Marine Policy 84 (2017) 110–118

Contents lists available at ScienceDirect

Marine Policy

journal homepage: www.elsevier.com/locate/marpol

Forecasting bivalve landings with multiple regression and data mining techniques: The case of the Portuguese Artisanal Dredge Fleet

Manuela M. Oliveira[a,b,f,*], Ana S. Camanho[c], John B. Walden[d], Vera L. Migués[c], Nuno B. Ferreira[e], Miguel B. Gaspar[f,g]

[a] INESC-TEC, Portugal
[b] Faculdade de Ciências da Economia e da Empresa, Universidade Lusíada de Lisboa, Portugal
[c] Faculdade de Engenharia da Universidade do Porto, Portugal
[d] NOAA Fisheries, Northeast Fisheries Science Center, USA
[e] IBS-ISCTE IUL, Lisboa, Portugal
[f] Instituto Português do Mar e da Atmosfera I.P./IPMA, Portugal
[g] Centro de Ciências do Mar, Universidade do Algarve, Portugal

ARTICLE INFO

Keywords:
Data mining
Random forests
Multiple regression
Forecasting
Small scale fisheries
Bivalve fisheries

ABSTRACT

This paper develops a decision support tool that can help fishery authorities to forecast bivalve landings for the dredge fleet accounting for several contextual conditions. These include weather conditions, phytotoxins episodes, stock-biomass indicators per species and tourism levels. Vessel characteristics and fishing effort are also taken into account for the estimation of landings. The relationship between these factors and monthly quantities landed per vessel is explored using multiple linear regression models and data mining techniques (random forests, support vector machines and neural networks). The models are specified for different regions in the Portugal mainland (Northwest, Southwest and South) using six years of data 2010–2015). Results showed that the impact of the contextual factors varies between regions and also depends on the vessels target species. The data mining techniques, namely the random forests, proved to be a robust decision support tool in this context, outperforming the predictive performance of the most popular technique used in this context, i.e. linear regression.

1. Introduction

The bivalve dredge fishery is considered one of the most important artisanal fisheries in mainland Portugal. It involves a large number of fishers and vessels, and the catch value is a large proportion of total revenue from fisheries of coastal communities. The sustainability of the bivalve fisheries has been at risk in the last few years. A contributing factor to this risk has been compulsory closure of the fishery due to phytotoxin episodes. With increased frequency of phytotoxin closures, the goal of this analysis is to develop a decision support tool that can help administrative/regulatory fishery authorities to forecast bivalve landings accounting for several contextual conditions.

Predicting trends has been an area of interest in several scientific fields, including fisheries. Forecasting landings and understanding their determinants are considered key issues for fishery managers. However, due to the uncertainties involved and the multitude of variables that influence fishing activity, forecasting landings is a very challenging task.

Despite the challenges, several techniques have been used to forecast fishery landings [1–7]. Multiple linear regression and discriminant analysis are two frequently used techniques for the construction of predictive landings' models. Yet, these models have some limitations. For example, the relationship between the dependent and independent variables is often non-linear, which prevents accurate modelling using approaches based on linear models. The analysis of landing series, even those from non-mobile species, often provides evidence of a dynamic behaviour, with random fluctuations of apparently chaotic nature. These fluctuations are more pronounced for short time windows and smoothed over time for long-term windows. However, increases in computing power have opened up the possibility of using different data mining techniques in forecasting models rather than traditional statistical models such as linear regressions.

Both data mining techniques and regression models have unique advantages. For example, an advantage of regression models over data mining techniques is the possibility to identify the individual impact of specific factors on the dependent variable, through the analysis of the regression coefficients and their significance. For certain data mining techniques, this is not possible. Instead, it is only possible to obtain a

* Corresponding author at: Faculdade de Ciências da Economia e da Empresa, Universidade Lusíada de Lisboa, Portugal.

http://dx.doi.org/10.1016/j.marpol.2017.07.013
Received 12 May 2017; Received in revised form 17 July 2017; Accepted 17 July 2017

measure of importance of each specific factor in the prediction performance of the model (see Section 3 for further details).

In fisheries, data mining techniques have generally been applied for classification purposes, particularly in the identification of fish species [8–16]. Joo et al. [17] extended data mining methods to the identification of fishing set positions from vessel monitoring system (VMS) data whereas Mendonza et al. [18] analysed the impacts of fishing inactivity due to closed seasons.

Despite the variety of studies applied to fisheries, the use of data mining for yield [19] or landings prediction is scarce – especially in the artisanal segment. Previous studies of the Portuguese artisanal dredge fleet have confirmed that factors such as the hydrodynamic conditions, phycotoxin episodes [20] and monthly seasonality [21] have an impact on vessel landings. The magnitude of this impact varies along the Portuguese coast. This study extends previous analysis by testing the influence on landings of a larger range of factors, including tourism and biomass stock indicators. An innovative methodology, applying both regression models and data mining techniques, was developed to forecast landings and to explore the impact of different external factors on landings using a system-wide approach. This means that all factors and their interrelationships are simultaneous considered in an overall model.

2. Portuguese artisanal dredge fishery

The dredge fleet is comprised of local and coastal vessels, where their classification depends on the area in which they are allowed to operate. Local vessels can only operate near their homeport or adjacent fishing ports, whereas coastal vessels can act within the fishing area for which they are registered. Currently, there are 84 active dredge vessels, which are distributed in three main fishing areas along the Portuguese coast (11 operating in the Northwest, 25 in the Southwest and 48 in the South). In both Northwest and Southwest fishing areas, only coastal vessels operate due to the distance of bivalves' beds from fishing ports and the adverse hydrodynamic conditions (i.e., high mean wave height (MWH)). The fleet targets five species. These are: surf clams (*Spisula solida*), which can be caught along the entire coast. Donax clams (*Donax trunculus*), striped venus clams (*Chamelea gallina*) and razor clams (*Ensis siliqua*) are caught between Lisboa and Sines, in the Southwest fishing area, and between Sagres and Vila Real de Santo António, in the South fishing area, and the smooth clam (*Callista chione*) which is only exploited in the Southwest due to its low abundance in the remaining fishing areas.

Seasonal closures, gear specifications and minimum landing sizes are management measures common to the three fishing areas. The quota regime differs between the Northwest and Southwest (maximum weekly quotas are currently in place) and the South (maximum daily quotas). Such differences are explained by the milder oceanographic conditions observed in the South coast compared to the West coast.

The Portuguese Institute for the Ocean and Atmosphere (IPMA) is the national entity responsible for managing and monitoring bivalve production zones (BPZ). European Union (EU) legislation established that legal controls of harmful algal blooms (HABs) would reside with the IPMA. For management purposes, the entire coast of mainland Portugal is divided into nine BPZs (Fig. 1), in which bivalve harvesting can be restricted to one or more species, to prevent the catch of species that contain levels of phycotoxins above prescribed EU limits. The Northwest fishing area comprises the L2, L3 and L4 BPZs. However, L4 is not explored, because of its distance from the main fishing ports. The Southwest fishing area includes L5 and L6 BPZs. Finally, the South fishing area encompasses the L8 and L9 BPZs. L1 and L7 are not explored due to the lack of bivalve beds.

3. Methodology

The purpose of this study is to predict dredge fishery landings, and to evaluate the impact of different external factors on predicted landings. This was accomplished through the use of the most popular technique in the fishery sector, i.e. multiple linear regression (MLR). For each fishing area, a MLR model was specified. This was followed by the use of a data mining techniques (random forests, support vector machines and neural networks) to refine the predictions of landings based on external factors.

The prediction performance of the linear regression and the data mining techniques was assessed using a 10-fold cross validation approach. This means that each dataset related to a vessel and a species was divided into 10 blocks. The models were trained with 9 of these blocks and evaluated with the other one, meaning that the observed landings were compared with the predicted landings. The process is repeated 10 times, once for each different block. In the end, all predictions were combined and the average performance is quantified using the predicted R-squared (R^2) and the root mean squared error (RMSE).

3.1. Multiple Linear Regression

Multiple linear regression is a statistical method that is used to analyse the relationship between a single response variable (dependent variable) with two or more controlled variables (independent variables). In this context, a model of MLR can be expressed as shown in (1).

$$Y_{ijk} = \alpha + \sum_{m} \beta_m X_{m,ijk} + \varepsilon_{ijk} \quad (1)$$

The regression models were specified for each region and species, in a total of 7 different models (surf clam in the North, surf clam, donax clam and smooth clam in the Southwest and surf clam, donax clam and striped venus clam in the South). Y_{ijk} is the response variable (quantity landed by vessel i, in month j from year k) for the species and region analysed.α represents the regression intercept and ε_{ijk} is the error. The parameters β_m (m = 1,…,7) represent the regression coefficients, and $X_{m,ijk}$ (m = 1,…,7) are the independent variables. $X_{1,ijk}$ is the type of vessel (binary variable equal to zero for coastal vessels and one for local vessels), with a fixed value for all j and k for a given vessel i. $X_{2,ijk}$ is the vessel tonnage, with a fixed value for all j and k for a given vessel i. $X_{3,ijk}$ corresponds to the fishing days of vessel i in month j of year k. $X_{4,ijk}$ is the biological stock indicator, $X_{5,ijk}$ is the mean wave height, $X_{6,ijk}$ represents the phytotoxins episodes and $X_{7,ijk}$ is the tourism indicator. The values of these variables change overtime, and were estimated on a weekly basis for the case of $X_{4,ijk}$ and $X_{5,ijk}$, and on a monthly basis for the case of $X_{6,ijk}$ and $X_{7,ijk}$.

3.2. Random forests

The random forests technique consists of an ensemble of multiple regression trees (see Breiman [22] for further details). Generally, a regression tree consists of recursively splitting the training set into several partitions and developing a model that specifies the relationship between the independent variables and the dependent variable for the corresponding partition.

Each regression tree used by the random forests algorithm is constructed based on different training sets which are drawn independently with replacement from the original dataset. Furthermore, each regression tree is generated considering a random sample of the independent variables. Usually, for prediction purposes, the number of variables selected is one third of the total number of independent variables and the number of trees generated is large [22].

Having generated the trees, the algorithm defines the response value of the dependent variable corresponding to each observation of the test dataset as the average prediction of the individual regression trees.

The measure of importance of the independent variables is usually