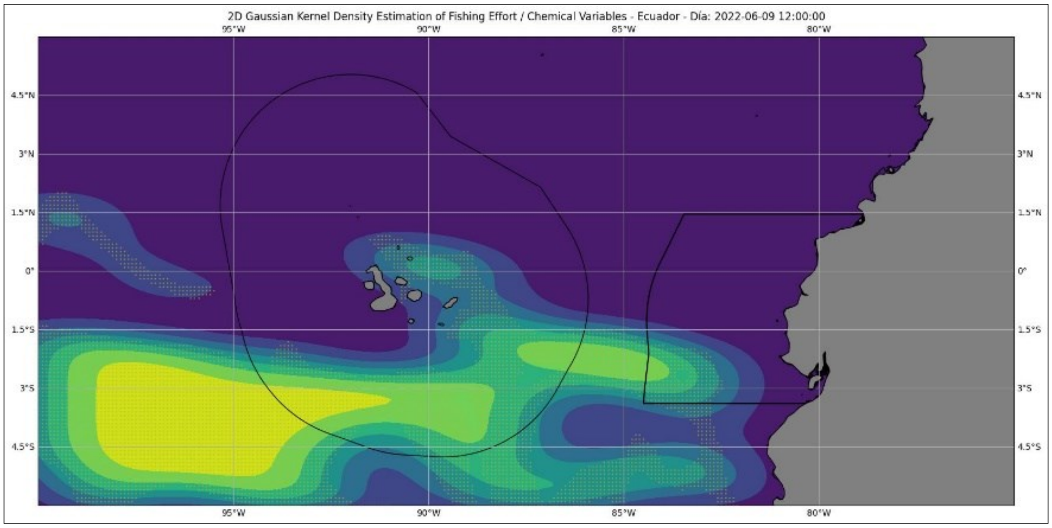


Figure 92. *Fishing hot spot – Type: Biological variables - June 9th, 2022*

Finally, with respect to the analysis of nutrients, it can be indicated that they are found in a more concentrated manner in the middle of the continental and insular economic zones, which corresponds to the high seas, and in the north of Perú. Figure 93 shows the above and other areas that are important but less intense, which are located to the west of the Galapagos Islands.

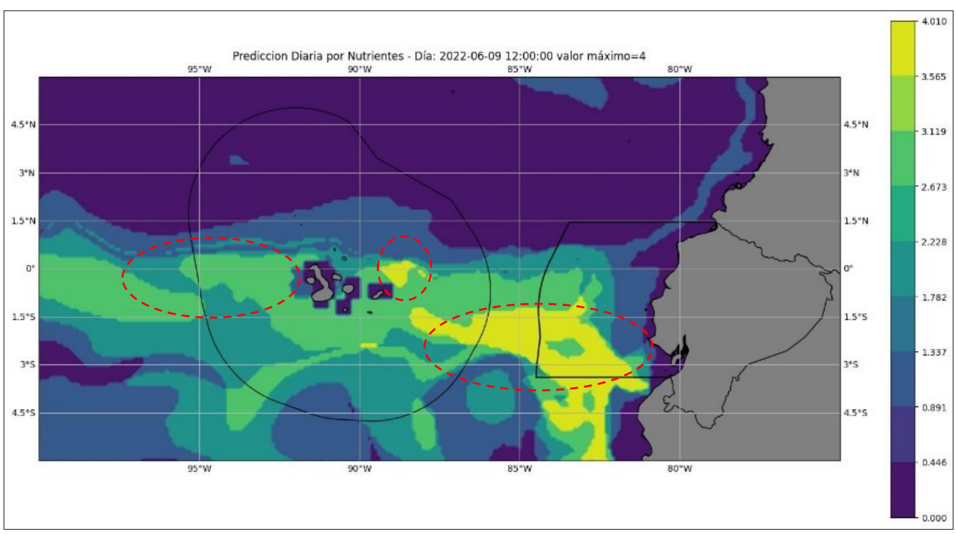


Figure 93. *Prediction of Fishing Areas – Type: Nutrient variables - June 9th, 2022*

In the case of the nutrient analysis, the hot spot varies its position with respect to what was identified when the physical and biological variables were analyzed; however, it has a certain relationship with the analysis carried out with the chemical variables, since there is a concentration in the middle of the exclusive economic zones. Figure 94 shows the above analysis.

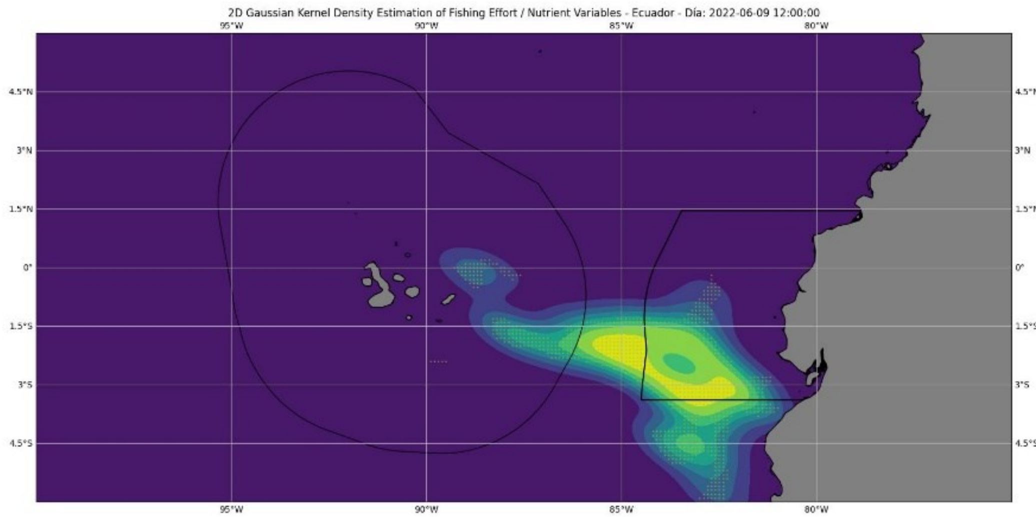


Figure 94. Fishing hot spot – Type: Biological variables - June 9th, 2022

Finally, figure 95 consolidates the results obtained with the partial analyses of the physical, chemical, biological and nutrient variables. Meanwhile figure 96 shows the areas that present better conditions for fishing considering partial results of the variables mentioned before. For these reasons, these areas mainly should be fully protected by the Coast Guard.

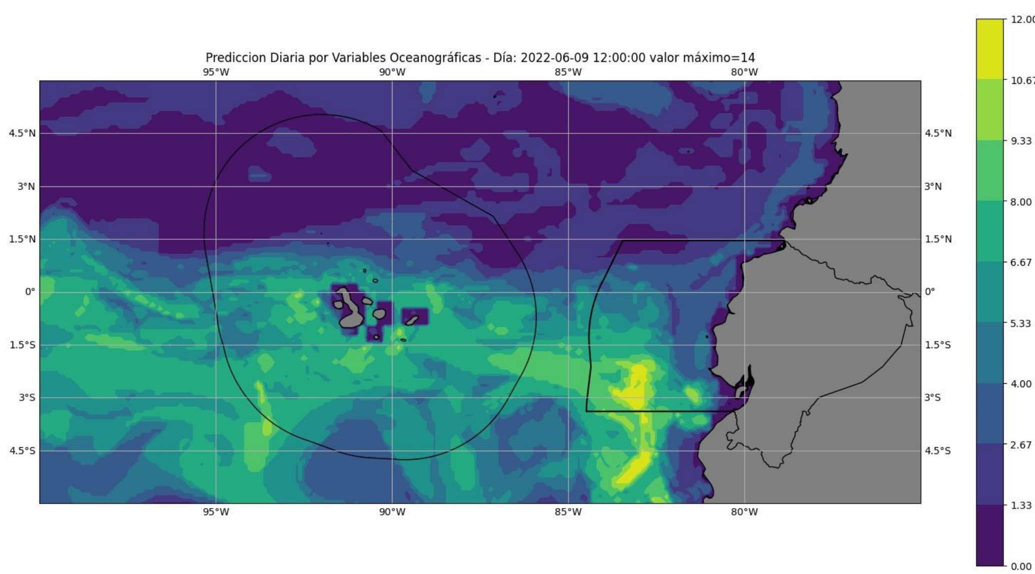


Figure 95. Prediction of Fishing Areas – Type: All oceanographic variables - June 9th, 2022

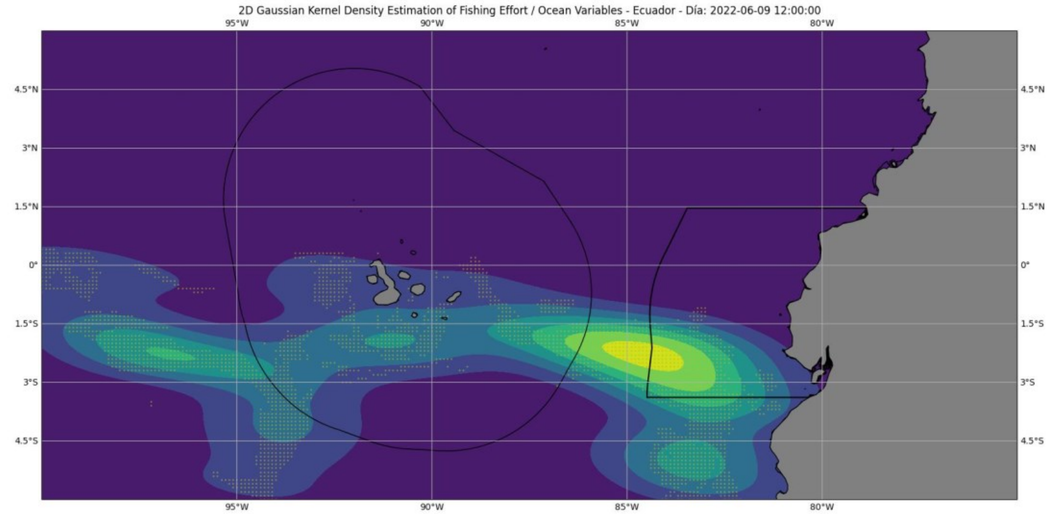


Figure 96. *Fishing hot spot – Type: All oceanographic variables - June 9th, 2022*

Finally, in figure 97, taken from the Global Fishing Watch platform, shows the fishing effort in the study area on June 9, 2022, showing that there was indeed a presence of vessels in the southwest of the Galapagos Islands, as determined by the analysis of physical, chemical and biological variables. Similarly, there was a presence of fishing vessels in the exclusive economic zones, insular and continental, which also coincides with the analysis of biological and nutrient variables.

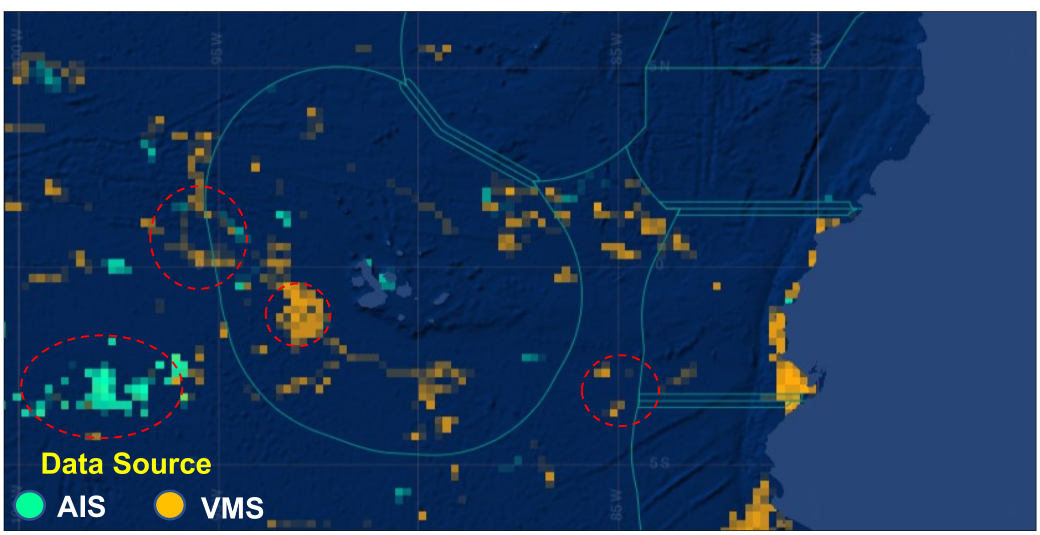


Figure 97. *Fishing activity on June 9th, 2022 – Source: Global Fishing Watch*

A limitation of this study, which will be the subject of further research, is to identify the range of physical and biogeochemical variables in relation to the type of fishing, i.e. the reference values for tuna, squid, hake, etc., so that the fishing authorities can prioritize the resources to be protected, taking into consideration the limitations of naval units available for the protection of jurisdictional waters.

It is worth mentioning that naval planners, to select patrol areas, must analyze the diagrams generated for both chlorophyll prediction and identification of potential fishing areas, depending on the type of variables, these tools work together. It must be noted that the daily identification diagrams also provide the routes that vessels must follow to approach the identified fishing hotspot. These tools should be used in conjunction with the other systems available to the Navy for monitoring the fishing fleet.

Likewise, it should be mentioned that the country's Fisheries Authority is currently implementing the digital registry of fishing and fish landing data, so this digitized information on the country's actual fishing is not available to further refine the model. Once Ecuador completes the implementation of this computer tool, the proposed model will need to be tuned up. Ecuador's fishing data are recorded manually and once the vessel arrives in port, so there is no real time data or certainty of its registration.

5.6 Conclusions

A daily prediction model of potential fishing zones in the continental and insular economic zone of Ecuador has been designed to provide the naval force with a tool to identify the areas that must be protected in order to combat illegal, unreported and unregulated fishing. A methodology is proposed to determine the range in which the physical and biogeochemical variables of the ocean should be to optimize the fish presence, taking into consideration past fishing events and oceanographic conditions, which must be identified on the day the prediction has to be made. This methodology also helps to determine the routes to be taken by coast guard and aeromaritime exploration aircraft to approach potential fishing areas, to further increase the probability of detecting illegal vessels. Using statistical tools such as kernel density estimation probability, maps are produced to help determine the hot spots where surveillance and control efforts should be focused.

As future work, it is required to examine the behaviour of the fishing fleet in other years, in order to determine a range of values of the oceanic variables that allow generalizing the daily prediction of the potential fishing zones.

CHAPTER VI

A PROPOSAL FOR IMPLEMENTING RESEARCH RESULTS

6.1. Abstract

The results obtained in the present research were the product of the development of several computer programs developed in Python, that it will have in the future, the name of PrediccionPesca_EC model. Around 24 scripts were designed to carry out the process of data acquisition, processing, analysis, prediction, and presentation of results. These scripts have made it possible to obtain the results presented mainly in chapter IV, V and VI of this thesis; however, it was considered necessary to group the results obtained in a web page that can be permanently accessed by Navy planners, in order to provide useful information in support to planning operations to combat illegal, unregulated and underregulated fishing. This web page is hosted on a private virtual server, as well as the scripts that are executed on a daily basis to keep the information on the page up to date.

6.2. Implementation Process

6.2.1. Virtual Server

For the implementation of the web site PrediccionPesca_EC, the choice was a virtual private server from the company InterServer, where the hostname of the web site that will host this server was configured. In figure 98 the information of virtual server is shown.

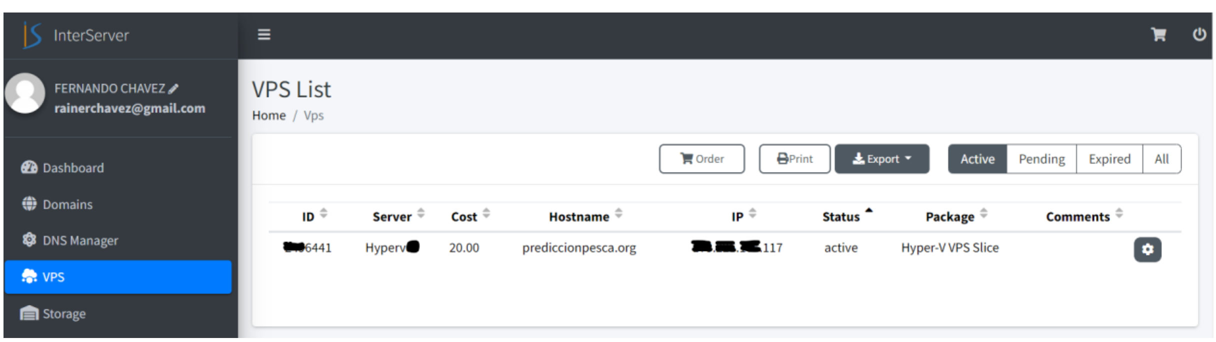


Figure 98. *Information of virtual server to develop PredicciónPesca_EC model*

The Remote Desktop Connection tool is used to log into the server and perform the necessary configurations. In figure 99 the access to remote desktop is shown.

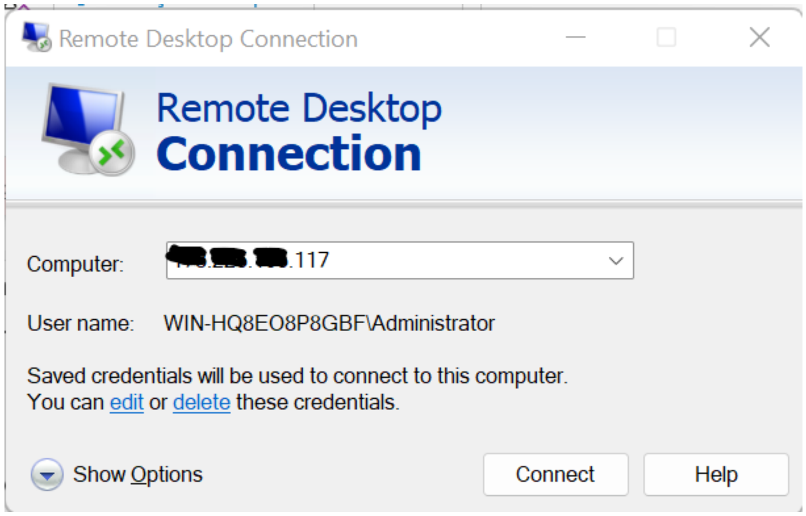


Figure 99. Remote desktop connection to virtual server for *PrediccionPesca_EC model*

When you log into the server, you can see that the virtual machine is configured with an AMD EPYC-7402P 24-Core processor, 4 GB Ram and 59.46 GB HDD.

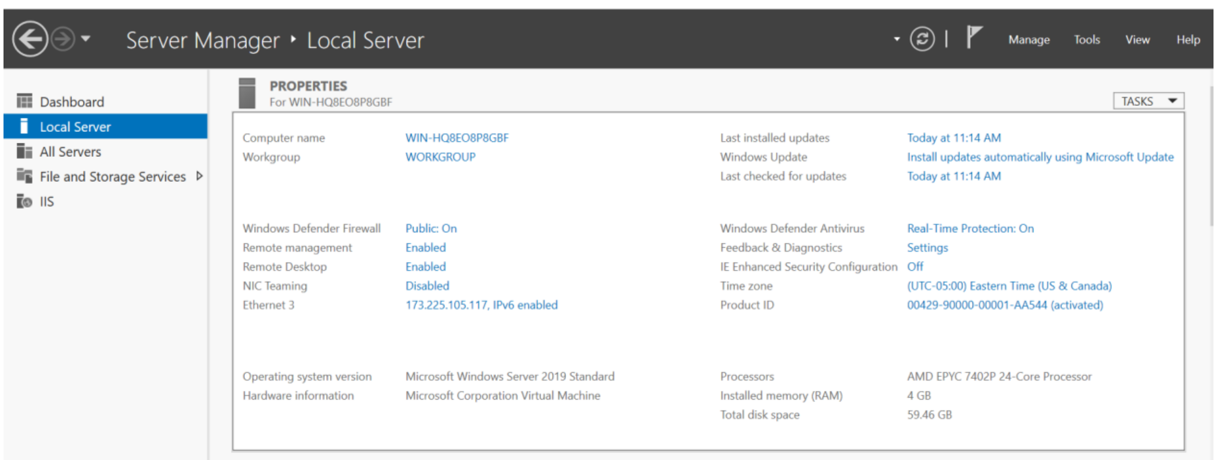


Figure 100. *Characteristic of virtual server for PrediccionPesca_EC model*

For the execution of the Python scripts, it was necessary to install several tools that help the execution of the scripts, such as: Anaconda and Visual Studio Code. The scripts that were developed, hosted in the Server folder in the virtual machine (Figure 101), are the following:

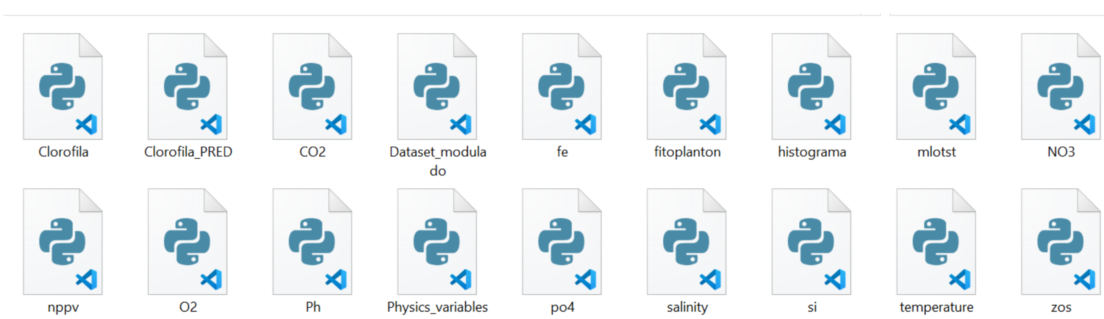


Figure 101. *Scripts hosted on virtual server for PrediccionPesca_EC model*

The connection with the Copernicus program database was essential to obtain the input oceanographic and biogeochemical data to reach the final products.. The pieces of code used for this purpose were as follows:

```
import getpass
import motuclient
class MotuOptions:
    def __init__(self, attrs: dict):
        super(MotuOptions, self).__setattr__("attrs", attrs)
    def __setattr__(self, k, v):
        self.attrs[k] = v
    def __getattr__(self, k):
        try:
            return self.attrs[k]
        except KeyError:
            return None
data_request_options_dict_manual = {
    "service_id": "GLOBAL_ANALYSIS_FORECAST_BIO_001_028-TDS",
    "product_id": "global-analysis-forecast-bio-001-028-daily",
    "date_min": today,
    "date_max": today,
    "longitude_min": -100.,
    "longitude_max": -75.,
    "latitude_min": -6.,
    "latitude_max": 6.,
    "depth_min": 21.5987,
    "depth_max": 21.599,
    "variable": ["chl","fe","no3","nppv","o2","ph","phyc","po4", "si","spco2"],
    "motu": "https://nrt.cmems-du.eu/motu-web/Motu",
    "out_dir": "C:\SERVER",
    "out_name": "quimica_pred.nc",
    "auth_mode": "cas",
    "user": "xxxxxxx",
    "pwd": "xxxxxxxx”
print(data_request_options_dict_manual)
motuclient.motu_api.execute_request(MotuOptions(data_request_options_dict_manual)
)
```

Figure 102. *Code to connect PrediccíónPesca_EC model to Copernicus project database*

Subsequently, it was necessary to process the information, both for the visualization of the oceanographic variables, as well as for the prediction procedures of the fishing zones, both by means of the procedure of Fourier-based optimization for multivariate spatial-temporal regression in chlorophyll-a presence prediction, as well as by the related method daily analysis of physical and biogeochemical variables of the ocean integrated with bottom morphology, oceanographic and geological setting. It was necessary to process the downloaded data, as well as the application of machine learning algorithms and statistical analysis in the second case. Since the data from Copernicus program is updated daily, the scripts are also executed every day to update the information on the web page.

6.2.2. Web page Description

The results provided by the scripts that were developed can be viewed on the page hosted on the virtual server, under the domain www.prediccionpesca.org, which was developed in HTML. It presents the graphs of the predictions made, the information related to the historical fishing effort around the Galapagos Islands, an the daily oceanographic information on physical and biogeochemical variables. A detailed description of the pages composing the web site follows.

6.2.2.1. Home Page

This page shows the options this web site has. Shortcuts are provided to consult fishing effort information, biomass prediction (Chlorophyll prediction), daily prediction of fishing areas and the oceanographic conditions. The figure 103 and figure 104 show the access to the web page.

Figure 103. *Home page for PredicciónPesca_EC model*

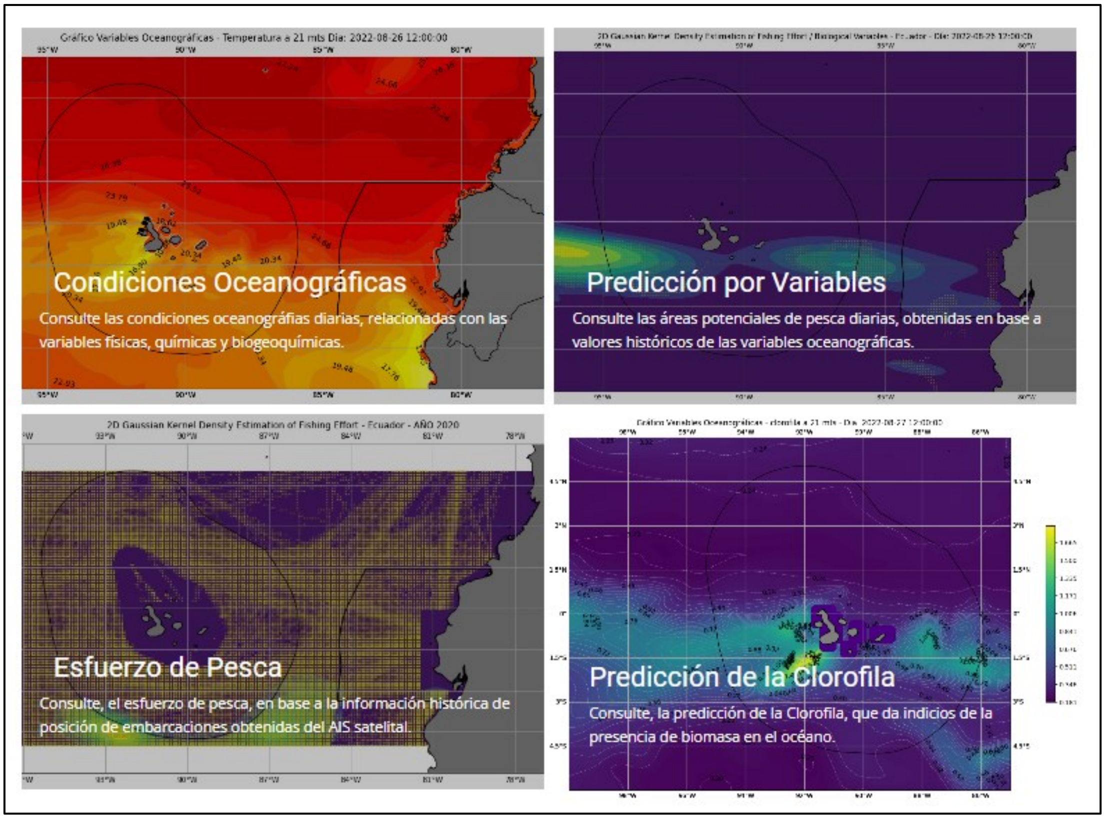


Figure 104. *Access links in home page for PrediccionPesca_EC model*

6.2.2.2. Fishing Effort

On this page, the user, also without any informatic skill, can select the behaviour of the fishing fleet to track, both in transit around the Galapagos Islands, as well as the fishing events that took place in this area. Figure 105 shows the links to consult the fishing effort by year and month.

Figure 105. *Fishing Effort for PredicciónPesca_EC model*

6.2.2.2.1. Annual Fishing Report

On this page, the user can view a summary of the fishing effort around the area of interest, as well as identify the hot spots, where fishing activities were identified. The entered information corresponds to the years from 2017 to 2020 (see Figure 106).

Figure 106. *Annual Fishing Report web page for PredicciónPesca_EC model*

6.2.2.2.2. Monthly Fishing Report

In the same way as in the previous point, the fishing effort around the Galapagos Islands can be consulted by month, as well as identifying the hot spots, where it was identified which fishing activities are most important for the Galapagos Islands, (see Figure 106).

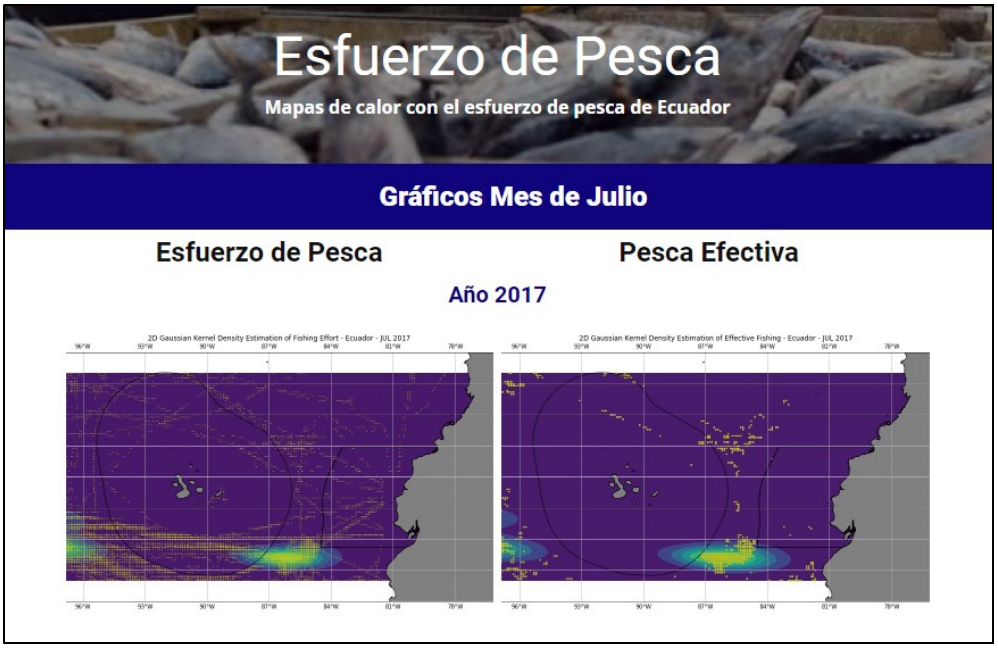


Figure 107. *Monthly Fishing Report web page for PrediccionPesca_EC model.*

6.2.2.3. Chlorophyll Prediction

This section shows the chlorophyll prediction around the area of interest for up to 7 days. It should be noted that if there is a high level of chlorophyll in the sea, it may indicate whether fish will be present, which will attract fishing activity in turn. Figure 108 shows the maps of the Chlorophyll predictions.

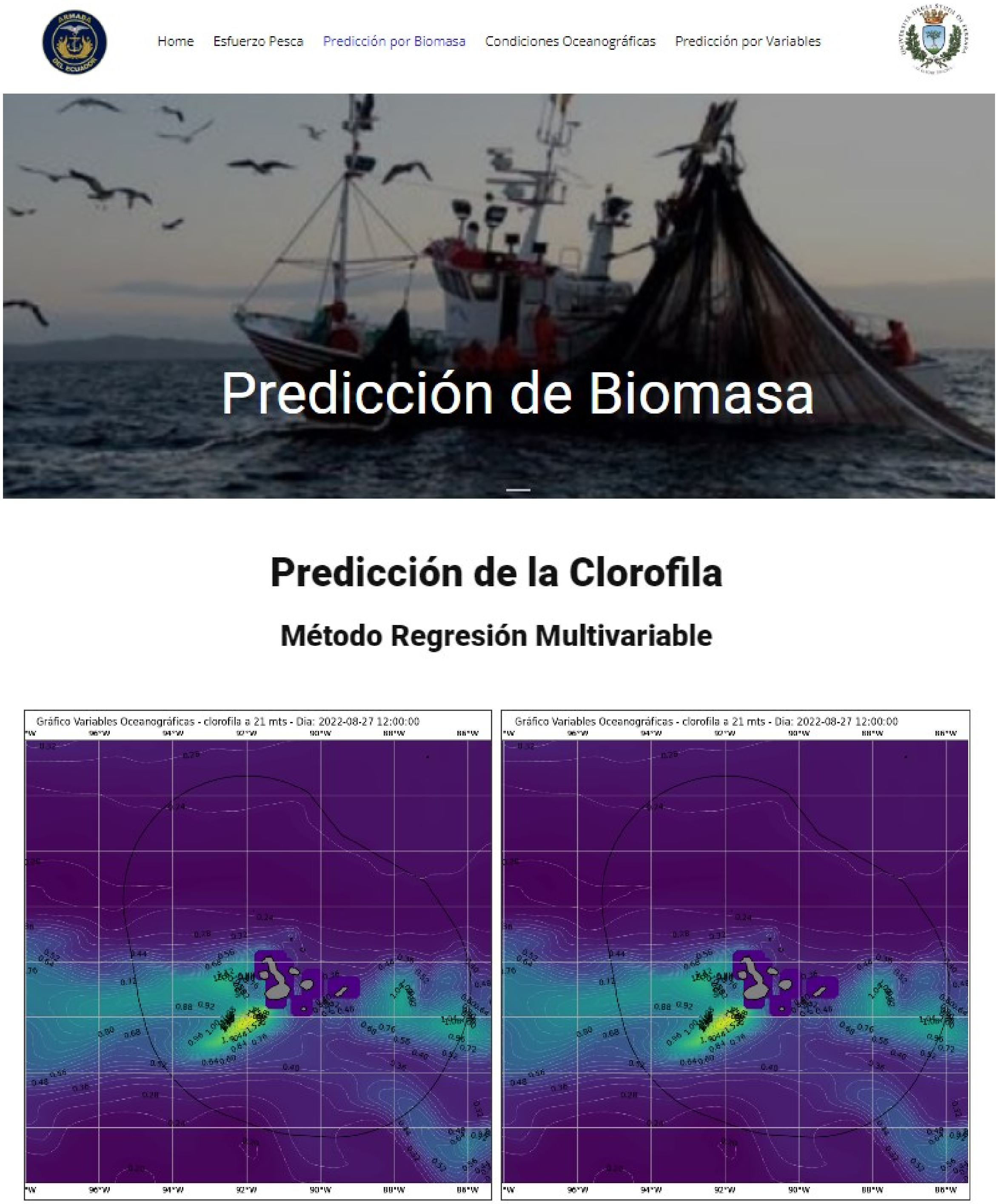


Figure 108. *Chlorophyll Prediction web page for PredicciónPesca_EC model*

6.2.2.4. Forecast of potential fishing areas

This section summarizes the results obtained for the prediction of potential fishing areas, taking into consideration the values of the different oceanographic variables, which were grouped in physical, geological, chemical, biological and nutrient variables. Figures 109 shows the access to the forecast of potential fishing areas.

Figure 109. *Forecast of potential fishing areas web page for PrediccIónPesca_EC model*

Figure 110 shows an example of the forecast maps of potential fishing areas considering only the physical variables. On the website it is possible to consult the forecast maps taking into account other variables such as chemical, nutrient and biological variables. A graph is made for each type of variable, which integrates the results of the identification of places where historical fishing conditions are repeated, as well as the identification of potential fishing zones for each type of variable.

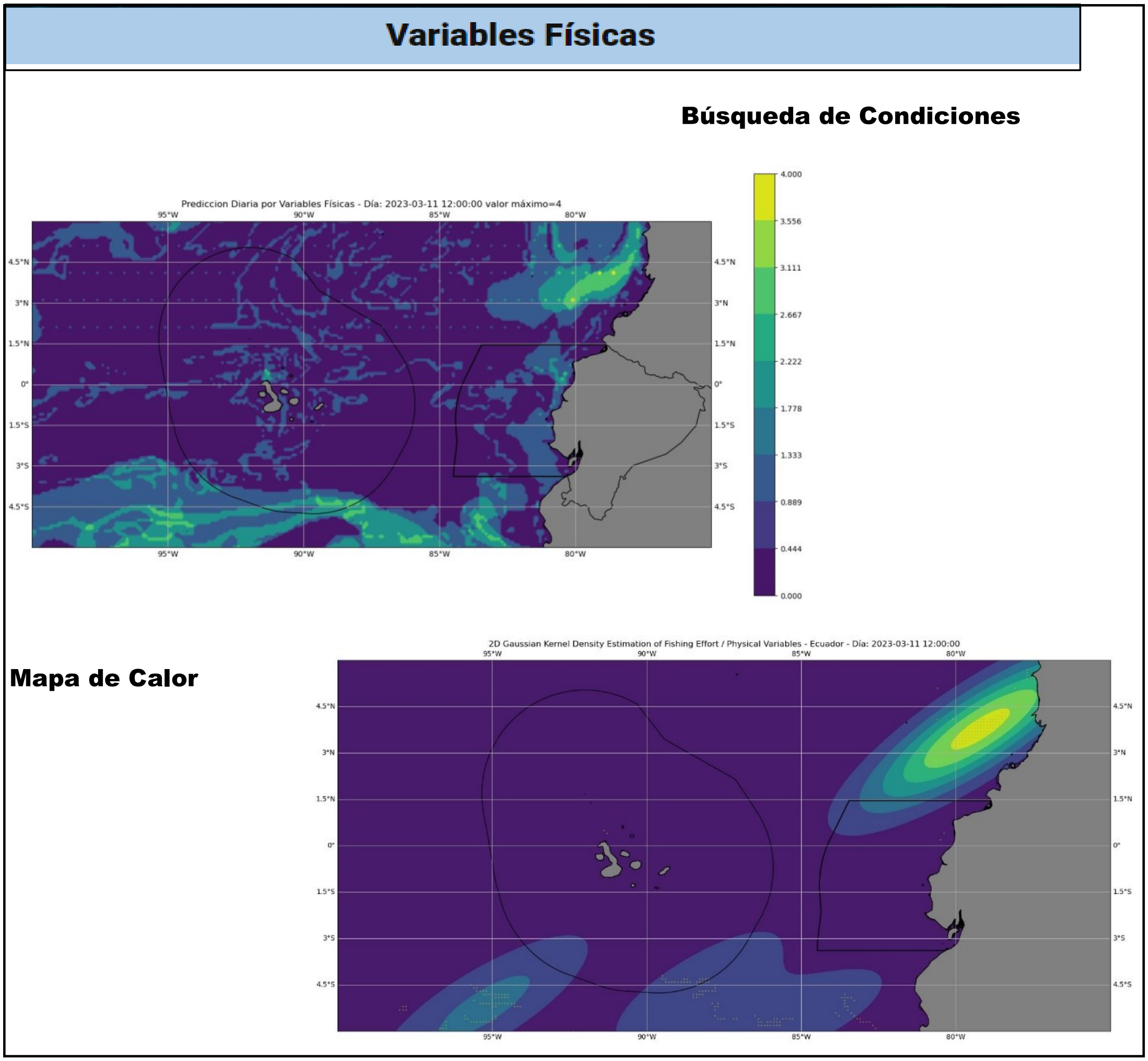
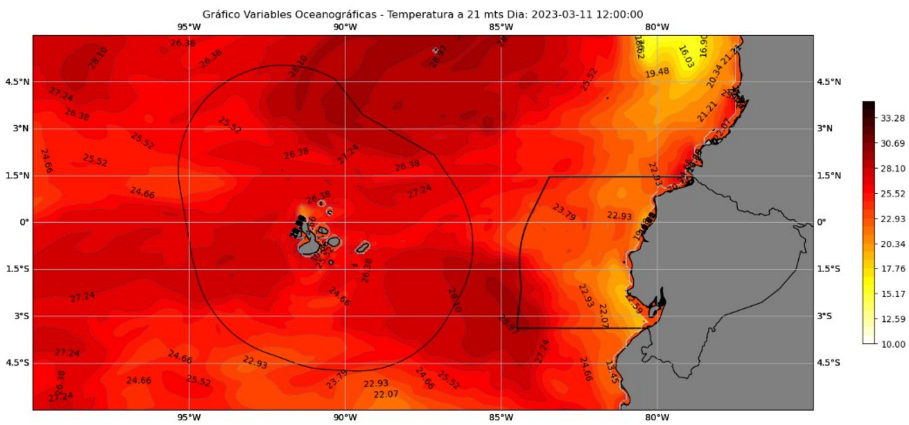


Figure 110. *Forecast of potential fishing using physics variables*

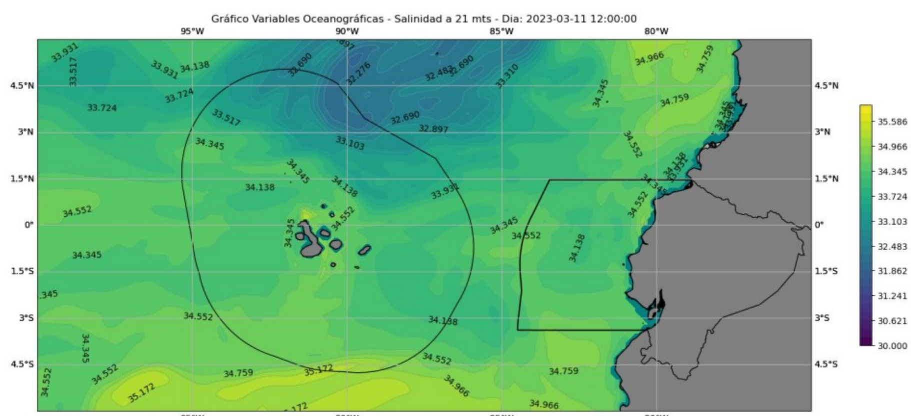
6.2.2.5. Oceanographic conditions

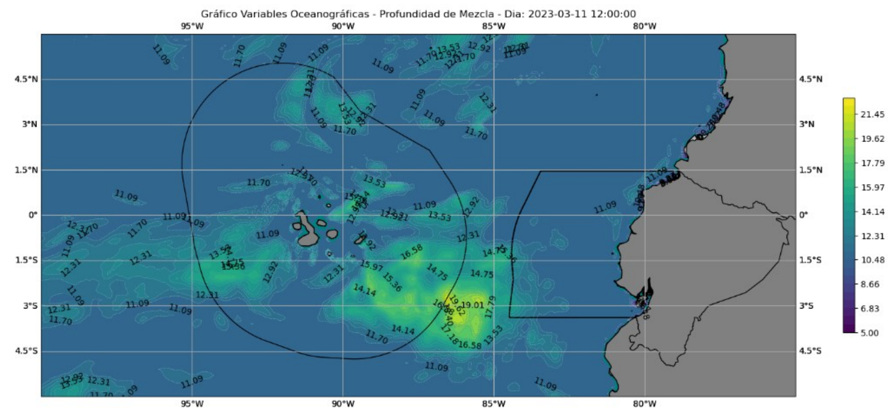
Likewise, a page has been prepared that allows daily consultation of oceanographic conditions in the study area. Similarly, the type of variables (physical, geological, chemical, biological and nutrients) have been grouped together, an example of these graphs can be seen in figure 111.



Temperatura

Salinidad





Profundidad De mezcla

Altura de Ola

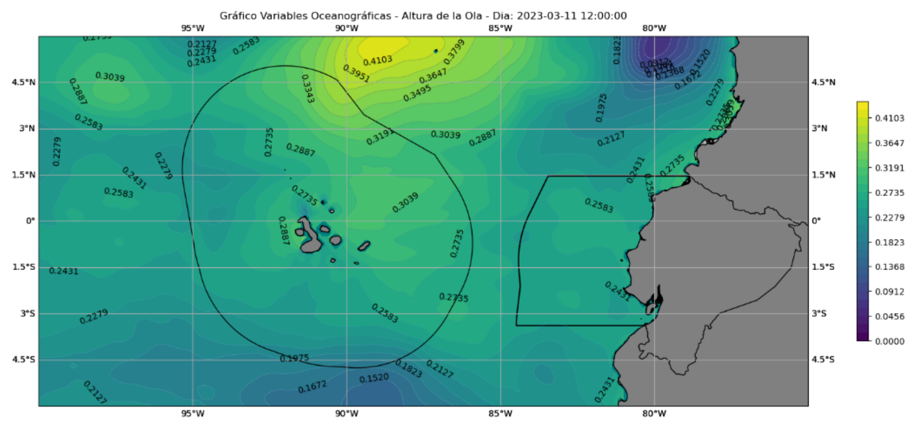


Figure 111. Daily Oceanographic conditions in Ecuador *for PredicciónPesca_EC model*

CAPTHER VII

SUMMARY AND CONCLUDING REMARKS

Illegal, unreported and underregulated fishing is threatening marine ecosystems, especially in the Galapagos Islands, which due to its geographical position, bottom morphology meteorology and the influence of currents and countercurrents, presents appropriate conditions for fishing. In this thesis, 2 alternatives were analyzed to predict the presence of the fishing fleet, in order to provide information to the Ecuadorian Navy to plan operations to combat illegal fishing, on potential fishing areas where to focus its attention.

The bottom morphology changes from an area to another, influencing the local ocean dynamics. Ridges drive currents and create the right conditions for upwellings and, consequently, for the presence of nutrients and then, fishes.

In this study, at first chlorophyll was taken as the variable that will indicate the highest level of biomass in the ocean and therefore the presence of fish, based on the marine food pyramid. Its prediction was performed using machine-learning techniques, such as neural networks and multivariate regression. Subsequently, this prediction was generalized and optimized using the discrete Fourier transform theory. During the chlorophyll prediction process, it was identified that the multivariate linear regression method presents a higher correlation between the real data and the predicted data, which is why this technique was chosen to predict chlorophyll around the Galapagos Islands. It has also been identified that temperature, mixed layer depth, sea surface height, oxygen and pH are inversely proportional to chlorophyll behaviour, while salinity, CO_2, current velocity and PO_4, NO_3, Fe and Si nutrients are directly proportional.

To increase the accuracy in the prediction of chlorophyll, the study area was divided into subareas. This was done considering bathymetry and geomorphology of the Galapagos seabed. Then Fourier was also used to increase the correlation between chlorophyll and the other biogeochemical variables used in the development of the model.

An optimization strategy of the multivariate linear regression method was proposed, by identifying the gap that exists between each of the independent variables in relation to chlorophyll. The dataset was restructured considering this gap and the prediction model was run using a 180-day dataset for the year 2020 - 2021, and its efficiency evaluated the following 180 days. During such an evaluation process we determined that the best correlation is given in a time gap of 3 days (>0.7), while it remains above 0.58 up to 10 days.

Another methodology was used to determine potential fishing areas through statistical analysis of the fishing effort around the Galapagos Islands in order to determine the ranges for the ocean variables in order to increase the fishing probability. It was determined that the fishing fleet does not have a stationary behaviour, and a methodology was developed to track the appropriate conditions in the ocean depending on the month of the day being analyzed.

Finally, a web page was developed. This web page presents the results of this research and it will be used by the Ecuadorian Navy in support of planning tasks. This web page consolidates the results on prediction by chlorophyll, by daily conditions of oceanographic variables in the study area. It is intended that naval operations planners consult this information on a permanent basis to identify potential fishing areas around the Galapagos Islands, where surveillance and control tasks should be intensified to protect fishery resources.

It will be necessary to begin a period of evaluation of the graphical interface developed together with the planners and executors of Navy operations, in order to obtain the evaluation of the convenience of use in conjunction with the other tools available for daily control of the fishing vessels. Once these methodologies are accepted by the Navy, the developed software will be uploaded to its application servers for consultation by naval operations planners.

It is worth mentioning that illegal vessels are not easily detected because they have disconnected their AIS. It should be noted that this study has tried to determine the areas where there are the best conditions for fishing to prioritize their protection with the naval units available in the country.

The analyses carried out in this thesis were performed thanks to the data freely provided by the European Union's Earth observation program coordinated and managed for the European Commission by the European Union Agency for the Space Program, in partnership with the European Space Agency (ESA), and the EU Member States and the data provided by Global Fishing Watch organization. It is worth mentioning that the algorithms developed to obtain the results are hosted in a virtual server and were coded in the Python programming language.

References

Agnew, D. J., Pearce, J., Pramod, G., Peatman, Watson, and Pitcher, T. J. (2013). Estimating the worldwide extent of illegal fishing. PloS one, 4(2):e4570. 12.

Ahmad, H. (2019). Machine learning applications in oceanography. . Aquatic Research 2(3), 161-169.

Alsterberg, C., Eklöf, J. S., Gamfeldt, L., Havenhand, J. N., & Sundbäck, K. (2013). Consumers mediate the effects of experimental ocean acidification and warming on primary producers. Proceedings of the National Academy of Sciences, 8603-8608. doi:https://www.pnas.org/doi/abs/10.1073/pnas.1303797110

Amini Pishro, A., Zhang, S., & Huang, D. (2021). Application of artificial neural networks and multiple linear regression on local bond stress equation of UHPC and reinforcing steel bars. Sci Rep 11, 15061. doi:https://doi.org/10.1038/s41598-021-94480-2

Anderson, T. (2009). Kernel density estimation and K-means clustering to profile road accident hotspots. Accident Analysis & Prevention,Volume 41, Issue 3, 359-364,ISSN 0001-4575. https://doi.org/10.1016/j.aap.2008.12.014.

Aumont, O., Ethé, C., Tagliabue, A., Bopp, L., & Gehlen, M. (2015). PISCES-v2: an ocean biogeochemical model for carbon and ecosystem studies. Geosci. Model Dev., 8 (pp. 2465–2513). doi:https://doi.org/10.5194/gmd-8-2465-2015

Barocio, O., Millan, R., Santamaria, E., & Gonzalez, A. (2007). Productividad primaria del fitoplancton en la zona eufótica del Sistema de la Corriente de California estimada mediante imágenes del CZCS. Cienc. mar [online]. 2007, vol.33, n.1, 59-72. http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0185-38802007000100006&lng=es&nrm=iso

Brunel, T., Van Damme, C. J., Samson, M., & Dickey-Collas, M. (2018). Quantifying the influence of geography and environment on the northeast atlantic mackerel spawning distribution. Fisheries oceanography, 27(2):159–173.

Carey, S., Fisher, C. R., de Leon, P. S., Roman, C., Raineault, N. A., Suarez, J., & Ballard, R. D. (2016). Exploring the undersea world of the Galápagos Islands. . In Ocean. Mag, 29, 32-34.

Castro, X. (2005). Analysis of the Current Socio Economic Situation of the 'Galapagos Artisanal Fishing Community'. Ecuador: Parque Nacional Galápagos/JICA (Japanese International

Cooperation Agency). https://www.jica.go.jp/project/ecuador/3185011E0/materials/pdf/analysis.pdf

Chang, C.-H. (2015). Development of ocean color algorithms for estimating chlorophyll-a concentrations and inherent optical properties using gene expression programming (gep). Optics Express. 23: 5417., 22.

Chassot, E., Bonhommeau, S., Reygondeau, G., Nieto, K., Polovina, J. J., Huret, M., . . . Demarcq, H. (2011). Teledetección satelital para un enfoque ecosistémico de la gestión pesquera. ICES Journal of Marine Science , Volumen 68, Número 4, 651-666. Retrieved from https://doi.org/10.1093/icesjms/fsq195

Chávez, F., Coltorti, M., Ivaldi, R., Sánchez, E., & Sciavicco, G. (2020). Temporal Aspects of Chlorophyll-a Presence Prediction Around Galapagos Islands. 6th International Conference on Technologies and Innovation. CITI 2020. doi:10.1007/978-3-030-62015-8_8

Chen, R., Wu, X., Liu, B., Wang, Y., & Gao, Z. (2022). Mapping coastal fishing grounds and assessing the effectiveness of fishery regulation measures with AIS data: A case study of the sea area around the Bohai Strait, China, Ocean & Coastal Management, Volume 223. https://doi.org/10.1016/j.ocecoaman.2022.106136

Christensen, V. (1996). Managing fisheries involving predator and prey species. Reviews in fish Biology and Fisheries. Rev Fish Biol Fisheries 6, 417– 442. doi:https://doi.org/10.1007/BF00164324

Cimino, M., Anderson, M., Schramek, T., Merri eld, S., Terrill, E., (2019). Towards a Fishing Pressure Prediction System for a Western Pacific EEZ. Sci Rep 9, 461, 1 10. doi:https://doi.org/10.1038/s41598-018-36915-x

COGMAR, Comandancia General de la Armada. (2013). Plan de Seguridad Integral y Protección de los Espacios Acuáticos - Libro IV. Armada del Ecuador. 4.

COPERNICUS, European Union's Earth observation programme. (2019). Product User Manual for Global Biogeochemical Analysis and Forecasting Product. Marine Environment Monitoring Service. Retrieved from https://doi.org/10.48670/moi-00015

COPERNICUS, European Union's Earth observation programme. (2020). Product User Manual for Global Physical Analysis Coupled System Forecasting Product Marine Enviroment Monitoring Service. Retrieved from https://resources.marine.copernicus.eu/

COPERNICUS, European Union's Earth observation programme. (2021). Product User Manual for Global Physical; Analysis and Coupled System Forecasting Product Marine Enviroment Monitoring Service. Retrieved from https://doi.org/10.48670/moi-00016

Cury, P., Shannon, L., & Shin, Y. (2003). The functioning of marine ecosystems: a fisheries perspective. Responsible fisheries in the marine ecosystem. 103–123. 18, 19.

De Souza, E., Boerder, K., Matwin, S., & Worm, B. (2016). Improving shing pattern detection from satellite ais using data mining and machine learning. PloS one, 11(7), e0158248.

Del Frate, F., Petrocchi, A., Lichtenegger, J., & Calabresi, G. (2000). Neural networks for oil spill detection using ers-sar data. IEEE Transactions on geoscience and remote sensing 38(5), 2282 - 2287.

Desortová, B. (1981). Relationship between Chlorophyll-α concentration and phytoplankton biomass in several reservoirs in Czechoslovakia. Internationale Revue der gesamten Hydrobiologie und Hydrographie 66(2), 153 - 169.

Detsis, E., Brodsky, Y., Knudtson, P., Cuba, M., Fuqua, H., & Szalai, B. (2012). Project catch: a space based solution to combat illegal, unreported and unregulated fishing: Part i: vessel monitoring system. . Acta Astronautica 80, 114–123.

Dunn, R. A., Lekić, V., Detrick, R. S., & Toomey, D. R. (2005). Three-dimensional seismic structure of the Mid-Atlantic Ridge (35 N): Evidence for focused melt supply and lower crustal dike injection. Journal of Geophysical Research: Solid Earth, 110(B9). doi:https://doi.org/10.1029/2004JB003473.

Dutta, S., Chanda, A., Akhand, A., & Hazra, S. (2016). Correlation of phytoplankton biomass (Chlorophyll-a) and nutrients with the catch per unit e ort in the PFZ forecast areas of northern bay of bengal during simultaneous validation of winter fishing season. . Turkish Journal of Fisheries and Aquatic Sciences, 1, 767 - 777.

Encyclopædia Britannica, (2007). Major surface currents of the oceans. https://www.britannica.com/science/ocean-current#/media/1/424354/60311

FAO, Food and Agriculture Organization of the United Nations. (2011). Subperfíles de pesca y acuicultura por países. Ecuador. https://www.fao.org/fishery/es/facp/ecu?lang=es

Filatov, N., Pozdnyakov, D., J., O. M., P., H., L., & Bobylev, L. P. (2007). White Sea: Its marine environment and ecosystem dynamics influenced by global change. Springer Science & Business Media.

Forryan, A., & Naveira, A. V. (2021). Galápagos upwelling driven by localized wind–front interactions. In Sci Rep 11, 1277. doi:https://doi.org/10.1038/s41598-020-80609-2.

Franklin, J., Sathish, T., Vinithkumar, N., & Kirubagaran, R. (2020). A novel approach to predict Chlorophyll-a in coastal-marine ecosystems using multiple linear regression and principal component scores. . Marine Pollution Bulletin, 152, 110902.

Garcia, S. M., & De Leiva Moreno, I. (2003). Global overview of marine fisheries. Responsible fisheries in the marine ecosystem, 1-24.

GEBCO. (2020). General Bathymetric Chart of the Oceans. Retrieved from https://www.gebco.net/

Gladju, J., Kamalam, B. S., & Kanagaraj, A. (2022). Applications of data mining and machine learning framework in aquaculture and fisheries: A review. Smart Agricultural Technology, 100061.

González Vilas, L., Spyrakos, E., & Torres Palenzuela, J. M. (2011). Neural network estimation of chlorophyll a from meris full resolution data for the coastal waters of galician rias (nw spain). Remote Sensing of Environment. 115(2):524 – 535. 20.

Guarderas Sevilla, M. J. (2019). Temporal and spatial patterns influence reef fish community structure along an upwelling gradient in the Galápagos Marine Reserve (Bachelor's thesis, Quito). http://repositorio.usfq.edu.ec/handle/23000/8482

Guardian. (2020). Alarm over discovery of hundreds of Chinese fishing vessels near Galápagos Islands. https://www.theguardian.com/environment/2020/jul/27/chinese-fishing-vessels-galapagos-islands

Guisande, C. (2010). Ipez: an expert system for the taxonomic identication of shes based on machine learning techniques. Fisheries Research 102(3), 240 - 247.

Hallman, J. (2019). A comparative study on Linear Regression and Neural Networks for estimating order quantities of powder blends. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1383464&dswid=1732

Harrison, J., & Morgera, E. (2012). Article 61 Conservation of the living resources. Washington Post. IOC-UNESCO. 2022. Multiple Ocean Stressors: A Scientific Summary for Policy Makers. P.W. Boyd et al. (eds). Paris, UNESCO. 20 pp. (IOC Information Series, 1404). doi:10.25607/OBP-1724.

Holland, E., Monier, M., Legeais, J.-F., Prado, S., Sathyendranath, S., & and Dupouy, C. (2019). The use of copernicus marine service products to describe the state of the tropical western pacific ocean around the islands: a case study. Journal of Operational Oceanography, 12:1–9. 20, 22.

IOC-UNESCO. (2022). Multiple Ocean Stressors: A Scientific Summary for Policy Makers. doi:10.25607/OBP-1724.

Jean-Yves, C. (2009). Geología y geofisica marina y terrestre del Ecuador: desde la costa continental hasta las Islas Galapagos. Guayaquil (ECU); Marseille (FRA); Guayaquil: CNDM; IRD; INOCAR, 269. 978-9978-92-737-3.

Kenchington Ellen, Francisco Javier Murillo, Camille Lirette, Mar Sacau, Mariano Koen-Alonso, Andrew Kenny, Neil Ollerhead, Vonda Wareham, Lindsay Beazley. (2014). Modelado de superficie de densidad de kernel como medio para identificar concentraciones significativas de indicadores de ecosistemas marinos vulnerables . PLoS ONE 9(10): e109365. https://doi.org/10.1371/journal.pone.0109365

Kennedy, C. M., & Erickson, A. S. (2017). China maritime report no. 1: China's third sea force, the people's armed forces maritime militia: Tethered to the plan 6.

Kwon, Y., Baek, S., L. Y., Pyo, J., Ligaray, M., Park, Y., & Cho, K. (2018). Monitoring coastal Chlorophyll-a concentrations in coastal areas using machine learning models. Water 10(8), 1020.

Learned-Miller, E. G. (2014). Introduction to supervised learning. Department of Computer Science, University of Massachusetts, 3. https://people.cs.umass.edu/~elm/Teaching/Docs/supervised2014a.pdf

Lindstrom, E., Gunn, J., Fischer, A., McCurdy, A., Glover, L., Alverson, . . . Checkley, D. (2012). A framework for ocean observing. . Proceedings of the Ocean Information for Society: Sustaining the Benefits, Realizing the Potential, Venice, Italy, 2125, 22, 23.

Lumini, A., & Nanni, L. (2019). Ocean ecosystems plankton classification. In Recent Advances in Computer Vision, 261–280. Springer, 18.

MacKenzie, B. R., Payne, M. R., Boje, J., Høyer, J. L., & Siegstad, H. (2014). A cascade of warming impacts brings bluefin tuna to greenland waters. Global change biology. 20(8):2484–2491.

Marañón, E., Cermeno, P., Latasa, M., & Tadonléké, R. (2015). Resource supply alone explains the variability of marine phytoplankton size structure. . Limnology and Oceanography, 60(5), 1848–1854.

Marzuki, M., Gaspar, P., Garello, R., Kerbaol, V., & Fablet, R. (2017). Fishing gear identication from vessel-monitoring-system-based shing vessel trajectories. IEEE Journal of Oceanic Engineering 43(3), 689 - 699.

Maslukah, L., Setiawan, R. Y., Nurdin, N., Helmi, M., & Widiaratih, R. (2022). Phytoplankton Chlorophyll-a Biomass and the Relationship with Water Quality in Barrang Caddi, Spermonde, Indonesia. Ecological Engineering & Environmental Technology, 23.

Medinets, V., Gazyetov, Y., Cherkez, E., Snihirov, S., Medinets, S., Kovalova, N., & Konareva, O. (2020). Geomorphological research as basis for spatial integrated marine ecosystems monitoring. In XIV International Scientific Conference “Monitoring of Geological Processes and Ecological Condition of the Environment” (Vol. 2020, No. 1, pp. 1-5). European Association of Geoscientists & Engineers.

Metuzals, K., Baird, R., Pitcher, T., Sumaila, R., U., & Ganapathiraju. (2010). P One fish, two fish, IUU, and no fish: unreported fishing worldwide. Handbook of marine fisheries conservation and management. Oxford University Press, New York. 165–181 6,7.

Monolisha, S., George, G., & Platt, T. (2017). Fisheries oceanography-established links in the Eastern Arabian Sea. 230 - 233. http://eprints.cmfri.org.in/12770/

Naselli-Flores, L., Padisák, J., & Dokulil, M. (2003). Phytoplankton and equilibrium concept: the ecology of steady-state assemblages. Developments in hydrobiology. 172:416. 19.

Nasteski, V. (2017). An overview of the supervised machine learning methods. Horizons. b, 4, , 51-62.

Payne, M. R., & Lehodey, P. (2019). Review of the use of ocean data in european fishery management and monitoring applications. Journal of Operational Oceanography, 12(supplement 1), S48–S52.

Pontavice, H. d. (2019). Chapter 12 - Changing biomass flows in marine ecosystems: from the past to the future. Predicting Future Oceans 121-128. Elsevier. doi:https://doi.org/10.1016/B978-0-12-817945-1.00012-5

Prince Albert II of Monaco, Joaquim Alves Gaspar, Joseph Aschbacher, Pierre Bahurel, Elaine Baker, Manuel Barange, Elżbieta Bienkowska, Pierre Boileau, Hans Bruyninckx, Gilberto Câmara, Richard Clavaud, Jean-Dominique Dallet, Claudia Delgado, Boris Dewitte, Cecilia Donati, Salvador Antonio De La Encina Ortega, Pedro Fidelman, Catherine Franche, Françoise Gaill, Rémi Gruet, Paul Holthus, Catherine Jeandel, Claire Jolly, Véronique Landes, Pierre-Yves Le Traon, Thomas Lesage, Davide Magagna, Iryna Makarenko, Maja Markovčić Kostelac, Angélique Melet, Claire Morel, Kobi Mosquera, Babacar Ndao, Tidiane Ouattara, Gratianne Quade, Emily Penn, Simon De Pietro, Alain Ratier, Jake Rice, Dominique Ristori, Vladimir Ryabinin, Francesca Santoro, Ricardo Serrão Santos, Virginie Tassin, Pierre Taverniers, Babette Tchonang, Peter Thomson, Jean-Noël Thepaut, Cécile Thomas-Courcoux, Maura Tronci, Romain Troublé, Peter Tuddenham, Philippe Vallette, Karmenu Vella, Martin Visbeck, Karina Von Schuckmann. Ao, R. B. (2019). Blue Book - Corpernicus for a sustaintable ocean. Mercator Ocean International. https://marine.copernicus.eu/services/user-learning-services/blue-book-copernicus-sustainable-ocean

Raven, J., & Maberly, S. (2009). Phytoplankton nutrition and related mixotrophy. Encyclopedia of Inland Waters. 192–196. 19. https://nora.nerc.ac.uk/id/eprint/4435.

Rentería, W., Valarezo, O. O., & García, G. L. (2019). Análisis del Esfuerzo Pesquero en el Territorio Marítimo Ecuatoriano. Revista de Ciencias de Seguridad y Defensa, 2019, vol. 4, no 6.

Reyes, L. M. (2007). Historia de la ecología. 14, 27. Trabajo de Grado. Maestría en Investigación. Universidad de San Carlos de Guatemala Facultad de Humanidades. Departamento de Postgrado. Guatemala. http://biblioteca. usac. edu. gt/tesis/07/07_1934. pdf.

Rodriguez, G. (2002). Hartley transform: Basic theory and applications in oceanographic time series analysis. Coastal Environment : Environmental Problems in Coastal Regions IV. 191 - 200.

Salgado Ugarte, I. H., Shimizu, M., Taniuchi, T., & Matsushita, K. (2000). Size frequency analysis by averaged shifted histograms and kernel density estimators. Asian Fisheries Science, 13(1), 1-12.

Santhanam, P., & Begum, A. (2019). Basic and Applied Phytoplankton Biology. Springer Singapore, 1st ed. edition. 25 - 61.

Santos, A. P. (2000). Fisheries oceanography using satellite and airborne remote sensing methods: a review, Fisheries Research, Volume 49, Issue 1. 1-20, ISSN 0165-7836. https://doi.org/10.1016/S0165-7836(00)00201-0

Schiller, L., Alava, J., Grove, J., Reck, G., & Pauly, D. (2015). The demise of darwin's fishes: evidence of fishing down and illegal shark finning in the galápagos islands. Aquatic Conservation: Marine and Freshwater Ecosystems, 25(3):431–446. 12.

Schowalter, T. D. (2016). Insect ecology: an ecosystem approach. Academic press, 14.

Schulz, M., Koschel, R., Reese, C., & Mehner, T. (2004). Pelagic trophic transfer efficiency in an oligotrophic, dimictic deep lake (lake stechlin, germany) and its relation to fisheries yield. Limnologica, 34(3): 264 – 273. 16.

Secretariat, G. (2010). Implementation plan for the global observing system for climate in support of the UNFCCC. Proceedings of the Conference of the Parties (COP), Copenhagen, Denmark, 7–18.

SENPLADES, Secretaría Nacional de Planificación y Desarrollo. (2017). Planificación y Ordenamiento del Espacio Marino Costero Ecuatoriano. Subsecretaría de Planificación Nacional - Dirección de Asuntos Marino Costeros DAMC.- Guayaquil. https://www.planificacion.gob.ec/wp-content/uploads/downloads/2018/07/Plan-de-Ordenamiento-del-Espacio-Marino-Costero.pdf

Smith, S. (2003). Linear Systems. Digital Signal Processing. Retrieved from https://doi.org/10.1016/B978-0-7506-7444-7/50042-X

Sumaila, U. R., Alder, J., & Keith, H. (2006). Global scope and economics of illegal fishing. Marine Policy. 30(6): 696–703. https://www.sciencedirect.com/science/article/pii/S0308597X05000795

Sundararajan, D. (2001). The discrete Fourier transform: theory, algorithms and applications. World Scientific. https://books.google.es/books?hl=es&lr=&id=54kTgFg5IVgC&oi=fnd&pg=PR7&dq=The+discrete+Fourier+transform:+theory,+algorithms+and+applications.&ots=T35VjaD1ia&sig=MyirHRhX7FrkJS3xus_jQuv7tBY#v=onepage&q=The%20discrete%20Fourier%20transform%3A%20theory%2C%20algorithms%20and%20applications.&f=false

Sunye, P. S., & Servain, J. (1998). Effects of seasonal variations in meteorology and oceanography on the Brazilian sardine fishery. Fisheries Oceanography, 7(2), 89-100. doi:https://doi.org/10.1046/j.1365-2419.1998.00055.x

Syah, Fahrudin, Ramdani, Laras, Suniada, & Iwan., K. (2020). Prediction of potential fishing zones for mackerel tuna (Euthynnus sp) in Bali Strait using remotely sensed data. In IOP

Conference Series: Earth and Environmental Science (Vol. 500, No. 1, p. 012070). IOP Publishing.

Thessen, A. (2016). Adoption of machine learning techniques in ecology and Earth science. One Ecosystem 1,e8621. https://oneecosystem.pensoft.net/article/8621/download/pdf/

Trenkamp, R., Kellogg, J. N., Freymueller, J. T., & Mora, H. P. (2002). Wide plate margin deformation, southern Central America and northwestern South America, CASA GPS observations. Journal of South American Earth Sciences, 15(2), 157-171. https://doi.org/10.1016/S0895-9811(02)00018-4.

Trites, A. W. (2003). Food webs in the ocean: who eats whom and how much. Responsible fisheries in the marine ecosystem. pages 125–141. 16, 17.

United Nations (2006). Sustainable fisheries, including through the 1995. Agreement for the Implementation of the Provisions of the United Nations Convention on the Law of the Sea of 10 December 1982 relating to the Conservation and Management of Straddling Fish Stocks and High. Retrieved from https://undocs.org/en/A/RES/61/105

United Nations (2019). World Population Prospects 2019: Press Release. Citation: United Nations, Department of Economic and Social Affairs, Population Division. Department of Economic and Social ; Population Division.

United Nations Educational (2021). The Ocean Decade at COP26 of the United Nations Framework Convention on Climate Change. París, UNESCO. 8. https://www.oceandecade.org/wp-content/uploads//2021/11/356287-The%20Ocean%20Decade%20at%20COP26

Valente, A., Sathyendranath, S., Brotas, V., Groom, S., Grant, M., Taberner, M., Baker, K. (2016). A compilation of global bio-optical in situ data for ocean-colour satellite applications. Earth System Science Data, 8(1):235. 21.

Valery, A. I. (2017). Posición de la fundación charles darwin ante pesca ilegal en Galápagos. 12. https://www.darwinfoundation.org/es/articulos-blog/126-posicion-de-la-fundacion-charles-darwin-ante-pesca-ilegal-en-galapagos

Von Schuckmann, K., Le Traon, P.-Y., Alvarez-Fanjul, E., Axell, L., Balmaseda, M., Breivik, L.-A., Drevillon, M. D. (2016). The copernicus marine environment monitoring service ocean state report. Journal of Operational Oceanography, 9(sup2): s235–s320. 21 - 25.

Von Schuckmann, K., Le Traon, P.-Y., Smith, N. P., Djavidnia, S. G., J.-P., G. M., & Nolan, G. A. (2019). Copernicus marine service ocean state report, issue 3. . Journal of Operational Oceanography, 12(sup1):S1–S123, 19 - 20.

Waikato, U. (2009, 09 17). Science Learning Hub. Retrieved from Marine trophic pyramid: https://www.sciencelearn.org.nz/images/144-marine-trophic-pyramid

Wang, S. C. (2003). Artificial neural network. In Interdisciplinary computing in java programming. Springer, Boston MA, 81-100.

Wang, S., Lee, Z., Shang, S., Li, J., Zhang, B., & and Lin, G. (2019). Deriving inherent optical properties from classical water color measurements: Forel-ule index and secchi disk depth. . Opt. Express, 27(5):7642–7655.

Wang, S., Zhao, J., Ding, X., & and Li, X. (2020). Nitric oxide and nitrous oxide production in anaerobic/anoxic nitrite-denitrifying phosphorus removal process: effect of phosphorus concentration. Environmental Science and Pollution Research. 19, 1–13.

Wang, Y., & Veluvolu, K. (2017). Time-Frequency Analysis of Non-Stationary Biological Signals with Sparse Linear Regression Based Fourier Linear Combiner. Sensors (Basel). https://doi.org/10.3390/s17061386

Wilmoth, J. (2019). Revision of world population prospects. 3. https://www.un.org/development/desa/publications/world-population-prospects-2019-highlights.html

Witt, M. J., & Godley, B. J. (2007). A step towards seascape scale conservation: using vessel monitoring systems (VMS) to map fishing activity. PloS one, 2(10), e1111. https://doi.org/10.1371/journal.pone.0001111

Wooster, W. S. (1966). The Oceanographic Setting of the Galápagos. . In In The Galapagos: Proceedings of the Symposia of Galapagos International Scientific Project (p. 100). Univ of California Press.

Zainuddin, M. (2011). Skipjack tuna in relation to sea surface temperature and Chlorophyll-a concentration of Bone Bay using remotely sensed satellite data. Journal Ilmu dan Teknologi Kelautan Tropis 3(1).

Zemp, M., Chao, Q., Han Dolman, A. J., Herold, M., Krug, T., Speich, S., & Yu, W. (2022). GCOS 2022 Implementation Plan. Global Climate Observing System GCOS, (244), 85.

Zhang B., Zhao, S., Wu, W., Zhou, X. (2018). Fast dynamic routing based on weighted kernel density estimation. In International symposium on artificial intelligence and robotics. 301-309. doi:https://doi.org/10.48550/arxiv.1805.10807

Zou, J. H. (2008). Overview of Artificial Neural Networks. D. J. Livingstone, Artificial Neural Networks: Methods and Applications vol 458 (pp. 14-22). Humana Press. https://doi.org/10.1007/978-1-60327-101-1_2.

ANNEX A

COMPARISON OF THE DATA PROVIDED BY THE COPERNICUS PROGRAM VERSUS THE DATA RECORDED IN SITU.

At the end of January 2023, Ecuador installed its first oceanographic buoy in position latitude -1.937867º and longitude -82.808933º; which records hourly data on the physical variables, temperature and salinity and the chemical variable dissolved oxygen. For this purpose, it has 5 sensors that are installed at the following depths: 0, 10, 20, 50, 100, 150, 200, 300, 500 meters. Although this buoy reports the measurements of each variables every hour, only we take the values recorded at 12:00:00 of each day, since that is the time of recording the measurements of these variables of the Copernicus program.

With 135 measurements recorded at 12:00:00 of each variable from January 31 to February 15, 2023, it has been possible to validate the information provided by the satellite in this area, since it has obtained a correlation of 98% for temperature, 88% for salinity and 97% for dissolved oxygen. Therefore, Ecuador previously conducted an annual cruise, in which only one value was recorded for each research point and there are not enough points available to perform the analysis of the correlation of satellite data with the data from the aforementioned cruises.

Although it is not feasible to validate all the variables analyzed in this study, with data available in situ, it can be determined that at least 2 physical variables, such as temperature and salinity, have a high correlation with the data recorded in the database of the “Global Analysis Forecast-PHY-001-024-TDS¨ (COPERNICUS, 2020), and one biogechemical variable such as oxygen, also has a high correlation with the information provided from the database “Global Analysis Forecast-BIO- 001-028-TDS¨ (COPERNICUS, 2019).

Below you can see the graphic correlation that exists between the data recorded by the satellite and the position in which the Ecuadorian oceanographic buoy is installed.

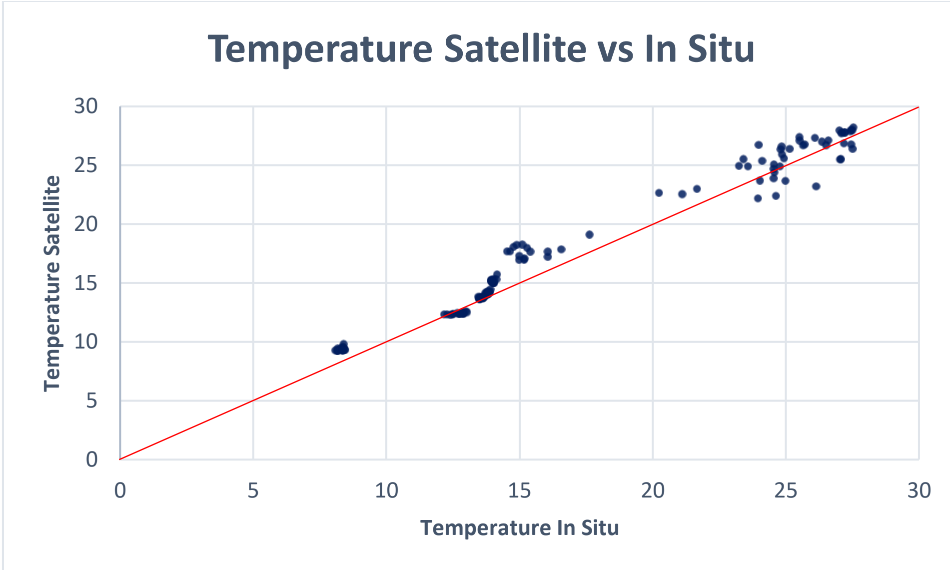


Figure 112. Correlation Temperature by Satellite vs In Situ

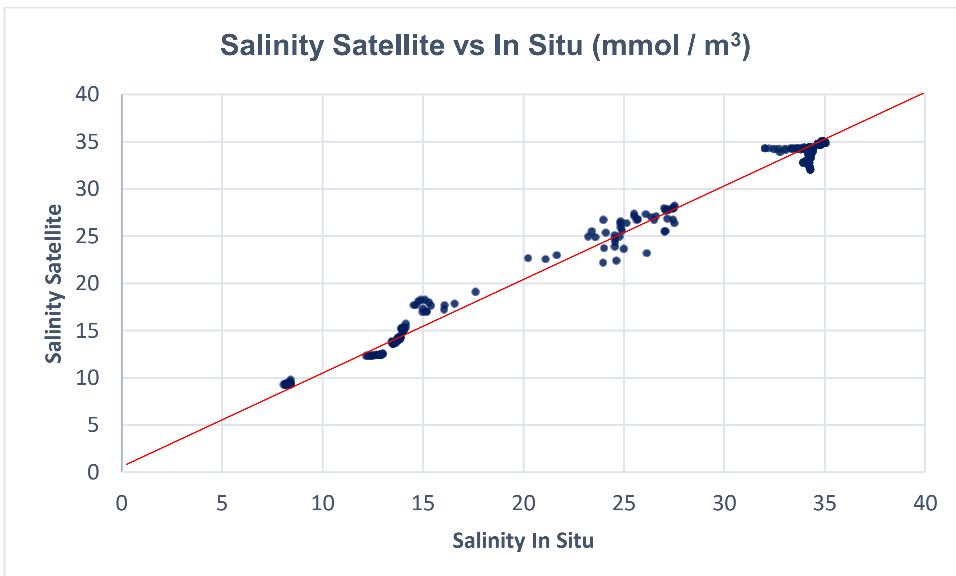


Figure 113. Correlation Salinity by Satellite vs In Situ

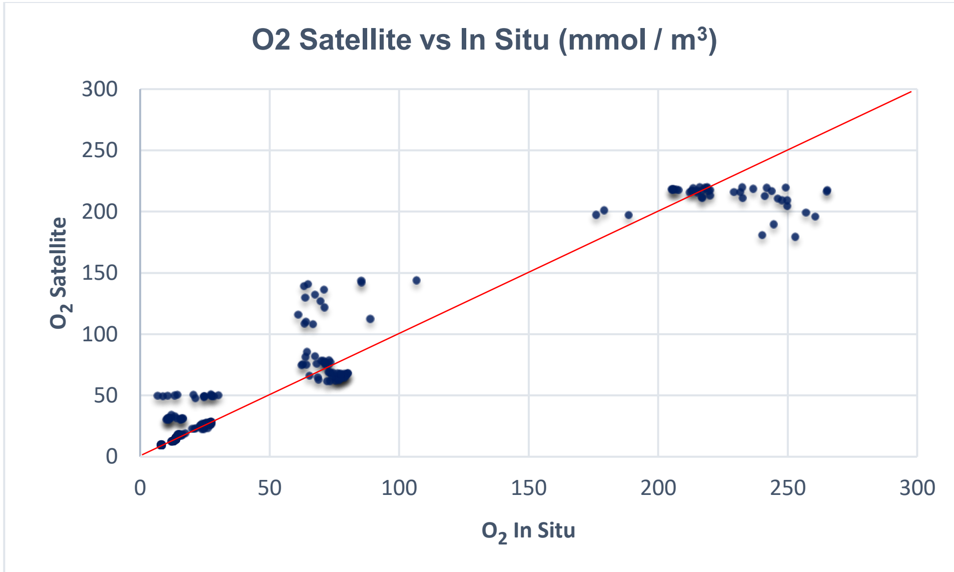


Figure 114. Correlation O_2 by Satellite vs In Situ