INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA

Área Departamental de Engenharia de Electrónica e Telecomunicações e de Computadores

LEARNING PORTUGUESE FISHING DATA PATTERNS

SERGE GASPAR AGUIAR FERNANDES LAGE

Master degree

Projecto Final para obtenção do Grau de Mestre
em Engenharia Informática e de Computadores

Orientadores : Prof. Doutora Iola Maria Silvério Pinto
Prof. Doutor João Carlos Amaro Ferreira

Júri:

Presidente: Prof. Doutor Nuno Miguel Soares Datia

Vogais: Prof. Doutor Artur Jorge Ferreira
Prof. Doutora Iola Maria Silvério Pinto

September, 2020

INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA

Área Departamental de Engenharia de Electrónica e Telecomunicações e de Computadores

LEARNING PORTUGUESE FISHING DATA PATTERNS

SERGE GASPAR AGUIAR FERNANDES LAGE

Master degree

Projecto Final para obtenção do Grau de Mestre
em Engenharia Informática e de Computadores

Orientadores : Prof. Doutora Iola Maria Silvério Pinto
Prof. Doutor João Carlos Amaro Ferreira

Júri:

Presidente: Prof. Doutor Nuno Miguel Soares Datia

Vogais: Prof. Doutor Artur Jorge Ferreira
Prof. Doutora Iola Maria Silvério Pinto

September, 2020

To my son and wife.

Acknowledgments

I would first like to thank my thesis supervisor, Professor Iola Maria Silvério Pinto for her support, guidance, and oversight along the writing of this dissertation.

I would like to thank Professor João Carlos Amaro Ferreira and the company Xsealence for providing real VMS data from Portuguese vessels.

I would also like to acknowledge my friend Bruno Miguel Carvalhido Lima from Faculdade de Engenharia da Universidade do Porto for his greatly appreciated comments and suggestions.

Finally, I cannot pass up the opportunity to thank my family — my wife in particular —, for her support and continuous encouragement throughout my years of study and through the process of researching and writing this dissertation. This accomplishment would not have been possible without them.

Thank you.

Serge Lage

Abstract

Portugal is a country historically linked to the sea, fishing being a very important activity for the Portuguese economy. On the other hand, tax fraud is usually present on fishery as other economic activity and harmful phenomenon for Portugal. For that reason there is a need to create ways to inspect this activity more efficiently. With the constant evolution of new technologies, this need is increasingly close to being met in a concrete way. With this motivation to contribute to the resolution of this problem the objectives of this dissertation analyze the data in order to derive patterns, which when compared to real data can generate alerts for the existence of unusual activities. Concretely, it will be possible to infer when the vessels are fishing, and when they fish in an area other than the usual one, this using only speed and location data as the first objective. The second objective is to classify the fishing activity with the same data, speed and location, and to compare with the vessel assigned license. There are several studies developed in this area all over the globe. What makes my work unique is the use of data created by a device on board, in which there is no human interference. The data used by this solution is produced by the system MONICAP, a bluebox system, mandatory for all vessels over 12 meters in the European Union. This system records speed, heading and location data. Concerning the first objective, a machine learning system will be used to identify whether the vessel is fishing using speed and using clustering algorithms, to identify whether the fishing zone is usual or not. This using a Hill Climbing algorithm and a Karnel estimator. This system is designed so that it can be integrated into MONICAP itself. For the second objective, we will use data mining methods, as Random Forests, Neural Networks an others, to analyze possible associations between the data provided by MONICAP and the type of fishing license. The models were tested and evaluated using well-established data mining techniques. The use of velocity turns out to be enough to create systems capable of satisfying the proposed objectives, so goals are all achieved.

Keywords: Vessel Monitoring System, Data Mining, Fishing

Resumo

Portugal é um país historicamente ligado ao mar, sendo a pesca uma atividade muito importante para a economia portuguesa. Por outro lado, a fraude fiscal está muito presente na pesca, e outras atividades económicas, é um fenómeno prejudicial para Portugal. É necessário inspecionar essa atividade com maior eficiência. Com a constante evolução das novas tecnologias, essa necessidade está cada vez mais próxima de ser atendida de maneira concreta. Para contribuir com a solução deste problema, os objetivos desta dissertação são classificar quando os navios estão a pescar e quando pescam numa área não habitual, apenas com uso de dados de velocidade e localização como primeiro objetivo. O segundo objetivo é classificar a atividade de pesca com os mesmos dados, velocidade e localização, e comparar com a licença atribuída. Existem vários estudos nessa área em todo o mundo. O que torna esse trabalho único é o uso de dados criados por um dispositivo a bordo sem interferência humana. Os dados utilizados por esta solução são produzidos pelo sistema MONICAP, um sistema bluebox, obrigatório para todos os navios com mais de 12 metros na União Europeia. Este sistema regista dados de velocidade, rumo e localização. Para o primeiro problema, criar um sistema de aprendizagem automatica para identificar se o navio está pescando usando velocidade e com algoritmos de agrupamento identificar se a zona de pesca é a habitual. Este sistema foi projetado para ser integrado no próprio MONICAP. Para o segundo problema, usamos métodos de mineração de dados para correlacionar os dados fornecidos pelo MONICAP e o tipo de pesca licença. Os modelos foram testados e avaliados usando técnicas de mineração de dados bem estabelecidas. O uso da velocidade acaba sendo suficiente para criar sistemas capazes de satisfazer objetivos propostos, para que todos os objetivos sejam alcançados.

Palavras-chave: Vessel Monitoring System, Mineralização de dados, Pescas

Contents

List of Figures **xvii**

List of Tables **xix**

Acronyms **xxi**

1 Introduction **1**

1.1 Motivation 1

1.2 Goals 2

1.3 Data 3

1.3.1 VMS Records 3

1.3.2 VMS Vessels 5

1.4 Problem Formulation and Solutions 7

1.4.1 Standalone Fishery Analysis 7

1.4.2 Joined Fishery Analysis 7

1.5 Thesis Structure 8

2 State of the Art **11**

2.1 Literature Review 11

2.1.1 Previous Work 11

2.1.1.1 U. C. report to the final project of course 11

2.1.1.2 Published Paper 12
2.1.2 Work in Literature 12
2.2 Data Analysis Methodologies 15
2.2.1 Distribution Fitting algoritms 15
2.2.1.1 Kernel density estimation 15
2.2.1.2 Hill-Climbing algorithm 15
2.2.2 Clustering algorithms 15
2.2.2.1 K-means clustering algorithm 15
2.2.2.2 Density-based clustering algorithms 16
2.2.2.3 DBSCAN clustering algorithms 16
2.2.3 Optimal number of clusters 17
2.2.3.1 Elbow method 17
2.2.3.2 The Silhouette Coefficient 17
2.3 Data Mining techniques 18
2.3.1 CRISP-DM 18
2.3.2 Data Mining Classification Algorithms 20
2.3.2.1 Decision Trees 20
2.3.2.2 Random Forests 21
2.3.2.3 Neural Network 22
2.3.2.4 Support Vector Machine 22

3 Standalone Fishery Analysis **23**
3.1 Fishing Velocity Patterns 24
3.2 Fishing Spots 30
3.3 SFA Library 32
3.3.1 Functionality 32
3.3.2 Architecture and Implementation 33
3.3.3 Deployment 35

4 Joined Fishery Analysis **37**

4.1 Implementation of CRISP-DM 37

4.1.1 Business Understanding 37

4.1.2 Data Understanding 37

4.1.3 Data Preparation 38

4.1.4 Modeling 43

4.1.5 Evaluation 44

4.1.6 Deployment 50

5 Experimental Evaluation **51**

5.1 Validation of Standalone Fishery Analysis 51

5.1.1 Validation and evaluation 52

5.2 Validation Joined Fishery Analysis 55

5.2.1 Validation and evaluation 56

6 Conclusions **57**

6.1 Overview 57

6.1.1 Conclusions about Standalone Fishery Analysis 58

6.1.2 Conclusions about Joined Fishery Analysis 58

6.2 Future Work 59

References **61**

List of Figures

1.1 Block Diagram of solution to goal 1. 8

1.2 Block Diagram of solution to goal 2. 8

2.1 The complete CRISP-DM Approach [37]. 19

3.1 MONICAP Blue Box. 24

3.2 SOG Histogram vessel 2. 25

3.3 Velocity histogram in nautical knots 26

3.4 Hill climbing algorithm step to find maximum 27

3.5 Hill climbing algorithm step to find minimum 27

3.6 Speed distribution after removing traveling occurrences 28

3.7 Estimated Kernel Density function 28

3.8 Estimated Cumulative Kernel distribution 29

3.9 Set margin and get limits 29

3.10 Vessel 2 geographic coordinates 30

3.11 Sum of squared error as function of the number of clusters 32

3.12 Representation of the SFALib architecture. 34

4.2 SOG minimum per license 39

4.1 SOG per license 39

4.3 SOG average per license 40

4.4 SOG maximum per license 40
4.5 The elbow method showing the optimal k 41
4.6 Silhouette analysis with 4 clusters 42
4.7 Silhouette analysis with 5 clusters 42
4.8 Silhouette analysis with 6 clusters 42
4.9 Confusion Matrix for Decision Tree using Entropy splitting criterion(C4.5) 46
4.10 Confusion Matrix for Random Forest using 200 estimators 47
4.11 Confusion Matrix for Neural Network using 5x500 hidden layers 48
4.12 Confusion Matrix for Support Vector Machine using Polynomial kernel coefficient 49

5.1 Representation of all VMS coordinates for vessel 2 52
5.2 Representation of VMS coordinates for vessel 2 with speeds inferior to 4 53
5.3 Representation of VMS coordinates for vessel 2, near Flores island, with speeds lower than 4 53
5.4 Representation of VMS coordinates for vessel 2 near Terceira island . . . 54
5.5 Confusion Matrix for Joined Fishery Analysis validation 56

List of Tables

1.1 VMS Records per vessel 4
1.2 VMS Records 5
1.3 VMS Vessels 6

3.1 Time of processing in miliseconds per model 31
3.2 The average value of the Silhouette coefficient for vessel 2 33

4.1 VMS Dataset 38
4.2 The average value of the Silhouette coefficient for all vessels 43
4.3 Confusion matrix for a two-class classifier 44
4.4 Cross-Validation results for Decision Trees models 46
4.5 Cross-Validation results for Random Forest models 47
4.6 Cross-Validation results for Neural Network models 48
4.7 Cross-Validation results for Support Vector Machine models 49

5.1 Precision per class and configuration 54
5.2 Recall per class and configuration 55
5.3 Confusion matrix for configuration with 0.001 for velocity 55

Acronyms

GDP Gross Domestic Product. 1
GPS Global Positioning System. 2
VMS Vessel Monitoring Systems. 2
UTC Coordinated Universal Time. 3
EGNOS European Geostationary Navigation Overlay Service. 3
COG Course Over Ground. 4
SOG Speed Over Ground. 4
HP Horse Power. 5
kW Kilowatt-hour. 5
SFA Standalone Fishery Analysis. 7
JFA Joined Fishery Analysis. 8

1

Introduction

This chapter introduces the motivation, context and goals of this work. Additionally, the formulation of the problem and the respective proposed solution is briefly described. For this purpose, a block diagrams is used to illustrate the proposed solution, identifying the input data and the results to be obtained by the applications. Finally, the structure of the work is presented, describing exactly the content of each chapter.

1.1 Motivation

The increased fishing activities mankind imposed on the marine ecosystems is a threat for the future sea economy and for the marine ecosystem's integrity [2].
Fisheries mapping is needed for implementing better ecosystem management and to secure a healthy marine population [3]. The fishing activity represents an important activity for the Portuguese economy. In the European Union, Portugal is the country that has the highest consumption of fish per person and the third worldwide needing 55,6 kg (per capita/year), [36]. Portugal is a country connected to the sea by its large coast, all along with its continental territory, almost all the coastal villages have a fishing community. An important issue of concern to the authorities is the occurrence of tax evasion that continues to cause damage to the Portuguese economy. In Portugal, the estimated tax evasion in all economic activities represents 21,9% of it's Gross Domestic Product (GDP) [6]. The use of statistical pattern recognition techniques to analyze the

data makes possible to identify who operates in the margin of the law more rapidly and methodically [32]. Reducing tax evasion will allow strong gains and potentially will develop the economy, making the fishing activity more fair for everyone involved in this activity.

MONICAP [41] is a monitoring system for the inspection of fishing using the Global Positioning System (GPS) for vessel location and Inmarsat-C [40] technology for satellite communications between ships and a ground control center, these devices commonly referred to as Blue Box. MONICAP was successfully introduced on the market by Xsealence [48]) and is currently installed or currently being installed on about 800 fishing vessels operating under the control of the authorities of Portugal, Spain, France, Ireland, and Angola. Within the scope of this Master's thesis, it is proposed to use Portuguese fishing data from the Vessel Monitoring Systems (VMS) to extract patterns of behavior related to the fishing zones, times, speeds, and directions of the course performed by the ships. The descriptive statistical analysis of these makes it possible to identify patterns of fishing activity that can be used for different proposes such as sustainable fishing, models for fuel efficiency, and models to detect illegal activities.

VMS provides a unique and independent method to derive patterns of spatially and temporally explicit fisheries activity. Such information may feed into ecosystem management plans seeking to achieve sustainable fisheries while minimizing potential risk to non-target species (e.g. cetaceans, seabirds, and elasmobranchs) and habitats of conservation concern. With multilateral collaboration, VMS technologies may offer an essential solution to quantifying and managing ecosystem disturbance, particularly on the high-seas.

1.2 Goals

Objective 1: Local tool
The first goal is to develop an application to be installed in the MONICAP, which allows better describing the fishing zones. At each one of the vessels, this application will work in real-time with the data from the fishing activity of each vessel. This tool will be local and will use unsupervised techniques of machine learning [30]. The derived application should be able to identify patterns in two strands:

- Speed: identify if the vessel is in fishing activity or not;
- Location: identify the usual fishing spots. These spots knowledge to cross-check in real-time if the current location of fishing is new to the vessel.

Objective 2: Centralized tool
Using VMS data and information on fishing licenses per vessel, our goal is to design models that are capable of classifying vessels by type of fishing only using VMS data. These models could be used to classify vessels by fishing activity, allowing crossing this classification with the information corresponding to the vessel license.

1.3 Data

1.3.1 VMS Records

VMS Data provided by the Xsealence [48] enterprise contains data generated by the MONICAP [41] "Blue Box". Information about the localization, direction, and velocity of the vessel, every 10 minutes is saved in a local database. VMS datasets contain a vessel identification code, a timestamp, the latitude and longitude positions, the speed and the direction. In this dataset, there are 769930 entries from thirty-eight vessels, between 2008-10-30 and 2016-11-04 Coordinated Universal Time (UTC). These data are from vessels operating in the Portuguese shore. This dataset is created automatically by the MONICAP system and follows the concept of integrity and confidentiality.

The variables registered in the dataset are:

- VesselID: Vessel identification;
- Utc: Date time of the log;
- Gps-id: identification of the GPS in use (0 = GPS with European Geostationary Navigation Overlay Service (EGNOS), 1 = MiniCs GPS);
- Fix/fix2: types of fix in the GPS:
  - 0 = invalid,
  - 1 = standard: valid, without integrity (without EGNOS),
  - 2 = differential: valid, with integrity (with EGNOS),
  - 3 = integrity: valid with integrity (with EGNOS);
- Lat/Lat2: latitude of GPS primary/secondary (in decimal);
- Lon/Lon2: longitude of GPS primary/secondary (in decimal);

- Course Over Ground (COG): Varies from 0 to 360 clockwise, being 0, facing north;

- Speed Over Ground (SOG): Velocity in knots;

Table 1.1 presents information, for each vessel, on the number of occurrences and the time range to which the records correspond.

Table 1.1: VMS Records per vessel

VesselID	Count	Time laps	VesselID	Count	Time laps
1	4767	2009-08-06 to 2016-12-04	20	6470	2014-07-06 to 2014-08-26
2	53465	2012-11-06 to 2014-12-09	21	3607	2009-10-26 to 2009-11-20
3	18887	2009-04-20 to 2009-09-13	22	65535	2014-02-21 to 2015-02-09
4	27970	2009-10-28 to 2010-09-24	23	46908	2015-05-28 to 2016-10-07
5	47870	2009-02-13 to 2014-10-14	24	2587	2012-08-24 to 2012-11-20
6	9071	2012-02-09 to 2012-04-22	25	23403	2013-10-09 to 2016-03-07
7	2050	2014-06-12 to 2015-07-31	26	23054	2012-04-11 to 2012-09-05
8	7192	2014-08-25 to 2014-10-06	27	1218	2011-05-27 to 2011-06-27
9	7577	2013-12-15 to 2014-06-02	28	29424	2008-10-30 to 2009-10-14
10	65530	2014-04-15 to 2015-02-09	29	6496	2013-04-26 to 2013-08-23
11	18949	2009-02-13 to 2010-11-23	30	41352	2012-09-25 to 2015-03-12
12	4367	2010-04-25 to 2010-06-01	31	51357	2015-01-27 to 2016-11-04
13	44476	2009-04-01 to 2010-10-04	32	4403	2015-10-08 to 2016-11-04
14	973	2010-02-15 to 2010-03-06	33	15722	2014-12-09 to 2015-07-15
15	3315	2010-04-09 to 2010-09-04	34	10315	2015-08-19 to 2016-06-02
16	290	2013-03-20 to 2013-04-08	35	17090	2016-06-05 to 2016-09-15
17	2632	2015-03-24 to 2015-04-13	36	16048	2015-04-27 to 2015-07-15
18	25516	2014-08-01 to 2015-02-09	37	3421	2015-11-30 to 2016-04-01
19	4441	2012-11-15 to 2013-02-20	38	52182	2015-02-08 to 2016-07-11

In Table 1.2 presents a descriptive summary of VMS records, considering all the vessels. Some cells of the table are empty due to some measures that do not apply to qualitative data. The P_{25} stands for the 1^{st} Quartile, P_{75} stands for the 3^{rd} Quartile and *SD* stands for Standard deviation.

Table 1.2: VMS Records

	Minimum	Maximum	Average	P_{25}	Median	P_{75}	SD
Lon	-52.706	35.965	-4.493	-9.811	-9.115	-7.986	21.156
Lat	-35.243	76.064	25.933	33.038	38.423	40.21	25.916
Sog	0	42	4.183	1.634	3.012	7.399	3.211
Cog	0	360	166.31	68.89	173.125	255.69	108.64
utc	2008-10-30	2016-11-04	-	-	-	-	-
gps_ id	0	1	-	-	-	-	-
fix	1	3	-	-	-	-	-
Fix2	0	2	-	-	-	-	-
Lon2	-52.706	155.977	-3.702	-9.726	-8.991	0	20.933
Lat2	-35.24	76.064	23.663	0	37.595	40.191	26.383
VesselId	1	38	-	-	-	-	-

1.3.2 VMS Vessels

The VMS records, as mentioned in subsection 1.3.1, includes observations from 38 vessels. The VMS vessels data comprises a total of 56 vessels. In order to enable an integrated analysis and also in accordance with the objectives of the present work, the common 38 vessels are considered in both registers.

VMS Vessels data is the vessel information that goes along with the VMS Records. These data contain information about vessels and fishing activities for which they are licensed. These data are created by the competent authority that process fisheries licensing.
The variables registered in the dataset are:

- ID: Vessel identification (VesselID/VMSRecords, foreign key);
- Name: Name of the vessel;
- Loa: Length Overall;
- GT: Gross Tonnage;
- HP: Vessel power Horse Power (HP);
- kW: Vessel power Kilowatt-hour (kW);
- License: Registration of the vessel's licenses;

- PriGearCode: FOA code of the principal fishery device;
- SecGearCode: FOA code of the secondary fishery device.

In Table 1.3 we can see the summary of the data regarding the vessels in the table VMS Records. Some cells of the table are empty due to some measures that do not apply to qualitative data.

Table 1.3: VMS Vessels

	Minimum	Maximum	Average	P_{25}	Median	P_{75}	SD
ID	1	38	-	-	-	-	-
Name	-	-	-	-	-	-	-
Loa	11.95	84.94	23.48	16.93	19.35	23.70	15.49
GT	22251	18.99	200.28	27.98	57.15	110.34	473.78
HP	3600	130	539	230	350	497	689.56
KW	2684.50	95.62	396.52	172.84	259.21	367.91	498.54
License	-	-	-	-	-	-	-
PriGC	-	-	-	-	-	-	-
SecGC	-	-	-	-	-	-	-

[1] P_{25} stands for 1st Quartile.

[2] P_{75} stands for the 3rd Quartile.

[3] *SD* stands for Standard deviation.

With regard to licensing for fishing, this data set considers the following fishing licenses:

Siege: The purse seine used on the mainland is characterized by the use of a catch at the bottom of the net - this allows the net to be closed like a bag in order to retain the catch.

Dragging:

- **Drag of Doors:** A bottom-trawl net towed by a single vessel, the horizontal opening of which is ensured by relatively heavy trawl doors, which may be fitted with a steel shoe designed to withstand a contact with the bottom.
- **Pole drag:** Rod trawling is characterized as a medium-sized trawl art where the mouth, devoid of wings, is held open by the action of two rods or a horizontal rod and rigid lateral structures.
- **Dredge:** Small and medium-sized trawling art in which the mouth is composed of a rigid structure and the bag is mesh or made up of a metal grid.

Gillnets and Trammel nets: Fishing method using a rectangular net with one, two or three rafts held upright by floatation cables and cables of used ballast insulated or in hunting.

Fishhook: A fishing method that uses lines and, in general, one or more hooks, ballasts and buoys. It can be practiced with gear that is integrated in the following groups: troll, cane and hand line, longline, tone and fishing nipple.

Traps:

- **Cage Traps:** Fishing method by which the prey is attracted or referred to a device that prevents leakage.
- **Shelter Traps:** Fishing method by which the prey is attracted or referred to a device, in this case the pots.

Sliding Enclosures: Fishing method that uses a net structure with pouch and large lateral wings that drag and, simultaneously or simultaneously, wrap or surround.

Catch: Uses several simple utensils. It can be practiced by an individual, using or not a support vessel and apnea diving equipment.

This data is available in https://www.dgrm.mm.gov.pt [35].

1.4 Problem Formulation and Solutions

1.4.1 Standalone Fishery Analysis

The proposed solution to meet objective 1 consists of a system that after receiving the VMS data from the Blue box returns information that indicates whether the vessel is fishing or not and if so, it also indicates that the vessel is fishing in a geographical area usual or in a new area. This system integrates an algorithm based on the adjustment of a probabilistic model to the observed data, using density estimates based on Kernel methods, hill climbing algorithm and Density Based Cluster.

In Figure 1.1 is presented the solution block diagram for the first purpose. We will call this solution Standalone Fishery Analysis (SFA).

1.4.2 Joined Fishery Analysis

The proposed solution to objective 2 is the creation of a model that can be used in a system that receives data from various vessels and classifies that data by the type of

Figure 1.1: Block Diagram of solution to goal 1.

fishing license. The objective is to understand whether the result of the classification algorithm is compatible or not with the vessel's fishing license. In this classification problem different methodologies are used, such as, Decision Trees, Random Forests, Neural Networks and Support Vector Machines algorithms. In Figure 1.2 is presented the solution blocks diagram for the second goal. We will call this solution Joined Fishery Analysis (JFA).

Figure 1.2: Block Diagram of solution to goal 2.

1.5 Thesis Structure

The remainder of this document is organized into six main chapters. Chapter 2 gives an overview of the State of the Art are presented the methods, approaches and tools of the state of the art and their main results and functionality. Reference articles are presented and the main basic methodologies used in each case are explained. In Chapter 3 we present two different methodologies to classify the data in real-time, only using the available data by the Blue Box:

- Using velocity data, to classify if the vessel is fishing;
- Using location data, to classify if the vessel is in activity in a new area.

For the solution the proposed approaches are presented in the form of an algorithm, clearly explaining the input and output parameters. All the actions of the applied techniques are described and presented, so that they can be reproduced by others. In chapter 4 we presented an approach to answer the second objective: using the data from all the vessels, how to identify fishing activities that are not under the vessel's fishing license. Different data mining methods will be used to derive predictive models. Corresponding results are compared through correct classification performance measures. Chapter 5 is presented with the results of the validation of the models and

methods presented to answer the proposed objectives. Chapter 6 contains the conclusions obtained during the elaboration of this work.

2

State of the Art

There exists a desire amongst the world's fisheries managers to coordinate their efforts so that the world's fish stocks - which recognize no national or regional boundaries - can be saved. (Food and Agriculture Organization of the United Nations, Rome, 1998)

2.1 Literature Review

2.1.1 Previous Work

2.1.1.1 U. C. report to the final project of course

My undergraduate final-year project entitled "Análise de Padrões para Encontrar Fraude nas Pescas" was developed in the same data analysis context. In that work I tried to solve an analogous problem with data coming from the VMS file, but with a different approach.
FPC work was focused on abnormalities regarding the declaration of fish caught, by quantities and type of fish. It was used the data provided by the Capitan with quantities caught per type of fish and used VMS Records data to consider standards, as the time of the year and fishing positions.

2.1.1.2 Published Paper

Fishing Monitor System Data: A Naïve Bayes Approach
Authors: Serge Lage, Iola Pinto, João Ferreira, Nuno Antunes
Book: Springer, Advances in Intelligent Systems and Computing volume 557
Date: 23 February 2017
DOI: 10.1007/978–3–319–43480–0—57
https://link.springer.com/chapter/10.1007/978–3–319–53480–0—57

In the paper we observed that it is possible to find patterns in the fishing data VMS and logbooks, to find outliers. The knowledge obtained about the VMS data will be used in this present work.

2.1.2 Work in Literature

There exists a desire amongst the world's fisheries managers to co-ordinate their efforts so that the world's fish stocks, which recognize no national or regional boundaries, can be saved. In order to do so, there must have to be an agreement concerning the procedures for implementing VMS. For example, when a South America fisheries manager agrees with a fisheries manager in Europe on VMS performance, security and data formats, it will be possible for a vessel to operate under the management of both, moving from one fishery to another, within legally and a maximum of transparency. Furthermore, only within such a context, can the two fisheries managers share data on vessel movements and activities, to improve operations on an international scale [14].
VMS is nowadays a standard tool of fisheries monitoring and control worldwide, but it was the EU that led the way, becoming the first part of the world to introduce compulsory VMS tracking for all the larger boats in its fleet. The EU legislation requires that all coastal EU countries should set up systems that are compatible with each other so that countries can share data and the Commission can monitor the respect of the rules. EU funding is available for the Member States to acquire state-of-the-art equipment and to train their people to use it [39]. If an international standard exists, the fisheries managers from all regions of the world would be able to set a common goal. However, there exists some consensus on VMS implementation, providing some welcome, but it will be temporary. This may not be enough to keep everyone on the same track but could be enough to keep them moving in the same direction.

There is some work being done using VMS data to reach very different objectives like:

- Illegal fishing: "Fishing Gear Recognition from VMS data to Identify Illegal Fishing Activities in Indonesia", [24]. The main propose of this study is to evaluate a novel method for the recognition of the fishing vessel gear type from VMS trajectories as a mean of detecting abnormal uses of undeclared fishing gear. The fishing gear list was trawl, longline, pole and line, purse seine. They reported mean correct recognition rates around 94.59%, which demonstrates the relevance of the proposed approach.;

- Illegal fishing: "Fishing Gear Identification From Vessel-Monitoring-System-Based Fishing Vessel Trajectories" [16]. The proposed approach combines the extraction of new VMS-derived features, issued from the nonsupervised identification and characterization of gear-specific movement patterns, and supervised machine learning, namely, random forest and support vector machine. They reach recognition performance greater than 97% for the considered Indonesian fisheries.

- Fuel efficiency: "Effects of fishing effort allocation scenarios on energy efficiency and profitability: An individual-based model applied to Danish fisheries",[5]. Using VMS data and data from logbooks create models to evaluate three scenarios. (A) preferring nearby fishing grounds rather than distant grounds with potentially larger catches and higher values, (B) shifting to other fisheries targeting resources located closer to the harbour, and (C) allocating effort towards optimising the expected area-specific profit per trip. The outcomes of scenarios A and B indicate a trade-off between fuel savings and energy efficiency improvements when effort is displaced closer to the harbour compared to reductions in total landing amounts and profit. Scenario C indicates that historic effort allocation has actually been sub-optimal because increased profits from decreased fuel consumption and larger landings could have been obtained by applying a different spatial effort allocation.;

- Sustainable fishing: "The importance of scale for fishing impact estimations",[26]. This study focused in the impact of a bottom trawl fishery on fish or benthos. This study shows the implication is that to determine the fishing-induced mortality of a particular species, the trawling frequency needs to be determined at those spatio-temporal scales that are appropriate considering the species' spatial processes (e.g., dispersion) or temporal processes described by life history characteristics.;

In terms of tools developed to analyze VMS data, we have two applications (VMStools and VMSbase).

- VMStools: is a package of open-source software, build using the freeware environment R, specifically developed for the processing, analysis, and visualization of landings (logbooks with information of the caught fish) and vessel location data (VMS) from commercial fisheries. Embedded functionality handles erroneous data point detection and removal, linking logbook and VMS data together to distinguish fishing from other activities, provide high-resolution maps of both fishing effort and landings, interpolate vessel tracks, calculate indicators of fishing impact as listed under the Data Collection Framework at different Spatio-temporal scales [12].

- VMSbase: is a R package derived to manage, process, and visualize information about fishing vessel activity (provided by the vessel monitoring system - VMS) and catches/landings (as reported in the logbooks). Standard analyses comprise: 1) tier identification (using a modified CLARA clustering approach on Logbook data or Artificial Neural Networks on VMS data); 2) linkage between VMS and Logbook records, with the former organized into fishing trips; 3)discrimination between steaming and fishing points; 4) computation of spatial effort concerning user-selected grids; 5) calculation of standard fishing effort indicators within Data Collection Framework; 6) a variety of mapping tools, including an interface for Google viewer; 7) estimation of trawled area [29].

The main difference between the present work, and those previously mentioned is that they combine VMS data with the logbooks (data of the type of fish captured and quantity) or the number of fishing licenses considered. In this work, we will use only VMS data. The main advantage is that VMS data is less subject to malicious changes than logbooks, taking into account that logbooks are filled by the ship owner. So, they are subject to misrepresentation of the truth. VMS data is generated automatically in a closed system like a black box. In addition, another difference between the present work and those already mentioned is that the 1st objective -to derive an application to be installed in the MONICAP, which allows better describing the fishing zones is not a topic addressed in the literature.

2.2 Data Analysis Methodologies

2.2.1 Distribution Fitting algoritms

2.2.1.1 Kernel density estimation

A density estimator is an algorithm which takes a D-dimensional dataset and produces an estimate of the D-dimensional probability distribution which that data is drawn from. A possible approach could be the Gaussian mixture models (GMM). This algorithm accomplishes representing the density as a weighted sum of Gaussian distributions.

Particularly, Kernel Density Estimation (KDE) is in some senses an algorithm which takes the mixture-of-Gaussians idea to its logical extreme: it uses a mixture consisting of one Gaussian component per point, resulting in an essentially non-parametric estimator of density. The free parameters of kernel density estimation are the kernel, which specifies the shape of the distribution placed at each point, and the kernel bandwidth, which controls the size of the kernel at each point. In practice, there are many kernels to use for a kernel density estimation: in particular, the Scikit-Learn KDE implementation supports one of six kernels, which you can read about in Scikit-Learn's Density Estimation documentation [18].

2.2.1.2 Hill-Climbing algorithm

Hill Climbing is a heuristic search used for mathematical optimization problems in the field of Artificial Intelligence. Given a large set of inputs and a good heuristic function, the algorithm tries to find the best possible solution to the problem in the most reasonable time period. This solution may not be the absolute best (global optimal maximum) but it is sufficiently good considering the time allotted[22].

The definition above implies that hill-climbing solves the problems where we need to maximize or minimize a given real function by selecting values from the given inputs.

2.2.2 Clustering algorithms

2.2.2.1 K-means clustering algorithm

K-means clustering algorithm [17] is a method of cluster analysis which aims the partition of n observations into k clusters, in which each observation belongs to the cluster

with the nearest mean. This results in a partitioning of the data space. K-means (Macqueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assuming k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed carefully because of different location causes a different result. So, the best choice is to place them as much as possible, far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed, and an early group is done. At this point, we need to recalculate k new centroids as bar centers of the clusters resulting from the previous step. After we have these k new centroids, a new binding must be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop, we may notice that the k centroids change their location, step by step, until no more changes are done.

2.2.2.2 Density-based clustering algorithms

Density-Based Clustering refers to unsupervised learning methods that identify distinctive groups/clusters in the data, based on the idea that a cluster in a data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density. The data points in the separating regions of low point density are typically considered noise/outliers. It can find arbitrarily shaped clusters, and handles noises, and yet is a one-scan algorithm that needs to examine the raw data only once. In density-based clustering algorithms, dense areas of objects in the data space are considered as clusters, which are segregated by low-density area (noise). The basic idea of density-based clustering is that clusters are dense regions in the data space, separated by regions of lower object density . The key idea of density-based clustering is that for each instance of a cluster, the neighborhood of a given radius (Eps) must contain at least a minimum number of instances (Min Pts).

2.2.2.3 DBSCAN clustering algorithms

DBSCAN (for density-based spatial clustering of applications with noise) is a data clustering algorithm proposed by Martin Ester, Hans-Peter Kriegel, Jorge Sander and Xiaowei Xu in 1996. It is a density-based clustering algorithm because it finds several clusters starting from the estimated density distribution of corresponding nodes. DBSCAN [19] is one of the most common clustering algorithms and most cited in the scientific literature.

2.2.3 Optimal number of clusters

2.2.3.1 Elbow method

The results the elbow method [20], for the within-cluster sum of squares. The within-cluster sum of squares refers to the distance between the vectors in each cluster are from their respective centroid. The goal is to get this number as small as possible. One approach to handle such an objective is to run the K-means clustering multiple times, raising the number of the clusters each time. Then, it is possible to compare the within-cluster sum of squares each time, stopping when the rate of improvement drops off. The better case corresponds to find a low withins while still keeping the number of clusters as small as possible.

The elbow method is visual. The idea is to start with K=2 and keep increasing it in each step by one unit, calculating the clusters and the cost that comes with the training. At some value for K, the cost drops dramatically, and after that, it reaches a plateau when you increase it further. At this moment, the value of K we are looking for is reached.

2.2.3.2 The Silhouette Coefficient

The Silhouette Coefficient is a metric based on the separation and compacting of the groups, starting this procedure by calculating the Silhouette index for the i th observation as in equation 2.1.

$$s(i) = \frac{b(i) - s(i)}{max\{a(i), b(i)\}} \tag{2.1}$$

where a (i) is the average distance between i^{th} observation and all other observations within the same group. For the calculation of b (i) a distance is first calculated between the observation i and the observations belonging to a group in which i is not inserted. After this calculation, such an average is carried out in those distances. An identical calculation is made for all groups to which the observation does not belong. At the final, b (i) represents the minimum distance between the calculated average distances [27].

The denominator of equality (2.4) only serves to normalize the result, so the values s (i) are represented between [-1,1], where -1 or negative values refers to observations wrong placed in the group, whereas for coefficients with a value of 1 or positive represent observations well placed in the group [27]. To obtain the performance index

for a given number of groups in each indicated group, the average of all indexes of Silhouette is calculated in equation 2.2.

$$SWC = \frac{1}{N}\sum_{j=1}^{N} s(j) \tag{2.2}$$

Finally, in order to decide which is the optimal number of groups to use, depending on the data, the criterion of choice favors the scenario that corresponds to the highest average value of the Silhouette coefficient [27].

2.3 Data Mining techniques

Data mining is the process of discovering interesting and useful patterns and relationships in large volumes of data. The field combines tools from statistics and artificial intelligence (such as neural networks and machine learning) with database management to analyze extensive digital collections, known as data sets. Data mining is widely used in business (insurance, banking, retail), science research (astronomy, medicine), and government security (detection of criminals and terrorists) [25].

2.3.1 CRISP-DM

In this work, it will be used the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology [9]. The CRISP-DM project proposed a comprehensive process model for carrying out data mining projects. The process model is independent of both the industry sector and the technology used [9]. The CRISP-DM reference model for data mining provides an overview of the life cycle of a data mining project. It contains the phases of a project, their respective tasks, and their outputs. The life cycle of a data mining project is broken down into six phases, which are shown in Figure 2.1. The sequence of the phases is not strict. The arrows indicate only the most important and frequent dependencies between phases, but in a particular project, it depends on the outcome of each phase, which phase, or which particular task of a phase, has to be performed next.

Figure 2.1: The complete CRISP-DM Approach [37].

In the following, we outline each phase briefly:

- Business Understanding
  This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, thus a preliminary project plan designed to achieve the objectives are drown.

- Data Understanding
  The data understanding phase starts with an initial data collection and proceeds with activities to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information. There is a close link between Business Understanding and Data Understanding. The formulation of the data mining problem and the project plan require at least some understanding of the available data.

- Data Preparation
  The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include a table, record, and attribute selection, data cleaning, construction of new attributes, and transformation of data for modeling tools.

- Modeling
  In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques require specific data formats. There is a close link between Data Preparation and Modeling. Often, one realizes data problems while modeling, or one gets ideas for constructing new data.

- Evaluation
  At this stage in the project, you have built one or more models that appear to have high quality, from a data analysis perspective. Before proceeding to the final deployment of the model, it is essential to more thoroughly evaluate the model, and review the steps executed to construct the model, to be sure it accurately achieves the business objectives. A key objective is to determine if there is some critical business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

- Deployment
  The creation of the model is generally not the end of the project. Usually, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases, it will be the user, not the data analyst, who will carry out the deployment steps. In any case, it is important to understand upfront what actions will need to be carried out to make use of the created models.

2.3.2 Data Mining Classification Algorithms

2.3.2.1 Decision Trees

While in data mining, a decision tree is a predictive model that can be used to represent both classifiers and regression models. In operations research decision trees refer

to a hierarchical model of decisions and their consequences[28]. The decision-maker employs decision trees to identify the strategy which will most likely reach its goal. When a decision tree is used for classification tasks, it is most commonly referred to as a classification tree. When it is used for regression tasks, it is called a regression tree [28]. Algorithms for constructing decision trees usually work top-down, by choosing a variable at each step that best splits the set of items [4].
Thus, when the target variable is a discrete set of values the model is called a classification tree; In the tree structure, leaves represent class labels and branches represent features conditions corresponding to those class labels. Tree models where the target variable assumes continuous values are called regression trees. ID3,CART and C4.5 are basically most common decision tree algorithms in data mining. They use different splitting criteria for splitting the node at each level to form a homogeneous node. In this work, it will be used to measure the quality of a split "gini" 2.3 for the Gini impurity (CART)[13] and "entropy" 2.4 for the information gain (C4.5)[13].

$$Gini = 1 - \sum_{i=1}^{n} (P_i)^2 \tag{2.3}$$

Where P_i denotes the probability of an element being classified for a distinct class.

$$Entropy = \sum_{i=1}^{n} -p_i log_2(p_i) \tag{2.4}$$

Where p denotes the probability that it is a function of entropy.

2.3.2.2 Random Forests

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest[7]. The generalization error for forests converges a.s. To a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. A random forest is a classifier consisting of a collection of tree-structured classifiers $\{h(\mathbf{x}, \theta_k), k = 1, ...\}$ where the $\{\theta_k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input $\mathbf{x}$[7]. The number of trees in the forest are called estimators, which the algorithm builds before taking the maximum voting or taking the averages of predictions. In general, a higher number of trees increases the performance and makes the predictions more stable, but it also slows down the computation.

2.3.2.3 Neural Network

Neural networks are a bio-inspired mechanism of data processing that enables computers to learn technically similar to a brain and even generalize once solutions to enough problem instances are tough [21]. An artificial neural network consists of simple processing units, the neurons, and directed, weighted connections between those neurons. A neural network is a sorted triple (N, V, ϖ) with two sets N, V and a function ϖ, where N is the set of neurons and V a set $\{(i, j)i, j \in \mathbb{N}\}$ whose elements are called connections between neuron i and neuron j. The function $\varpi : V \rightarrow \mathbb{R}$ defines the weights, where $\varpi(i, j)$, the weight of the connection between neuron i and neuron j, is shortened to $\varpi_{i,j}$. Depending on the point of view, it is either undefined or 0 for connections that do not exist in the network [21].
In this work, we will train models with different hidden layer sizes. The solver for weight optimization used is BFGS[11] and Adam[1] for large sizes of hidden layers.

2.3.2.4 Support Vector Machine

The folklore view of SVM is that they find an optimal hyperplane as the solution to the learning problem. The simplest formulation of SVM is the linear one, where the hyperplane lies in the space of the input data x. In this case, the hypothesis space is a subset of all hyperplanes of the form: $f(x) = w \cdot x + b$. In their most general formulation, SVM finds a hyperplane in a space different from that of the input data x. It is a hyperplane in a feature space induced by a kernel K (the kernel defines a dot product in that space)[33].
In this work, we will create models using the following kernel algorithms: Linear[34], Polynomial[34], and RBF[34].

3

Standalone Fishery Analysis

This chapter explains the approach used to reach the first goal of this work. It describes an application that implements the work done in this chapter with respect to its functionality, architecture, implementation details and usage.
In Section 4.1 we learn how can we obtain the fishing velocity patterns.
In Section 4.2 we learn how can we obtain the fishing spots patterns.
In Section 4.3 we demonstrate an approach to address goal 1.

The first objective is to develop a locally implemented tool that registers whether the vessel is engaged in fishing, and if so, whether the fishing area is new or is habitual. This solution must be implemented by the vessel. Therefore each vessel will only have access to its data. That is, each vessel only will know its own data. This data consists of VMS data described in Section 1.3.1 VMS Records.

The solution developed to meet this objective consists of a machine learning application to analyze data in real-time. Doing this analysis is done by vessel allows avoiding bias in the results, since each vessel has a different power, size and its suitable for a specific fishing activity.
This solution could be implemented and used as a library by the MONICAP [41] system shown in Figure 3.1.

Figure 3.1: MONICAP Blue Box.

As MONICAP systems are installed on ships, they can, in real-time, send alerts to the authorities whenever an abnormal change is detected concerning the standard.

3.1 Fishing Velocity Patterns

To know whether a vessel is fishing, we can use its velocity patterns, given that the speed of the vessel differs when it is traveling or when it is fishing. We can verify this fact in the histogram shown in Figure 3.2, corresponding to a vessel velocity.

In Figure 3.2, the histogram allows us to recognize two different velocity patterns, identified by two distinct distributions. They are visible when we graphically represent the velocity's data of each of the vessels. The distribution characterized by lower average speeds corresponds to fishing activity, and the other speed distribution corresponds to the movement of the vessel between the port and the fishing sites [23].
So, it is needed to isolate the first distribution's range to be able to classify the upcoming future velocity's as being fishing associated or not.

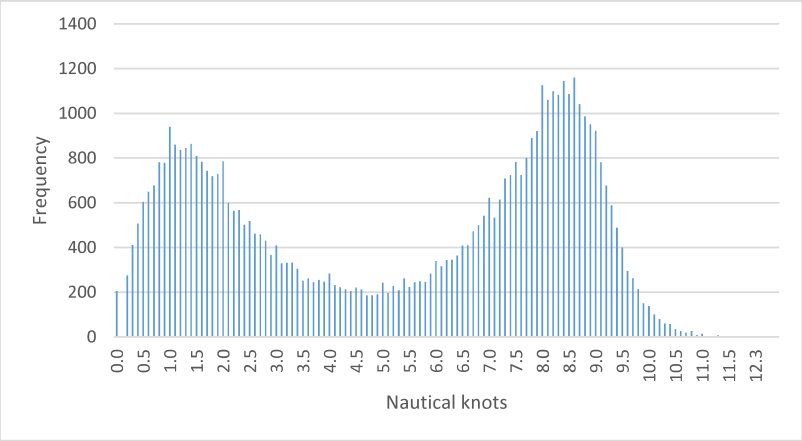

Figure 3.2: SOG Histogram vessel 2.

For the purpose of isolating the first distribution we considered and analyzed the potential of two different procedures:

- **Kernel Density Estimation:** Method explained in Sub Chapter 2.2.1.1.

  The implementation used was KernelEstimator from the WEKA library [47].

- **Filter:** Use a Hill-Climbing algorithm explained in Sub Chapter 2.2.1.2. With this algorithm, the first maximum is found, and then the algorithm identifies the next minimum. Then remove all the velocity occurrences that happen to be less than 10 % of the maximum occurrence and isolating the occurrences that are followed. A clean distribution of the fishery speed for each vessel is derived. With this, the minimum and the maximum values of this distribution are used to classify the new inputs.

After the experiment and the study of all these different methods, the chosen procedure can be described in two steps: it starts by using the method based on the **Filter** to isolate the fishery speed occurrences from the remaining ones. Then ,the Kernel distribution method was applied.

1. **Filter:** In the first step, it retrieves all velocity data from the database to create a histogram like it is shown in Figure 3.3.

   In the next step, all observations corresponding to zero velocities were removed, because we do not want to consider when the vessel is completely stopped. Then it uses the hill-climbing algorithm to get the minimum and the maximum value

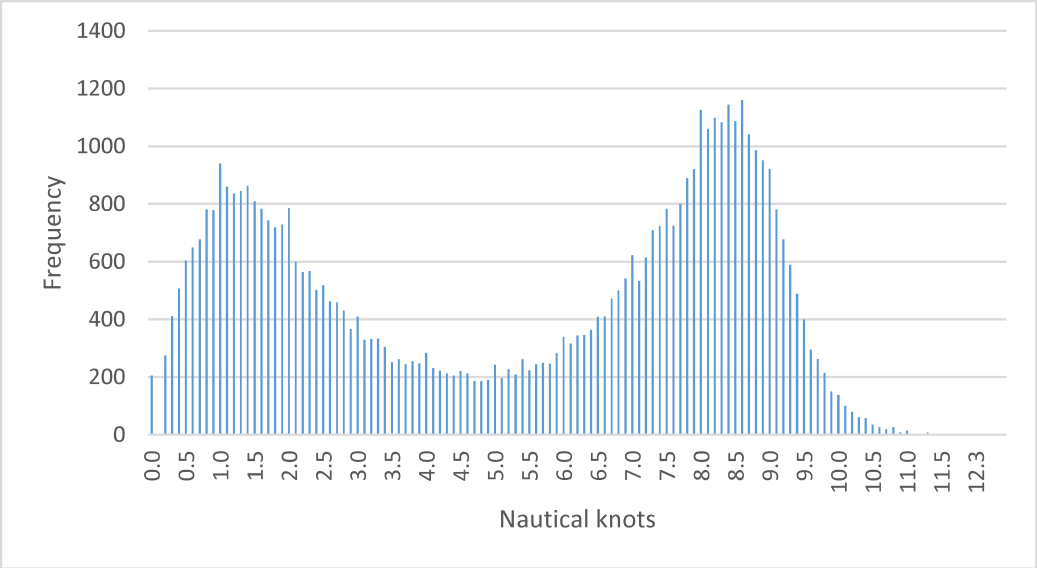

Figure 3.3: Velocity histogram in nautical knots

of the first distribution. The implementation of the hill climbing algorithm used consisted of identifying the maximum, then continuing the search, to determine the distribution limit. To obtain this solution, the algorithm searches for the first local maximum that has not have a higher value in the following three points, as shown in Figure 3.4.

In this way, we can find the maximum value of the fishing speed range.
To find the end of the fishing speed range, the algorithm continues to sweep the histogram until the next three points are not lower than the current point as shown in Figure 3.5.

This way, we can end up with a histogram of the intended distribution, as we can observe in Figure 3.6.

2. **Kernel method :** It was applied a kernel distribution method in the filtered histogram to have the distribution represented in orange on 3.7. Then, it was created a dictionary with the velocities and the cumulative percentage of velocity. This way, we end up with a representation like the one presented in Figure 3.8. Then, a range across quantiles is defined for some probability. Considering this last distribution, a confidence area is defined through a probability. The speeds within this area correspond to fishing activity. Thus, two-speed limits are identified and used to classify the new data. For the example in Figure 3.9, the output corresponds to the following values: minimum= 0.6, maximum= 3.5.

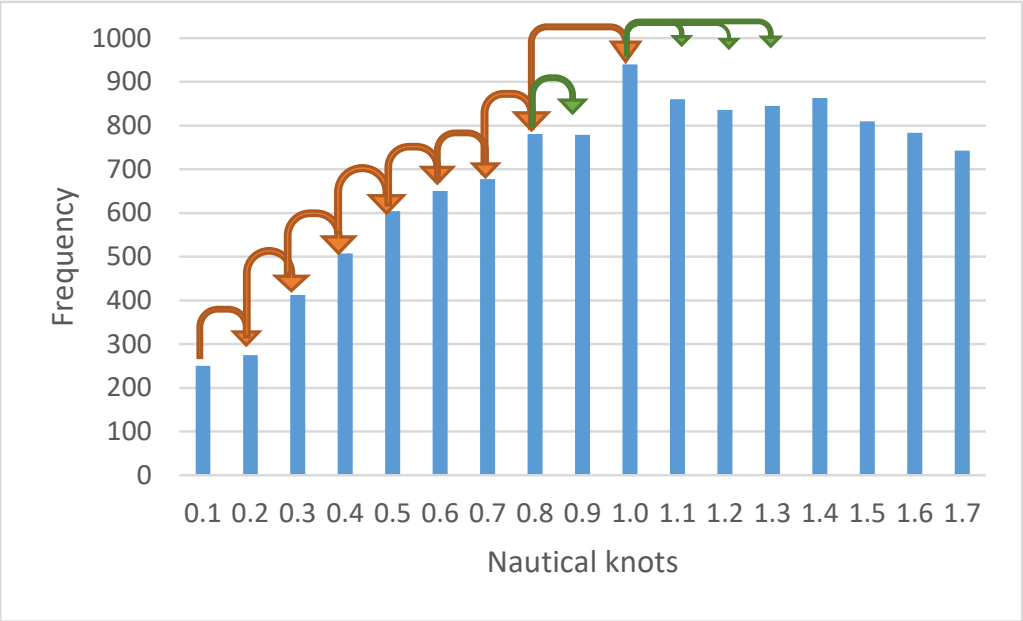

Figure 3.4: Hill climbing algorithm step to find maximum

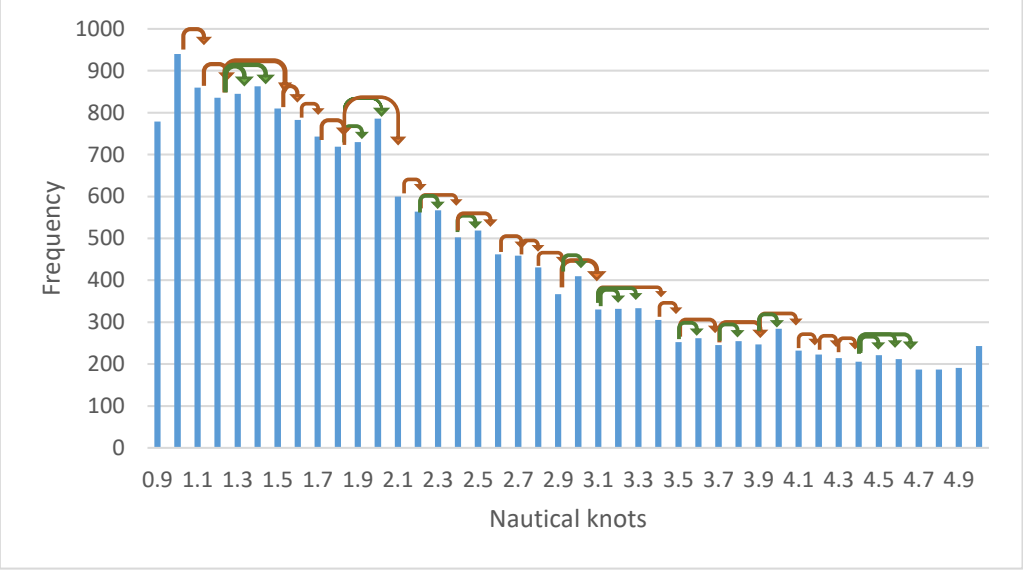

Figure 3.5: Hill climbing algorithm step to find minimum

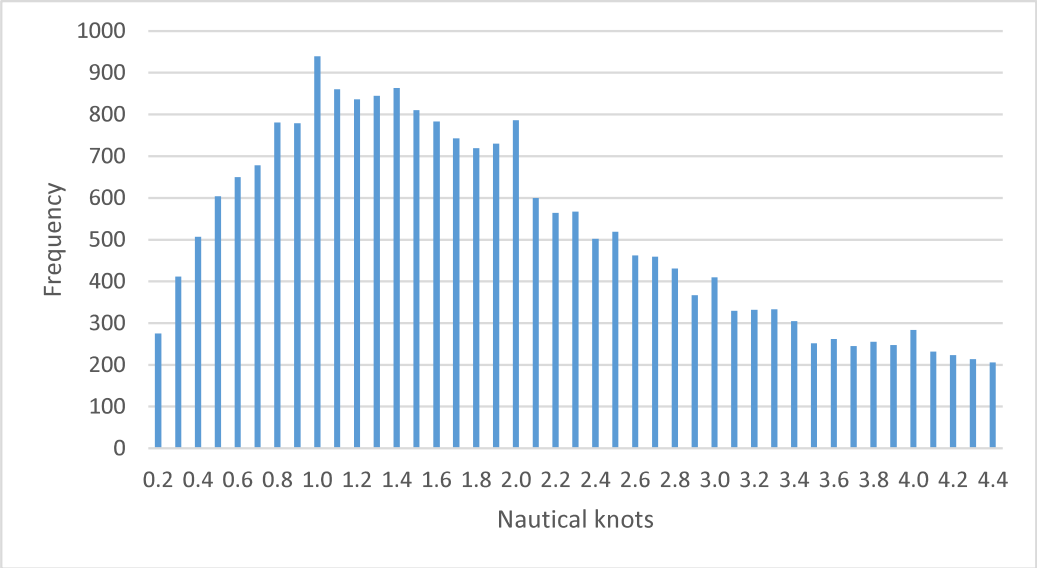

Figure 3.6: Speed distribution after removing traveling occurrences

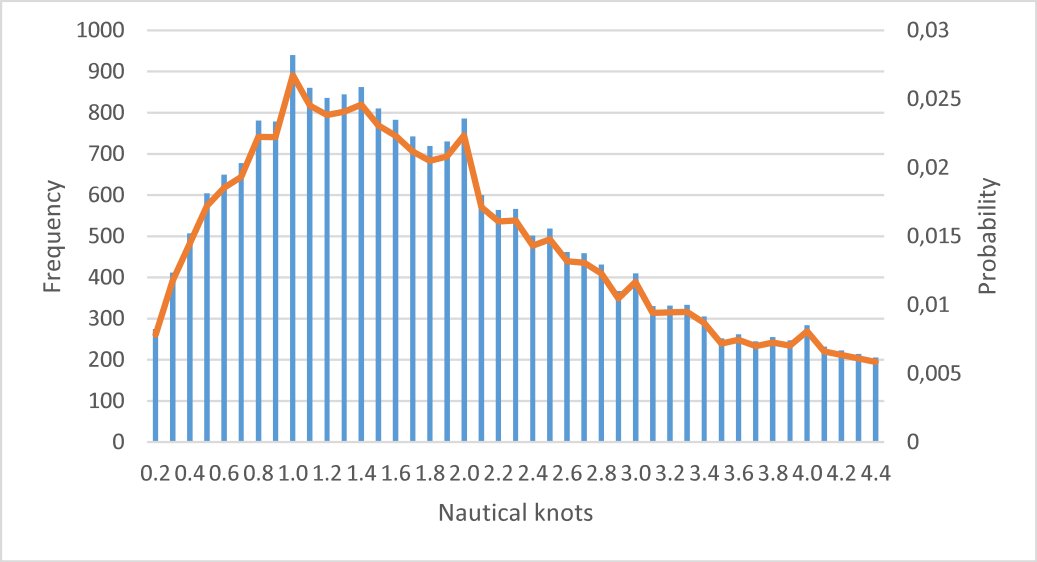

Figure 3.7: Estimated Kernel Density function

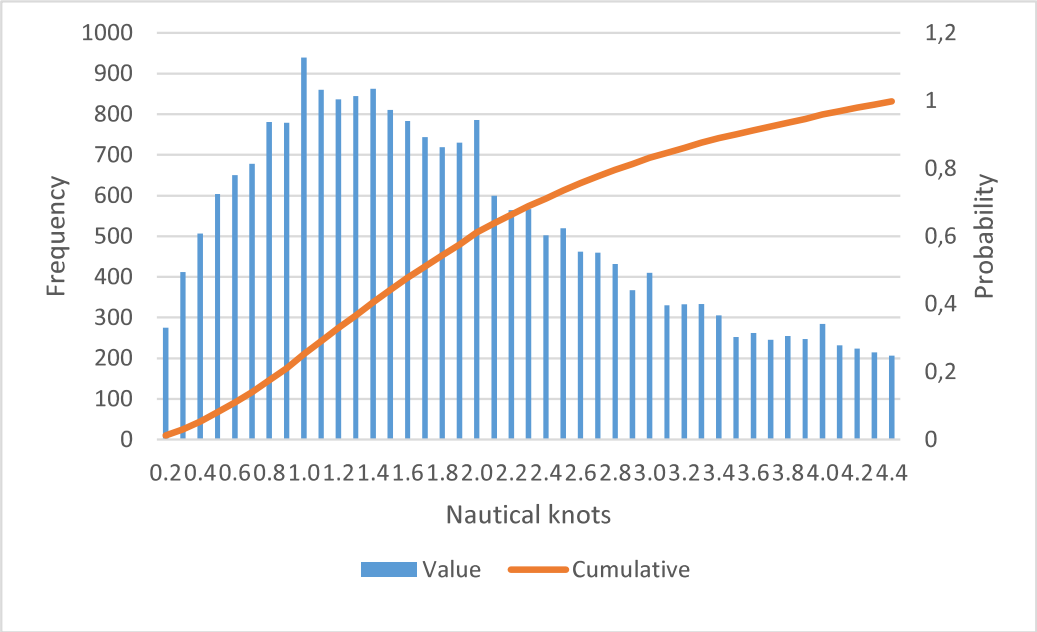

Figure 3.8: Estimated Cumulative Kernel distribution

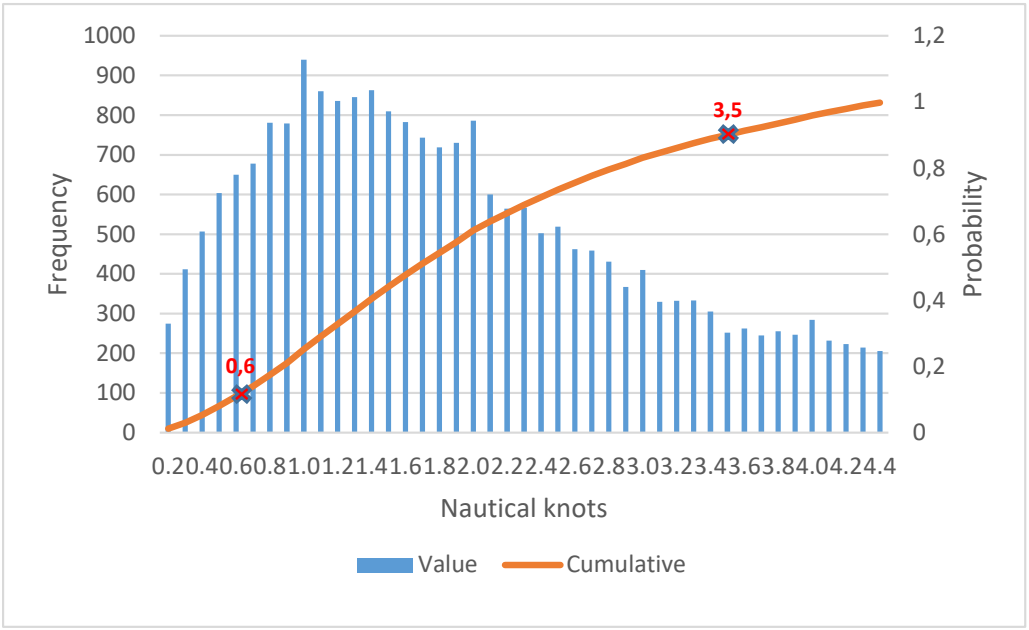

Figure 3.9: Set margin and get limits

Now, we can compare the new data with the established limits. If the new data is within limits, we classify as fishing, and if not, we classify as not fishing.

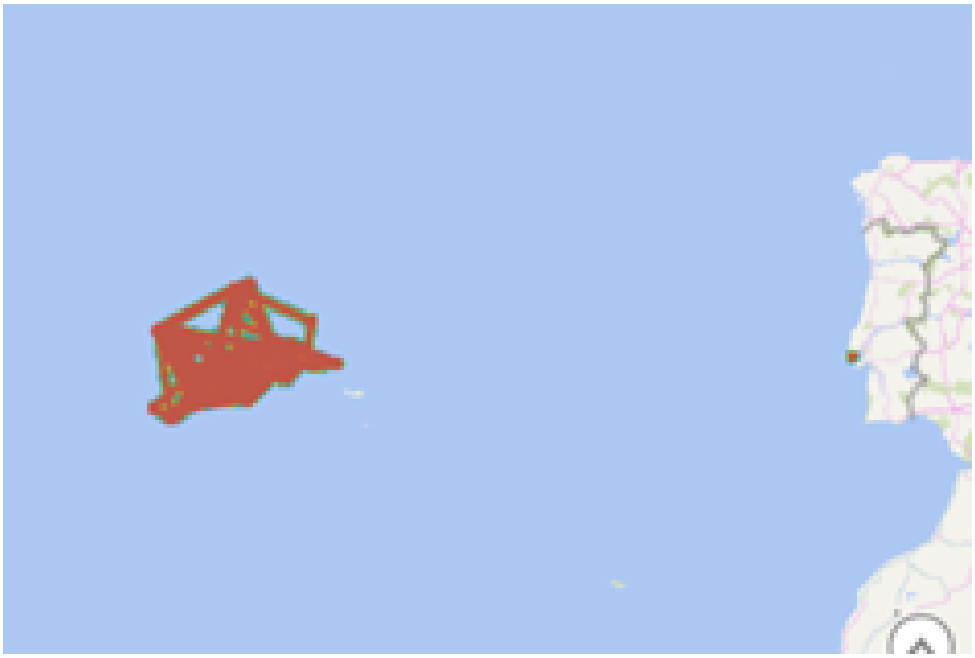

Figure 3.10: Vessel 2 geographic coordinates

3.2 Fishing Spots

To discover whether the vessel is fishing in its fishing zone, or in a new location, the history of GPS locations by vessel was used. Fishing in a new zone may mean that the vessel has changed its type of fishing or is engaging in an activity that is not licensed.

In Figure 3.10, we can see the geographic coordinates points for the vessel 2. Using methods based on clustering, it is possible to identify, by vessel, several areas that are the standard fishing zones of this vessel. When the vessel is outside that standard zone, a flag should occur.

Using the fishing velocity range encountered in the previous point, we get the GPS points of the vessel within that range, so we can work only with the positions where the vessel was fishing. The next step is to use a clustering algorithm to define the fishing areas so that we can compare it with the new GPS points.

For this purpose, several data mining algorithms were performed in order to choose the best results:

- **K-Means:** Method explained in Sub Chapter 2.2.2.1. The implementation used was SimpleKMeans from WEKA library.
- **Density Based Cluster:** Method explained in Sub Chapter 2.2.2.2. The implementation used was MakeDensityBasedClusterer from WEKA library.

Table 3.1: Time of processing in miliseconds per model

	K-Means	**Density Based Cluster**	**DBSCAN**
Initializing	862	923	25848
New data	25	45	35

- **DBSCAN:** Method explained in Sub Chapter 2.2.2.3. The implementation used was DBSCAN from WEKA library.

After some tests, it was decided that Density-Based Cluster is the best approach for this case. It was excluded DBSCAN, because as we can observe in Table 3.1, this model needs much processing power to estimate the clusters. These values were retrieved using a computer with an Intel i5 (2.5 GHz) processor and 8 GB of RAM. Considering that the Blue Box has a lot less processing power, it was decided that this model is not a good solution for this problem.

The choice between K-Means and Density-Based Cluster algorithms was based on the fact that Density-Based Cluster represents a great advantage because it estimates the probability of the new geographic coordinates belonging to a cluster-based in the specific cluster probabilistic distribution. This way, the user can choose the most suitable configuration. The resulting clusters themselves are equal when K-Means or Density-Based Cluster were applied since Density-Based Cluster uses K-Means to define the centroids, so they only differ by adding a layer to define the area of density per cluster.

One of the important steps when performing the cluster analysis is the determination of the number groups. To fix the number of groups we can use some indicators such as: coefficient of Silhouette as seen in Sub Chapter 2.2.3.2, the coefficient of Davies Bouldin, the coefficient of Calinski Harabasz and the Elbow method as seen in Sub Chapter 2.2.3.1.

In Figure 3.11 is presented the value of the within sum of squares as function of the number of clusters, using the geographic coordinates data for vessel 2. As we can observe, six clusters seems to be a good number as the error is not decreasing much as the number of clusters increases.

The average value of the Silhouette coefficient for vessel 2 is as described in Table 3.2 with nine and six having good values.

Considering the two methods I chose six as the number of clusters for the vessel 2.

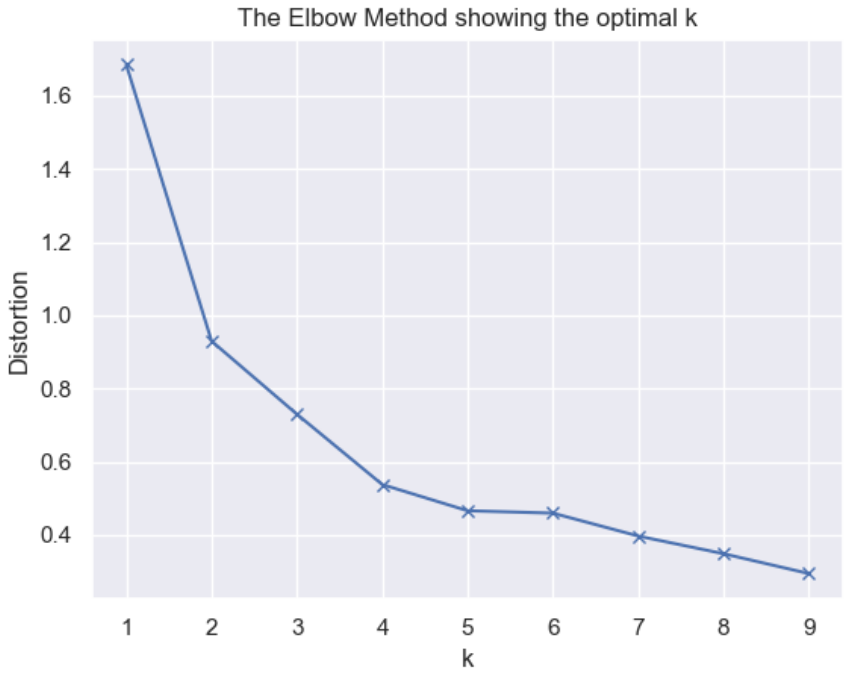

Figure 3.11: Sum of squared error as function of the number of clusters

3.3 SFA Library

It was created a software application called SFALib for (Standalone Fishery Analysis Library that can be find in a online repository [31]. In this application, were developed the solutions described in this chapter to help with the elaboration and tests for this project. For a market solution, this library could be used by the main application of a Blue Box, to send alerts to support decision making or to simply classify each VMS data entry into two categories:

1. Is fishing (yes/no).
2. Is fishing in a new area (yes/no).

3.3.1 Functionality

This application allows to:

- Test new data
  Send and receive VMS data, if it is considered to be fishing and if it is in a new area.

Table 3.2: The average value of the Silhouette coefficient for vessel 2

Number of clusters	Average silhouette coefficient
2	0.5933
3	0.5950
4	0.5510
5	0.6233
6	0.6635
7	0.6319
8	0.6514
9	0.6907

- Test new velocity
  Send sog data and receive true if it is considered to be fishing.
- Test new location
  Send GPS data and receive true it is in a new area.
- Restart models
  Request to create new models. It can be used if the objective is running for a long time and want to renew models with new data.
- Get limits
  Request the velocity limits. Receive a tuple with two doubles (item1 = low-speed limit, item2 = high-speed limit). It can be used for analysis like described in Chapter 4.

To make this possible, it is necessary to configure the data access layer to get the VMS data from the local data repository. The data repository used in this work to save and access VMS data was a database but could be other type like text files. The type of data repository used by MONICAP was not revealed to me. Currently, the application supports connection to SQL Server [46] and PostgreSQL [43].

3.3.2 Architecture and Implementation

To develop the software, it was decided to use Java 8 [42] because it is a powerful, full object-oriented, and cross-platform programming language. MONICAP uses Linux, so using a JRE (Java Runtime Environment) application is a good choice. The architecture is depicted in Figure 3.12. In this architecture, it is possible to distinguish three main modules: One that is the core of the SFALib, create the modules and use them. Another

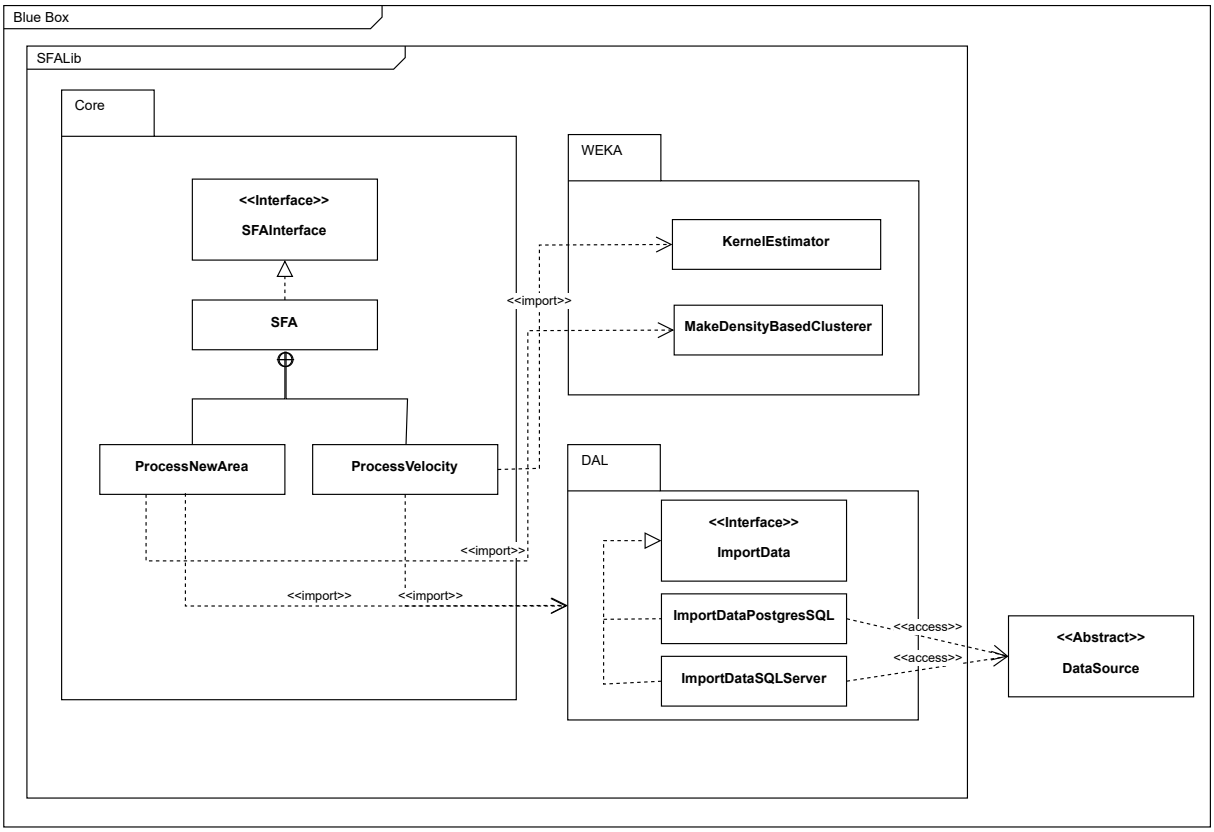

Figure 3.12: Representation of the SFALib architecture.

one is WEKA [47] that creates the cluster modules for locations and implements the Kernel density estimation for velocity. The last one is the data access layer that is responsible for getting the VMS data from the local repository so SFALib can create the modules.

Core module

The core module is responsible for initializing the models and using them the way described in this chapter. The procedure starts with the creation of an instance, called "ProcessVelocity", whose objective is to use the historical speed data of the ship to classify whether or not it is fishing. After that, is created another instance called, "ProcessNewArea", whose objective is to identify if a new data (geographic coordinates) is or is not in a usual fishing location for that vessel, using for that purpose the history of the vessel's GPS locations.

WEKA module

WEKA is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization. In this application, WEKA is used as a tool to create the

modules.

DAL module
The data access module was implemented in a way not only to get data but also to filter data in the database engine. Filtering data (where) is optimized on the database engine, and so we gain some performance.

The application starts by initializing two objects:

1. ProcessVelocity: This object is responsible for doing the process explained in section 4.2 of this document. It will request to the static class ImportData to retrieve all SOG (Speed Over Ground) data from the database. Then, the process will end with the limits (minimum speed of fishing and the maximum speed of fishing).

2. ProcessNewArea: This object is responsible for doing the process explained in section 4.3 of this document. This object is only initialized after ProcessVelocity because it needs the velocity fishing limits necessary to create the clusters of the fishing areas. With these limits, the object requests to the ImportData instance to obtain the latitude and longitude values where the vessel was in between the velocity limits. With this, the object ends up with the clusters of the fishing areas.

3.3.3 Deployment

We start by initializing SFALib with the doubles limitVelocity and limitArea. These doubles range between 0 and 1:

- limitVelocity: used to get the maximum and minimum speed by reducing the speed range. This limit will reduce the maximum speed and increase the minimum speed by setting the maximum velocity as the velocity that as (1-limit) percentage of the cumulative kernel distribution and the minimum velocity as the velocity that has a (limit) percentage of the cumulative kernel distribution.

- limitArea: used to compare with the probability to belong in a cluster given to the new points. If the limit is smaller than the given probability, then the vessel is classified as fishing in a new area.

These limits could be defined by the user. This possibility allows to configure the application according to the preference in obtaining more false positive or false negative classifications. A false positive (type I error) is when the classifier rejects a true hypothesis. A false negative (type II error) is when the classifier accepts a false hypothesis.

After the SFALib is ready, we need to send a new velocity data and GPS coordinates to receive an object with an "isFishing" as true if the vessel is fishing. "isNewArea" as true if the vessel is in an area that is not a normal fishing area and it's in a fishing velocity.

The methods that can be used are:

- newData: method that receives VMS data and returns isFishing(boolean) and isNewArea(boolean).
- isFishing: method with SOG (double) as input and a boolean as output with, True = is fishing, False = not fishing.
- isNewArea: method with geographic coordinates (double longitude, double latitude) as input and a boolean as output with, True = is fishing in a new area, False = not fishing in a new area.
- restart: this method restarts the models. It can be used to create models with new data.
- GetLimit: the method used to get SOG limits used by the SFALib to classify velocity. Returns SOG limits as doubles.

4

Joined Fishery Analysis

This chapter explains the approach used to reach the second goal of this work.

In Section 4.1 we implement the methodology learn in Section 2.3.1 to approach goal 2.

4.1 Implementation of CRISP-DM

4.1.1 Business Understanding

In the fishing sector, vessels operating in the various fishing techniques must be licensed. A common problem is that there is a likelihood that vessels will be fishing for which they are not licensed. The objective is to obtain a predictive model, capable of receiving VMS data and proceed to its classification in order to predict which type of fishing is being carried out by the vessel. In this way, it will be possible to ascertain whether or not a given vessel is carrying out a legal fishing activity, that is, according to the license it has.

4.1.2 Data Understanding

To answer this goal, the data used as input are the same VMS Records as used in Chapter 3, initially presented in Chapter 1, Section 1.3.1 and VMS Vessels presented in Chapter 1, Section 1.3.2. The output variable is nominal whose categories correspond to the labels of the different fishing licenses.

4.1.3 Data Preparation

To use data mining models, a first step is to build a dataset with all the data needed to feed the models. So, it was created a dataset from VMS Vessels and VMS Records to end with Table 4.1. The correlation between the license and the HP data (vessel power), Loa (Boat length), Gt (vessel weight) are quite significant. This is expected as different fishing activities require specific types of vessels. This does not mean that the type of vessel is only capable of entering a type of fishing activity. For these reasons, we will not use these variables in the model so as not to create a problem with bias.

Table 4.1: VMS Dataset

Name	Description	From	Why
ID	Key	Native	Identify the row
VesselID	Vessel Identifier	VMS Records	Identify the vessel
UTC	Date Time	VMS Records	Identify the time of the entry
LAT	Latitude	VMS Records	Discriminated by fishing areas
LON	Longitude	VMS Records	Discriminated by fishing areas
COG	Direction	VMS Records	Course Over Ground
SOG	Velocity	VMS Records	Discriminated by fishing velocity
License	Vessel's Linceses	VMS Vessels	Objective

When analyzing the speed distribution by license, we can observe in Figure 4.1 that there is no obvious difference between licenses. For example the value 1 of SOG can be assigned to any license.

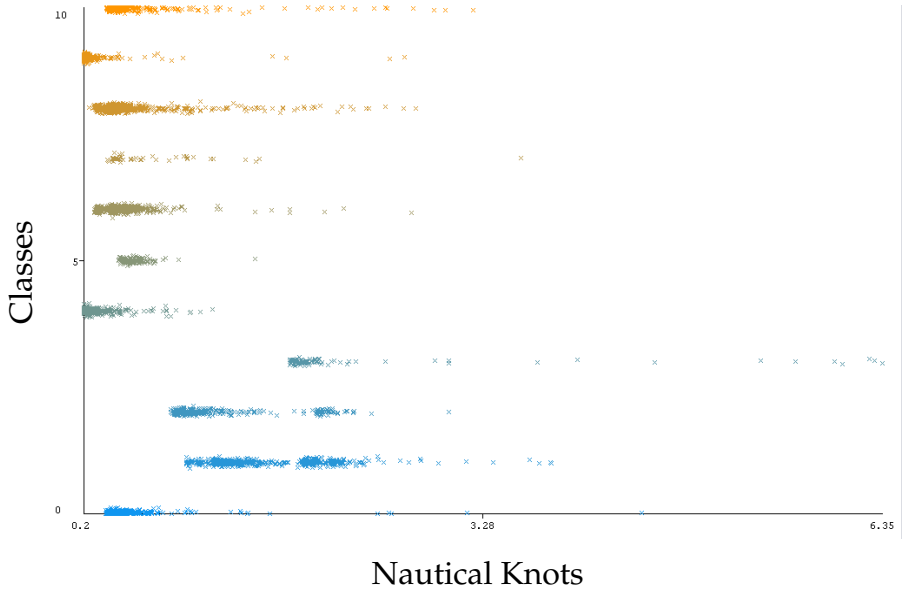

Figure 4.2: SOG minimum per license

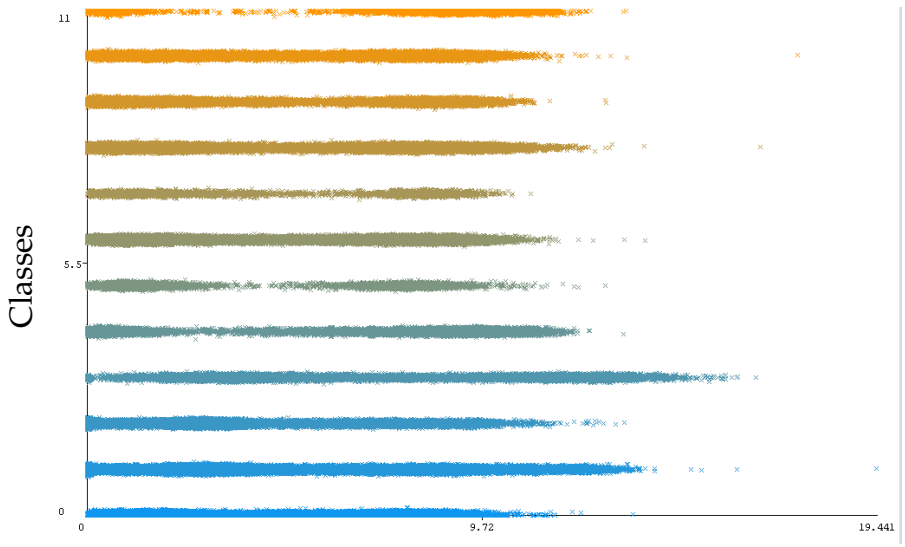

Figure 4.1: SOG per license

To try to help the predictive model to have good results, some pre-processing is necessary. A simple and effective way is to group data by vessel and activity day. Then apply the SFA to the data to have, per day, per vessel, the minimum fishing speed, maximum fishing speed and also remove the average speed. With these three variables, it is possible to better distinguish the behavior of vessels by license through speed.

The method used to transform the data consists of the following steps:

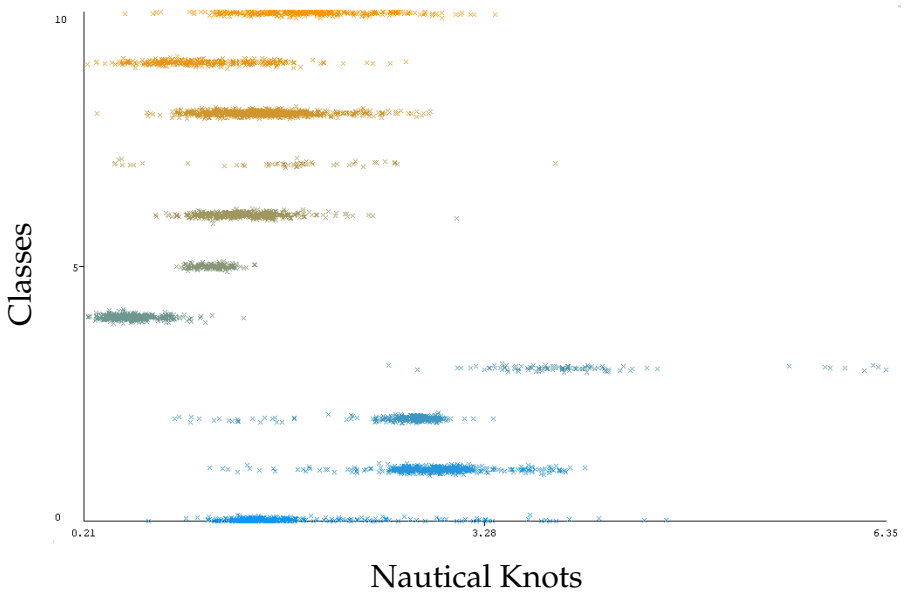

Figure 4.3: SOG average per license

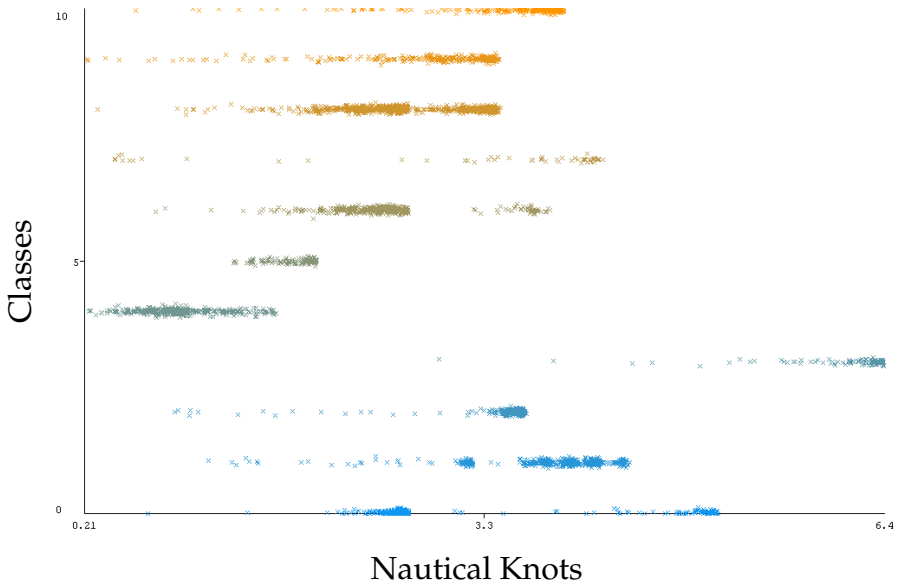

Figure 4.4: SOG maximum per license

- Create a dataset in which data are grouped per day, per vessel.
- Use DSALib to get the minimum and maximum speeds of fishing per vessel. From this data, apply a filter to obtain observations for which the velocity varies between its minimum and maximum value.

Regarding location data, clustering techniques to discretize the data were used. First, the best number of clusters is determined. For that, we used the same techniques as in Section 4.3, resulting the Figure 4.5 for the elbow method, for Silhouette method Table 4.2 and Figures 4.6, 4.7 and 4.8. The data used was the dataset filtered, so we have only the positions of fishing. The chosen number of clusters was 6.

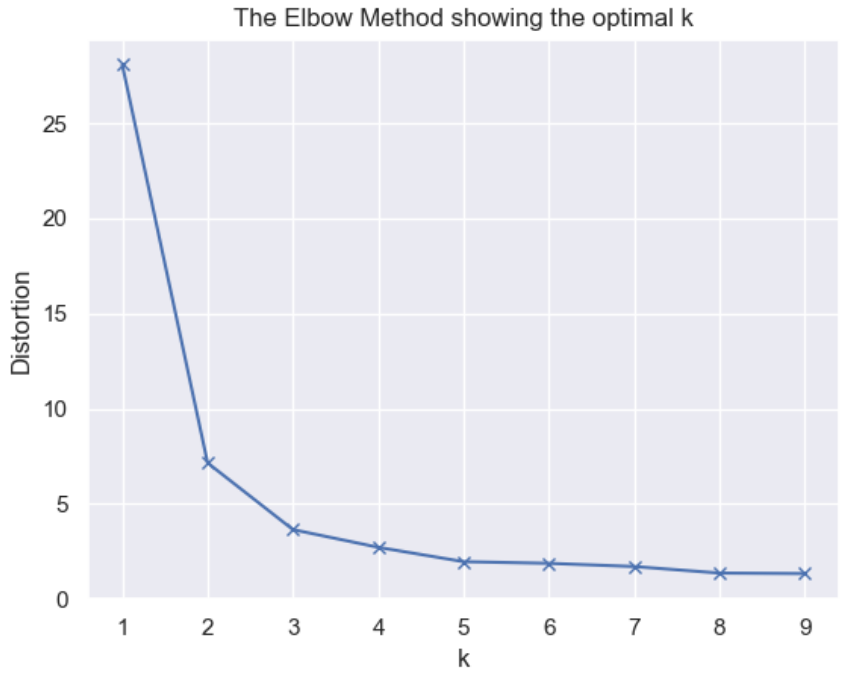

Figure 4.5: The elbow method showing the optimal k

Now, is possible to create data mining models based on the location and on the velocity patterns.

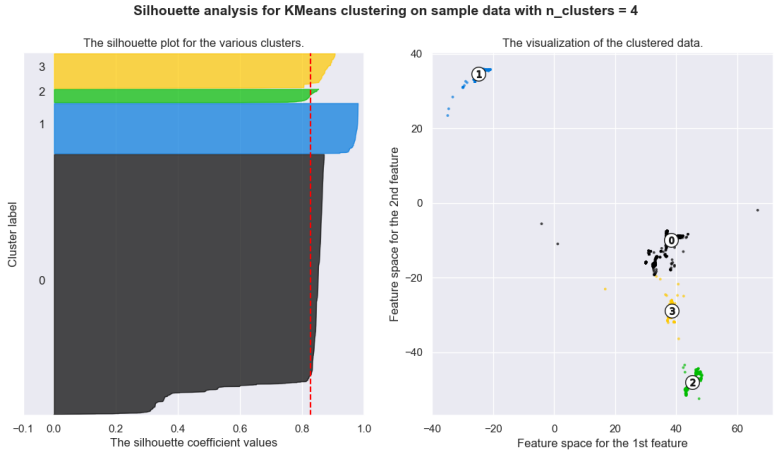

Figure 4.6: Silhouette analysis with 4 clusters

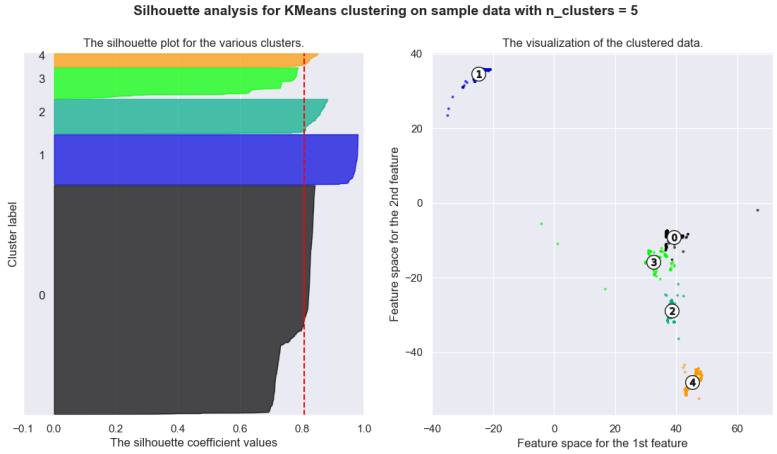

Figure 4.7: Silhouette analysis with 5 clusters

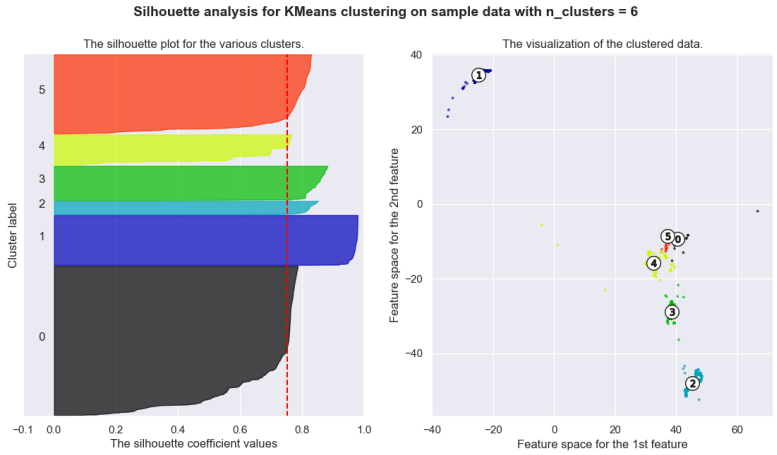

Figure 4.8: Silhouette analysis with 6 clusters

Table 4.2: The average value of the Silhouette coefficient for all vessels

Number of clusters	Average silhouette coefficient
2	0.8997
3	0.8301
4	0.8276
5	0.8078
6	0.7528
7	0.7527
8	0.7502
9	0.6947

4.1.4 Modeling

In order to classify a new type of license using VMS data, several data mining algorithms were tested, which we now list:

- **KMeans-** This model was used to organize GPS data in clusters to improve the result of the data mining algorithms. The operating mode of this algorithm has already been covered in the Section 2.2.2.1. The implementation used was KMeans from the sklearn library [45].
- **Decision Trees-** This model was used to classify the fishing license from VMS data. The operating mode of this algorithm has already been covered in the Section 2.3.2.1. The implementation used was DecisionTreeClassifier from the sklearn library.
- **Random Forests-** This model was used to classify the fishing license from VMS data. The operating mode of this algorithm has already been covered in the Section 2.3.2.2. The implementation used was RandomForestClassifier from the sklearn library.
- **Neural Network-** This model was used to classify the fishing license from VMS data. The operating mode of this algorithm has already been covered in the Section 2.3.2.3. The implementation used was MLPClassifier from the sklearn library.
- **Support Vector Machine-** This model was used to classify the fishing license from VMS data. The operating mode of this algorithm has already been covered in the Section 2.3.2.4. The implementation used was SVC from the sklearn library.

4.1.5 Evaluation

Cross –validation [10] provides a simple and effective method for both model selection and performance evaluation, widely employed by the machine learning community. Under f –fold cross –validation, the data are randomly partitioned to form f disjoint subsets of approximately equal size. In the ith fold of the cross-validation procedure, the ith subset is used to estimate the generalization performance of a model trained on the remaining f –1 subset. The average of the generalization performance observed overall f folds provides an estimate (with a slightly pessimistic bias) of the generalization performance of a model trained on the entire sample. To evaluate these models a 10-fold, f=10, cross - validation was performed.
To evaluate the classification results customized with the different algorithms, a confusion matrix, as described in table 4.3 were created and interpreted according ours goals.

Table 4.3: Confusion matrix for a two-class classifier

		Predicted	
		Positive	Negative
Actual	Positive	True Positive(TP)	False Negative(FN)
	Negative	False Positive(FP)	True Negative(TN)

In addition, the following performance indicators were used:
Precision, also called as positive predictive value attempts to identify what proportion of positive identifications was actually correct, is given by 4.1, being f the index corresponding the folds.

$$\sum_{n=0}^{f-1} \frac{TP}{TP + TF} \tag{4.1}$$

Recall, also called as sensitivity attempts to identify what proportion of actual positives was identified correctly 4.2, being f the index corresponding the folds.

$$\sum_{n=0}^{f-1} \frac{TP}{TP + FN} \tag{4.2}$$

A confusion matrix [8] illustrates the accuracy of the solution to a classification problem. Given n classes, a confusion matrix has general element $C_{i,j}$ corresponding to the number of tuples from D that were assigned to class $C_{i,j}$, but where the correct class is C_i. The best solution will have only zero values outside the diagonal. A confusion matrix contains information about actual and predicted classifications done by a classification system. The performance of such systems is commonly evaluated using the data in the matrix. Table 4.3 shows the confusion matrix for a two-class classifier.

For the purpose of this work, the classes corresponding to the different types of licenses are:

0 Armadilhas / De abrigo / Alcatruzes
1 Arrasto / De fundo de portas
2 Arrasto / De fundo de portas / Crustáceos
3 Arrasto / Pelágico / Com portas
4 Cerco / para bordo / Tipo americano
5 Emalhar de 1 pano / De deriva / Grandes Pelágicos
6 Emalhar de 1 pano / De fundo
7 Pesca à linha / Cana e linha de mão
8 Pesca à linha / Palangre de fundo / Espécies demersais
9 Pesca à linha / Palangre de Fundo + Cana e linha de mão
10 Pesca à linha / Palangre de superfície / Grandes Migradores

The model's test results are:

- **Decision Trees:**

  We can observe in Table 4.4 the usage of velocity parameters (SogMin, SogAVG, and SogMax) and location (clustering result of K-Means) having the best result. The algorithm with the best result is Entropy(C4.5), with a precision of 0.8142 and an recall of 0.8142. In Figure 4.4 the confusion matrix shows that only the classes 0 and 7 have a low prediction rate. The max depth of the trees was tested as 200, 300, and 400, with 300 giving the best results.

Table 4.4: Cross-Validation results for Decision Trees models

Splitting criteria	**Velocity and locations**		**Velocity**	
	Precision	Recall	Precision	Recall
Gini	0.8047	0.8047	0.7479	0.7479
Entropy	0.8142	0.8142	0.7659	0.7659

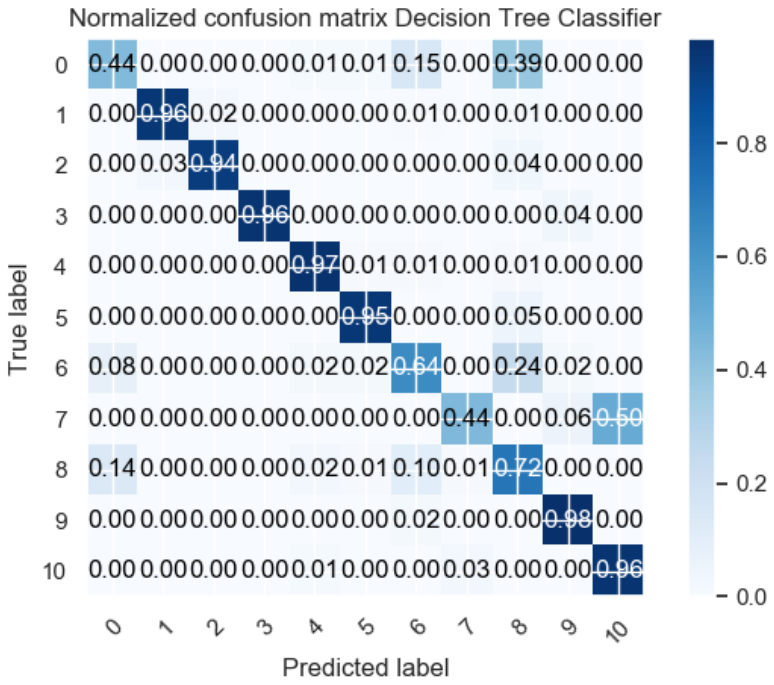

Figure 4.9: Confusion Matrix for Decision Tree using Entropy splitting criterion(C4.5)

- **Random Forest:** Concerning Random forest use, the model was trained again considering Giny impurity and Entropy Information Gain as splitting criteria. The results mentioned in this document for Random Forest models are using Entropy(C4.5) as the algorithm to measure the quality of a split. The max depth of the trees was tested with the values of 200, 300, and 400, being the best results attained with the max depth value of 300. Table 4.5, describes the cross-validation results obtained with the number of estimators at 200. In Figure 4.5 is shown the confusion matrix corresponding to the Random forest classifier. The results are satisfactory given that only the classes 0 and 7 have a lower correct prediction rate.

Table 4.5: Cross-Validation results for Random Forest models

No. of estimators	**Velocity and locations**		**Velocity**	
	Precision	Recall	Precision	Recall
50	0.8389	0.8389	0.8	0.8
100	0.8398	0.8398	0.7896	0.7896
200	0.8474	0.8474	0.7953	0.7953
300	0.8445	0.8445	0.7943	0.7943

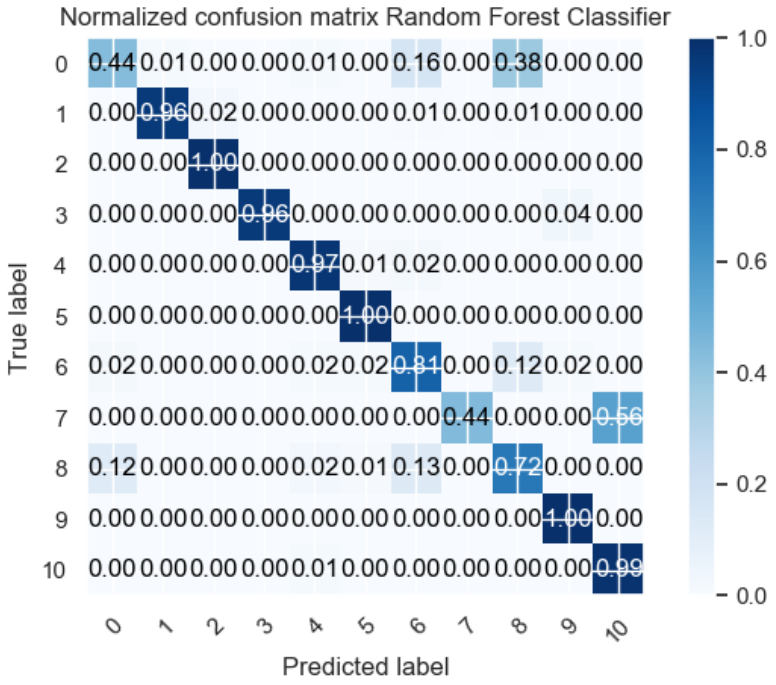

Figure 4.10: Confusion Matrix for Random Forest using 200 estimators

- **Neural Network:**
  For these models, it was trained in various configurations of hidden layers. It was used from 1 to 8 hidden layers with 11, 100, 250, 500, and 750 neurons. Some of the results are in Table 4.6. The best result, as we can observe in Table 4.6 corresponds to the case with six layers with 500 neurons for which the confusion matrix is presented in Figure 4.11. For some models, it was used the Adam solver and BFGS, but with BFGS having better results. So all reported results are using BFGS solver. As in Decision Trees models, the usage of locations has a good impact on the results of the Neural Network models.

Table 4.6: Cross-Validation results for Neural Network models

Hidden Layers	**Velocity and locations**		**Velocity**	
	Precision	Recall	Precision	Recall
100	0.7507	0.7507	0.70426	0.7043
4 (500)	0.7536	0.7536	0.7715	0.7715
5 (500)	0.7479	0.7479	0.7659	0.7659
6 (500)	0.7867	0.7867	0.7583	0.7583

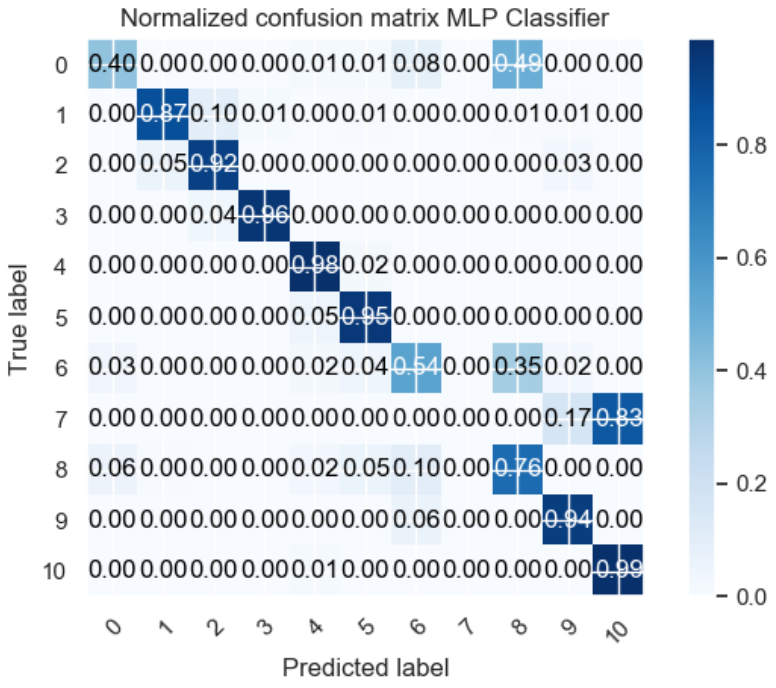

Figure 4.11: Confusion Matrix for Neural Network using 5x500 hidden layers

- **Support Vector Machine:**
  For the built of support vector machine models, it was used the type of kernel of Polynomial, RBF, and linear. With the best result as we can observe in Table 4.7 is the Polynomial, with its confusion matrix represented in Figure 4.12.

Table 4.7: Cross-Validation results for Support Vector Machine models

Kernel coefficient	**Velocity and locations**		**Velocity**	
	Precision	Recall	Precision	Recall
Polynomial	0.7422	0.7422	0.672	0.672
RBF	0.7374	0.7374	0.6986	0.6986
Linear	0.6569	0.6568	0.545	0.545

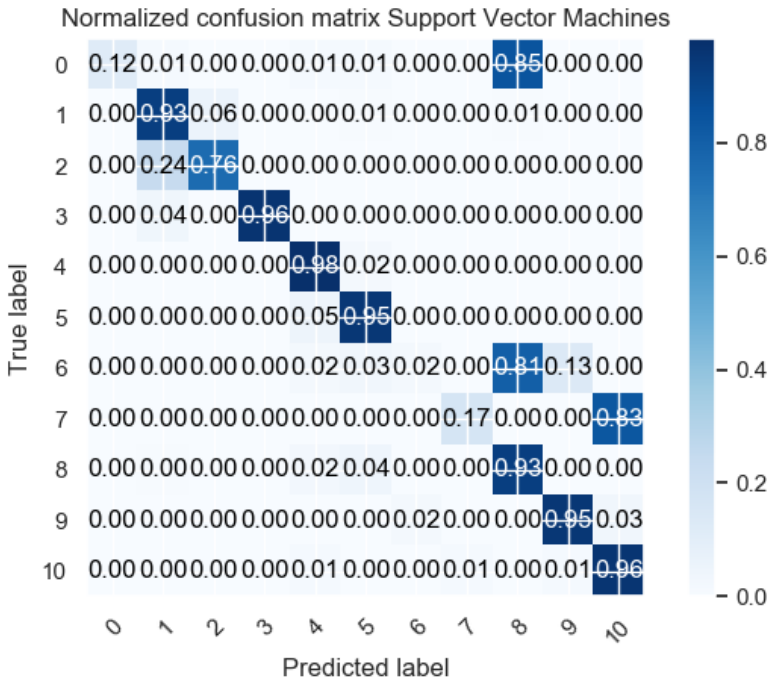

Figure 4.12: Confusion Matrix for Support Vector Machine using Polynomial kernel coefficient

The model with the best result corresponds to the use of a random forest method, with the configuration of 200 estimators. For this model was obtained a precision of 84.74%. With this, we can determine that it is possible to create a good model to classify the VMS data as the fishing license.
As seen in Figure 4.10, the classes 0 and 7 have a low true positive rate. The usage of more data and data that are certified that the VMS data of a vessel is only for the purpose of the fishing license can improve the precision of the model.

4.1.6 Deployment

To use the trained model, we separated JFA into 4 steps:

1. **Receive VMS data and register in a BD.** JFA needs to be able to receive VMS data so it can classify it, but it is also important to persist the data in a way that we can update the model with more recent data.

2. **Pre process the data.** If SFA is installed in the blue box that is filling the VMS data, the data entries can be already classified as fishing or not fishing. If not, JFA needs to collect data and run its local SFA for each vessel that does not have SFA. If the data is received in one time when the vessel arrives at the port, convert all the data received, as explained in Subsection 4.1.3. If the blue box is broadcasting data when the vessel is on activity, we can classify the data as fishing or not (if SFA not installed in the blue box) and have a subsystem that detects when a vessel fails to send data or have stopped fishing for more than some defined time. This will allow time as the competent authorities will prepare an inspection as marked vessels even before reaching the port.

3. **Use the model to classify.** With the data obtained in the last step, use the model tested chosen in Subsection 4.1.5.

4. **Validate vessel license comparison, send an alert.** Compare the class attributed by the model with the license of the vessel and, if different, register and send a message to the designated authority.

To keep the model up to date, it is essential to train a new model with recent data from time to time. If possible, to know that data is verified by a competent authority, when training the model gives more relevance to this verified data. This is important because the training data that is not verified, can be given by vessels that are not respecting their fishing license and this way corrupting the model.

5

Experimental Evaluation

This chapter describes the process used to evaluate the work done in this thesis.

5.1 Validation of Standalone Fishery Analysis

The VMS data are not classified, for the validation and evaluation of SFALib presented in Chapter 3. The first step consists in the classification of the data.

The classification is realized using three categories, as follows:

- Class 0 = Fishing in a known area;
- Class 1 = Not Fishing;
- Class 2 = Fishing in a new area.

It was chosen the vessel with id two because it has the most entries in the database provided to this work. In Figure 5.2, the circles represent locations given by the VMS data to vessel two, and the diameter of the circle, the velocity. This analysis was done using Windows Power Bi [44].

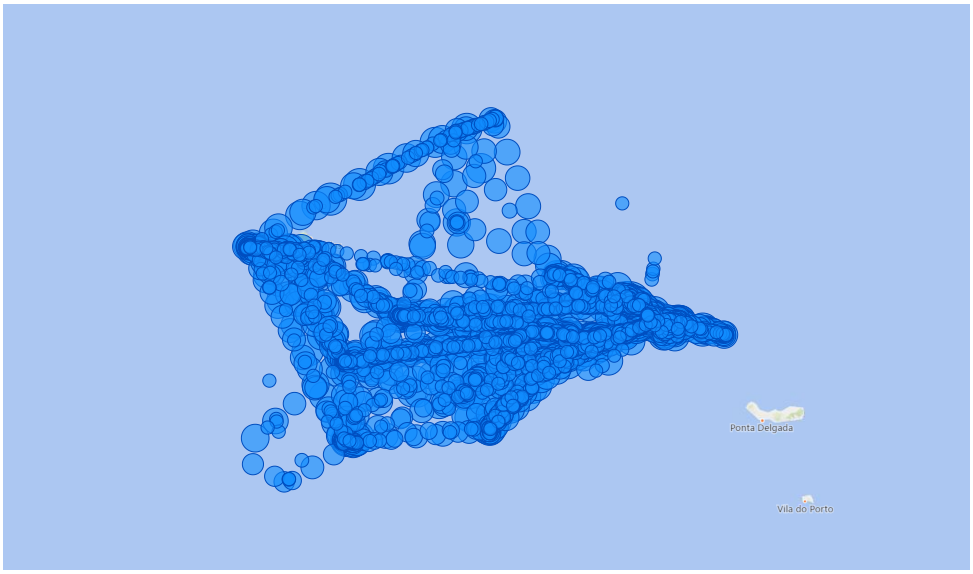

Figure 5.1: Representation of all VMS coordinates for vessel 2

To get data to test SFALib, we need to classify data into three classes.

- Class 0 (Fishing in a known area): To get data to classify as fishing in an area, the data was filtered with a SOG between 0 and 4. Number 4 was chosen because of the analysis made in Section 3.3, which concludes that the fishing activity done by this vessel is at speeds below four nautical miles. With this, we end up with the data represented in Figure 5.2. To extract the text data, it was chosen the fishing spot near the island of Flores represented in Figure 5.3. 500 random VMS entries were collected at this location without speed restriction. This means that we may have collected travel data, but no speed bias was passed to this test data.
- Class 1 (Not fishing): For this class, it was extracted 500 random VMS entries with the location near Terceira island represented in figure 5.4. It was chosen this spot because most points appear to be on a well-defined trajectory. This pattern contrasts with the random-looking and under-lapping points in Figure 5.3.
- Class 2 (Fishing in a new area): To have data from this class, it was created new data. So 500 new VMS entries were created. The data was created using the data from class 0 but changing to a location near mainland Portugal that is placed with no entries for that vessel.

5.1.1 Validation and evaluation

To evaluate the classification accuracy of SFALib, it was used precision and recall measures.

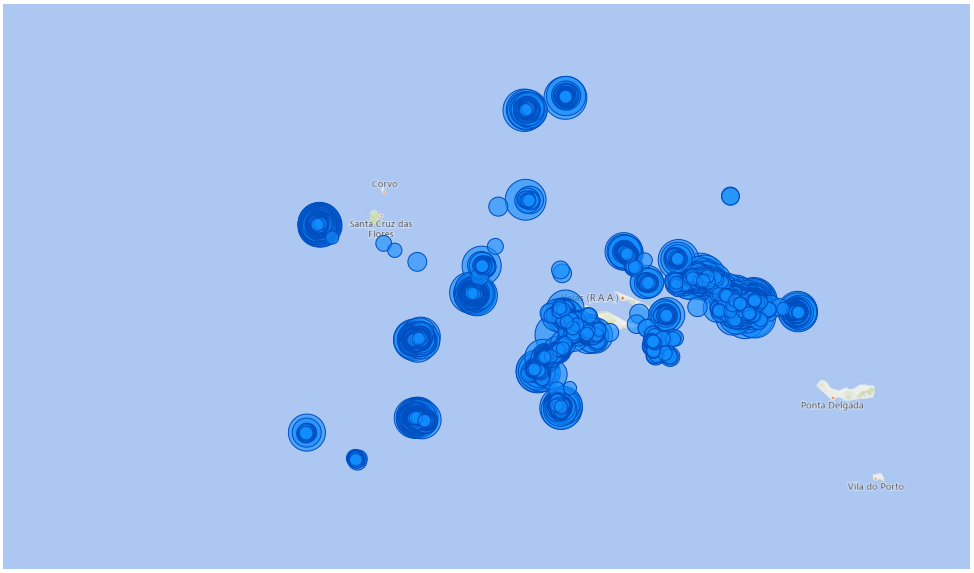

Figure 5.2: Representation of VMS coordinates for vessel 2 with speeds inferior to 4

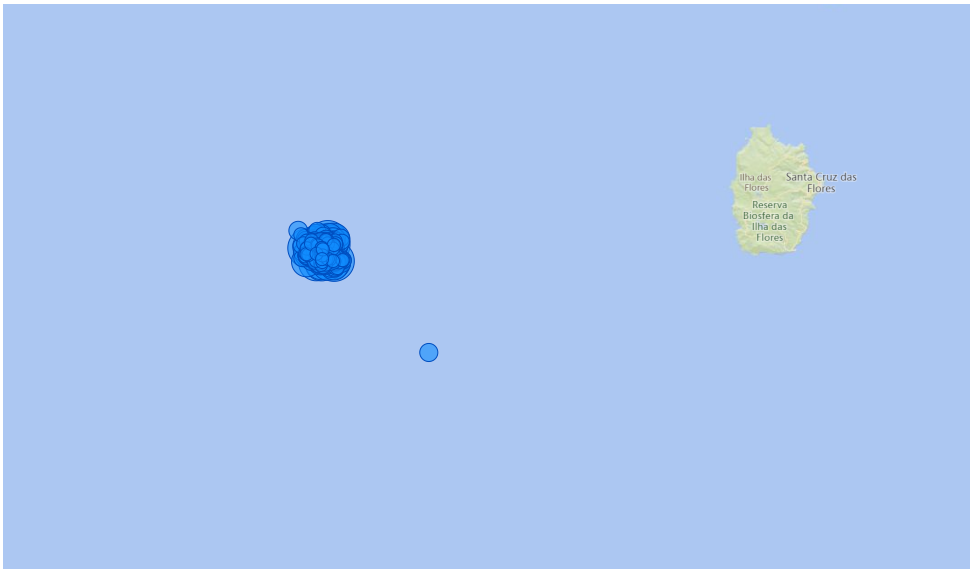

Figure 5.3: Representation of VMS coordinates for vessel 2, near Flores island, with speeds lower than 4

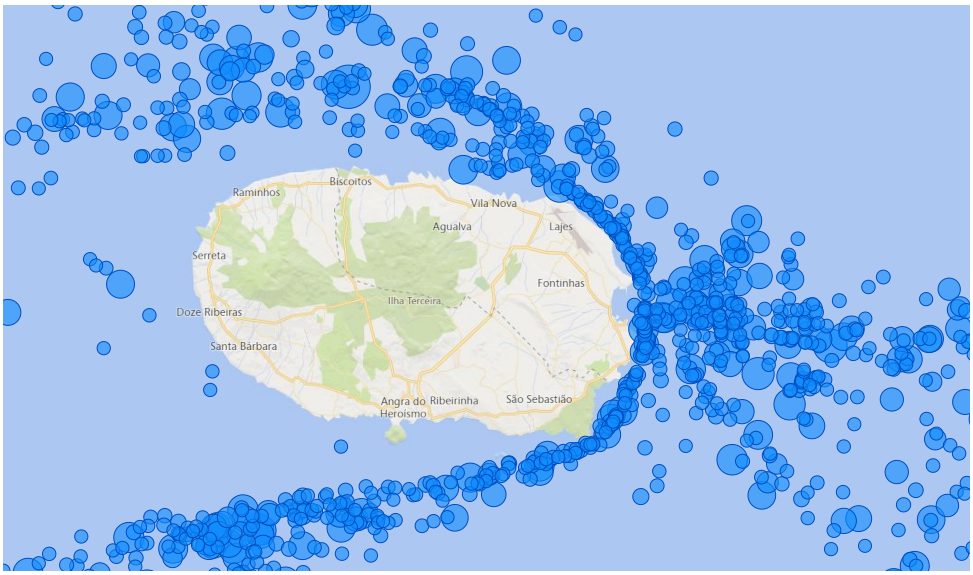

Figure 5.4: Representation of VMS coordinates for vessel 2 near Terceira island

Precision: In table 5.1 we can see that the lower the configuration value for velocity, the better are the results. This can mean that the gap in speed between fishing and traveling is big. The configuration value for velocity is the parameter limitVelocity explained in Sub Chapter 4.1.6.

The best result has an average precision of 0.9.

Table 5.1: Precision per class and configuration

Configuration for velocity	**Class 0**	**Class 1**	**Class 2**
0.1	0.67	1	0.67
0.05	0.738	1	0.738
0.01	0.838	1	0.838
0.001	0.85	1	0.85
0	0.85	1	0.85

Recall: In Table 5.2 we can confirm that, like in precision, the recall is better with lower configuration values for velocity.

The best result has an average recall of 0.9231.

Table 5.2: Recall per class and configuration

Configuration for velocity	**Class 0**	**Class 1**	**Class 2**
0.1	1	0.6024	1
0.05	1	0.6562	1
0.01	1	0.7553	1
0.001	1	0.7692	1
0	1	0.7692	1

In Table 5.3, we can see that some fishing data(Class 0 and Class 2) was classified as not fishing(Class 1). However, considering that the classified data can have some errors in classification, we cannot assure that the predicted not fishing(Class 1) of fishing classes(Class 0 and Class 2) is not actually, not fishing(Class 1).

Table 5.3: Confusion matrix for configuration with 0.001 for velocity

Prediction/Real	**Class 0**	**Class 1**	**Class 2**
Class 0	0.85	0	0
Class 1	0.15	1	0.15
Class 2	0	0	0.85

5.2 Validation Joined Fishery Analysis

The model used for Joined Fishery Analysis decided in subsection 4.1.5 as being the Random Forest, was validated and to evaluated using the same data used in the subsection 4.1.5.

In subsection 4.1.5 we use cross-validation [10] that trains the model with approximately the same percentage of samples of each target class as the complete set. In this validation, it was only used one vessel of each license for the train and all of the tests. In this case, we will train the model with 11 vessels but test with 30. This model will have a strong bias to these 11 vessels, but we will see the degradation of the results when testing with these unknown vessels by the model. This is useful to understand the importance of having the model up to date and how badly it performs with new vessels.

5.2.1 Validation and evaluation

To evaluate the classification accuracy of Joined Fishery Analysis model it was used a confusion matrix and the model precision and recall.

In figure 5.5 we can observe that this model is less accurate than the model tested in Figure 4.10. The precision in this model is 0.6727, and the recall is 0.6379.

This means that different vessels in the same license can operate at different speeds, and so for the model to work without differentiation of vessel power, length, or gross tonnage, it requires that the training dataset is varied in the type of vessels. The other solution is to create different models by categorizing vessels by vessel information (Gross tonnage, length, and power). This way, the models can by specific and achieve better results.

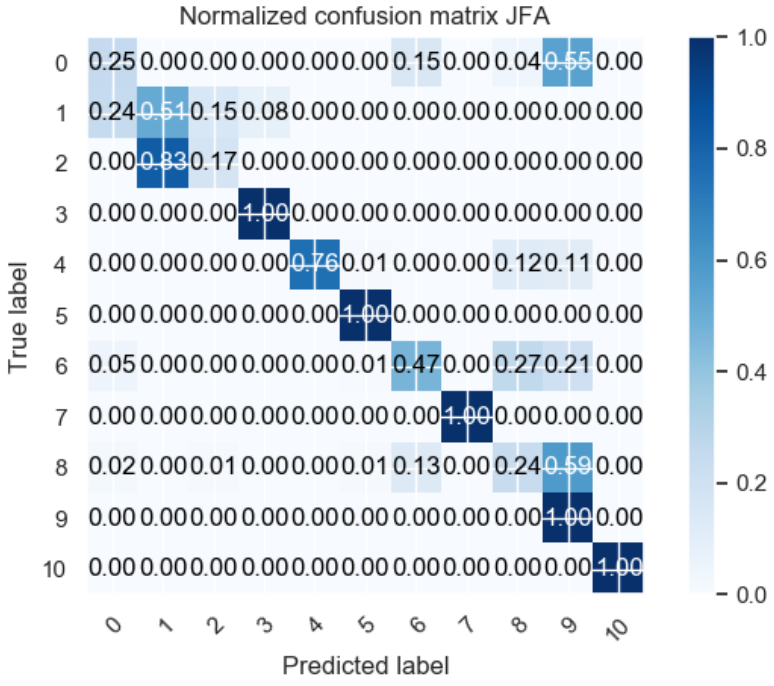

Figure 5.5: Confusion Matrix for Joined Fishery Analysis validation

6

Conclusions

This chapter presents the conclusion of this research, starting with an overview of the work covered in this document and concluding with a set of possible routes for future work.

6.1 Overview

There is great concern about fisheries fraud as demonstrated by the Food and Agriculture Organization of the United Nations [15]. In the European Union through Directorate-General for Maritime Affairs and Fisheries in the European commission, studies are being done and discussing the best way to act on this illegal action using new technologies like VMS [38].

The objective of this dissertation was to find ways to detect tax fraud in fishing activity efficiently and quickly to be able to detect when a vessel had an operation with abnormal behavior, to be able to notify the authorities while the vessel is still unloading the fish at the dock.

We were able to demonstrate two ways to classify and validate VMS data. The importance of being able to evaluate this type of data will be a great weapon against the tax fraud that occurs in the fishing sector. Another potential return from this work was the possibility to understand the fishing patterns to be able to create plans for environmental protection.

For this work, it was used real VMS data from Portuguese fisheries.

All the objectives proposed in this document have been achieved.

6.1.1 Conclusions about Standalone Fishery Analysis

In SFA, we were able to demonstrate that it is possible to classify in real-time two crucial aspects. If the vessel is fishing and if it is fishing in a new area.
Considering the work done we conclude that classifying if the vessel is fishing at a given moment, taking into account the historic of its speed, it is possible since we demonstrate that the speed of each vessel has well defined the distribution of fishing speeds, thanks to the fact that the boats spend much of their time fishing. With this information and using clustering algorithms, it is also possible to define fishing areas.

This type of classification is advantageous to understand the fishing patterns in a given area. Another impressive result is the possibility of over the years understand if there are variations in the level of hours of fishing and fishing zones, trying to understand the temporal evolution of the fishing and by consequence of its raw material. With this to understand if the boats spend more time or less inactivity by each time, they leave (it can mean that it is becoming easier or more difficult to catch fish) if there is a movement of the activity by type of license (can infer if certain types of fish are disappearing in certain areas and emerging in new areas).

The main weakness of SFA is the lack of classified VMS data. Since it is not possible to test the SFA results with data classified on the ship, it is not possible to measure the real accuracy of the classifier. With classified data, it would also be possible to adjust the classification parameters of the SFA better.

6.1.2 Conclusions about Joined Fishery Analysis

In the second solution, we want to show that it is possible to classify the fishing license by taking into account the VMS data, more precisely speed and position data. The treatment of the data was rewarding since it was possible to find correlations between the type of fishing and the actions of the vessels in fishing activity.

It still takes much work to have variables of enough quality to create a good classification model. We created different types of data mining algorithms to determine which best fits this problem.

The main weakness of JFA is the prospect of having fraudulent data used in training the model. This could be resolved by training the model with data on which the on-board inspection took place or giving this data more weight in the model than the un-inspected.

6.2 Future Work

In future work, it is imperative to create classified VMS data to analyze the SFA accuracy better. Also, to have data from fisheries activity that was inspected by a competent authority to test the models better and if with enough data, to train them with only inspected data or giving more weight to this data so we can have more accurate models.

References

[1] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization", *International Conference on Learning Representations*, Dec. 2014.

[2] T. Agardy, "Effects of fisheries on marine ecosystems: A conservationist's perspective", *ICES Journal of Marine Science*, pages 761–765, 2000.

[3] K. S. Alfred M. Duda, "A new imperative for improving management of large marine ecosystems", *Ocean and Coastal Management*, pages 797–833, 2002.

[4] L. Rokach and O. Maimon, "Top-down induction of decision trees classifiers - a survey", *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 35, no. 4, pages 476–487, 2005.

[5] J. R. N. B. S. A. O. R. E. Francois Bastardie, "Effects of fishing effort allocation scenarios on energy efficiency and profitability: An individual-based model applied to danish fisheries", in *Fisheries Research*, vol. 106, 2010, pages 501–516.

[6] C. Pimenta, "Esboço de quantificação da fraude em portugal", in. Edições Húmus, 2009, pages 1–44.

[7] Leo Breiman, "Random forests", *Machine Learning*, vol. 45, no. 1, pages 5–32, 2001, ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. [Online]. Available: https://doi.org/10.1023/A:1010933404324.

[8] T.R. Patil and Swati Sherekar, "Performance analysis of naive bayes and j48 classification algorithm for data classification", *Int. J. Comput. Sci. Appl.*, vol. 6, pages 256–261, Jan. 2013.

[9] R Wirth and Jochen Hipp, "Crisp-dm: Towards a standard process model for data mining", *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, Jan. 2000.

[10] M. Stone, "Cross-validatory choice and assessment of statistical predictions", *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, no. 2, pages 111–147, 1974, ISSN: 00359246. [Online]. Available: http://www.jstor.org/stable/2984809.

[11] Yu-Hong Dai, "A perfect example for the bfgs method", *Mathematical Programming*, vol. 138, no. 1, pages 501–530, 2013, ISSN: 1436-4646. DOI: 10.1007/s10107-012-0522-2. [Online]. Available: https://doi.org/10.1007/s10107-012-0522-2.

[12] C. Ulrich N. T. Hintzen F. Bastardie and N. Deporte, "Vmstools: Open-source software for the processing, analysis and visualization of fisheries logbook and vms data", in *Fisheries Research*, 2012, pages 31–43.

[13] Bhumika Gupta, Aditya Rawat, Akshay Jain, Arpit Arora, and Naresh Dhami, "Analysis of various decision tree algorithms for classification in data mining", *International Journal of Computer Applications*, vol. 163, no. 8, pages 15–19, 2017, ISSN: 0975-8887. DOI: 10.5120/ijca2017913660. [Online]. Available: http://www.ijcaonline.org/archives/volume163/number8/27414-2017913660.

[14] FAO, *Fishing Operations*. Food and Agriculture Organization of the United Nations, 1998. [Online]. Available: availableathttp://www.fao.org/3/a-w9633e.pdf.

[15] Alan Reilly, "Overview of food fraud in the fisheries sector", in *FAO Fisheries and Aquaculture Circular No. 1165*, 2018.

[16] Marza Marzuki, Philippe Gaspar, Rene Garello, Vincent Kerbaol, and Ronan Fablet, "Fishing gear identification from vessel-monitoring-system-based fishing vessel trajectories", *IEEE Journal of Oceanic Engineering*, vol. PP, pages 1–11, Jul. 2017. DOI: 10.1109/JOE.2017.2723278.

[17] Jinxin Gao and David B. Hitchcock, "James–stein shrinkage to improve k-means cluster analysis", *Computational Statistics & Data Analysis*, vol. 54, no. 9, pages 2113–2127, 2010, ISSN: 0167-9473. DOI: https://doi.org/10.1016/j.csda.2010.03.018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167947310001209.

[18] Bin Liu, Ying Yang, Geoffrey I. Webb, and Janice Boughton, "A comparative study of bandwidth choice in kernel density estimation for naive bayesian classification", in *Advances in Knowledge Discovery and Data Mining*, Thanaruk Theeramunkong, Boonserm Kijsirikul, Nick Cercone, and Tu-Bao Ho, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pages 302–313, ISBN: 978-3-642-01307-2.

[19] Slava Kisilevich, Florian Mansmann, and Daniel A. Keim, "P-dbscan: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos", in *COM.Geo*, 2010.

[20] Trupti M. Kodinariya and Prashant R. Makwana, "Review on determining number of cluster in k-means clustering", in *Review on determining number of Cluster in K-Means Clustering*, 2013.

[21] David Kriesel, *A Brief Introduction to Neural Networks*. dkriesel.com, 2007. [Online]. Available: http://www.dkriesel.com.

[22] Vladimir Kvasnicka, Martin Pelikan, and Jirí Pospíchal, "Hill climbing with learning (an abstraction of genetic algorithm)", in *Hill Climbing with Learning*, 1995.

[23] Alfredo Alessandrini Fabrizio Natale Maurizio Gibin, "Mapping fishing effort through ais data", *PLOS ONE*, 2015.

[24] M. I. Marzuki, R. Garello, R. Fablet, V. Kerbaol, and P. Gaspar, "Fishing gear recognition from vms data to identify illegal fishing activities in indonesia", in *OCEANS 2015 - Genova*, 2015, pages 1–5. DOI: 10.1109/OCEANS-Genova.2015.7271551.

[25] Raphael Obi Okonkwo and Francis O. Enem, "Combating crime and terrorism using data mining techniques", in *COMBATING CRIME AND TERRORISM USING DATA MINING TECHNIQUES*, 2011.

[26] G. J. Piet and F. J. Quirijns, "The importance of scale for fishing impact estimations", *Canadian Journal of Fisheries and Aquatic Sciences*, vol. 66, no. 5, pages 829–835, 2009. DOI: 10.1139/F09-042. eprint: https://doi.org/10.1139/F09-042. [Online]. Available: https://doi.org/10.1139/F09-042.

[27] Roelofsen, "Time series clustering", Vrije Universiteit Amsterdam Faculty of Science De Boelelaan, 2018.

[28] Lior Rokach and Oded Maimon, *Data Mining With Decision Trees: Theory and Applications*, 2nd. River Edge, NJ, USA: World Scientific Publishing Co., Inc., 2014, ISBN: 9789814590075, 981459007X.

[29] Tommaso Russo, Lorenzo D'Andrea, Antonio Parisi, and Stefano Cataudella, "Vmsbase: An r-package for vms and logbook data management and analysis in fisheries ecology", *PLOS ONE*, vol. 9, no. 6, pages 1–18, Jun. 2014. DOI: 10.1371/journal.pone.0100195. [Online]. Available: https://doi.org/10.1371/journal.pone.0100195.

[30] Reza Sadoddin and Ali A. Ghorbani, "A comparative study of unsupervised machine learning and data mining techniques for intrusion detection", in *Machine Learning and Data Mining in Pattern Recognition*, Petra Perner, Ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pages 404–418, ISBN: 978-3-540-73499-4.

[31] Serge Lage, *Standalone-fishery-analysis*, https://github.com/SergeLage/Standalone-Fishery-Analysis, 2019.

[32] N. S. A. J. T. P. S. S.-P. Yashashwita Shukla, "Big data analytics based approach to tax evasion", *International Journal of Engineering Research in Computer Science and Engineering*, pages 56–59, 2019.

[33] Theodoros Evgeniou and Massimiliano Pontil, "Support vector machines: Theory and applications", in *Machine Learning and Its Applications: Advanced Lectures*, vol. 2049, Jan. 2001, pages 249–257. DOI: 10.1007/3-540-44673-7_12.

[34] Michael Ringgaard Yin Wen Chang Cho Jui Hsieh, "Training and testing lowdegree polynomial data mappings via linear svm", *Journal of Machine Learning Research*, vol. 11, 1471-1490, 2010. [Online]. Available: http://www.jmlr.org/papers/volume11/chang10a/chang10a.pdf.

[35] DGRM, *Artes de pesca*, https://www.dgrm.mm.gov.pt/artes-de-pesca, Accessed on 2020-07-06, 2018.

[36] J. e. C. S. Pires, *Consumo de peixe em portugal*, https://sites.google.com/site/docapescacreative/consumo-de-peixe-em-portugal, Accessed on 2018-04-06, 2013.

[37] Matheus Gonzalez, *Crisp-dm na prática*, https://medium.com/@mgonzalez/crisp-dm-na-pr%C3%A1tica-65be0ee92ada, Accessed on 2020-03-05, 2019.

[38] European Commission, *Control technologies*, https://ec.europa.eu/fisheries/cfp/control/technologies_en, Accessed on 2020-02-01, 2014.

[39] ec.europa.eu, *European commission*, https://ec.europa.eu/fisheries/cfp/control/technologies/vms_en, Accessed on 2019-01-21, 2018.

[40] Inmarsat, *Inmarsat c*, https://www.inmarsat.com/services/safety/inmarsat-c/, Accessed on 2019-09-11, 2019.

[41] Xsealence, *Xsealence*, http://www.xsealence.pt/portfolio/monicap/, Accessed on 2019-07-31, 2018.

[42] Oracle, *Java 8 central*, https://www.oracle.com/technetwork/java/javase/overview/java8-2100321.html, Accessed on 2019-08-13, 2018.

[43] PostgreSQL, *Postgresql*, https://www.postgresql.org/, Accessed on 2019-08-14, 2019.

[44] Windows, *Power bi*, https://powerbi.microsoft.com/pt-pt/, Accessed on 2019-10-06, 2019.

[45] scikit-learn developers, *Scikit-learn machine learning in python*, https://scikit-learn.org/stable/index.html, Accessed on 2020-08-07, 2020.

[46] Microsoft, *Sql server*, https://www.microsoft.com/pt-pt/sql-server, Accessed on 2019-08-14, 2017.

[47] Waikato University, *Weka 3*, https://www.cs.waikato.ac.nz/ml/weka/, Accessed on 2019-08-14, 2019.

[48] Xsealence, *Xsealence*, http://www.xsealence.pt, Accessed on 2019-09-11, 2018.