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

(Grau do candidato)

Relatório intercalar 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

June, 2019

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

(Grau do candidato)

Relatório intercalar 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

June, 2019

Abstract

MONICAP is a monitoring system for the inspection of fishing using the Global Positioning System (GPS) for vessel location and Inmarsat-C technology for satellite communications between ships and a ground control center. MONICAP was successfully introduced on the market by Xsealence (www.xsealence.pt) 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 with the objective of extracting 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, as well as the identification of outliers. The identification of outliers when performed in real time will lead to the consequent generation of alarms. The present study represents the first comprehensive approach with the objective of detecting and identifying the potential behavior of fishing activities for the main types of equipment based on the tracking of Portuguese fishing fleet data.

Keywords: Vessel Monitoring System, Data Mining, Fishing

Resumo

O MONICAP é um sistema de monitorização para a inspeção da atividade pesqueira que utiliza o Sistema de Posicionamento Global (GPS) para a localização da embarcação e a tecnologia Inmarsat-C para as comunicações via satélite entre navios e um centro de controle terrestre. O MONICAP foi introduzido com sucesso no mercado pela empresa Xsealence (www.xsealence.pt) e está atualmente instalado, ou em fase de instalação em cerca de 800 navios de pesca que operam sob o controlo das autoridades de Portugal, Espanha, França, Irlanda e Angola. No âmbito desta tese de mestrado propõe-se a utilização dos dados de pesca portugueses provenientes dos Sistemas de Monitorização de Embarcações com o objetivo de extrair padrões de comportamento relacionados com as zonas de pesca, os tempos, as velocidades e as direções do percurso efetuado pelos navios. A análise estatística destes viabiliza a identificação de padrões de atividade de pesca, bem como a identificação de outliers. A identificação de outliers quando realizada em tempo real, levará à consequente geração de alarmes os quais podem ser utilizados pelas autoridades competentes potenciando assim o desenvolvimento de meios de controlo de atividades ilícitas. O presente estudo representa a primeira abordagem abrangente com o objetivo de detetar e identificar o potencial comportamento das atividades de pesca, para os principais tipos de equipamento, baseado no rastreamento de dados de frota pesqueira portuguesa.

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

Contents

List of Figures

List of Tables

1

Introduction

This chapter introduces the motivation, context and the goals of this work. Finally, it presents the overall structure of this document.

1.1 Motivation

1.1.1 Fishing Activity

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 [1]. Fisheries mapping is needed for implementing a better ecosystem management and to secure a healthy marine population [4]. Fishing Activity represents an important activity for the Portuguese economy. In European Union, Portugal is the country that has the highest consumption of fish per person and the third world wide needing 55,6 kg (per capita/year), [5]. Portugal is a country connected to the sea by its great coast all along its continental territory, almost all the coastal villages has a fishing community. Another 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 represents 21,9The use of statistical pattern recognition techniques to analyse the data, makes possible to identify who operates in the margin of the law in a more rapid and methodical way [20]. Reducing tax evasion will allow strong gains and potentially will develop the

economy, making the fishing activity more just for everyone involved in this activity.

MONICAP is a monitoring system for the inspection of fishing using the Global Positioning System (GPS) for vessel location and Inmarsat-C technology for satellite communications between ships and a ground control center. MONICAP was successfully introduced on the market by Xsealence (www.xsealence.pt) 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) with the objective of extracting 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 possible to identify patterns of fishing activity.

VMS provides an 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 important solution to quantifying and managing ecosystem disturbance, particularly on the high-seas.

1.1.2 Analytics

The concept of Analytics refers to the ability to use data, perform predictive analytics, and systematic reasoning to improve performance in key business domains and lead to a more efficient decision-making process. There is a extensive use of mathematics and statistics, like descriptive techniques and predictive models which allow to gain valuable knowledge from the data. The insights from data are used to recommend action or to guide decision making rooted in business context.

1.2 Objectives

This project proposes two methods to better map the fishing activities.

Objective 1: Local tool
In real time, will be developed an application to be installed in the MONICAP

which allow to describe better the fishing zones. At each one of the vessels this application will work with the data from the fishing activity of this vessel. This tool will be local and will be used unsupervised techniques of automatic learning. 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. Use these spots 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 and compare with the license of the vessel.

1.3 Work Structure

The work is divided into four main chapters. Chapter 2 gives an overview of the State of the Art concerning the usage of technologies in the fishing industry to detect outliers' behaviors such activity. In Chapter 3 we focus on the data analysis, pre-processed treatment needed to find possible solutions for the goals described previously. The data are divided into two categories:

- VMS Records: data generated by MONICAP system;
- VMS Vessels: data coming from the vessel's captains, manually inserted.

In Chapter 4 were presented two possible processes 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;

In Chapter 5 is presented an approach to answer to the second objective: using the data from all the vessels, how to identify fishing activities that are not in accordance with the vessel's fishing license. Data mining methodologies will ve user to obtain the predictive models. The Chapter 6 conteins a description of the software developed for the purpose presented in the chapters 4 and 5.

1.4 Publications

1.4.1 U. C. report to final project of course (FPC)

My FPC 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 time of the year and fishing positions.

1.4.2 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

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)

In order to do this, 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 a vessel operate under the management of both, moving from one fishery to another, within legally and 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.
VMS is nowadays a standard tool of fisheries monitoring and control worldwide, but it was the EU which 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 that the rules are respected. EU funding is available for Member States to acquire state-of-the-art equipment and to train their people to use it. [6]
If an international standard exists, the fisheries managers from all regions of the world would be able to set a common goal. However, some consensus on VMS

implementation, allow to provide 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", [12];
- Fuel efficiency: "Effects of fishing effort allocation scenarios on energy efficiency and profitability: An individual-based model applied to Danish fisheries",[2];
- Sustainable fishing: "The importance of scale for fishing impact estimations",[15];

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 visualisation 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, métier identification through the use of clustering techniques, linking logbook and VMS data together in order 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 [13].
- VMSbase: is an R package devised to manage, process and visualize information about fishing vessels 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 with respect to 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[18].

The main difference between my work and this previously mentioned is that they combine VMS data with the logbooks (data of type of fish captured and quantity). In this work it will only be used VMS data. The main advantage is that VMS data is less subject to malicious changes that logbooks taking into account that logbooks are filled by the ship owner and so subject to misrepresentation the truth. VMS data is generated automatically in a close system like a black box.

3

Data

This chapter provides the information about the used data and some analysis on it in how to use it for my work.

3.1 VMS Records

VMS Data provided by the Xsealence enterprise contained data generated by the MONICAP "Blue Box". Information about the localization, direction and velocity of the vessel at each 10 minutes, is saved in a local database. VMS datasets contained a vessel identification code, a timestamp, the latitude and longitude positions, the speed and direction. In this dataset there are 537138 entries from thirty vessels. These data is 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 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);
- Cog: Course Over Ground. Varies from 0 to 360 clockwise, being 0, facing north;
- Sog: Speed Over Ground (velocity in knots);

3.2 VMS Vessels

VMS Vessels 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.
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 (HP);
- kW: Vessel power (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 Appendix A a detailed descriptive analysis of these variables is provided.

3.3 Data Analysis

The data used as input to the models is VMS Records data. Within which the speed and location data are the ones that could have some demarking between the types of fishing.
About the locations, certain fishing types only ocourre in certain depth. So the locations can help in this cases.
In order to meet the objectives we need to understand the velocity patterns. The first feature we find when studying speed is its separation into two distributions.

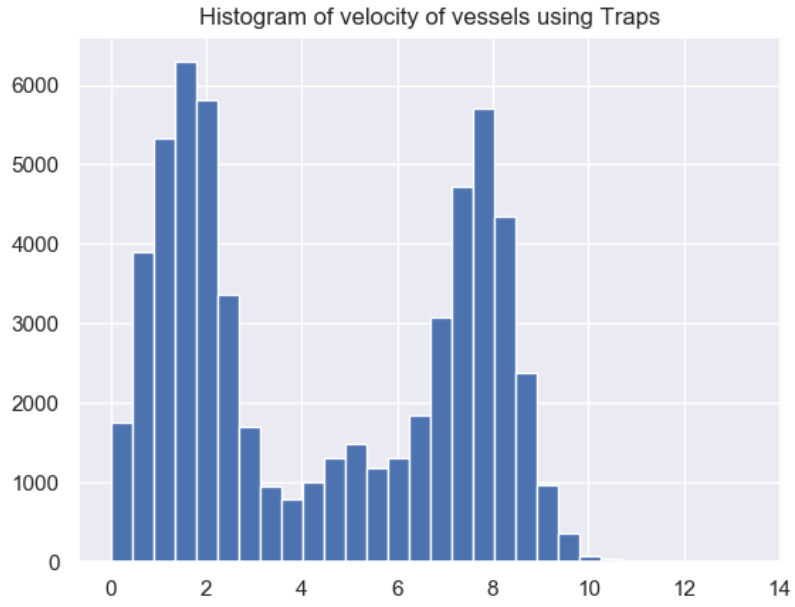

Figure 3.1: Histogram of velocity of traps license.

As we can see in 3.1 that at lower speeds we have the speeds at which the vessel is fishing and the higher speeds represent the movements of the vessel from the port to the fishing grounds and back to the port.

Knowing this we can notice that different types of fishing have different fishing speeds as we can see the differences between the histogram 3.1 containing data on fishing vessels using traps and **??** concerning fishing vessels using siege or 3.2 concerning fishing vessels using fishhook.

It is important to separate the inputs representing fishing speeds from the others in the data as the input to the models used in Chapter 5 it is necessary to have the data filtered out only with the fishing activity data. Data from other operations

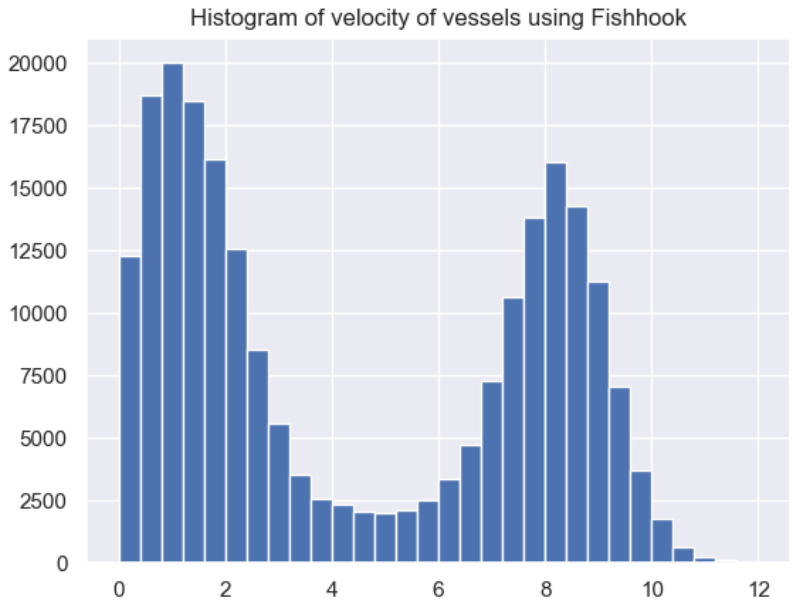

Figure 3.2: Histogram of velocity of fishhook license.

that are not fishing must be discarded because there are not discriminatory for the fishing type. For example, speed 0 nautical miles is not interesting as all vessels, regardless of their type of fishing, stop at the fishing port. Hence the importance of Chapter 4.

4

Blue Box

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.

4.1 Introduction

The first objective is to develop a locally implemented tool that informs whether the vessel is engaged in fishing, and if so, whether the fishing area is new or is habitually.
One of the solutions developed to meet this objective consists in a machine learning application to analyze data in real time, in order to determine whether the vessel is fishing, and if so, whether it is fishing in its fishing zone or in new location.
This solution must be implemented by vessel, therefore each vessel will only have access to its own data that is, each vessel only will know it's one data. The fact that this analysis is done by vessel allows avoiding bias in results, in fact each vessel has different power, size and its suitable for certain fishing activity.
This solution could be implemented and used as a library by the MONICAP system shown in 4.1 .[25].

Figure 4.1: 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 in relation to the standard.

4.2 Fishing Velocity Patterns

In order to know whether a vessel is fishing, we can use it's velocity patterns, given that the speed of the vessel differs where it is travelling or when it is fishing. We can verify this fact in the plot shown in 4.2 , corresponding to vessel with identification number 2.

On 4.2 the histogram allows us to recognize two different velocity patterns, identified by two distinct distributions that 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. So, it is needed to isolate the first distribution's range to be able to classify the upcoming future velocity's as fishing associated or not.

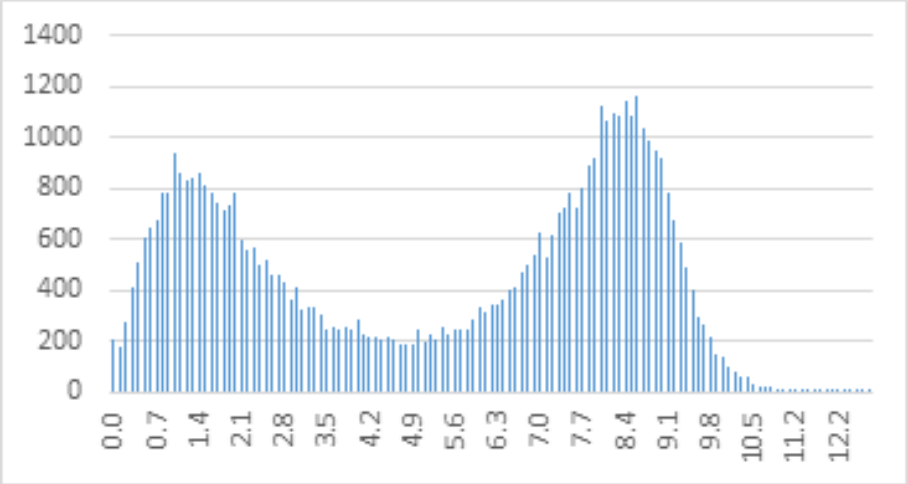

Figure 4.2: SOG Histogram vessel 2.

In 4.2 we have the speed in x axes in nautical knots and in the y axes the quantity of records per speed value.

To isolate the first distribution, it was used the Hill Climbing algorithm [10]. This algorithm is a local optimization algorithm which provides a direct search. The Stochastic Hill Climbing algorithm works supported in an iterate process of randomly selecting a neighbor for a candidate solution. The acceptance of the solution is conditioned by a criterion of improvement with respect to the previous solution.
The implementation used was altered in a way that when the algorithm converges to the maximum, it will continue to find the limit of the distribution. To obtain this solution, the algorithm searches for the first local maximum that does not have a higher value in the following three points (current key + 0.5, current key + 1 and current key + 2), 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. This way we can end up with a histogram of the intended distribution as we can observe in 4.3.

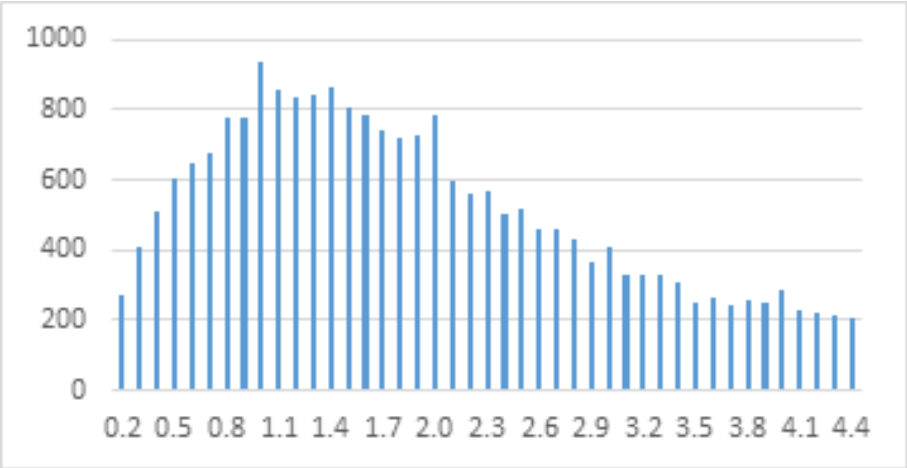

Figure 4.3: Velocity distribution of vessel 2 after the application to the Hill Climbing algorithm

Once the distribution corresponding to the fishing speeds has been identified the next step is to define a range to classify the new data, a minimum speed and a maximum speed. Inside this range we classify the vessel as fishing. With this purpose were considered three optional procedures:

- **Standard devariation:** This solution assume that the velocity distribution is a normal distribution. Using the distribution of the fishing velocity we find the fishing velocity with more occurrences and with the standard devariation we can chose a distance from mean and so we can get the minimum population desired (explained by Chebyshev's inequality [19]) to be within the fishing range.
- **Kernel Density Estimation:** Kernel Density Estimation method estimates the probability density function by imposing a model function on every data point and then adding them together. The function applied to each data point is called a kernel function [11].
- **Filter:** Using a filter that remove all the velocity occurrences that happens less than 10 % of the maximum occurrence and isolating the occurrences that are fallowed we can retrieve a clean distribution of the fishery speed of that vessel. With this we can use the first and last values to classify the new inputs. To isolate the fishery speed, I use a hill climb algorithm, assuming that the first distribution is the fishery speed.

The used procedure was performed in two parts: it starts by using the method

based in the **Filter** to isolate the fishery speed from the remain. 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 is shown in the 4.2 . In the next step it uses the hill climbing algorithm to get the minimum and the maximum value of the first distribution. The velocity 0 is removed because we don't want to consider when the vessel is completely stopped.

2. **Kernel:** It was applied a kernel distribution method in the filtered histogram to have the distribution represented in orange on 4.4. The it was created a dictionary with the velocity's and the cumulative percentage of velocity. This way we end up with a histogram like 4.5. Then a range across quantiles is defined for some probability. As in the estimation of confidence intervals, a confidence level is also defined here, to which will be associated the two speed limits that correspond to the fishing activity.

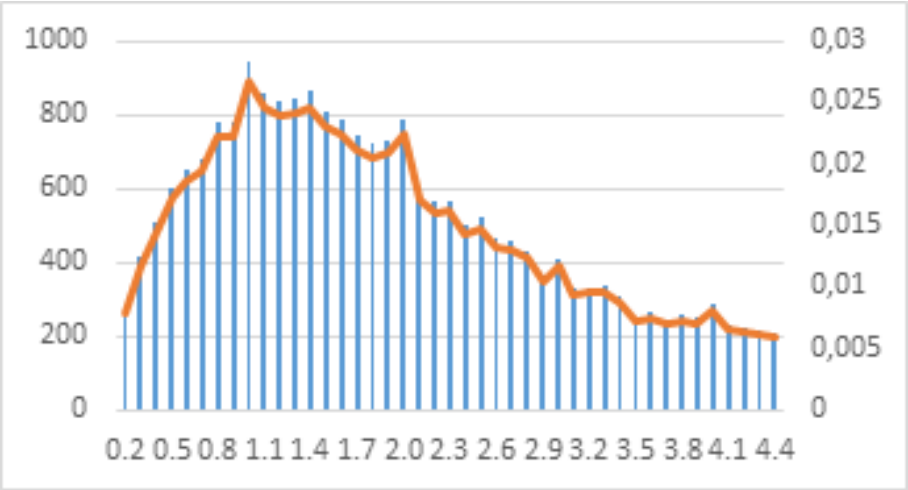

Figure 4.4: Kernel distribution filtered data

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

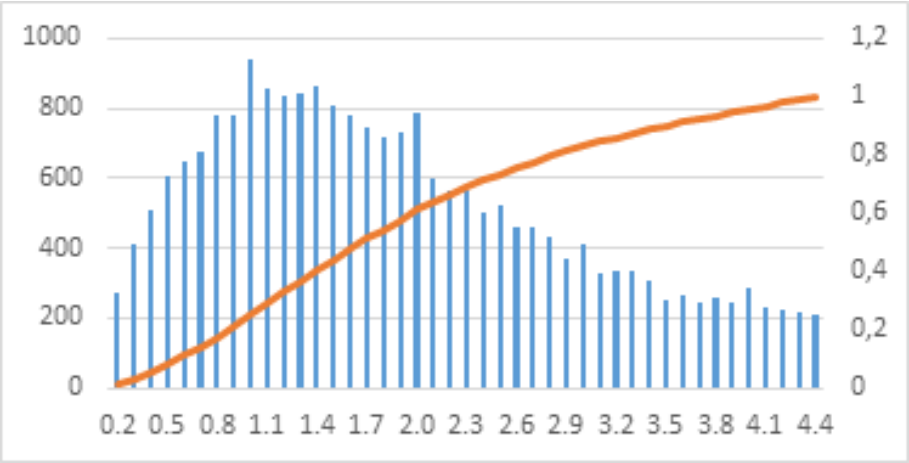

Figure 4.5: Filtred histogram with comulative kernel distribution

4.3 Fishing Spots

Using the history of GPS locations by vessel, in this point to decide whether the vessel is fishing in its fishing zone or in a new location. Fishing in a new zone may mean that the vessel has change its type of fishing or is engaging in an activity that is not licensed.

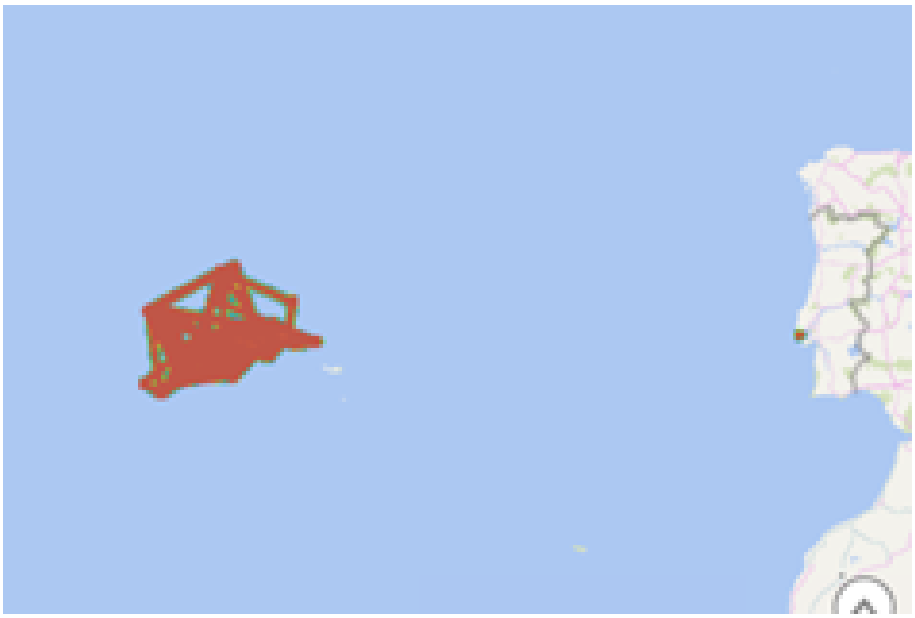

Figure 4.6: Vessel 2 GPS points

In 4.6 we can see the GPS points of a vessel. Using methods based in clustering it is possible to identify several areas by vessel that are the normal fishing zones of

the vessel. When the vessel is outside that 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 we can compare with the new GPS points.

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

- **K-Means:** K-means clustering algorithm [7] 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 into 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 (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better 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.

- **Density Based Cluster:** Density-based clustering algorithms [22] try to find clusters based on the estimation of the density of data points in a region. 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 clusters are dense regions in the data space, separated by regions of lower object density [23]. 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).

- **DBSCAN:** 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 a number of clusters starting from the estimated density distribution of corresponding nodes. DBSCAN [8] is one of the most common clustering algorithms and most cited in scientific literature.

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 4.3, this model needs a lot of processing power to estimate the clusters. These values were retrieved using a computer with an Intel i5 (2.5 GHz) and 8 GB of RAM. Considering that the Blue Box as a lot less processing power it was decided that this model is not a good solution for this problem.

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

Table 4.1: Milliseconds per model

The choice between K-Means and Density Based Cluster algorithms was based in the fact that Density Based Cluster represents a great advantage because it estimates the probability of the new GPS point belonging to a cluster based in the cluster probabilistic distribution. So that way the user can set what is the best configuration. The clusters themselves are equal between K-Means and Density Based Cluster since Density Based Cluster uses K-Means to define the centroids, only differ by adding a layer to define the area of density per cluster.
In order to decide the number of clusters it was used the elbow method [9], for the within cluster sum of squares, as could be seen in 4.7. The within means the distance the vectors in each cluster are from their respected centroid. The goal is to get this number as small as possible. One approach to handling such objective is to run the kmeans clustering multiple times, raising the number of the clusters each time. Then it is possible to compare the withinss each time, stopping when the rate of improvement drops off. The better case corresponds to find a low

withinss while still keeping the number of clusters low.
The elbow method is a visual method. The idea is that 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. This is the K value we want. We can observe that six clusters are a good number as the error is not decreasing much as the number of clusters increases. To exemplify the data used in 4.7 was the GPS points of all vessels.

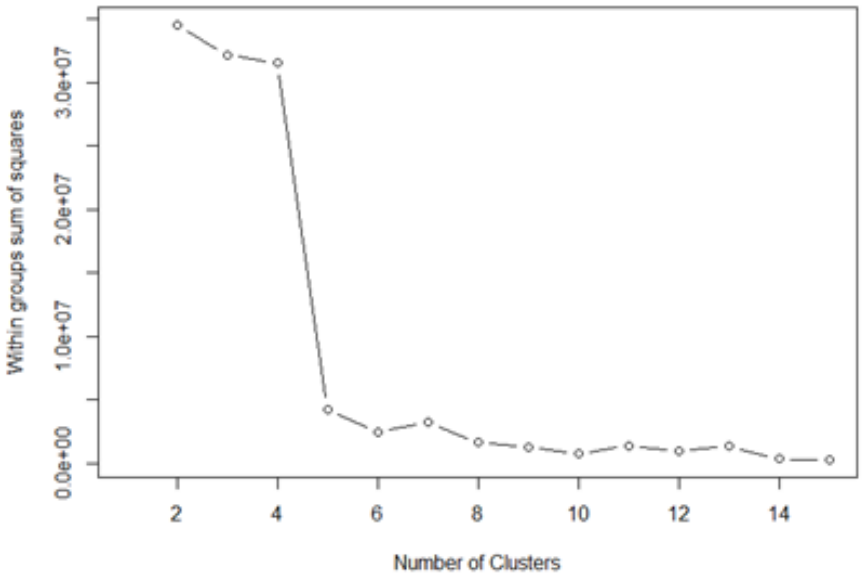

Figure 4.7: Sum of squared error by number of clusters

4.4 DSA Library

It was created a software application called DSALib for (Decision Support Alerts Library). In this application it was applied 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 MONICAP to send alerts to support decision making.
The application starts by initializing two objects:

1. ProcessVelocity: This object is responsible to do the process explained in the point 4.2 of this document. This object 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 (minimal speed of fishing and the maximal speed of fishing).

2. ProcessNewArea: This object is responsible to do the process explained in the point 4.3 of this document. This object is only initialized after ProcessVelocity because it needs the velocity fishing limits to only create the clusters of the fishing areas. With these limits the object request to ImportData only the Lat and Lon where the vessel was in between the velocity limits. With this the object ends up with the clusters of the fishing areas.

To use we start by instantiate DSA with the doubles limitVelociry and limitArea. This doubles range between 0 and 1:

- limitVelociry: used to get the maximum and minimum speed by reducing the speed range. This limit will reduce de maximum speed and increase the minimum speed by setting the maximum velocity as the velocity that as (1-limit) percentage of the cumulative karnel distribution and the minimum velocity as the velocity that as (limit) percentage of the cumulative karnel distribution.
- limitArea: used to compare with the probability to belong in a cluster given to the new points. If the limit is smaller that the given probability, then the vessel is classified as fishing in a new area.

These limits are important so we can configure if we prefer to have more false positive or false negative classifications. A false positive (type I error) is when the classifier reject a true hypothesis. A false negative (type II error) is when the classifier accept a false hypothesis.

After to have the two objects 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 and an “isNewArea” as true if the vessel is in an area that is not a normal fishing area and it’s in a fishing velocity. In Appendix B there is more information about the software developed.

To develop the software, I decided to use Java 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.

5

Server

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

5.1 Introduction

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 large 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) [14].

In this work it will be used the CRISP-DM (CRoss Industry Standard Process for Data Mining) methodology [24]. 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 [24]. 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 in six phases which are shown in 5.1. 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.

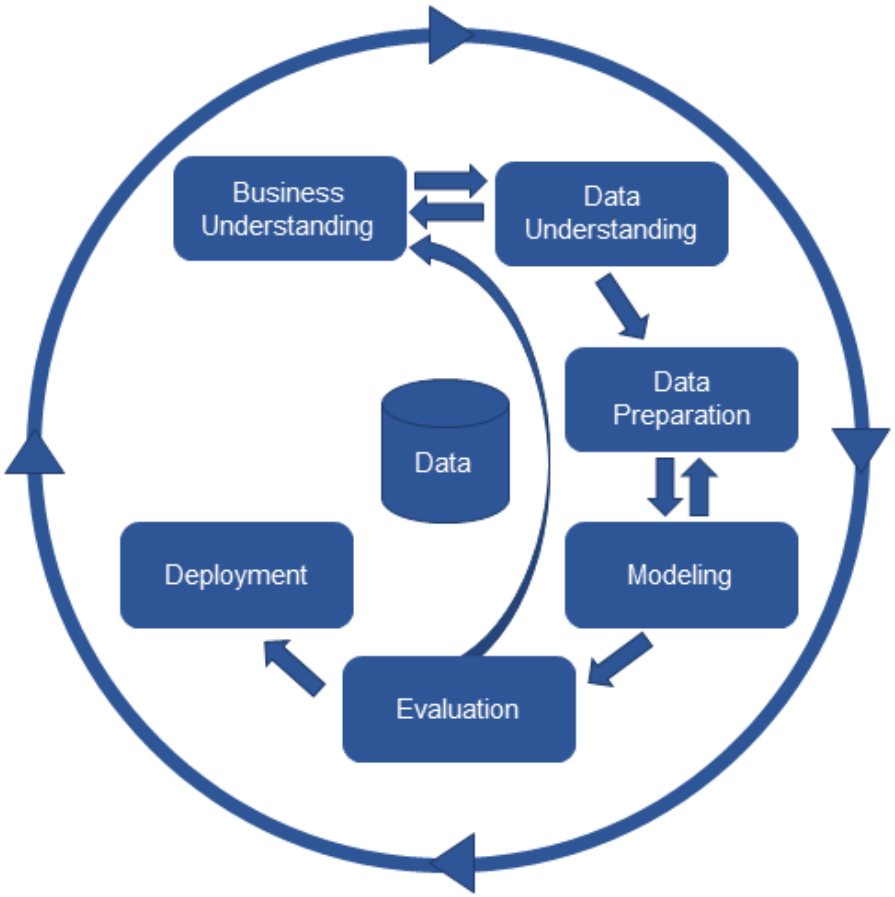

Figure 5.1: Complete CRISP-DM Approach.

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, and a preliminary project plan designed to achieve the objectives.

- Data Understanding
  The data understanding phase starts with an initial data collection and proceeds with activities in order 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 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 final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important 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
  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 up front what actions will need to be carried out in order to actually make use of the created models.

5.2 Business Understanding

In the fishing industry it is necessary that vessels have licenses to use fishing techniques. The problem is that there is a probability that vessels are using fishing techniques or gear that are not licensed to do so. The objective is to classify the VMS data by fishing type. In this way we can try to confirm if each boat is carrying out the type of fishing for which it has the license payed.

5.3 Data Understanding

The data we use for this objective is the same VMS Records as used in chapter 4 and VMS Vessels explained in chapter 3.2. Let us focus on the use of speed, but not discriminating any of the remaining columns a prior without determining its potential in the contribution to a solution

5.4 Data Preparation

To use data mining models, we need a dataset with all the data needed to feed the models. So, I created a dataset from VMS Vessels and VMS Records to end with Table 5.4.

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
LOA	Length Overall	VMS Vessels	Discriminated by vessel type
GT	Gross Tonnage	VMS Vessels	Discriminated by vessel type
HP	Vessel Power	VMS Vessels	Discriminated by vessel type
License	Vessel's Linceses	VMS Vessels	Objective

Table 5.1: VMS Dataset

The License that occurs in the dataset are:

- Armadilhas / De abrigo / Alcatruzes
- Arrasto / De fundo de portas
- Arrasto / De fundo de portas / Crustáceos
- Arrasto / Pelágico / Com portas
- Cerco / para bordo / Tipo americano
- Emalhar de 1 pano / De deriva / Grandes Pelágicos
- Emalhar de 1 pano / De fundo
- Pesca à linha / Cana e linha de mão
- Pesca à linha / Palangre de fundo / Espécies demersais
- Pesca à linha / Palangre de Fundo + Cana e linha de mão
- Pesca à linha / Palangre de superfície / Grandes Migradores
- Tresmalho / De fundo

The first thing to test is the correlation of the data to check if licenses are strongly correlated to some of the other variables. In Figure 5.2 (the method used to find correlation was Pearson Correlation [3] that measures the degree of correlation and the direction of this correlation - whether positive or negative) between two metric scale variables). We can see that we can't use the data as it is because de correlation between License and sog are very weak so we need to do some pre-processing.

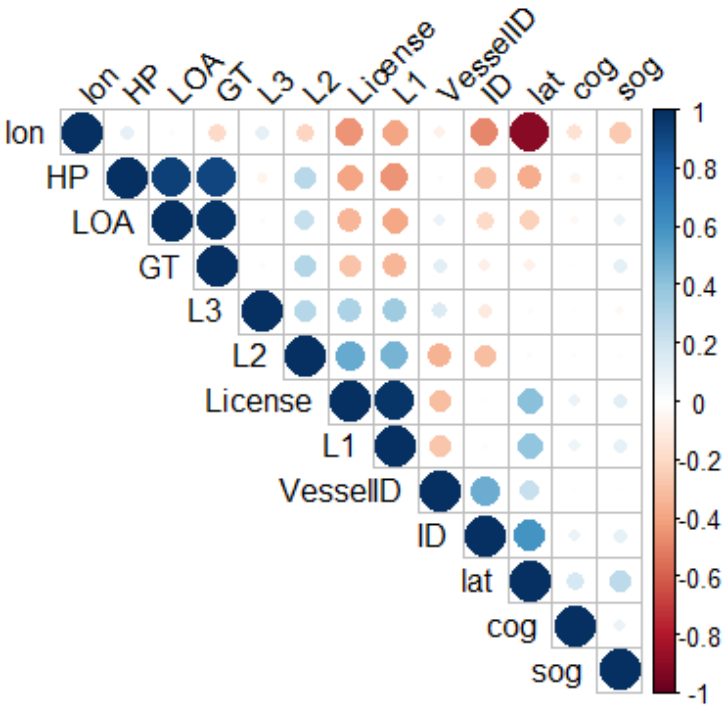

Figure 5.2: Data Correlation

The best way to achieve greater correlation between fishing speeds and licenses was the transformation of data to storage fishing average, maximum and minimum per day per vessel. This way it presents the results shown in Figure 5.3.

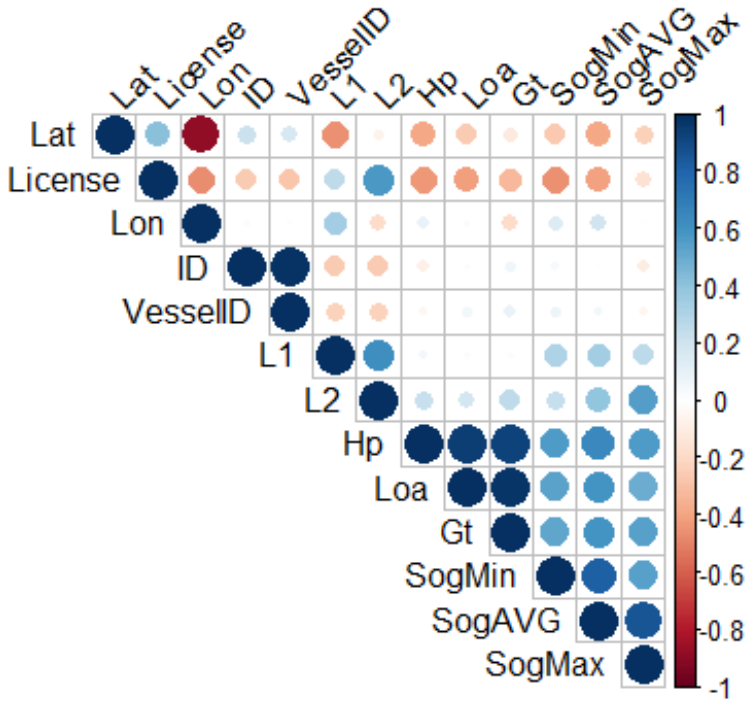

Figure 5.3: Data coorelation after preprocessing

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

- Create dataset in which data is grouped per day, per vessel.
- Use DSALib to get the minimum and maximum speeds of fishing for poop vessel. With this data, filter the dataset to be only with data in which the speed of the vessel is between the minimum and maximum obtained.

With this data we can also observe that the correlation between the license and the HP data (vessel power), Loa (Boat length) Gt (vessel weight) are quite significant. This is normal 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 I will not use these variables in the model so as not to create a problem with bias.
Regarding location data, used clustering techniques to discretize the data. First, we need to know what the best number of clusters is. For that we used the same technique used in chapter 4.3 with the output in Figure 5.4. The data used was the dataset filtered so we have only the positions of fishing. The chosen number of clusters was 4.

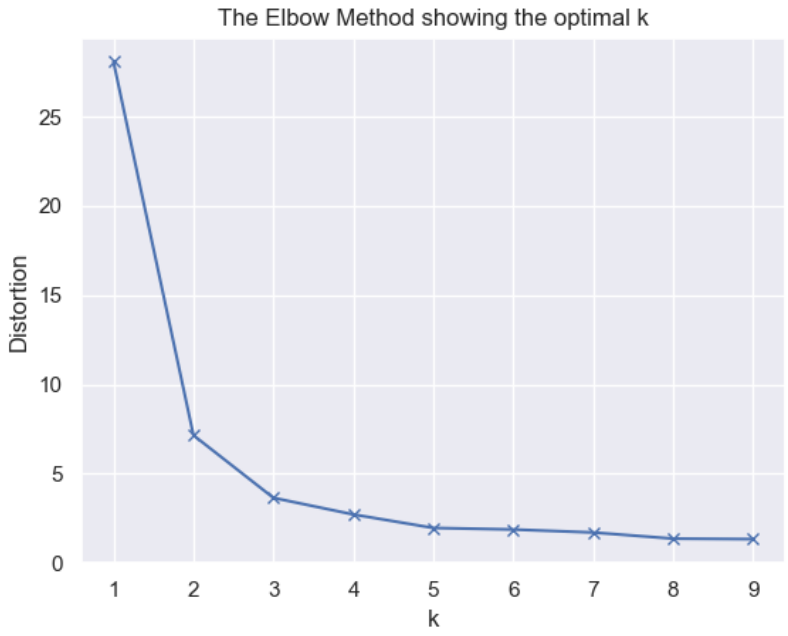

Figure 5.4: The elbow method showing the optinal k

Now we can create data mining models with location and velocity parameters.

5.5 Modeling

Several data mining algorithms were used to create the necessary models to classify the license type with the VMS data.

- **KMeans:** This model was used in the previous step with the GPS data, so it was classified in well-defined clusters in order to improve the operation

of the data mining algorithms. The operation of this algorithm is described in chapter 4.3.

- **DecisionTrees:** While in data mining a decision tree is a predictive model which 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. 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 [17]. Algorithms for constructing decision trees usually work top-down, by choosing a variable at each step that best splits the set of items [16].
- **Neural Network:**
- **Support Vector Machine:**

Cross –validation [21] provides a simple and effective method for both model selection and performance evaluation, widely employed by the machine learning community. Under k –fold cross –validation the data are randomly partitioned to form k 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 k –1 subsets. The average of the generalization performance observed over all k folds provides an estimate (with a slightly pessimistic bias) of the generalization performance of a model trained on the entire sample. The k used to test this models is 10.

5.6 Evaluation

The model's test results are:

- **DecisionTrees:**

Algorithm	Gini	Entropy
Velocity and locations	0.7771544	0.7780856
Velocity	0.7264879	0.730252

Table 5.2: Cross-Validation results for Decision Trees models

- **Neural Network:**

Layers	15	30 3	100 3	500 4
Velocity and locations	0.6820688	0.7032381	0.733293	0.7433809
Velocity	0.6614556	0.7178236	0.7214212	0.7295752

Table 5.3: Cross-Validation results for Neural Network models

- **Support Vector Machine:**

Kernel coefficient	auto	scale	linearSVC
Velocity and locations	0.6970205	0.694583	0.601798
Velocity	0.6610622	0.664224	0.5191799

Table 5.4: Cross-Validation results for Support Vector Machine models

Evaluate Results Assessment of Data Mining Results w.r.t. Business Success Criteria Approved Models

Review Process Review of Process

Determine Next Steps List of Possible Actions Decision

5.7 Deployment

Plan Deployment Deployment Plan

Plan Monitoring and Maintenance Monitoring and Maintenance Plan

Produce Final Report Final Report Final Presentation

Review Project Experience Documentation

6

Conclusion

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

With the work done we were able to demonstrate several 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 the developer work was the possibility to understand the fishing patterns in order to be able to create plans of environmental protection.

6.1.1 Blue Box

In the first solution we were able to demonstrate that it is possible to classify in real-time two important 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 historical speed of the same, 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 very useful to understand the fishing patterns in a given area. Other interesting 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 in activity 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).

6.1.2 Server

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 a lot of work to have variables of enough quality to create a good classifying model. I intend to create different types of data mining algorithms to determine which best fits this problem. Finally, I intend to create an informatic system capable of receiving, classifying and verifying the VMS data of the fishing vessels.

6.2 Future Work

References

[1] T. Agardy. Effects of fisheries on marine ecosystems: a conservationist's perspective. *ICES Journal of Marine Science*, pages 761–765, 2000. (p. 1)

[2] 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*, volume 106, pages 501–516, September 2010. (p. 6)

[3] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. *Pearson Correlation Coefficient*, pages 1–4. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-642-00296-0. doi: 10.1007/978-3-642-00296-0_5. URL https://doi.org/10.1007/978-3-642-00296-0_5. (p. 27)

[4] K. S. Alfred M. Duda. A new imperative for improving management of large marine ecosystems. *Ocean and Coastal Management*, pages 797–833, 2002. (p. 1)

[5] J. e. C. S. Pires. Consumo de peixe em portugal. https://sites.google.com/site/docapescacreative/consumo-de-peixe-em-portugal, 2013. Accessed on 2018-04-06. (p. 1)

[6] ec.europa.eu. European commission. https://ec.europa.eu/fisheries/cfp/control/technologies/vms_en, 2018. Accessed on 2019-01-21. (p. 5)

[7] Jinxin Gao and David B. Hitchcock. James–stein shrinkage to improve k-means cluster analysis. *Computational Statistics & Data Analysis*, 54 (9):2113 – 2127, 2010. ISSN 0167-9473. doi: https://doi.org/10.1016/

j.csda.2010.03.018. URL http://www.sciencedirect.com/science/article/pii/S0167947310001209. (p. 19)

[8] 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. (p. 20)

[9] 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. (p. 20)

[10] Vladimir Kvasnicka, Martin Pelikan, and Jirí Pospíchal. Hill climbing with learning (an abstraction of genetic algorithm). In *Hill Climbing with Learning*, 1995. (p. 15)

[11] 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 Thanaruk Theeramunkong, Boonserm Kijsirikul, Nick Cercone, and Tu-Bao Ho, editors, *Advances in Knowledge Discovery and Data Mining*, pages 302–313, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. ISBN 978-3-642-01307-2. (p. 16)

[12] 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*, pages 1–5, May 2015. doi: 10.1109/OCEANS-Genova.2015.7271551. (p. 6)

[13] 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*, pages 31–43, September 2012. (p. 6)

[14] 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. (p. 23)

[15] G. J. Piet and F. J. Quirijns. The importance of scale for fishing impact estimations. *Canadian Journal of Fisheries and Aquatic Sciences*, 66(5):829–835, 2009. doi: 10.1139/F09-042. URL https://doi.org/10.1139/F09-042. (p. 6)

[16] L. Rokach and O. Maimon. Ieee transactions on systems, man, and cybernetics—part c: Applications and reviews publication information. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35(4): c2–c2, Nov 2005. ISSN 1094-6977. doi: 10.1109/TSMCC.2005.859799. (p. 31)

[17] Lior Rokach and Oded Maimon. *Data Mining With Decision Trees: Theory and Applications*. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2nd edition, 2014. ISBN 9789814590075, 981459007X. (p. 31)

[18] 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*, 9(6):1–18, 06 2014. doi: 10.1371/journal.pone.0100195. URL https://doi.org/10.1371/journal.pone.0100195. (p. 7)

[19] John G. Saw, Mark C. K. Yang, and Tse Chin Mo. Chebyshev inequality with estimated mean and variance. *The American Statistician*, 38(2):130–132, 1984. ISSN 00031305. URL http://www.jstor.org/stable/2683249. (p. 16)

[20] 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. (p. 1)

[21] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147, 1974. ISSN 00359246. URL http://www.jstor.org/stable/2984809. (p. 31)

[22] Monika Jena Swasti Singhal. A study on weka tool for data preprocessing, classification and clustering. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 2(6):250–253, 2013. ISSN 2278-3075. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.687.799&rep=rep1&type=pdf. (p. 19)

[23] Prajwala Talanki. Comparative analysis of em clustering algorithm and density based clustering algorithm using weka tool. *International Journal of Engineering Research and Development*, 9:2278–800, 02 2014. (p. 19)

[24] 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*, 01 2000. (p. 23)

[25] Xsealence. Xsealence. http://www.xsealence.pt/portfolio/monicap/, 2018. Accessed on 2019-07-31. (p. 13)