UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE ESTATÍSTICA
E INVESTIGAÇÃO OPERACIONAL

STOCHASTIC FRONTIER ANALYSIS APPLIED TO THE FISHERIES

Nuno Madeira Veiga

MESTRADO EM ESTATÍSTICA

2011

UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE ESTATÍSTICA
E INVESTIGAÇÃO OPERACIONAL

STOCHASTIC FRONTIER ANALYSIS APPLIED TO THE FISHERIES

Nuno Madeira Veiga

Dissertação orientada pela Prof. Doutora Maria Lucília Carvalho
e supervisionada pela Doutora Ivone Figueiredo

MESTRADO EM ESTATÍSTICA

2011

Aknowledgments

A presente tese foi desenvolvida no âmbito do projecto DEEPFISHMAN, FP7-KBBE-2008-1-4-02, Management and Monitoring of Deep-sea Fisheries and Stocks.

Quero agradecer...

À Dra. Ivone,
por ter acreditado no meu valor, pelo que me fez evoluir e aprender durante este tempo.

À Prof. Lucília,
por ter apostado em mim, pelo que me proporcionou e pelo que me transmitiu.

À Prof. Isabel,
pelo apoio e pela ajuda para a apresentação do poster.

Aos meus colegas do Ipimar,
pela calorosa recepção, pela fácil integração e pela inesgotável simpatia.

Aos meus amigos,
por serem amigos no verdadeiro sentido da palavra.

À minha namorada,
por todo o apoio, amor e paciência que tiveste e tens para comigo e por seres a fonte da minha força.

Ao meu pai, irmã e cunhado,
pelo apoio incondicional ao longo destes seis anos.

À minha mãe,
por não me ter deixado desistir ...

Contents

Abstract

In fisheries world the knowledge of the state of the exploited resource, is vital to guarantee the conservation of the resource and the sustainability of the fishery itself. The present study is focused on the Portuguese longline deep-water fishery that targets black scabbardfish. This fish is a deep-water species and its landings have an important economical value for Portugal. The fleet that explores the species is composed by 15 vessels with a mean overall length of 17 m.

In the first part of this work Generalized Linear Model was used to standardize the Capture-per-unit-effort, so the first aim is to improve the estimate of CPUE, which is widely used as an index of stock abundance. This is done by reanalyzing the data stored at Portuguese General Directorate from fishery industry and in particularly the logbooks, which are used to record catch data as part of the fisheries regulation.

The second part focused on Technical Efficiency, which refers to the ability to minimize the production inputs or the ability to obtain the maximum output. In this study TE estimates were obtained through Stochastic Frontier Analysis. This methodology embraces two science fields, Economy and Statistics, and has been the subject of studies in various areas but there are few applications to fisheries and the available ones are often studied from the economic point of view rather than a statistical one.

This work aimed to analyze the quality of the logbooks and identify the relevant factors to the CPUE estimation as the theoretical evaluation of SFA approach and the identification of the statistical differences between several models. TE of each vessel was estimated and was verify if the black scabbardfish fishery operating in Portugal mainland can be considered efficient.

Keywords: Black scabbardfish, Catch-per-unit-effort, Generalized Linear Models, Stochastic Frontier Analysis, Technical Efficiency.

Resumo

Portugal é um país costeiro com cerca de 1200 km de costa, fazendo da pesca uma das actividades mais importantes, económica e culturalmente. Uma das espécies mais pescadas em Portugal é o peixe-espada preto, fazendo desta espécie uma das mais estudadas devido ao seu impacto socioeconómico. Desde o século XVII que na Madeira, o peixe-espada preto é pescado, mas só em 1983 foi iniciada esta pesca em Portugal continental, sendo Sesimbra a principal zona pesqueira. Assim sendo, foi de Sesimbra que vieram grande parte dos dados que foram usados neste trabalho.

A regulação e a gestão da actividade pesqueira continuam a ser um dos maiores desafios, sendo assim essencial a avaliação do estado dos recursos explorados (neste caso o peixe-espada preto). Tal avaliação é vital para procurar medidas que garantam a sustentabilidade do recurso e da pesca.

Um dos índices de abundância mais utilizados é o CPUE (captura-por-unidade-esforço), que é definido como a razão entre o total capturado e o total de esforço aplicado nessa mesma captura. Apesar do seu frequente uso é sabido que o CPUE é influenciado por outros factores para além do nível de abundância. Assim, para minimizar essa influência, o CPUE é estandardizado de forma a diminuir ou até remover os eventuais factores de confusão. Para tal foram aplicados Modelos Lineares Generalizados (GLM), que não são mais do que uma generalização dos Modelos Lineares. Essa generalização permite que a distribuição da variável resposta pertença à família exponencial (para além da Normal), e permite que a função de ligação entre a variável resposta e as variáveis explicativas seja uma função monótona diferenciável.

Para estimar tal índice, a fonte de dados é frequentemente o diário de bordo. Na União Europeia e desde a introdução de Política Comum das Pescas, que reúne várias medidas para garantir a sustentabilidade da pesca europeia, é obrigatório registar toda a viagem desde a partida do porto até ao desembarque. Além disso, dado que não há dados independentes da pesca, ou seja, não há estudos dirigidos para a recolha de dados através de amostragem, a estimação deste tipo de índices acaba por depender quase exclusivamente dos diários de bordo. Assim acabam por assumir uma importância vital quer na monitorização quer na regulamentação da actividade pesqueira.

O preenchimento destes diários de bordo é feito pelos mestres das embarcações no mar e é posteriormente introduzido numa base de dados pela Direcção Geral das Pescas e da Aquicultura. Contudo há erros ou más interpretações no preenchimento dos diários de bordo que podem de alguma forma enviesar quer os resultados quer as conclusões de estudos neles baseados. Além de que os dados retirados dos diários de bordo reflectem sempre imensa variedade nas espécies capturadas além da espécie alvo. Apesar disto, os diários de bordo são a fonte de dados de vários trabalhos que visam estimar níveis de abundância.

Desta forma, é necessário medir e quantificar o impacto que uma base de dados menos cuidada pode ter na qualidade e na veracidade dos trabalhos que nela se baseiam. É este objectivo que visa a primeira parte deste trabalho (chapter 2), usando os dados contidos nos diários de bordo da frota que opera em Sesimbra e que tem como espécie alvo o peixe-espada preto. Os factores e variáveis relevantes para a estimação do CPUE também foram identificadas, assim como a respectiva influência.

Desta primeira parte do trabalho resultou uma análise extensiva e detalhada dos diários de bordo, permitindo identificar os erros e até nalguns casos corrigi-los através do conhecimento de trabalho anteriores e da comunidade pesqueira de Sesimbra. Análise essa que recorreu a várias ferramentas estatísticas (p.e. Análise de Clusters, Tabelas de Contingência, e Testes de Significância) e que foi suportada por análise gráfica (p.e. Scatter-plots, QQ-plots e Histogramas). Foi possível então comparar os resultados obtidos entre duas bases de dados, uma mais cuidada do que outra no que toca ao registo de observações. Diferença essa que foi bem visível na percentagem de explicação do modelo, onde houve um decréscimo de 20 pontos percentuais.

Inspirada nestes resultados, surgiu a ideia de aplicar outra abordagem e usar outra fonte de dados que não os diários de bordo. A sustentabilidade do recurso, para além de outros factores, passa pela utilização eficiente de recursos de modo a garantir a renovação constante do peixe para níveis óptimos. Tal eficiência só pode ser atingida minimizando o desperdício dos recursos gastos durante a actividade pesqueira e maximizando o proveito socioeconómico dessa mesma actividade.

Apesar deste conhecimento geral, nem todos os produtores (neste caso embarcações) são bem sucedidos em atingir níveis satisfatórios de eficiência. Existem várias abordagens para estimar e avaliar a eficiência duma actividade económica, em particular Análise de Fronteiras Estocásticas (SFA), que combina dois campos da ciência, a Estatística e a Economia. Esta metodologia foi desenvolvida por Aigner and Schmidt [1977] e por Meeusen and van den Broeck [1977], e tem sido aplicada em vários campos e sido objecto de várias pesquisas, sendo até considerada por alguns autores como a melhor abordagem na presença da ineficiência. Dentro desta metodologia podem ser consideradas três tipos

de eficiência: Técnica (Technical Efficiency), Custo (Cost Efficiency) e Lucro (Profit Efficiency).

Neste estudo apenas foi estimada a Eficiência Técnica que pode ser descrita como a habilidade de, dado um resultado fixo (output), minimizar a quantidade de variáveis (inputs) necessárias para obter tal resultado, ou a habilidade de maximizar o resultado obtido de um conjunto de variáveis fixas. O conceito é simples e até tem havido um crescente interesse em aplicar esta metodologia à actividade pesqueira, no entanto são poucos os trabalhos realizados sobre este tema, e os poucos que há são estudados duma perspectiva económica e não estatística. Assim este trabalho vem, de alguma forma, tentar preencher esse vazio realizando esta abordagem do ponto de vista estatístico.

A segunda parte deste trabalho (chapter 3) tem então o propósito de avaliar esta abordagem teoricamente e verificar se é na prática uma ferramenta útil e de fácil aplicação. Assim, dentro deste estudo, a eficiência técnica de todas as embarcações que compõem a frota de peixe-espada preto de Sesimbra foram estimadas. Para tal foram recolhidos dados através de inquéritos aos envolvidos nesta actividade, sendo obtido dados relativos aos anos de 2009 e 2010.

Dos resultados foi possível identificar diferenças entre várias abordagens e modelos, avaliar a evolução da eficiência no tempo, procurando tendência e/ou sazonalidade e finalmente verificar que a pesca do peixe-espada preto desenvolvida em Sesimbra pode ser considerada eficiente.

Palavras-chave: Peixe-espada preto, Captura-por-unidade-esforço, Modelos Lineares Generalizados, Análise de Fronteiras Estocásticas, Eficiência Técnica.

Chapter 1

Introduction

On the Portuguese continental slope, in the south of ICES Division IXa, the longline fishery targeting black scabbardfish was initiated in 1983 at fishing grounds around Sesimbra. In Madeira Island there is also a fishery targeting this species which dates back to the 17^{th} century. At present, the fleet targeting black scabbardfish in Portuguese waters is composed by small vessels that still display artisanal features (see Figueiredo and Bordalo-Machado [2007] for detailed description).

Longline fishery is a commercial fishing technique which uses (as the term indicates) a long line, called mainline, with several branch-lines attached as Figure 1.1 shows. Fishing operations usually start at dusk and two manoeuvres generally occur: the newly baited longline gear is deployed into the sea and another longline gear, previously set in the last 24-48 hours is recovered, usually with the aid of a hauling winch. Thus the soaking time of the fishing gear in sea is more than 24h and on average 46h. The preparation of one single gear takes some time, since it can last more than half a day. According to the stakeholders, in Sesimbra to preserve and guarantee the freshness of the fish, only one fishing haul is made by trip.

At beginning, longlines had 3600-4000 hooks, however this number has been largely increased over time, since in 2004 number of hooks ranged from 4000 to 10000. Fishing activity takes place on hard bottoms along the slopes of canyons at depths normally ranging from 800m to 1200m; though 1450m has been reached in the last years. This fishery is also characterized by the fact that the fishing grounds are specific for each vessel, i.e. each fishing vessel around Sesimbra has a specific and unique place to fish. This fishery takes other deepwater species as a by-catch, i.e. during the fishery while attempting to catch the target fish, they unintentionally end up capturing other species, being the Portuguese dogfish and Leaf-scale gulper shark the principal species caught [Figueiredo and Gordo, 2005].

In the process of data collection, to evaluate the species abundance and the fishing impact, there has been in EU, since the introduction of the Common Fisheries Policy (CFP) in 1983, a requirement to record fish catches in a standard community format. This is done by skippers that record the activity at sea and such information is contained on the logbooks that might become an integral tool for monitoring and enforcement. In fact, since there is no independent data from the fishery, as the one commonly collected during directed surveys, the abundance index of black scabbardfish relies on information collected from the fishery itself.

Therefore this knowledge, detailed in logbooks, is vital to define fishery policies and this way to ensure a sustainable activity. Because of this importance is necessary to know, through the logbooks (and other data sources), which variables and factors are important in the performance of the fishery, being fundamental to this end, establish a correct measurement for that performance (CPUE) and estimate the efficiency of the vessels involved and what variables it depends.

This way in chapter 2 the quality of the logbooks data was analyzed in detail and the significant factors for the estimation of the CPUE were identified. Whereas chapter 3 aimed apply Stochastic Frontier Analysis to estimate the Technical Efficiency of the vessels involved in this fishery.

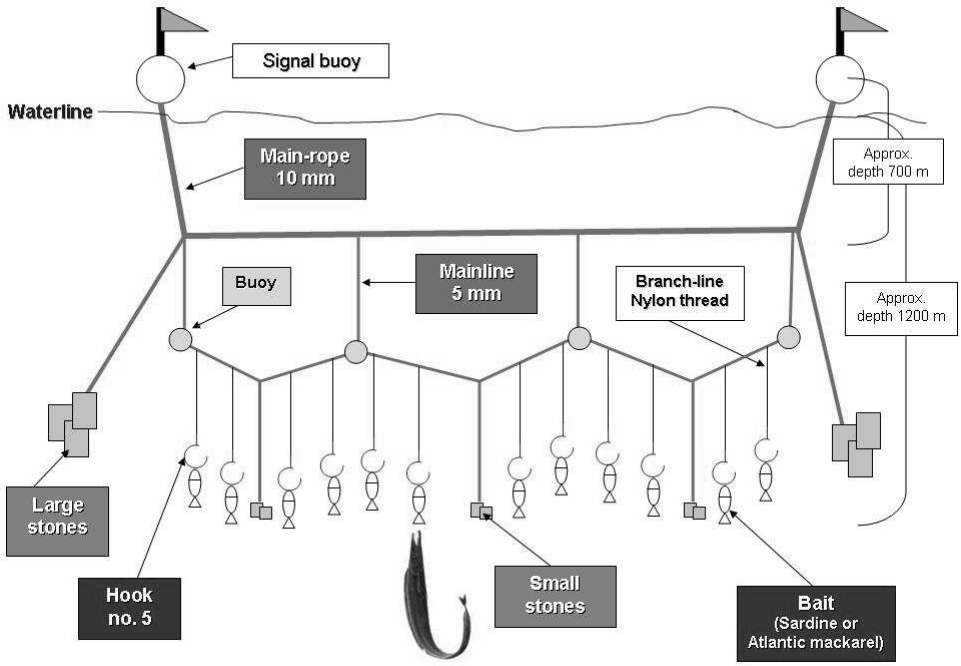

Figure 1.1: Longline scheme.

Chapter 2

CPUE study based on information contained in logbooks

2.1 Introduction

Portugal is a coastal country with about 1200 km of coastline. Therefore the fishery have been throughout history, a present activity in the culture and in the economy of this country. This activity has become of crucial economic importance reinforcing the trade and the related arts. For any fishery the knowledge of the state of the exploited resource is vital for the evaluation of the fishing impact, as well as, for the proposal of management rules that guarantee the sustainability of the resource and consequently of the fishery.

These were the motivations for this study focused on the black scabbardfish fishery, which is the one of the most important fisheries ongoing in Portugal.

As mentioned above the data source used was the logbook. There are however errors or misinterpretations on how to fulfil these logbooks that might hinder its use and purposes related to stock status evaluation. Moreover data in logbooks sampled directly in the field, often reflect the presence of a variety of other species or habitats targeted by the fishermen, even within a single fishing trip. Consequently, some of the records in data may not be relevant to evaluate the stock status of only one target species. Despite this fact, the data contained in logbooks have been used in several working papers to calculate measures of effort like Catch-per-unit-effort (CPUE).

CPUE is defined as the total catch divided by the total effort spent to obtain that catch and is commonly used as an abundance index over time. That effort, in this case fishing effort, may be measured by several variables (e.g. number of vessels, soaking time and number of hooks) and in the recent years considerable energy has been applied by researchers to develop reliable measures of fishing effort. Despite the frequent use, it is

known that CPUE is influenced by many factors other than abundance. Thus to minimize that unwanted influence CPUE is standardized, through this process the effect of confounding factors is reduced or even removed [Maunder and Punt, 2004].

In statistic the fitted models have two main objectives, estimation of the model parameters and the prediction of the study variable values. In CPUE standardization the appropriate modeling strategy is to build an estimation model, rather than a predictive. To do so, it was used the Generalized Linear Model (GLM), which is recognized as a valuable tool for the analysis of fisheries data [Maunder and Punt, 2004].

Linear Models (also known as Regression Model) are used when it is assumed that the study variable (known as the response or dependent variable) has a linear relationship ($Y = \beta X + \varepsilon$) with other variables (denoted as independent or explanatory variables) and the distribution of the response variable is assumed to be Normal. However these assumptions are rarely encountered in the real world and to overcome these restrictions the GLM, which are a flexible form of linear models, were built.

The GLM generalizes Linear Models by allowing two new possibilities: the distribution of the response variable may come from any member of the exponential family other than the Normal (e.g. Gamma, Poisson, Binomial...) and the link function (the link between response variable and the independent variables) may come from any monotonic differentiable function (e.g. inverse function, log function...) as detailed in McCullagh and Nelder [1989]. Despite the limitations still imposed, the GLM have been acquiring an increasingly important role in statistical analysis.

Summarizing, the first part of this work critically analyzes the data contained in the logbooks from the Portuguese fleet operating with longline in Portugal mainland (Sesimbra). The quality and mainly the reliability of the logbooks and the consequences of the absence of carefully collected data, were assessed and analyzed in detail. Finally after being found the best way to set the CPUE, the factors relevant for the estimation of the CPUE of black scabbardfish fishery were identified as well as their influence on the CPUE.

2.2 Materials and Methods

2.2.1 Data and Variables

Two different sets of logbook data were available: one covering the period from 2000 to 2005 and the second one covering the period from 2000 to 2008.

The first data set (covering five years) was, prior to this work, reviewed in detail. This set included trip data on the following variables: vessel identification code (ID); fishing gear; port and date of departure; port and date of arrival; number of fishing hauls

(NHAUL); soaking time (ST); ICES rectangle where fishing haul took place (ERECTAN); ICES subarea; caught species (SP); catch weight by species in kilogram (CATCH) and number of hooks used in each fishing haul (HOOKS). This last variable was obtained by detailed revision, so it was absent in the second data set.

This set had 9330 trips and since each trip had multiple records of different species, they produce a total of 32136 records from 31 vessels. This means that, for the variables SP and CATCH there were altogether 32136 observations (records) and for other variables, since they are unique for the each trip, there were 9330 observations (trips).

The data set was then restricted to trips in which deep-water longline (LLS) was used. This restriction was essential since the studied fishery only uses such fishing gear. The restriction resulted in 7095 trips with 24235 records (around 75% of initial number of records) and 28 vessels. Among these, positive catches of black scabbardfish were only reported for 22 in a total of 5507 records, which in this case coincided with the total number of trips, because a single species was being considered (about 60% of initial number of trips). However information on the number of hooks used was available only for 2514 trips (unfortunately, the fishermen do not usually fill this field in logbooks).

The second set included the data stored at the Portuguese General Directorate for Fisheries and Aquaculture (DGPA) database. This information covered data on a trip basis of all the variables mentioned before, except the HOOKS. In total the data set had 14319 trips with 77483 records (representing 102 vessels) but only 8764 trips with positive catches of black scabbardfish where LLS was employed (around 61% of initial number of trips).

Additionally information on the daily landings of vessels that landed in the Portuguese ports were also available. However in this database each record contained only information about the ID, port and date of arrival, fishing gear, SP and weight and selling price of the fish landed. In this case the number of records regarding positive catches of black scabbardfish was 52734, however due to multiple landings (in different ports) this number was actually 52051 (see Table 2.1 for summarized information).

Table 2.1: Summary of Database about missing variables (x means present)

Database	Period	No of Records	NHAUL	ST	HOOKS	ERECTAN
1^{st} Data set	2000-2005	5507	x	x	x	x
2^{nd} Data set	2000-2008	8764	x	x		x
Daily Landings	1989-2008	52051				

2.2.2 Exploratory Data Analysis

As previously stated, the analysis of both data sets was based on data restricted to the trips where the longline was the fishing gear used (LLS) and the quantity caught of black scabbardfish (BSF) was positive. Posteriorly three extra variables were considered. The first one called TOTAL was added to the two data sets and corresponds to the total weight caught per trip, i.e. the sum of the weight of all species caught in each trip.

As mentioned before the main by-catch species of the Portuguese black scabbardfish fishery are the sharks Portuguese Dogfish - CYO and Leafscale Gulper Shark - GUQ. Therefore the relationships between the CYO and GUQ catch values and the BSF catch values were evaluated. To do so, catch values of CYO and GUQ were considered as well as two new variables: i) PERC which corresponded to the percentage of BSF in the TOTAL; ii) RATIO which gives the percentage of BSF catches in the sum of catches of BSF, GUQ and CYO, i.e. CATCH of BSF /(CATCH of BSF + CATCH of CYO + CATCH of GUQ). These two last variables were taking into consideration due to the fact that the weight of the two deepwater sharks are very different from the weight of BSF.

Additionally there was also information on vessels technical characteristics, namely length-over-all (XCOMP), gross registered tonnage (XTAB) and power of the engine in horse power (XPOW). These features summarizes the main characteristics of the vessels and are invariant throughout time according to stakeholders.

1^{st} Data set

Data contained in the 1^{st} set was analyzed to identify possible discrepancies on each variable values, particularly on soaking time (ST), number of hauls (NHAUL) and number of hooks (HOOKS). The analysis included i) graphical analysis (e.g. boxplots, histograms and scatter plots) and ii) confronting the data with the knowledge on the exploitation regime of the BSF fishery. The graphical analysis was made by plotting the CATCH of BSF versus each of the these three variables. To clarify some of the identified discrepancies, inquiries to stakeholders and to DGPA authorities responsible for database maintenance were made.

The analysis continued by defining criteria to distinguish vessels with a regular activity targeting BSF from those for which the capture of BSF could be considered sporadic. Such restriction was critical to eliminate confounding vessels and consequently confounding observations in the data. This analysis was based on comparing the cumulative sum of CATCH of BSF (per vessel) with the cumulative sum of total catch (of all species) and in the estimation of the proportion of BSF in that sum.

The data set was then restricted to the subset of vessels considered as having a con-

stant activity targeting BSF (15 vessels with 5440 records). To evaluate the relationship between CATCH and the variables ST, NHAUL and HOOKS, Pearson's correlation coefficients were estimated sustained by a graphical analysis. To exclude the potential confounding effect of the factor vessel, similar analysis was applied separately to a subset of three vessels selected using three criteria: i) they had the longest records; ii) they did not have problematic observations in variables HOOKS and ST; and iii) together they represented the majority of total records (51%).

The relationship between the two main by-catch species (GUQ and CYO) and the target species (BSF) was also evaluated using the two variables previously described (PERC and RATIO). This analysis was done by estimating the Pearson's correlation coefficient between PERC (same for RATIO) and CATCH of CYO, CATCH of GUQ and CATCH of CYOGUQ (i.e. CATCH of CYO + CATCH of GUQ).

The relation between the geographical location of fishing grounds (ERECTAN) and the catch of BSF was also investigated. To this end, since ERECTAN is a categorical variable, contingency tables were used to test the independence between the two variables. In this analysis two spatially adjacent rectangles 05E1 and 05E0 were joined, because they are next to each other and 05E1 is obviously an error since it is in the mainland (Fig. 2.1). The total catch of BSF (in kg) was discretized into the following levels: 0 – 500; 500 – 1000; 1000 – 1500; 1500 – 2000; 2000 – 2500; > 2500, which were defined taking into account the minimum and maximum catches and to prevent further problems in the application of independence tests. Particularly, the Pearson chi-square independency test which requires that all expected frequencies have to be at least one and no more than 20% of the expected frequencies can be less than 5 [Zar, 1996].

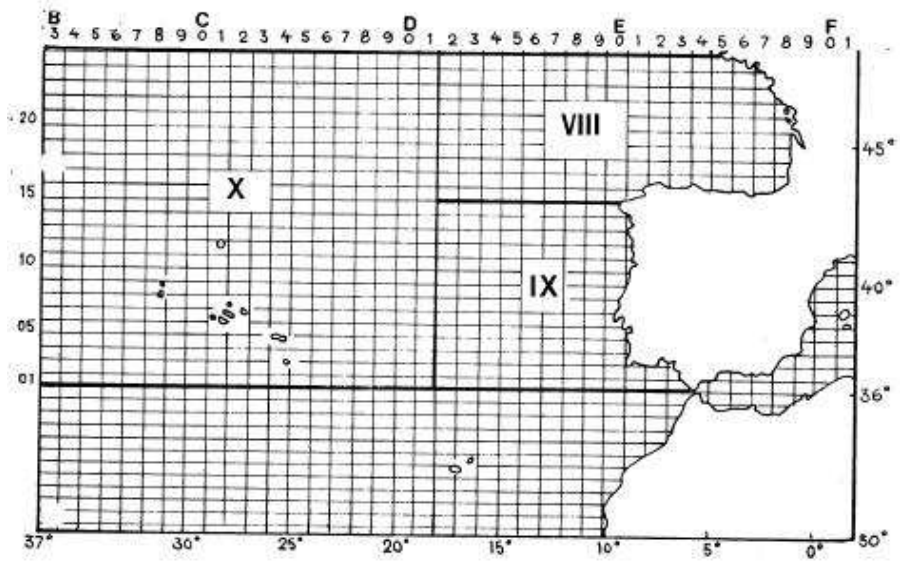

Figure 2.1: ICES statistical rectangles map.

2^{nd} Data set

Through a crude analysis it was verified that the second data set contained a high number of errors, as for example: i) trips with more than 10 fishing hauls (NHAUL), such situation is impossible due to the duration of a fishing operation, when compared with the duration of a fishing trip; ii) more than 30 times of the median value of black scabbardfish caught per trip (CATCH of BSF), which is about 1 ton; iii) different soaking times (ST) assigned for different species caught in the same haul and in the same trip and iv) in some cases ST was swapped with the NHAUL (e.g. in the same trip, 12 hauls with 1 hour of soaking time). These cases are just examples of the complexity and type of errors that were present in a careless database. The procedure for the inspection and correction of data was the same applied for the 1^{st} data set, however the final result of this correction was not so effective and efficient due to the data dimension and due to the long time that was required for such correction.

Since this data set contained a lot of conflicting and less reliable observations, a cross-checking was performed by comparing the BSF catches values recorded in the DGPA database (from hereon denoted as LBSF) with the BSF catches values recorded in the logbooks (2^{nd} set and from hereon denoted as CBSF). Trips with extremely high discrepancies were excluded from the database.

The procedure applied to this data set was similar to the one applied to the 1^{st} set, either in the treatment of the variables related to the by-catch species as well as in the selection of vessels and statistical rectangles (ERECTAN).

2.2.3 CPUE standardization using Generalized Linear Model

Standardization of commercial catch and effort data is important in fisheries where standardized abundance indices based on fishery-dependent data are a fundamental input to stock assessments [Bishop, 2006]. In the standardization of the CPUE through GLM, the variables to include in the model should be selected if there is an a priori reason to suppose that they may influence catchability. However this selection must be careful, because the inclusion of explanatory variables that are correlated should be avoided. To avoid this problem, estimation of correlation measures and corresponding graphical analysis were performed between some of the explanatory variables.

In GLM adjustment different combinations of explanatory variables were used and several output models were tested to understand the relationship between the CATCH of BSF (response variable) and the others variables. Because the 1^{st} set contained more detailed information on several variables, this set was used to evaluate which variables

contribute more to explain the CATCH of BSF and to select the variables to enter in the model adjustment of the 2^{nd} set. The GLM can be expressed through the following expressions:

- The response variable $\mathbf{Y}$ has a distribution that comes from a member of exponential family, with $E(\boldsymbol{Y}) = \boldsymbol{\mu}$ and constant variance σ^2;
- The explanatory variables $\boldsymbol{x}_1, ..., \boldsymbol{x}_p$ produce a linear predictor $\boldsymbol{\eta} = \sum_1^p \boldsymbol{x}_j \beta_j$, with the $\boldsymbol{\beta}$ parameters to be estimated;
- The link function g between the $\boldsymbol{\mu}$ and $\boldsymbol{\eta}$ may come from any monotonic differentiable function $\eta_i = g(\mu_i)$, $i = 1, \ldots, n$ individual.

Several GLMs were adjusted to the final subset of data using a stepwise procedure and this procedure can be summarized in the following steps:

- Step 1 - Selection of the distribution (under exponential family) that best fits to the response variable. Graphical analysis was performed and the distributions were adjusted via the maximum likelihood method;
- Step 2 - Selection of the variables to enter in the model. Maunder and Punt [2004] suggest to always include in the model the factor year. In this case, since the temporal aspect is the major goal of the abundance analysis and given that both the year as the quarter were available, these two variables (YEAR and QUARTER) were always included in the models. The following explanatory variables were also considered: HOOKS, ERECTAN, XCOMP, XTAB, XPOW, PERCCYOGUG (which represented the percentage of Leafscale Gulper Shark and Portuguese Dogfish on the total weight caught, i.e. (CYO + GUQ) / TOTAL). The absolute values of CATCH of CYO and GUQ were not used because, as mentioned before, their weights are very different in scale from the weight of BSF. In the construction of this last variable the missing values of CATCH of CYO and GUQ were replaced by zero;
- Step 3 - Choice of a link function compatible with the distribution of the proposed error for the data. This choice must be based on a set of considerations made a priori [Turkman and Silva, 2000]. For the Gamma distribution the logarithmic link function is recommended, whereas the identity link is recommended for the Lognormal distribution;
- Step 4 - Selection of the best model adopting a parsimonious criterion (model with the smallest number of explanatory variables but a high fit to the data). The deviance function and the generalized Pearson χ^2 statistic were estimated to assess

the models quality of adjustment. Both statistics follow an approximate χ^2 distribution with n - p degrees of freedom, where n is the sample size and p the number of parameters. However asymptotic results may not be specially relevant even for large samples [McCullagh and Nelder, 1989]. The information criterion of Akaike, denoted as AIC and based on the log-likelihood function, was also used. The lower the value of AIC is, the better is the models adjustment. AIC is a flexible likelihood-based approach, which is commonly used in model selection, having the advantage of allowing the comparison of non-nested models. However has the disadvantage of usually choose a complex model (with more variables) instead of a simpler one. To measure the goodness of fit the adjusted coefficient of determination, which corresponds to the ratio of the residual deviance with the null deviance and its respective degrees of freedom (ρ^2), was also used [Turkman and Silva, 2000];

- Step 5 - Model checking by residual graphical analysis. Plots of residuals against different functions of the fitted values, as well as residuals against an explanatory variable in the linear predictor were performed (as suggested by McCullagh and Nelder [1989]). Three residuals were considered and in the following expressions the Turkman and Silva [2000] notation was used:

  Standardized Pearson Residual:

$$R_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{var(Y_i)}(1 - h_{ii})}}, \tag{2.1}$$

  where h_{ii} are the diagonal elements of the 'hat' matrix, which describes the influence of each observed value on each fitted value.

  Anscombe Residual:

$$R_i^A = \frac{A(y_i) - A(\hat{\mu}_i)}{\sqrt{\widehat{var(Y_i)}}A'(\hat{\mu}_i)}, \qquad A(x) = \int \frac{1}{V^{1/3}(x)}dx, \tag{2.2}$$

  where $V(x)$ is the variance function.

  Standardized Deviance Residual:

$$R_i^D = \frac{sign(y_i - \hat{\mu}_i)\sqrt{d_i}}{\sqrt{\hat{\phi}(1 - h_{ii})}}, \tag{2.3}$$

where $\hat{\phi}$ is the dispersion parameter estimate and d_i is the contribution of the $i-th$ observation for the deviation of the GLM.

Both the *Pearson* and *Anscombe* residuals are expected to have a distribution close to Normal, however generally the distribution of the *Pearson* residuals is very asymmetric for non Normal models. In the case of *Deviance* residuals, is recommended by McCullagh and Nelder [1989] to plot against fitted values or transformed fitted values (for each distribution family there is one specific transformation). It is expected that the distribution of these residuals occurs around zero with constant variance.

- Step 6 - Identification of conflicting observations which can be categorized in three different ways: *leverage*, *influence* and *consistency*.

An indicator of the *influence* of the $i-th$ observation can be calculated by the difference $\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}_{(i)}$ denotes the estimates without the extreme point i and $\hat{\boldsymbol{\beta}}$ with it. If this difference is high, the observation i can be considered influential and its exclusion can produce significantly changes in the parameters estimates.

An isolated point of high *leverage* may have a value of h_{ii} such that $\frac{nh_{ii}}{p} > 2$ [McCullagh and Nelder, 1989], where h_{ii} are the diagonal elements of the 'hat' matrix and p is the trace of the 'hat' matrix (i.e. the sum of diagonal elements). The 'hat' matrix describes the influence of each observed value on each fitted value (i.e. the influence of $\boldsymbol{Y}$ in $\boldsymbol{\mu}$), therefore the leverage measures the effect of the observation in the matching fitted value.

For the last kind of conflicting observation, an *inconsistency* observation can be considered as an outlier. Williams [1987] suggests plotting the likelihood residuals (detailed below) against i or h_{ii} to study the consistency of observation i.

$$R_i^L = sign(y_i - \hat{\mu}_i)\sqrt{(1-h_{ii})(R_i^D)^2 + h_{ii}(R_i^P)^2}. \tag{2.4}$$

Note that R_i^D and R_i^P are respectively the Deviance and Pearson residuals detailed before.

2.3 Results

2.3.1 Exploratory data analysis

The knowledge already available for the longline fishery operation, allowed to identify the major inconsistencies both in the 1^{st} and the 2^{nd} data sets. After a crude analysis the most obvious inconsistencies corresponded to null soaking time (ST) and to more than 10 fishing hauls per trip. Other discrepancies consisted on dates of arrival earlier than date of departure, however fortunately some of the discrepancies found were later corrected by logbooks scrutiny and through enquiries to the fishermen. As mentioned previously, the exploratory data analysis began to be made to the 1^{st} set.

1^{st} Data set

In this set the variables HOOKS was the first to be analyzed. The histogram of the number of hooks (HOOKS) used per trip, showed the existence of a group of trips in which the number of hooks was much smaller than the number commonly used. Note that despite this fact, the quantity of fish caught was similar in both groups (as can be seen in the scatter plot of Fig. 2.2). As a result it was considered only the trips in which it was used more than 3000 hooks (taking into consideration the knowledge of the stakeholders and the previous works on this matter).

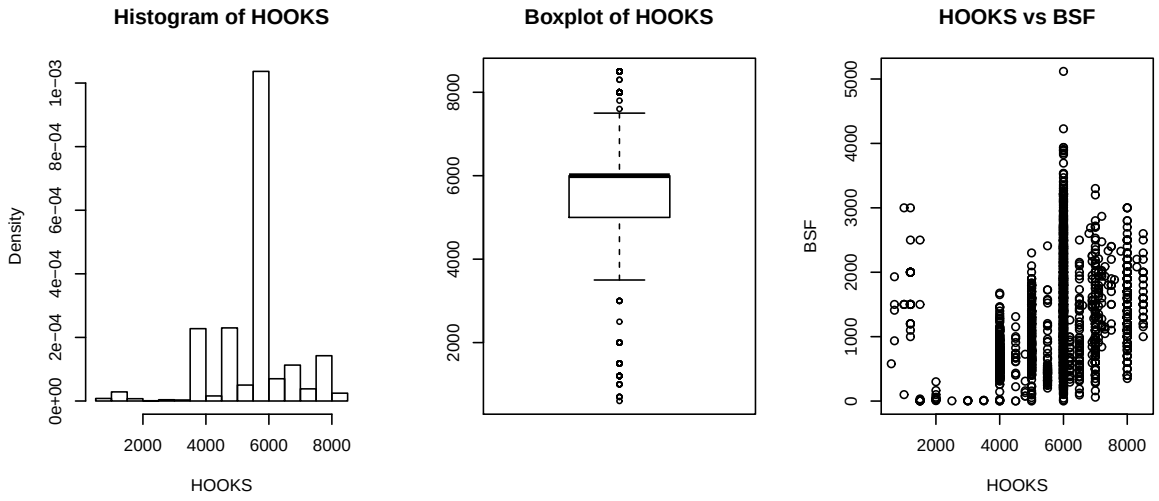

Figure 2.2: Histogram, Boxplot and Scatter plot of CATCH of BSF versus HOOKS.

As mentioned previously, before analyzing the other variables, it is important to distinguish between vessels with a regular activity targeting BSF and those for which the capture of BSF can be considered sporadic. There was no value set a priori, but this

selection was based on two variables: sum of CATCH of BSF of each vessel (Tab. 2.2) and proportion of BSF catch values on the total catch considering the whole time period (i.e. sum of CATCH of BSF / sum of TOTAL, for each vessel and for all trips made). In this table vessels numbered as 2, 3, 9, 11, 14, and 19 (all in bold) had proportions of CATCH of BSF lower than 1.6%, which is very low compared with the remaining vessels. The vessel 5 (in bold), despite having 100% of CATCH of BSF only landed 300 kg of BSF, which was very low when compared with other vessels. Based on these results the subset of 15 vessels was considered for the remaining analysis resulting in a loss of only 0.5% of observations.

Table 2.2: Proportion of CATCH of BSF in the TOTAL catch from 2000 to 2005.

Vessel	Total	Catch of BSF	Proportion
Vessel 1	507416	406747	0,802
Vessel 2	28106	44	0,002
Vessel 3	19551	129	0,007
Vessel 4	235603	204745	0,869
Vessel 5	300	300	1
Vessel 6	418782	387834	0,926
Vessel 7	1438811	1245804	0,866
Vessel 8	197534	151408	0,767
Vessel 9	6730	100	0,015
Vessel 10	232752	156885	0,674
Vessel 11	1396512	233	0,0002
Vessel 12	1050712	925657	0,881
Vessel 13	484794	457478	0,944
Vessel 14	139252	552	0,004
Vessel 15	774293	607795	0,785
Vessel 16	394740	339338	0,860
Vessel 17	436599	345385	0,791
Vessel 18	158184	132950	0,841
Vessel 19	109800	1750	0,016
Vessel 20	259081	165732	0,640
Vessel 21	862399	759973	0,881
Vessel 22	40065	23520	0,587

To evaluate the relation between CATCH of BSF and the variables ST, NHAUL and HOOKS (potential measures of effort), Pearson's correlation coefficients were estimated (Tab. 2.3) sustained by a graphical analysis. All the correlations obtained were relatively low even when different combinations of the three variables were considered (e.g. NHAUL $\times$ ST).

Beginning with the evaluation of the variable NHAUL, for most of the fishing trips only one fishing haul was performed, remaining only four trips in which two hauls were

recorded (Fig. 2.3), and yet when two fishing hauls were performed the catch value of BSF did not increased. This lack of variability did not allow to consider the number of hauls as a variable, so NHAUL was not taken into account in the remaining analysis. Notice that this variable was the only one, among the three variables, for which the independency hypothesis with CATCH of BSF was not rejected, with *p-value* ≈ 0.8.

As for the variable ST, this had a quite large range, however using the knowledge available on fishery, ST $<$ 24h are almost impossible since the fishing gear stays at the fishing ground at least 24h. Thus the values of ST lower than 24h were considered as errors, such errors could probably resulted from a generalized misinterpretation of the variable by fishermen. Instead of including the soaking time they introduced the travel time to the fishing ground. Thus this variable may loose its utility in this study, however through the Pearson's independency test, the independency was rejected with *p-value* $\approx$ 6e-06.

In the analysis of the ST it was verified, through a graphical analysis, the existence of two main groups of records (ST $<$ 24h and ST $\geq$ 24h). However in Table 2.4 the trips with ST $\geq$ 24h were only registered in 107 trips and it was not possible to identify a vessel or a group of vessels that systematically reported ST $\geq$ 24h. This way this variable was not taken into account for the remaining analysis, since ST did not correspond to the soaking time of fishing haul in sea.

The analysis of HOOKS showed that, among the three variables, this one was the most significant (null hypothesis rejected with *p-value* ≈ 0), in sense that it had the highest value on Pearson's coefficient (Tab. 2.3) and the plot showed a slight positive trend (Fig. 2.3). Despite these facts, the variable did not achieve high indices of linear correlation with CATCH of BSF (only 0.31).

Table 2.3: Pearson's coefficient between CATCH of BSF and the variables HOOKS, ST and NHAUL.

Pearson's Correlation	HOOKS	ST	NHAUL
Catch of BSF	0.31	-0.09	0.005

Next it was considered the subset of three vessels (the choice was based on three criteria detailed before) and it was considered only the variables HOOKS and ST (Tab. 2.5). For this subset the Pearson's correlation coefficient, between CATCH of BSF and HOOKS, decreased when compared with the global value. For ST the correlation coefficient increased for Vessel 3, but the improvement was not significant nor regular among the vessels (with positive and negative values). Therefore since neither of the variables showed significant differences in correlation with CATCH of BSF, the 15 vessels were again considered.

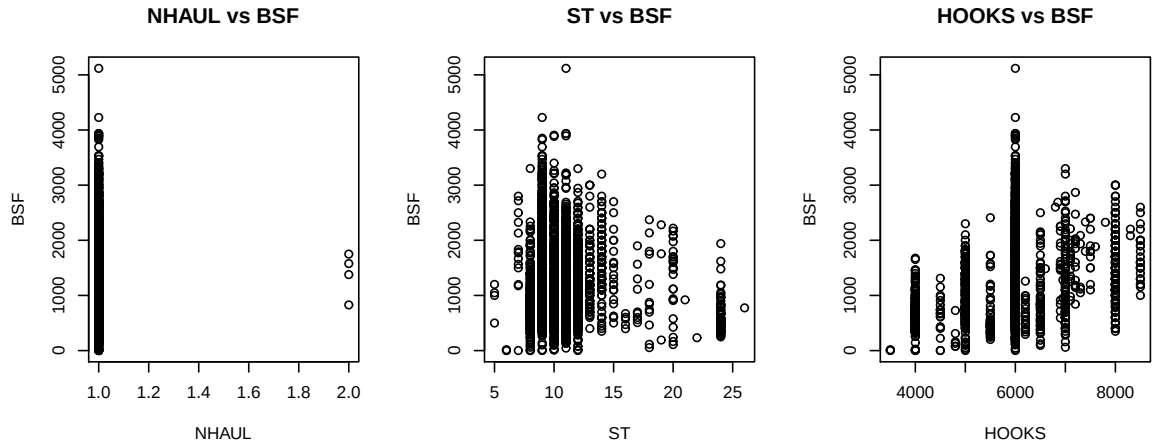

Figure 2.3: Plot of CATCH of BSF against the variables: NHAUL, ST and HOOKS.

Table 2.4: The total number of records and the number of records with at least 24h of ST per Vessel.

Vessel	No of records	No of records with ST $\geq$ 24h
Vessel 1	510	3
Vessel 2	299	0
Vessel 3	682	72
Vessel 4	693	0
Vessel 5	120	0
Vessel 6	246	1
Vessel 7	468	0
Vessel 8	316	0
Vessel 9	695	0
Vessel 10	265	30
Vessel 11	270	0
Vessel 12	90	0
Vessel 13	328	1
Vessel 14	426	0
Vessel 15	31	0

Table 2.5: Pearson's coefficient between CATCH of BSF and HOOKS / ST for a subset of three vessels.

Vessel / Variable	HOOKS	ST
Vessel 1	0.14	-0.16
Vessel 2	-0.08	-0.09
Vessel 3	0.19	0.09

To study the correlation between catches of BSF (represented by RATIO and PERC) and the main by-catch species (CYO and GUQ), the estimates of Pearson's correlation coefficient were estimated and are presented in Table 2.6. All the estimates were significant (p-*value* $\ll 0.01$ for all estimates) and greater than 0.5 (in modulus). Therefore

catch levels of sharks affects the catch levels of BSF, particularly catches of CYO as can be observed in the variable RATIO. This analysis supported the fact that the catches of the two deep-water sharks have significant negative correlation with CATCH of BSF.

Finally regarding the ERECTAN variable, the null hypothesis of independence between ERECTAN and the catches of BSF was rejected ($X^2 \approx 1035$ and $p - value \approx 0$).

Table 2.6: Pearson's coefficient between PERC/RATIO and CATCH of: CYO; GUQ and CYO+GUQ.

Pearson's Coefficient	CATCH of CYO	CATCH of GUQ	CATCH of CYOGUQ
PERC	-0.53	-0.52	-0.68
RATIO	-0.75	-0.65	-0.70

For the adjustment of the GLM model a new factor associated with the vessels characteristics was created. It is important to note that there are different vessels, both in characteristics and in the total catch of BSF, therefore it is necessary to quantify the weight and the significance of these differences in characteristics on the total catch of BSF.

Considering each vessel as a factor is clearly an exaggeration, when it comes to degrees of freedom and because there are vessels that are similar in their main features. Therefore the vessels were grouped by the variables that best describes them: XTAB; XPOW and XCOMP. The levels of these factors correspond to the groups identified after a cluster analysis was applied to the matrix of vessel's characteristics. As those characteristics were found to be highly correlated (Tab. 2.7), one of them should be enough to characterize the vessels, consequently four different cases were considered and groups of vessels were defined based on the results from the following cluster analysis (Fig. 2.4):

- In the first case all the three vessels characteristics were considered at once. Due to the high correlation between them, it was used for clustering the Mahalanobis distance, which is the most appropriate distance function for these cases. This way five clusters were identified with the complete-linkage approach, which resulted in the assembly of a new discrete variable: CLUSTER-ALL with five levels.

- Then it was considered a feature at a time. For the cluster analysis on XTAB, the results were added as CLUSTER-XTAB, and for the variables XCOMP and XPOW the procedure was similar. In all three analyzes the Euclidean distance and the average-linkage approach were used. For all approaches four groups were identified.

Table 2.7: Pearson's correlation coefficient between vessels characteristics.

Variables	XTAB	XCOMP
XCOMP	0.91	
XPOW	0.93	0.85

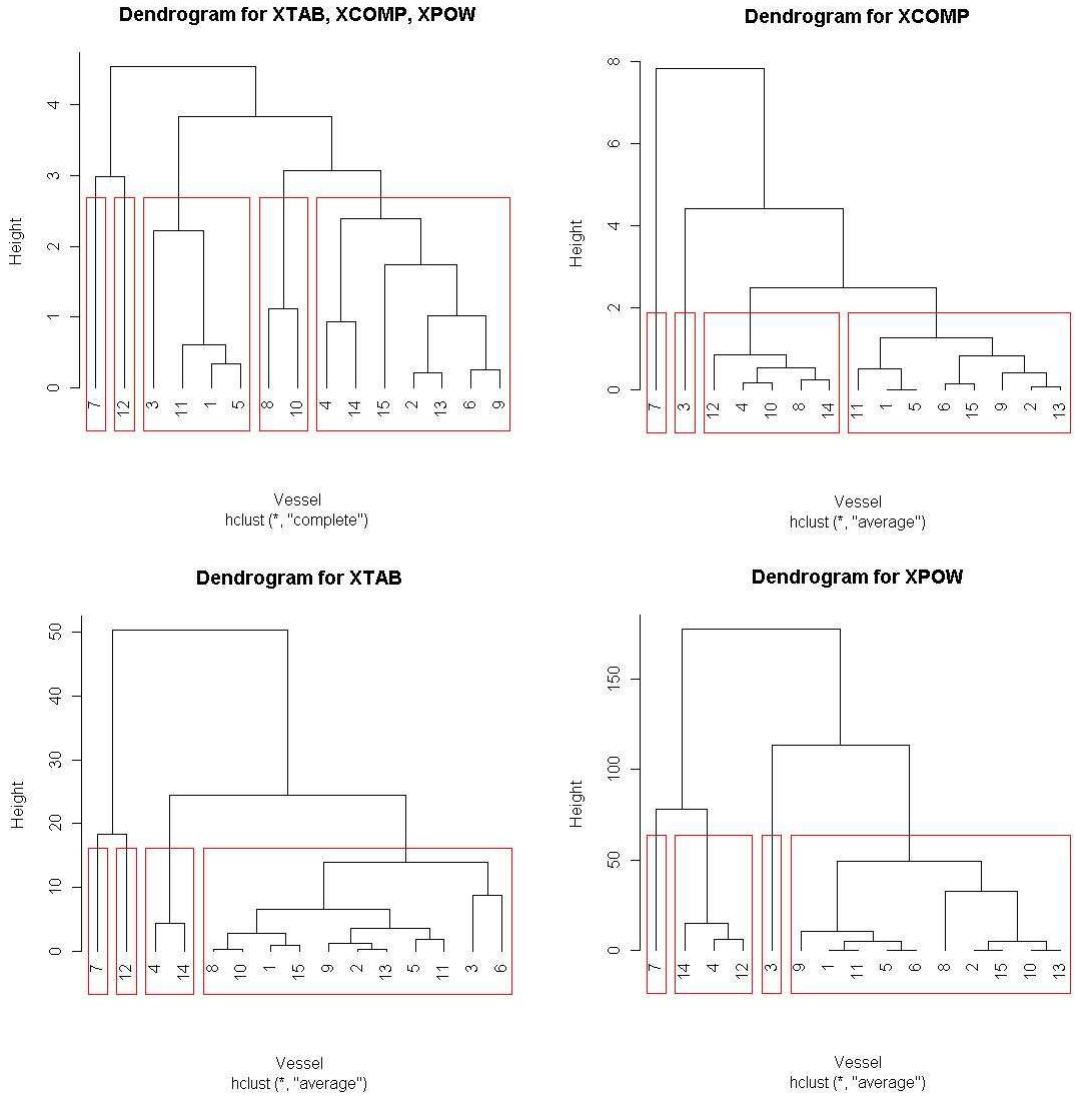

Figure 2.4: Dendrogram for all characteristics (left above), for XCOMP (right above), for XTAB (left below) and for XPOW (right below).

Finally the empirical distribution of CATCH of BSF was analyzed through a graphical analysis. In Figure 2.5 it appears that the variable had a positively skewed distribution (Lognormal and Gamma characteristics) and the same figure suggests the Gamma as the distribution that best fits to the data instead of the Lognormal (by the QQ-plot). Despite this fact, both distributions were later considered in the models adjustment.

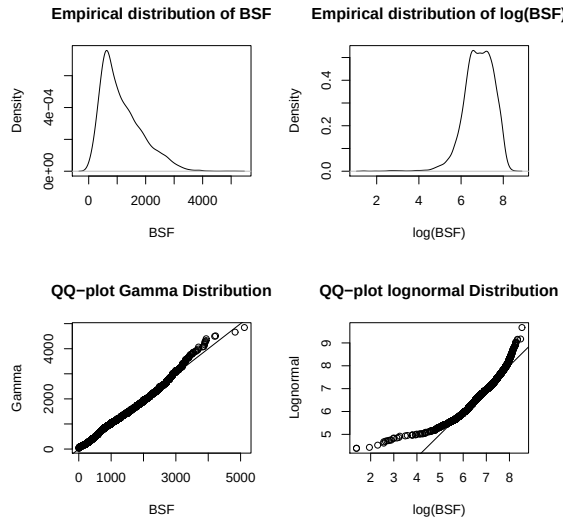

Figure 2.5: Graphical analysis of CATCH of BSF (left) and logarithm of CATCH of BSF (right).

2^{nd} Data set

This set contained much more errors than the first one, therefore the catch data on black scabbardfish reported at the logbooks (from hereon denoted as CBSF) was compared with the corresponding data from daily landings (from hereon denoted as LBSF). The analysis of their empirical distributions (Fig. 2.6) clearly indicated the existence, in the former, of extreme values, while the distribution of the LBSF seems to be much more in agreement with the empirical distribution observed for the 1^{st} data set (Fig. 2.5).

The corresponding difference between BSF catches registered in logbooks and daily landings was further assessed by computing the linear regression between them (Fig. 2.7). Although a close agreement was expected, high discrepancies were observed (Pearson's correlation coefficient around 0.65). To trim the data the 99% quantile of the absolute differences between CBSF and LBSF was determined and all the observations which exceeded that quantile were removed from the 2^{nd} set, excluding this way the higher differences between the two data sets. Figure 2.8 plots the empirical distribution of the new restricted data set which become quite similar and the variability of points around the regression line is much lower. Pearson's coefficient correlation was higher than 0.95, which indicates a strong linear relation between them.

Note however that the unmatched observations, because of discrepancies in dates, were not taken into account in this procedure. Therefore it was not possible to calculate the differences for all trips recorded in 2^{nd} set, this way conflicting observations still remained in this set. After a detailed analysis of these observations, it was decided to remove the CBSF above the 99.5% quantile and below the 0.5% quantile.

Summarizing, were applied two criteria, the first one excluded the higher discrepancies between CBSF and LBSF, and the second one removed the extreme values of CBSF. At the end, comparing the two empirical densities, the improvement is clearly visible and after these restrictions 78 vessels remained in the data set (Fig. 2.9).

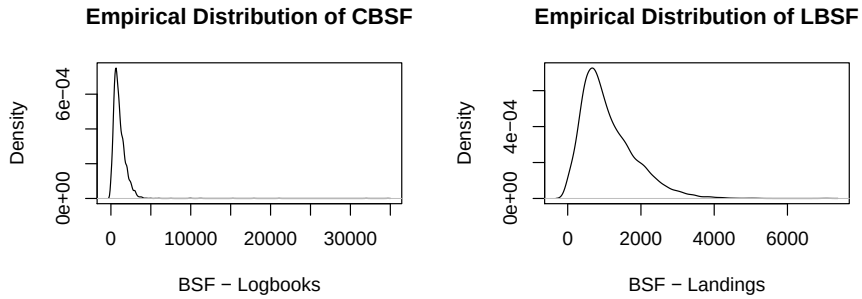

Figure 2.6: Empirical Distribution of BSF from logbooks (left) and from daily landings (right).

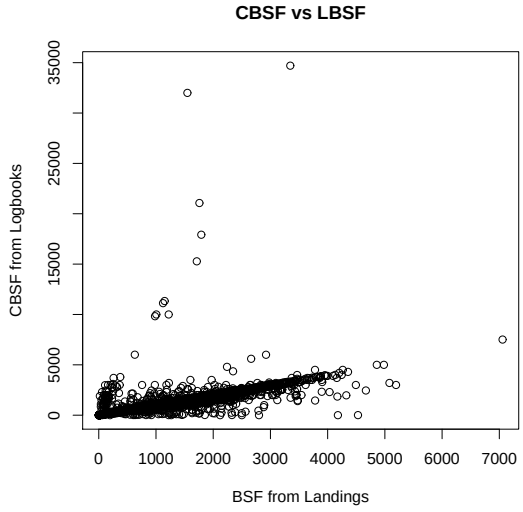

Figure 2.7: Catch of BSF from logbooks plotted against Catch of BSF from landings records.

The criteria adopted for the 1^{st} data set, to differentiate vessels with a regular activity targeting BSF, was also applied to the 2^{nd} data set. However an additional criterium was included, since this set had a higher number of vessels. This way, the total number of trips was also used to avoid vessels with a very short time activity. To identify these vessels, for each vessel the number of trips was plotted against total CBSF (Fig. 2.10).

In the left plot it was easily identified 16 vessels, based on catches values and number of trips (inside the superior ellipse). However on the right side (which is a zooming of the left side of the figure) the selection becomes more difficult, because the number of trips and the catches values are lower. Nevertheless 11 vessels were distinguished in the dashed ellipse, which started the activity recently. Applying these criteria only 27 vessels from 78 vessels remained in the data, although this reduction just reflects a decrease of about 2.5% of number of observations (trips) and a loss of 3% of total of CBSF (sum of catch of BSF for all vessels).

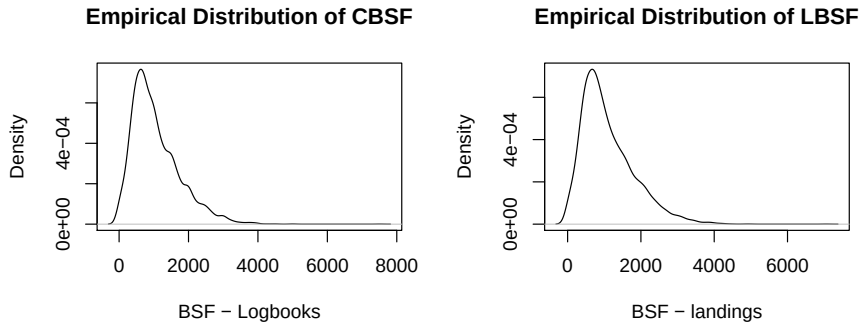

Figure 2.8: Empirical Distribution of BSF from logbooks (left) and from landings records (right), using the observations with differences below 99% quantile.

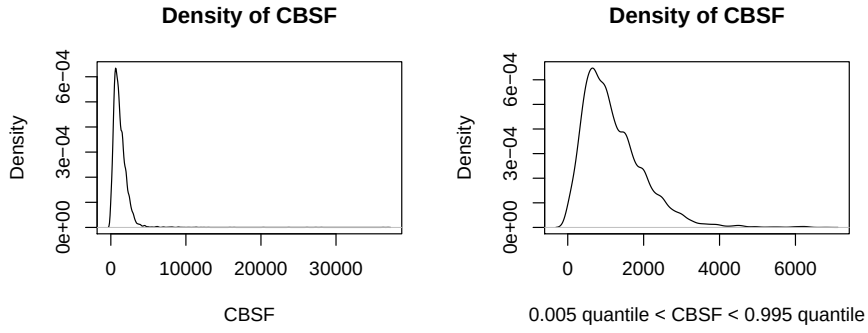

Figure 2.9: Density of all observations and observations between 0.5% and 99.5% quantiles of CBSF.

Although the categorical variable ERECTAN was also available for the 2^{nd} set, the level of detail was much lower than in the 1^{st} set. In fact there were about 16% of records under the category IX, which encompass all the ERECTAN commonly frequented by the vessels and which results in an undoubtedly great loss of information, even further in such an important variable. This way, after this loss only 22 vessels were considered for the application of the GLM.

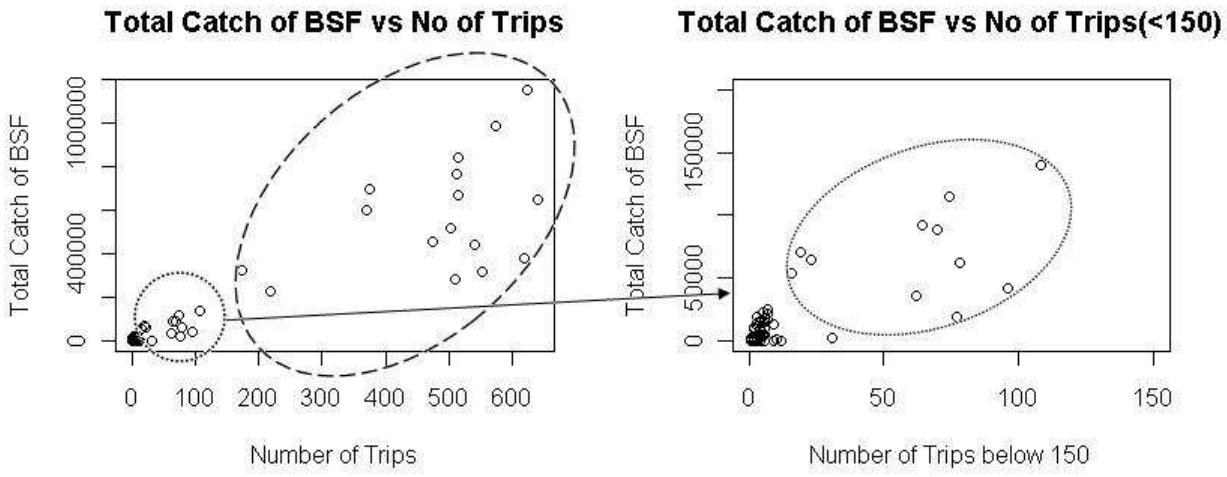

Figure 2.10: Total catch of BSF versus number of trips (left) and the same plot zoom in (right).

2.3.2 Generalized Linear Model

1^{st} Data set

The 1^{st} set is a subset of the 2^{nd} set, that was fully scrutinized in a previous work. So the GLM procedure was used first on this set as a way to identify the most relevant explanatory variables. In the model the response variable was CATCH of BSF (not be confounded with CBSF from 2^{nd} set). In this method was also considered the factors YEAR and QUARTER, the interaction between them, the factor ERECTAN, the variables HOOKS and PERCCYOGUQ and finally the group index levels identified on cluster analysis (CLUSTER). For detailed information about the models applied in this section see Annex 1.

The adjustment of the GLM was done through a stepwise procedure; which select the best model by AIC criterion (minimum), which tends to choose complex models with many variables. Several explanatory variables were essayed and the adequacy of the fit was evaluate based on the estimated generalized Pearson statistic and on the Deviance statistic. The *p-value* in both statistics was always 1, so the selected model was never rejected and Table 2.8 summarizes the results for all models tested for this data set.

Information criterion (AIC) should not be compared across different data sets, thus the models used for this set should all have the same response variable. Therefore it was not possible to compare the models with Gamma distribution with model 6, which uses the Lognormal distribution. However a substantial advantage in using information-theoretic criteria is that they are valid for nonnested models, so it was possible to compare all models with Gamma distribution since they have the same data set.

Using the Gamma distribution the best model was the number 2, because it had the lowest AIC and dispersion parameter, and the highest ρ^2. In this model the variable HOOKS, which is missing from the 2^{nd} data set, was included. But despite this, the model 5 (forcibly without HOOKS) showed that in fact the HOOKS was not so influential, because the values of dispersion parameter and the ρ^2 remained the same and the increase in AIC is very slight.

With Lognormal distribution, model 6 presented the highest ρ^2 (0.62) among all models (including the Gamma models) but this was not greatly different from the one obtained in model 2 (0.61). Since the two models came from different distributions and the ρ^2 were identical, the comparison relied on the dispersion parameter; which showed that model 6 (0.184) deviates much more than model 2 (0.115). This is a strong indicator of the difference in goodness of fit.

Table 2.8: Resume of GLMs applied for 1^{st} data set.

	Model 1	**Model** 2	Model 3	Model 4	Model 5	**Model** 6
Distribution Family	Gamma	Gamma	Gamma	Gamma	Gamma	Lognormal
Link Function	Log	Log	Log	Log	Log	Identity
Null Deviance	423.69	423.69	423.69	423.69	423.69	538.51
Residual Deviance	174.91	161.36	170.00	166.99	163.06	202.40
ρ^2	0.58	0.61	0.59	0.60	0.61	0.62
Dispersion Parameter ϕ	0.131	0.115	0.126	0.123	0.117	0.184
AIC	16645	16554	16612	16594	16564	1291.7
Selected Variables:						
YEAR	X	X	X	X	X	X
QUARTER	X	X	X	X	X	
YEAR × QUARTER						
ERECTAN	X	X	X	X	X	X
HOOKS		X	X	X		X
PERCCYOGUQ	X	X	X	X	X	X
CLUSTER	X	X	X	X	X	X

The residual analysis of models 2 and 6 (Fig. 2.11) presented a better fit for model 2 in relation to the hypothesis of normality of the residuals (mean around zero and constant variance). Two normality test were applied, the Lilliefors (test 1) and the Pearson test (test 2). For model 2, the normality hypothesis was not rejected (*p-value* ≈ 0.1 for test 1 and *p-value* ≈ 0.5 for test 2), whereas for model 6 the both tests rejected it with *p-value* ≈ 0. Thus according to the normality test, model 2 gave a better fit than the model 6.

Standardized deviance residuals were plotted against fitted values for the two models. McCullagh and Nelder [1989] said that if the data are extensive, which happened in this case, no analysis can be considered complete without this plot. The null pattern of this

plot is a distribution of residuals with zero mean and constant range, i.e. no trend, which is verified in Figure 2.12.

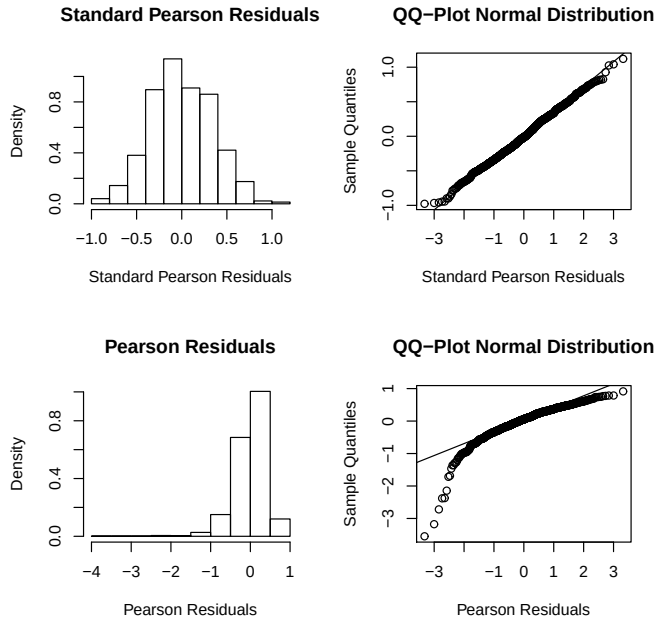

Figure 2.11: Histogram of Pearson's Residuals (left), QQ-plot of Pearson's Residuals (right) from Model 2 (above) and from Model 5 (below).

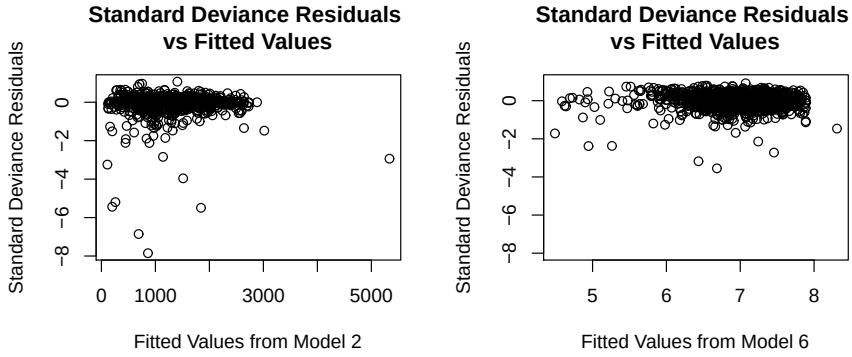

Figure 2.12: Standard Deviance Residuals plotted against Fitted Values from Model 2 (left) and from Model 6 (right).

The residuals were also plotted against the explanatory variable PERCCYOGUQ for both models 2 and 6 (Fig. 2.13). No trend was observed in the linear predictor for both models, which once again was a good indicator [McCullagh and Nelder, 1989], since the residuals are suppose to be uncorrelated with explanatory variables. Note nevertheless greater dispersion on the residuals from model 6.

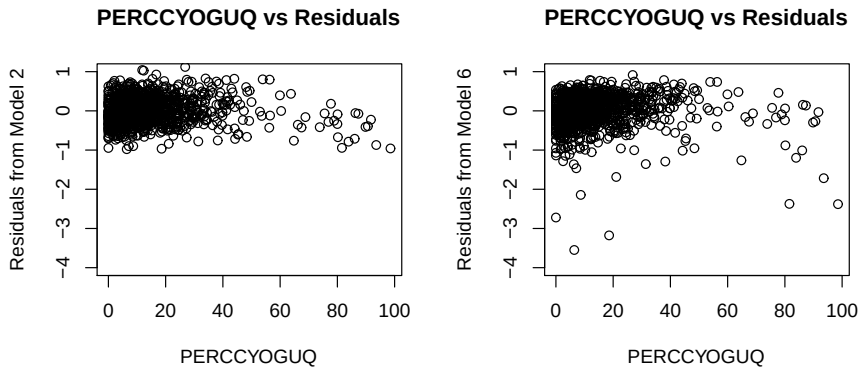

Figure 2.13: Deviance Residuals plotted against PERCCYOGUQ from Model 2 (left) and from Model 6 (right).

Therefore, according to all the goodness of fit indices (AIC, dispersion parameter and ρ^2), considering the residual graphical analysis and the normality test, and following a parsimonious criterion, the chosen model was model 2.

2^{nd} Data set

As the best model from the 1^{st} set was obtained with the vessels grouped by XCOMP, the same cluster analysis was applied to the 2^{nd} set. The repetition of this analysis was necessary due to the fact that the number of vessels in this data set was higher than in the previous one. Unfortunately, it was not possible to have access to this variable in one vessel, however as this vessel was one of the less influential in the data set (represented 1% of total trips), it was removed, resulting on 21 vessels with 6976 observations. For this restricted data set, the cluster analysis resulted in the identification of three groups (Fig. 2.14).

The remaining set of explanatory variables selected by the GLM model adjusted to the 1^{st} set were then used in the adjustment of GLM model to the 2^{nd} set. Both Gamma distribution (*Log link function*) and Lognormal distribution (*Identity link function*) were considered. The model based on Lognormal distribution was considered to verify that the adjustment results were always worse for the 2^{nd} set, independently of the family distribution (Tab. 2.9).

The percentage of explanation (ρ^2) declined about 33%, and the dispersion parameter (for the model 1, with Gamma distribution) doubled, which clearly shows the significance of this worse adjustment.

Considering from hereon only the model 1, for both estimates of the generalized Pearson statistic and Deviance statistic, the *p-value* was equal to 1. However the graphical

analysis of the residuals suggested that the Normal assumption was not fulfilled (Fig. 2.15). Compared to the 1^{st} set, Pearson residuals deviates a lot from a normality hypothesis, while the Anscombe Residuals did not do as bad, however both residuals failed in the normality test, i.e. the normality hypothesis was rejected with $p\text{-}value \approx 0$.

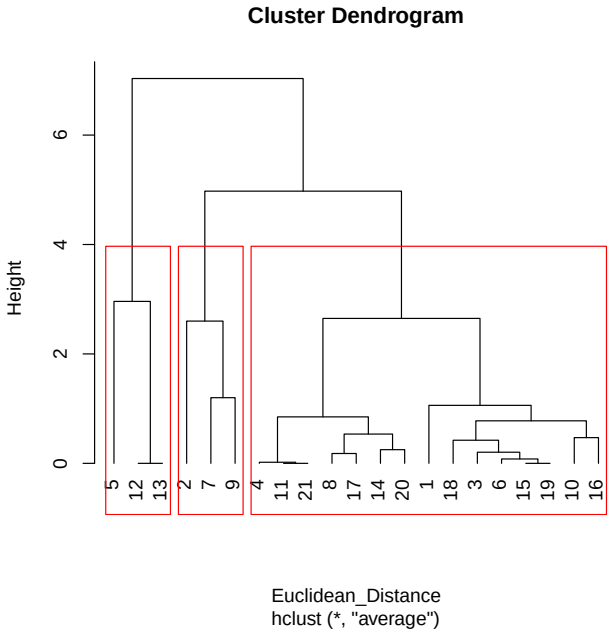

Figure 2.14: Dendrogram of XCOMP from cluster analysis.

Table 2.9: Resume of GLMs of 2^{nd} data set.

	Model 1	Model 2
Distribution Family	Gamma	Lognormal
Link Function	Log	Identity
Null Deviance	2833.6	3314.5
Residual Deviance	1662.5	1896.3
ρ^2	0.412	0.426
Dispersion Parameter ϕ	0.227	0.273
AIC	106361	10754

Standardized deviance residuals against fitted values ($\hat{\mu}$) showed a wide variation of the residuals around zero (Fig. 2.16). Since the Gamma distribution was applied, the transformation $2\log(\hat{\mu})$ suggested by McCullagh and Nelder [1989] was also tried, however the plot did not improve.

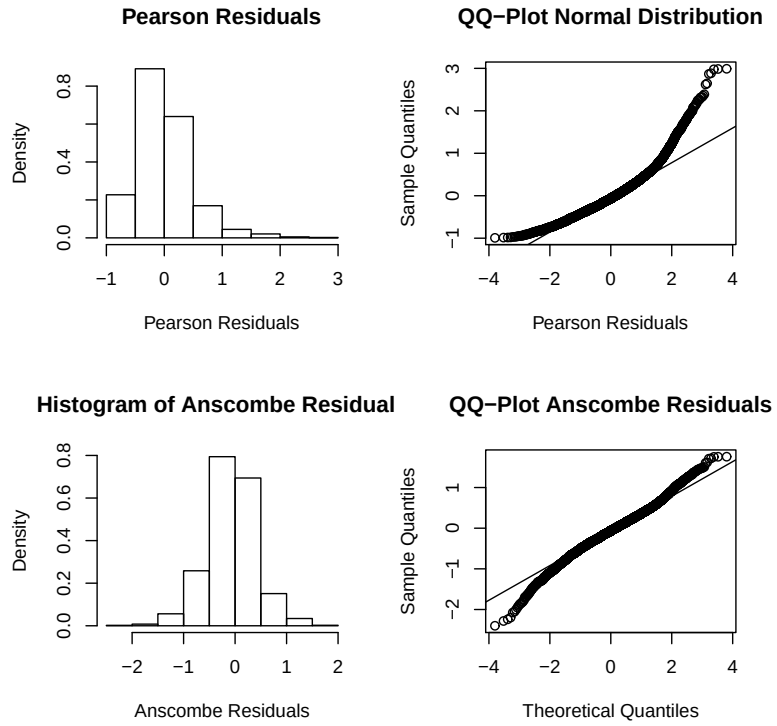

Figure 2.15: Histogram (right above) and QQ-plot (left above) of Pearson Residuals. Histogram (right below) and QQ-plot (left below) of Anscombe Residuals from Model 1

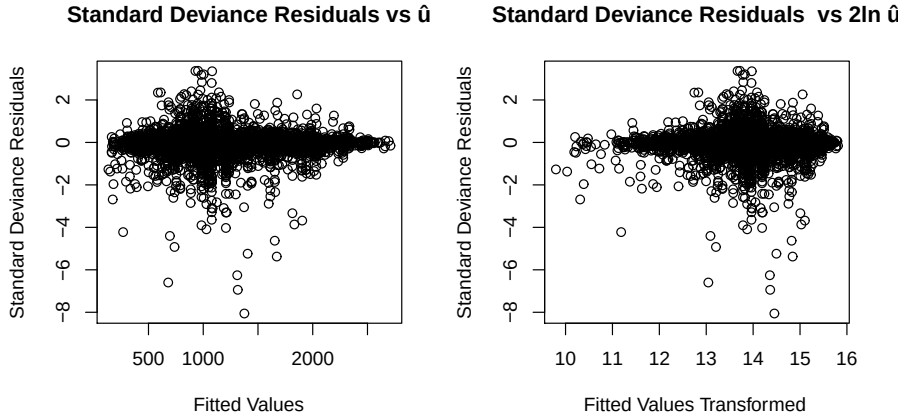

Figure 2.16: Density of the Standard Deviance Residuals versus Fitted Values (left) and versus Transformed Fitted Values (right)from Model 1

The analysis proceeded with the identification of isolated departures (conflicting observations). The measures suggested by McCullagh and Nelder [1989] and by Turkman and Silva [2000] were followed. In Table 2.10 are presented the absolute frequencies for all discordant observations for each vessel. It was chosen to focus only on influential observations, since these are the observations that can change the coefficients. Unfortunately

were detected 600 influential observations, which represents almost 10% of the data. So it was verify and identify the more influential vessels and there was one that stood out from the others, the vessel 4 (in bold). This vessel had almost 100% of the influential observations, so it was tried the same model but this time without vessel 4. However, the improvement was insignificant (Tab. 2.11).

Table 2.10: No of conflicting observations per Vessel.

Vessel	No of Observations	Leverage	Influential	Consistency	% of Influential
Vessel 1	528	65	48	18	9%
Vessel 2	94	0	5	5	5%
Vessel 3	470	2	21	24	4%
Vessel 4	104	99	100	2	**96%**
Vessel 5	146	0	4	8	3%
Vessel 6	502	25	74	41	15%
Vessel 7	402	1	5	4	1%
Vessel 8	415	38	50	18	12%
Vessel 9	208	1	21	30	10%
Vessel 10	425	4	11	7	3%
Vessel 11	31	12	14	4	45%
Vessel 12	615	22	21	2	3%
Vessel 13	61	1	0	0	0%
Vessel 14	376	23	28	9	7%
Vessel 15	62	2	18	16	29%
Vessel 16	501	13	18	8	4%
Vessel 17	513	12	18	9	4%
Vessel 18	443	2	72	103	16%
Vessel 19	412	3	30	19	7%
Vessel 20	321	30	33	12	10%
Vessel 21	347	2	9	10	3%
Total	6976	357	600	349	9%

Table 2.11: Resume of GLM without Vessel 4.

	Model 3
Distribution Family	Gamma
Link Function	Log
Null Deviance	2817.8
Residual Deviance	1646.9
ρ^2	0.414
Dispersion Parameter ϕ	0.229
AIC	104767

2.4 Discussion

This part of the work had the ultimate purpose to identify the most important and useful variables and information to assess stock abundance. Prior to that, it was necessary to identify the errors and misunderstandings regarding the filling of logbooks, which are currently the most important source of data. Such errors can lead to erroneous and biased conclusions and the consequences are quite visible when the results from both data sets are compared. For the first set (which was subject to revision and reintroduction of data) the chosen model explains almost 62% whereas the same model explains less than 42% in the second set. So the first conclusion is that a good filling of the logbooks is an essential starting point to a good statistical analysis and in fact the data contained in logbooks is far from the desired quality.

Regarding to the purposes related to the CPUE, the logbooks are record on trip basis and some of the variables (such as ERECTAN and PERCYOGUQ) are trip dependent, so this way the CPUE was defined by catch per trip. To identify the variables relevant for the estimation of the CPUE, is important to find which variables should be considered in the GLM procedure. After a detailed analysis, based on graphical and cluster analysis, correlation coefficients, contingency tables and the knowledge of the stakeholders, it was concluded that the variables such as the temporal (YEAR and QUARTER) and spatial indicators (ERECTAN) are essential in understanding and assessing the abundance. Moreover the vessels characteristics (economical variables), although many of them are strongly correlated and biological variables, such as the presence of natural predators (PERCCYOGUQ), are also important to assess the stock status.

All of these variables and additionally the HOOKS, which is correlated with catch values, were considered in GLM. These same variables, based on GLM results, should be considered in CPUE standardization. In fact, if the logbooks were correctly filled, variables such as HOOKS and ST could be more significant for the evaluation and estimation of the stock. Nevertheless we conclude that the variables where the filling should be more careful, is the capture of both the target species and accessory species.

As seen throughout this work, by the nature of the variables, fishing activity is a process very complex that encompasses many branches of science (Biology, Economy, Geology). Hence, other variables should be explored such as: skipper skills and education; number of workers in sea and in land; occurrence of technical problems and presence or absence of marine mammals.

However, although these variables are qualitative, there is a strong possibility that they will be erroneously filled, so the defenders of the logbooks as a valid and a reliable source of data, must be aware of this errors and possible misleading analysis. To account

for that source of errors, it would be important to invest in guiding stakeholders in order to explain the importance of a proper filling of logbooks.

No less important is the filtering that the fishery regulators should do when entering the data in databases, in order to detect discordant observations and then to assess the true facts. Taking into consideration that it might interfere with the fishing process, it is especially important to instruct fishermen and skippers about the importance of a correct filling of the logbooks and only combining the work of scientists with the fishing community will the sustainability of sea and of artisanal fishing be achieved.

Chapter 3

Fishery technical efficiency through stochastic frontier analysis

3.1 Introduction

The management and regulation of fisheries continues to be one of the challenges of the marine world. These issues are particularly important for Portugal, one of the countries with the highest fish consumption in the world. The sustainable management of fish stocks and the efficient utilization of resources must guarantee the renewal of the fish resource to optimum levels, minimize waste and maximize the social and the economic benefits of the fishing activity [Flores-Lagunes and Schnier, 1999].

The maximization of social and economic benefits from fisheries requires the production to be optimized, which involves maximizing the profit and minimizing the expense associated with the exploitation. Despite this, it is known that not all producers are equally successful in solving the optimization problems by utilizing the minimum inputs required to produce the maximum output, i.e. not all producers are succeed in achieving a high level of efficiency [Kumbhakar and Lovell, 2000].

Several approaches are available for the evaluation of the efficiency of an economical activity, in particularly, Stochastic Frontier Analysis (SFA). This approach is commonly used, since in the presence of inefficient producers, SFA emerges as the best theoretical approach. This procedure was developed in the 70´s by Aigner and Schmidt [1977] and by Meeusen and van den Broeck [1977] and since then has been subject to considerable econometric research in several fields such as health, agriculture and industry.

SFA allows to estimate the efficiency of each producer, as well as, the average efficiency of all producers involved in the production process and can be applied to estimate and analyze Technical Efficiency (TE), as well as Cost and Profit Efficiency.

In fisheries there has been a growing interest on estimating the efficiency of fishing vessels using SFA [Flores-Lagunes and Schnier, 2007], however relatively few applications of SFA to fisheries are available. Such deficiency may result from the complex relationship between resources and their exploitation. Underlying dynamics of the resource is taking place at the same time as the fishing process and changes on fish resource status might play an important role on efficiency of the fishing vessels.

To study a process so complex and dynamic, as is the artisanal fishing, the Stochastic Frontier Analysis embraces two science fields, Economy and Statistics. This methodology has been applied in fishing activity, however studies are often conducted from the economic view point, rather than a statistical one. Therefore, this work aims somehow to fill that void and analyze the results from this perspective. To this end, several approaches were tested and several comparisons were made, both from the theoretical and practical perspectives.

The second part of this study aimed to theoretically evaluate the SFA approach as well as the statistical properties of their estimators. Under this study, the Technical Efficiency (TE) of each vessel, that compose the Sesimbra black scabbardfish fleet was estimated and the efficiencies were compared between vessels. This chapter had also the purpose of evaluating the evolution of technical efficiency in time, compare the results from logbooks with the results from daily landings, identify the differences between several models and finally verify if the black scabbardfish fishery in Sesimbra can be considered efficient.

3.1.1 Technical Efficiency

In the present study due to the type of data available, only Technical Efficiency was analyzed. Theoretical aspects of TE will be presented using the work of Kumbhakar and Lovell [2000] as main reference. According to them, Technical Efficiency (TE) refers to the ability to minimize the production inputs for a given output vector, or the ability to obtain the maximum output from a given input vector. The chapter 2 of the same book, presents a detailed review of TE properties.

If applied to fisheries, TE can be interpreted as a way to measure the relationship between the inputs related to fishing operation and the outputs (usually the weight of fish caught). Several input variables have been considered in studies on fishing efficiency. For example Pascoe and de Wilde [2001] found that characteristics of vessels can directly affect the efficiency of individual fishing vessels. In fact, characteristics such as age and size of vessel, have a significant impact on the level of technical efficiency according to Tingley and Coglan [2005]. Squires and Kirkley [1999] suggested that much of the difference

between vessels may be due to differences in skipper skill, which is one of the most difficult variable to quantify and measure.

The TE is defined as the ratio of the observed output ($\boldsymbol{y}$ - response variable), to the maximum feasible output ($f(\boldsymbol{x};\boldsymbol{\beta})$) the production frontier, which is a function of the inputs (explanatory variables):

$$\boldsymbol{TE} = \frac{\boldsymbol{y}}{f(\boldsymbol{x};\boldsymbol{\beta})}. \tag{3.1}$$

This way, since $f(\boldsymbol{x};\boldsymbol{\beta})$ is the maximum feasible, TE ≤ 1. Two different approaches are commonly used to estimate the parameters of the $f(\boldsymbol{x};\boldsymbol{\beta})$:

- Deterministic Envelopment Analysis (DEA):

$$\boldsymbol{y} = f(\boldsymbol{x};\beta) \cdot \boldsymbol{TE}. \tag{3.2}$$

- Stochastic Frontier Analysis (SFA):

$$\boldsymbol{y} = f(\boldsymbol{x};\boldsymbol{\beta}) \cdot \exp\{\boldsymbol{v}\} \cdot \boldsymbol{TE}. \tag{3.3}$$

The first method (DEA) ignores the effect of random errors in the model, so under this method the variation in the output is entirely attributed to the lack of efficiency, i.e. inefficiency.

In the SFA two sources of variation on output are considered; one associated with random noise and the other related to technical efficiency. Under SFA the stochastic production frontier consists of two parts: a deterministic part $f(\boldsymbol{x};\boldsymbol{\beta})$ common to all producers and a producer-specific part $\exp\{\boldsymbol{v}\}$, which captures the effect of random variation produced by the environment on each producer.

Comparing the two methods the latter method (SFA) is preferred, because when applying DEA there is a high risk of improperly attributing unmodeled random variation to technical efficiency variation.

Two main groups of data can be used in SFA: Cross-sectional data and Panel data. In Cross-sectional data there is only one observation of each producer and provide a snapshot of producers and their efficiencies. Panel data provide more reliable evidence, allowing the monitoring of each producer performance over time, since more than one observation for producer is available. With panel data, SFA can be deal either by assuming that the TE is time-invariant or by considering the TE time-variant.

Cross-Sectional Data

For this type of data the model is expressed as:

$$y_i = f(x_i;\boldsymbol{\beta}) \cdot \exp\{v_i\} \cdot TE_i, \tag{3.4}$$

where y_i is the output (can be a vector) of producer i and x_i is a vector of N inputs $(n = 1,\ldots,N)$, for $i = 1,\ldots,I$. The production frontier $f(x_i;\boldsymbol{\beta})$ is a function of the inputs, $\boldsymbol{\beta}$ is a vector of parameters to be estimated while TE_i is the technical efficiency of producer i, which is usually expressed as $exp\{-u_i\}$. Both $exp\{-u_i\}$ and $exp\{v_i\}$, which represents the random error (statistical noise), are producer-specific.

To estimate TE_i, the production frontier takes the log-linear Cobb-Douglas form (log transformation). The Cobb-Douglas form is widely used in economic studies and through it the stochastic production frontier model can be written as:

$$\ln y_i = \beta_0 + \sum_n \beta_n \cdot \ln x_{ni} + v_i - u_i. \tag{3.5}$$

Considering $\varepsilon_i = v_i - u_i$, the same model can be expressed as:

$$\ln y_i = \beta_0 + \sum_n \beta_n \cdot \ln x_{ni} + \varepsilon_i. \tag{3.6}$$

Inefficiency and random errors are multiplicative, that for simplicity appear as exponential functions since with logarithm transformation the errors become additive. The inefficiency error component u_i has to be nonnegative since $TE_i \leq 1$, thus all producers operate under or at their stochastic production frontier, according as $u > 0$ or $u = 0$.

The estimation of TE_i and u_i is performed in two steps, the first one involves the estimation of all parameters of the model and in the second step the technical efficiency is estimated for each producer. For the first step there are two methods, the maximum likelihood method (MLE) and a modified ordinary least squares (MOLS). In MOLS procedure the first step is divided into two parts, in the first one OLS is applied to generate consistent estimates of all parameters of the model, apart from the intercept. In second part of the estimation, consistent estimates of the intercept and the parameters describing the structure of the two error components are obtained. For both methods it is necessary to impose the following assumptions:

- Noise error component v_i is assumed to be iid $N(0,\sigma^2)$, which is an assumption commonly imposed in other approaches.
- Inefficiency error u_i is assumed to be iid and can be:

  - Truncated Normal Model, particulary the Half-Normal Model ($\mu = 0$);
  - Gamma Model, particulary the Exponential Model ($\theta = 1$).

  These distributions are selected because they are flexible and appropriate for non-negative and positively skewed variables, as in case of u_i error.

- The errors v_i and u_i are independently distributed of each other and of the regressors (inputs).

Despite the different distributions that can be consider, there are evidences that the producers TE rankings are insensitive to the distribution assumed. This derives from the fact that the error distribution affects all producers and so the change in the distribution is the same for all producers. Thus it is recommended to use a relatively simple distribution rather than a flexible and a complex one. With this kind of data Schmidt and Sickles [1984] noted two main drawbacks:

- I) Maximum likelihood estimation of the stochastic production frontier model and the subsequent separation of technical inefficiency from statistical noise, both require strong distributional assumptions.
- II) Maximum likelihood estimation also requires an assumption that the technical inefficiency error component be independent of the regressors, although it is not unlikely that inefficiency be related with the regressors.

These limitations are avoidable if the type of data is panel instead of cross section. A panel (repeated observations on each producer) contains a lot more information than does a single cross section. Therefore, it is to expected that access to panel data will enable some of the strong assumptions to be relaxed or result in estimates of technical efficiency with more desirable properties.

Panel Data

The structure of the model with panel data is similar to the cross-sectional model, but in addition a index time t is associated with the output, inputs and random error ($\boldsymbol{v}$). So using the same notation as for the cross-sectional data, the SFA model can be written as:

$$\ln y_{it} = \beta_0 + \sum_n \beta_n \cdot \ln x_{nit} + v_{it} - u_i, \tag{3.7}$$

if technical efficiency (associated with the $\boldsymbol{u}$ error term) is time-invariant, or

$$\ln y_{it} = \beta_0 + \sum_n \beta_n \cdot \ln x_{nit} + v_{it} - u_{it}, \tag{3.8}$$

if technical efficiency is time-variant for I producers indexed by i and by $t = 1, \ldots, T$ time periods, with T fixed for all producers.

The assumption that TE is time-invariant (i.e. constant over time) is strong and the longer the panel is, the more desirable it is to relax this assumption. However for a production process where the technical changes are rare or unlikely in the time period considered, the time-invariant approach is more suitable, since under this approach the number of parameters to be estimated are less.

In the Time-Invariant Technical Efficiency model, the parameters can be estimated by three different methods. Two of them do not impose any distributional assumption for the inefficiency error term and are designated as *fixed* and *random* effects model. The third method uses MLE and can be considered as a generalization of the method used in cross-sectional data.

- Fixed-Effects Model: In this model is allowed that the u_i be correlated with the inputs and with v_i. Thus the requirements are $u_i \geq 0$ and as usually v_i are iid $(0, \sigma^2)$ and uncorrelated with regressors. The inefficiency errors are treated as fixed effects and thus are producer-specific, consequently can be considered $\beta_{0i} = (\beta_0 - u_i)$ as producer-specific intercepts, and the model can be expressed as:

$$\ln y_{it} = \beta_{0i} + \sum_n \beta_n \ln x_{nit} + v_{it}. \tag{3.9}$$

  In this approach, OLS is used to estimate the parameters for any of three ways: suppressing β_0 and estimating I producer-specific intercepts; retaining β_0 and estimating (I - 1) producer-specific intercepts; or applying the within transformation, in which all data are expressed in terms of deviations from producer means and the I intercepts are recovered as means of producer residuals. Then β_0 is determined as:

$$\hat{\beta}_0 = \max_i \{\hat{\beta}_{0i}\}, \tag{3.10}$$

  and the u_i as:

$$\hat{u}_i = \hat{\beta}_0 - \hat{\beta}_{0i}. \tag{3.11}$$

This estimator guarantees that all $\hat{u}_i$ are nonnegative. The estimates of technical efficiency are obtained as:

$$\widehat{TE_i} = \exp\{-\hat{u}_i\}. \tag{3.12}$$

- Random-Effects Model: The u_i (still nonnegative) are treated as random. Under this method u_i are now assumed to be uncorrelated with the regressors and with v_i, the assumptions made on v_i remain. No distributional assumption is made on u_i. The model may be expressed as:

$$\ln y_{it} = \beta_0^* + \sum_n \beta_n \ln x_{nit} + v_{it} - u_i^*, \tag{3.13}$$

where $u_i^* = [u_i - E(u_i)]$ and $\beta_0^* = [\beta_0 - E(u_i)]$. This random-effects model fits exactly into the one-way error components model in the panel data literature, be estimated by the standard two-step generalized least squares. Once β_0^* and $\beta_n s$ have been estimated, the u_i^* can be estimated from the residuals by means of:

$$\hat{u}_i^* = \frac{\sum_t \left(\ln y_{it} - \hat{\beta}_0^* - \sum_n \hat{\beta}_n \ln x_{nit}\right)}{T}. \tag{3.14}$$

Estimates of the u_i are obtained by:

$$\hat{u}_i = \max_i\{\hat{u}_i^*\} - \hat{u}_i^*. \tag{3.15}$$

- Maximum Likelihood: In this method the same assumptions as those assumed for the cross-sectional are also imposed. The methodology to estimate the parameters is identical to the one expressed for the cross-sectional data, which is obtained from the present one when T=1.

Despite these different approaches to estimate parameters, comparisons on the basis of Monte Carlo method showed that the three techniques generate similar results and are likely to generate similar efficiency rankings (Kumbhakar and Lovell [2000]).

In Time-Variant Technical Efficiency model, as with the time-invariant model, two estimation approaches are available. An approach in which time-variant technical efficiency is modeled using fixed or random effects and a MLE approach. As in other models, the first objective is to obtain estimates of the parameters describing the structure of production technology, and the second objective is to obtain producer-specific estimates of TE. With an $I \times T$ panel it is not possible to obtain estimates of all intercepts β_{it}, the N

slope parameters and σ_v^2. This way what is usually done, is write u_{it} in a special form. In this work it was followed the Coelli [1996] model specification, known as Efficiency Effects Frontier (EEF) which can be expressed as:

$$y_{it} = \boldsymbol{x_{it}}\boldsymbol{\beta} + v_{it} - u_{it}, \tag{3.16}$$

where y_{it} represents the logarithm of the output, $\boldsymbol{x_{it}}$ the logarithm of inputs, $\boldsymbol{\beta}$ and v_{it} are defined as earlier. The inefficiency error term is assumed to be time-variant and independently distributed as truncations at zero of the $N(m_{it}, \sigma_u^2)$, where $m_{it} = \boldsymbol{\delta z_{it}}$.

$$u_{it} = \boldsymbol{z_{it}}\boldsymbol{\delta} + w_{it}. \tag{3.17}$$

The $\boldsymbol{\delta}$ parameter is a vector of parameters to be estimated, $\boldsymbol{z_{it}}$ is a vector of variables which may influence the efficiency and w_{it} is defined by the truncation of the Normal distribution with zero mean. It is easy to note that this model encompasses the others models, by setting the $\boldsymbol{\delta}$ parameters equal to zero. Then in this model the major purpose is to test if the $\boldsymbol{\delta}$ parameters are zero, to know if the $\mathbf{z}$ variables affect or not the producer efficiency.

Despite the $\mathbf{z}$ variables may influence directly the producer inefficiency, has to be noted that these variables are, by the construction of the model, hierarchically below the explanatory variables ($\mathbf{x}$ variables). This approach was considered to verify if the efficiency shows seasonality, i.e., to verify if the quarter is, as in the last chapter, a significant factor.

Although it is assumed that there are T time periods for which N observations are available, it is not necessary that all the producers were observed for all time periods.

3.1.2 Estimation of Technical Efficiency

Cross-Sectional Data

For cross-sectional data it was said that the estimation procedure was divided in two steps. For the first step were defined two methods (MLE and MOLS), now we describe the second step. Note that the $\boldsymbol{u}$ and $\boldsymbol{v}$ are independently distributed and their distributions are already known.

- Step 1: Density function of $\boldsymbol{v}$ and $\boldsymbol{u}$ are considered and based on them:
  - The joint density function of $\boldsymbol{u}$ and $\boldsymbol{v}$ is obtained as the product of the two density functions, since they are assumed to be independently distributed;

- The joint density function of $\boldsymbol{u}$ and $\boldsymbol{\varepsilon}$ ($\boldsymbol{\varepsilon} = \boldsymbol{v} - \boldsymbol{u}$) is then obtained by replacing $\boldsymbol{v}$ by $\boldsymbol{\varepsilon}$;

- The marginal density function of $\boldsymbol{\varepsilon}$ is obtained by integrating the previous function in order to $\boldsymbol{u}$.

- Step 2: Estimation of the expected value of technical efficiency:

  - $E(\exp\{-\boldsymbol{u}\})$ (Lee and Schmidt [1978]), which is in agreement with the definition of TE.

  - $1 - E(\boldsymbol{u})$ (Aigner and Schmidt [1977]), which is an approximation of the previous estimate, since it includes only the first term of the Taylor series.

- Step 3: Based on the joint density function of $\boldsymbol{u}$ and $\boldsymbol{\varepsilon}$ and the marginal density function of $\boldsymbol{\varepsilon}$, can be calculated:

  - $f(u|\varepsilon_i) = \frac{f(u,\varepsilon)}{f(\varepsilon)}$.

- Step 4: There are two point estimators for u_i:

  - $\hat{u_i} = E(u|\varepsilon_i)$ (Conditional Mean);

  - $\hat{u_i} = M(u|\varepsilon_i)$ (Conditional Mode).

- Step 5: Finally the technical efficiency of each producer i is estimated as:

  - $TE_i = \exp\{\text{-}\hat{u_i}\}$, where $\hat{u}_i$ can be $E(u|\varepsilon_i)$ or $M(u|\varepsilon_i)$;

  - $TE_i = E(\exp\{-u\}|\varepsilon_i)$, as proposed by Battese and Coelli [1988].

The expected value estimator (step 2) proposed by Lee and Schmidt [1978] is preferable since the estimator suggested by Aigner and Schmidt [1977] is an approximation. The two point estimators of TE_i (step 5) provide different results, being the second preferable, based on the same grounds that support the choice on step 2. Unfortunately for this type of data, TE estimators produces unbiased but inconsistent estimates of technical efficiency.

As example we describe the estimation procedure for two cases, assuming Half-Normal distribution and assuming Truncated Normal distribution, for the inefficiency error term (u_i).

- Cobb-Douglas production frontier using cross-sectional data and assuming Half-Normal distribution for the inefficiency error:
  i) $v_i \sim$ iid $N(0, \sigma_v^2)$,

ii) $u_i \sim$ iid $N^+(0, \sigma_u^2)$.

Given the independence assumption between $\boldsymbol{u}$ and $\boldsymbol{v}$, the joint density function of both errors is:

$$f(u,v) = \frac{2}{2\pi\sigma_v\sigma_u} \cdot \exp\left\{-\frac{u^2}{2\sigma_u^2} - \frac{v^2}{2\sigma_v^2}\right\}. \tag{3.18}$$

Since $\boldsymbol{\varepsilon} = \boldsymbol{v} - \boldsymbol{u}$, the joint density function takes now the following expression:

$$f(u,\varepsilon) = \frac{2}{2\pi\sigma_v\sigma_u} \cdot \exp\left\{-\frac{u^2}{2\sigma_u^2} - \frac{(\varepsilon+u)^2}{2\sigma_v^2}\right\}. \tag{3.19}$$

Thus the marginal density function of $\boldsymbol{\varepsilon}$, which is given by integrating $f(u,\varepsilon)$ in order to $\boldsymbol{u}$, can be written as:

$$f(\varepsilon) = \frac{2}{\sigma} \cdot \phi\left(\frac{\varepsilon}{\sigma}\right) \cdot \Phi\left(-\frac{\varepsilon\lambda}{\sigma}\right), \tag{3.20}$$

where $\sigma = \sqrt{\sigma_u^2 + \sigma_v^2}$ and $\lambda = \frac{\sigma_u}{\sigma_v}$.
The $\Phi(\cdot)$ and $\phi(\cdot)$ represent the standard Normal cumulative distribution and density functions. This way the conditional distribution of $\boldsymbol{u}$ given $\boldsymbol{\varepsilon}$ takes the following expression:

$$f(u|\varepsilon) = \frac{\exp\left\{-\frac{(u-\mu_*)^2}{2\sigma_*^2}\right\}}{\sqrt{2\pi}\sigma_*\left[1 - \Phi\left(-\frac{\mu_*}{\sigma_*}\right)\right]}, \tag{3.21}$$

which results on the density function of a variable distributed as $N^+(\mu_*, \sigma_*^2)$.
The parameters can be rewritten as $\mu_* = -\frac{\varepsilon\sigma_u^2}{\sigma^2}$ and $\sigma_*^2 = \frac{\sigma_u^2\sigma_v^2}{\sigma^2}$, which determines that, the mode of the distribution can be used as an estimator of $\boldsymbol{u}$:

$$\hat{u}_i = M(u|\varepsilon_i) = \begin{cases} -\varepsilon_i\left(\frac{\sigma_u^2}{\sigma^2}\right) & \text{if } \varepsilon_i \leq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.22}$$

The TE estimator of each producer can be obtained from:

$$\widehat{TE}_i = E(\exp\{-u\}|\varepsilon_i) = \frac{1 - \Phi\left(\sigma_* - \frac{\mu_{*i}}{\sigma_*}\right)}{1 - \Phi\left(-\frac{\mu_{*i}}{\sigma_*}\right)} \cdot exp\{-\mu_{*i} + \sigma_*^2/2\}, \tag{3.23}$$

where $\mu_{i*} = -\frac{\varepsilon_i\sigma_u^2}{\sigma^2}$ and $\sigma_*^2 = \frac{\sigma_u^2\sigma_v^2}{\sigma^2}$.

- Cobb-Douglas production frontier using cross-sectional data and assuming Truncated-Normal distribution:
  i) $v_i \sim$ iid $N(0, \sigma_v^2)$,
  ii) $u_i \sim$ iid $N^+(\mu, \sigma_u^2)$.

  Passing some steps which are analogous to those presented in the previous model, the conditional distribution can be expressed as:

$$f(u|\varepsilon) = \frac{\exp\left\{-\frac{(u-\tilde{\mu})^2}{2\sigma_*^2}\right\}}{\sqrt{2\pi}\sigma_* \left[1 - \Phi\left(-\frac{\tilde{\mu}}{\sigma_*}\right)\right]}, \tag{3.24}$$

  where $\tilde{\mu}_i = \frac{(-\sigma_u^2\varepsilon_i + \mu\sigma_v^2)}{\sigma^2}$.
  The u_i and TE_i estimators are:

$$\hat{u}_i = M(u|\varepsilon_i) = \begin{cases} \tilde{\mu}_i & \text{if } \tilde{\mu}_i \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.25}$$

$$\widehat{TE}_i = E(\exp\{-u\}|\varepsilon_i) = \frac{1 - \Phi\left(\sigma_* - \frac{\tilde{\mu}_i}{\sigma_*}\right)}{1 - \Phi\left(-\frac{\tilde{\mu}_i}{\sigma_*}\right)} \cdot \exp\{-\tilde{\mu}_i + \sigma_*^2/2\}. \tag{3.26}$$

Panel Data

For the panel data, we describe now the other procedure of estimating TE, when there is no fixed or random effects. Although there are two different methodologies (time variant and invariant) the steps in both procedures are identical. For both only the maximum likelihood method was performed and the steps detailed below corresponds to the time-invariant approach. The procedure is similar to one made in the cross-sectional data and includes:

- Step 1: The density functions of $\boldsymbol{v} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_T)$ and $\boldsymbol{u}$ are used to estimate:
  - The joint density function of $\boldsymbol{u}$ and $\boldsymbol{v}$, which are independently distributed;
  - The joint density function of $\boldsymbol{u}$ and $\boldsymbol{\varepsilon_i}$ ($\boldsymbol{\varepsilon_i} = \boldsymbol{v_i} - \boldsymbol{u}$);
  - The marginal density function of $\boldsymbol{\varepsilon}$.
- Step 2: The conditional distribution $f(u|\boldsymbol{\varepsilon})$ is calculated based on the joint density function $f(u, \boldsymbol{\varepsilon})$ and the marginal density function $f(\boldsymbol{\varepsilon})$.

- Step 3: Then the u estimator corresponds to: $M(u|\boldsymbol{\varepsilon_i})$ (Conditional Mode).
- Step 4: The estimator of the technical efficiency is: $TE_i = E(\exp\{-u\}|\boldsymbol{\varepsilon}_i)$.

In this case the inefficiency error term $\boldsymbol{u}$ is a vector of I dimension, that corresponds to the number of producers. For each producer i the random error $\boldsymbol{v_i}$ is a vector of T dimension, being T the number of time periods considered (the same applies to $\boldsymbol{\varepsilon}$).

For this kind of data were performed two approaches. In the time-invariant approach, the same cases considered in cross-sectional data are now described, and in time-variant approach the EEF model was described.

- Cobb-Douglas production frontier using panel data and assuming Half-Normal distribution:
  i) $v_{it} \sim$ iid $N(0, \sigma_v^2)$,
  ii) $u_i \sim$ iid $N^+(0, \sigma_u^2)$.

  The density function of $\boldsymbol{v}$, which is now time dependent, is given by the following expression:

$$f(\boldsymbol{v}) = \frac{1}{(2\pi)^{T/2}\sigma_v^T} \cdot \exp\left\{\frac{-\boldsymbol{v'v}}{2\sigma_v^2}\right\}. \tag{3.27}$$

  Given the independence assumption between u and $\boldsymbol{v}$, the joint density function is:

$$f(u, \boldsymbol{v}) = \frac{2}{(2\pi)^{(T+1)/2}\sigma_v^T\sigma_u} \cdot \exp\left\{-\frac{u^2}{2\sigma_u^2} - \frac{\boldsymbol{v'v}}{2\sigma_v^2}\right\}. \tag{3.28}$$

  The joint function of u and $\boldsymbol{\varepsilon} = (v_1 - u, \ldots, v_T - u)$ is given by:

$$f(u, \boldsymbol{\varepsilon}) = \frac{2}{(2\pi)^{(T+1)/2}\sigma_v^T\sigma_u} \cdot \exp\left\{-\frac{(u-\mu_*)^2}{2\sigma_*^2} - \frac{\boldsymbol{\varepsilon'\varepsilon}}{2\sigma_v^2} + \frac{\mu_*^2}{2\sigma_*^2}\right\}, \tag{3.29}$$

  where $\mu_* = -\frac{\sigma_u^2 T\bar{\varepsilon}}{\sigma_v^2 + T\sigma_u^2}$, $\bar{\varepsilon} = \frac{1}{T}\sum_t \varepsilon_{it}$ and $\sigma_*^2 = \frac{\sigma_u^2\sigma_v^2}{\sigma_v^2 + T\sigma_u^2}$.

  Thus the marginal density function of $\boldsymbol{\varepsilon}$, which is given by integrating $f(u, \boldsymbol{\varepsilon})$ in order to u, can be written as:

$$f(\boldsymbol{\varepsilon}) = \frac{2[1 - \Phi(-\mu_*/\sigma_*)]}{(2\pi)^{T/2}\sigma_v^{T-1}(\sigma_v^2 + T\sigma_u^2)^{1/2}} \cdot \exp\left\{-\frac{\boldsymbol{\varepsilon'\varepsilon}}{2\sigma_v^2} + \frac{\mu_*^2}{2\sigma_*^2}\right\}. \tag{3.30}$$

  The conditional distribution of u given $\boldsymbol{\varepsilon}$ is given by:

$$f(u|\boldsymbol{\varepsilon}) = \frac{1}{(2\pi)^{1/2}\sigma_*[1-\Phi(-\mu_*/\sigma_*)]} \cdot \exp\left\{-\frac{(u-\mu_*)^2}{2\sigma_*^2}\right\}, \tag{3.31}$$

which results on the density function of a variable distributed as $N^+(\mu_*, \sigma_*^2)$.
Considering the mode of this distribution as the point estimator of technical inefficiency, results on:

$$\hat{u_i} = M(u|\boldsymbol{\varepsilon}_i) = \begin{cases} \mu_{*i} & \text{if } \boldsymbol{\varepsilon}_i \leq 0; \\ 0 & \text{otherwise} \end{cases} \tag{3.32}$$

with $\mu_{*i} = -\frac{\sigma_u^2 T\bar{\varepsilon}_i}{\sigma_v^2 + T\sigma_u^2}$.
The point estimator μ_{*i} has to be nonnegative, which means $-\frac{T\bar{\varepsilon}_i\sigma_u^2}{\sigma_v^2+T\sigma_u^2} \geq 0$. This condition is verified if $\sum_t \varepsilon_{it} \leq 0$, i.e. if for all i: $\boldsymbol{\varepsilon}_i \leq 0$.
The estimator of TE_i for each producer i takes the following expression:

$$\widehat{TE_i} = E(\exp\{-u\}|\boldsymbol{\varepsilon}_i) = \frac{1-\Phi[\sigma_* - (\mu_{*i}/\sigma_*)]}{1-\Phi(-\mu_{*i}/\sigma_*)} \cdot \exp\left\{-\mu_{*i} + \sigma_*^2/2\right\}. \tag{3.33}$$

- Cobb-Douglas production frontier using panel data and assuming Truncated-Normal distribution:
  i) $v_{it} \sim$ iid $N(0, \sigma_v^2)$,
  ii) $u_i \sim$ iid $N^+(\mu, \sigma_u^2)$.

Passing some steps similar to those presented above, the conditional distribution of u given $\boldsymbol{\varepsilon}$ is given by:

$$f(u|\boldsymbol{\varepsilon}) = \frac{1}{(2\pi)^{1/2}\sigma_*[1-\Phi(-\tilde{\mu}/\sigma_*)]} \cdot \exp\left\{-\frac{(u-\tilde{\mu})^2}{2\sigma_*^2}\right\}, \tag{3.34}$$

where $\tilde{\mu} = \frac{\mu\sigma_v^2 - \sigma_u^2 T\overline{\varepsilon}}{\sigma^2 + T\sigma_u^2}$ and $\sigma_*^2 = \frac{\sigma_u^2\sigma_v^2}{\sigma^2}$.
This conditional distribution is the density function of a variable with $N^+(\tilde{\mu}, \sigma_*^2)$ distribution. Thus the mode of the distribution, used as point estimator of the inefficiency error u, corresponds to:

$$\hat{u_i} = M(u|\boldsymbol{\varepsilon}_i) = \begin{cases} \tilde{\mu}_i & \text{if } \tilde{\mu}_i \geq 0; \\ 0 & \text{otherwise} \end{cases} \tag{3.35}$$

with $\tilde{\mu}_i = \frac{\mu\sigma_v^2 - \sigma_u^2 T\overline{\varepsilon}_i}{\sigma^2 + T\sigma_u^2}$.
The point estimator of TE_i can be expressed as:

$$\widehat{TE}_i = E(\exp\{-u\}|\boldsymbol{\varepsilon}_i) = \frac{1-\Phi[\sigma_* - (\tilde{\mu}_i/\sigma_*)]}{1-\Phi(-\tilde{\mu}_i/\sigma_*)} \cdot \exp\left\{-\tilde{\mu}_i + \sigma_*^2/2\right\}. \tag{3.36}$$

- Efficiency Effects Frontier using the expressions detailed in Coelli and Battese [1993]:
  i) $v_{it} \sim$ iid $N(0, \sigma_v^2)$
  ii) $u_{it} \sim$ iid $N^+(z_{it}\delta, \sigma_u^2)$

  The u_{it} density function is:

$$f(u) = \frac{\exp\left(-\frac{(u-\boldsymbol{z\delta})^2}{2\sigma_u^2}\right)}{\sqrt{2\pi}\sigma_u \Phi(\boldsymbol{z\delta}/\sigma_u)} \tag{3.37}$$

  The indexes i and t are omitted and Φ represents the standard Normal distribution function. Given the independency between v and u the joint density function of ε and u (replacing v by $\varepsilon + u$) is given by:

$$f(u,\varepsilon) = \frac{\exp\left(-\frac{1}{2}\left[\frac{\varepsilon^2}{\sigma_v^2} + \frac{(\boldsymbol{z\delta})^2}{\sigma_u^2} + \frac{(u-\mu_*)^2}{\sigma_*^2} - \frac{\mu_*^2}{\sigma_*^2}\right]\right)}{2\pi\sigma_u\sigma_v\Phi(\boldsymbol{z\delta}/\sigma_u)}, \tag{3.38}$$

  where $\mu_* = \frac{\sigma_v^2 \boldsymbol{z\delta} - \sigma^2\varepsilon}{\sigma_v^2+\sigma_u^2}$ and $\sigma_*^2 = \frac{\sigma_v^2\sigma_u^2}{\sigma_v^2+\sigma_u^2}$.
  The marginal density function of ε is given by integrating $f(u,\varepsilon)$ in order to u:

$$f(\varepsilon) = \frac{\exp\left(-\frac{1}{2}\left[\frac{\varepsilon^2}{\sigma_v^2} + \frac{(\boldsymbol{z\delta})^2}{\sigma_u^2} - \frac{\mu_*^2}{\sigma_*^2}\right]\right)}{\sqrt{2\pi(\sigma_u^2+\sigma_v^2)}} \frac{\Phi\left(\frac{\mu_*}{\sigma_*}\right)}{\Phi\left(\frac{\boldsymbol{z\delta}}{\sigma_u}\right)}. \tag{3.39}$$

  The conditional distribution of u given ε is obtained by the quotient between the two expressions above:

$$f(u|\varepsilon) = \frac{\exp\left(-\frac{1}{2}\left[\frac{(u-\mu_*)^2}{\sigma_*^2}\right]\right)}{\sqrt{2\pi}\sigma_*\Phi\left(\frac{\mu_*}{\sigma_*}\right)}, u \geq 0. \tag{3.40}$$

  The technical efficiency estimator is given by the conditional expectation of $\exp(-u)$:

$$E(\exp(-u)|\varepsilon) = \exp\left(-\mu_* + \frac{1}{2}\sigma_*^2\right) \cdot \frac{\Phi\left(\frac{\mu_*}{\sigma_*} - \sigma_*\right)}{\Phi\left(\frac{\mu_*}{\sigma_*}\right)}. \tag{3.41}$$

3.2 Materials and Methods

3.2.1 Variables

The estimator of TE of fisheries requires data on inputs and outputs from fishing process. In this study, 9 explanatory variables (inputs) were analyzed. Among these 8 of them were collected through inquires to the stakeholders and the other collected through the daily landings or the logbooks.

The explanatory variables (inputs) considered, can be divided into three categories: vessels characteristics, skipper skill level and fishing activity features. The variables in the first categoric include: XCOMP (vessel length-over-all), which was highly significant in the previous GLM approach; vessel's age in years (AGE) and the construction material of the vessel (MAT). The other category is related to the skipper of fishing vessels and the variables considered were: the skippers experience in years (XP) and the education level (SCHOLAR), which was further divided into the following levels: primary - 1, first cycle - 2 and so on. The last category, which is more directly related to the fishing process, include: HOOKS (number of hooks) because, according to the stakeholders, is constant throughout the year (and consequently throughout the quarters), and according to the results obtained from the last chapter is a significant variable; number of workers in land (NLAND); number of fishermen at sea (NSEA) and number of trips during the time period considered.

Additionally the variable PERCCYOGUQ, which corresponds to the ratio of deep-water sharks in the total catch, was also considered. This variable was only considered in the EEF procedure detailed before as a variable which may influence the technical efficiency (z variable), because the sharks are a by-catch and their quantities cannot be controlled by the skipper.

Black scabbardfish catches (the output) was presented into two different ways, one per year and another per quarter. The first option is due to the fact that all variables are constant throughout the year (except number of trips), whereas for the quarter, is due to the fact that the seasonality affects the catch values (as could be seen in the GLM results).

The quality and the consistency of the data collected at the logbooks was once again evaluated. Daily landings stored at DGPA (database used before in GLM in the 2^{nd} data set - LBSF) and catches registered in logbooks (CBSF) were used as output ($\boldsymbol{Y}$). The variable representing the number of trips depends on the database used, thus there are two variables: LTRIPS - number of trips recorded in daily landings and CTRIPS - number

of trips recorded in logbooks.

Since many of these variables have been collected through inquires to the stakeholders, the older the data is, the harder is the access to this information and less reliable is the information collected. So to avoid misunderstandings and biased data, both in the panel data as in the cross-sectional data, only the most recent years were considered. In this exercise only the consecutive and recent years (2009 and 2010) were considered. This way it was possible to compare the results and developments of the fishery process from 2009 to 2010.

3.2.2 Computer Routines

The R package *frontier* was used to implement the models. This package allows to consider two of the four possible distributions previously mentioned: Half and Truncated Normal for the inefficiency error component u. Therefore the analysis was restricted only to these two distributions. This routine follows the approach detailed in Coelli [1996] and can be summarized in the following steps:

- 1) Firstly OLS estimates of the function $f(\boldsymbol{x}, \boldsymbol{\beta})$ are obtained and all $\boldsymbol{\beta}$ estimators with the exception of the intercept (β_0) are unbiased.
- 2) A two-phase grid search of $\gamma = \frac{\sigma_u^2}{\sigma^2}$ is conducted, with the $\boldsymbol{\beta}$ parameters set to the OLS values. Possible values of γ varies from 0.1 to 0.9 with increments of 0.1. The β_0 and $\sigma^2 = \sigma_u^2 + \sigma_v^2$ parameters are adjusted according to the corrected ordinary least squares formula. At this phase other parameters such as μ or $\boldsymbol{\delta}$ are set to zero.
- 3) Finally the values resulting from the grid search are used as starting values in an iterative maximization procedure. Davidon-Fletcher-Powell (DFP) Quasi-Newton algorithm is applied.

The Davidon-Fletcher-Powell (FPD) method has been successfully used in a wide range of econometric applications and was also recommended for the estimation of the stochastic frontier production function [Coelli, 1996]. DFP belongs to the quasi-Newton methods group, which attempt to locate a local minimum of a function f.

3.2.3 Models

The TE estimation, as mentioned before, was restricted to two distributions: Half-Normal and Truncated Normal. Under TE estimation several models were adjusted to the

2009 and 2010 years, to the different response variables (LBSF and CBSF) and considering two types of data (Cross and Panel). The models adjusted were:

- Model 1: Cross - LSBF - 2009 - Half Normal
- Model 2: Cross - LBSF - 2009 - Truncated Normal
- Model 3: Panel - LBSF - 2009 - Half Normal
- Model 4: Panel - LBSF - 2009 - Truncated Normal
- Model 5: Panel - CBSF - 2009 - Best distribution resulted with LSBF
- Model 6: Panel - LBSF - 2010 - Efficiency Effects Frontier
- Model 7: Panel - LSBF - 2009 - Efficiency Effects Frontier
- Model 8: Panel - LBSF - 09/10 - Efficiency Effects Frontier

For a complete description of the models, see Annex 1. Based on the results of these models, several comparisons were performed:

- Model 1 vs Model 2 and Model 3 vs Model 4:
  The two pairs of models aimed to analyze the impact of the two distributions used. In the first pair the approach followed admitted that the data was of cross sectional type, whereas in the second pair the data was in panel type. If the confidence interval of μ did not included zero for some significance level, the appropriated distribution should be the Truncated Normal (with $\mu \neq 0$).
- Model 1 vs Model 3 and Model 2 vs Model 4:
  With these two pairs of models the purpose was analyze the differences in results between the cross-sectional and panel approaches.
- Model 5:
  This model was performed to evaluate the differences between the two databases (logbooks vs daily landings). This evaluation was done by comparing the results of model 5 with the results obtained with the best fitted model among the first four.
- Model 6 vs Model 7:
  This comparison aimed to analyze the main differences between 2009 and 2010.
- Model 8:
  This model was considered to evaluate the seasonality and the trend over the two years and to compare this approach with models 6 and 7 together.

For these comparisons was evaluated the variables selected from the initial ones, the estimates of technical efficiencies (and the respective rankings) and the mean of the technical efficiency. To verify if the different approaches gave different estimates and rankings of technical efficiencies, it was performed, besides graphical analysis, the Wilcoxon signed-rank test. This is a non-parametric statistical test used when two related samples are compared (i.e. it's a paired difference test). The null hypothesis is that the difference between the two samples is zero.

A backward stepwise procedure based on the highest *p-value* was adopted to select the variables to be included in the models. It has begun with the saturated model (all variables included) and step by step the not significant variables were removed. This was done until only remained in the model the significant ones and was assessed by likelihood ratio test, comparing the fit of two models, to verify if the model without the removed variable was statistically different from the model with it.

3.3 Results

We now present the results for models and we recall that $\sigma^2 = \sigma_u^2 + \sigma_v^2$ and $\gamma = \frac{\sigma_u^2}{\sigma^2}$.

Model 1 - Cross Sectional data - Half Normal - LBSF 2009 - Tab. 3.1

Table 3.1: Cross data and $u_i \backsim N^+(0, \sigma_u^2)$

	Estimate	Std. Error	z value	p-value
Intercept	0.500	0.984	0.508	0.611
XCOMP	1.922	0.869	2.212	**0.027**
LTRIPS	1.317	0.528	2.495	**0.013**
σ^2	0.198	0.741	0.267	0.780
γ	0.992	0.674	1.472	0.141

- Variables selected: XCOMP and LTRIPS (using 0.1 as significance level);
- Inefficiency parameters: σ^2 and γ were not statistical significant, i.e. they were statistically null.
- Mean of technical efficiency estimates: 0.731.
- Conclusions: As the large value of γ indicates, the variation in the composed error term ε is mainly due to the inefficiency error and so it should not be ignored. However, one should note, that the standard error of γ is so large that the null

hypothesis for this parameter is not rejected. This uncertainty about the real relative weight of the variation of u compared to that of v, must be originated on the small number of observations (only 16).

Model 2 - Cross-Sectional data - Truncated Normal - LBSF 2009 - Tab. 3.2

Table 3.2: Cross data and $u_i \backsim N^+(\mu, \sigma_u^2)$

	Estimate	Std. Error	z value	p-value
Intercept	-0.169	0.633	-0.267	0.789
XCOMP	1.821	0.233	7.823	**5e-15**
LTRIPS	1.518	0.112	13.57	**< 2e-16**
σ^2	0.285	0.103	2.750	**0.006**
γ	0.999	0.0001	19387.4	**< 2e-16**
μ	-0.190	0.343	-0.552	0.581

- Variables selected: XCOMP and LTRIPS.
- Inefficiency parameters: σ^2 and γ were statistical significant, i.e. they were statistically different from zero.
- Mean of technical efficiency estimates: 0.727.
- Conclusions: In this case, the inefficiency error term was clearly present in the data and constituted a large part of the variation of ε (due to the high significance of gamma). Moreover, the parameter which defines the assumed distribution (μ) was not statistically different from zero, thus the more suitable distribution for u was the Half distribution instead of the Truncated on μ.

Model 3 - Panel data - Half Normal - LBSF 2009 - Tab. 3.3

Table 3.3: Panel data and $u_i \backsim N^+(0, \sigma_u^2)$

	Estimate	Std. Error	z value	p-value
Intercept	3.123	1.292	2.417	**0.016**
XCOMP	1.167	0.493	2.365	**0.018**
NLAND	0.510	0.266	1.921	**0.055**
LTRIPS	0.979	0.121	8.096	**5e-16**
σ^2	0.222	0.079	2.805	**0.005**
γ	0.769	0.094	8.137	**4e-16**

- Variables selected: XCOMP, LTRIPS and NLAND (using 0.1 as significance level).

- Inefficiency parameters: σ^2 and γ were statistical different from zero and highly significant.
- Mean of technical efficiency estimates: 0.74.
- Conclusions: As the cross sectional data results already indicated, the inefficiency error should be present and a large parte of the error variation should be attributed to it.

Model 4 - Panel data - Truncated Normal - LBSF 2009 - Tab. 3.4

Table 3.4: Panel data and $u_i \frown N^+(\mu, \sigma_u^2)$

	Estimate	Std. Error	z value	p-value
Intercept	3.113	1.367	2.278	**0.023**
XCOMP	1.170	0.518	2.261	**0.024**
NLAND	0.511	0.269	1.901	**0.057**
LTRIPS	0.978	0.123	7.936	**2e-15**
σ^2	0.217	0.261	0.831	0.406
γ	0.763	0.289	2.637	**0.008**
μ	0.017	0.885	0.019	0.984

- Variables selected: XCOMP, LTRIPS and NLAND (using 0.1 as significance level).
- Inefficiency parameters: γ was statistical different from zero while the σ^2 parameter was not. This difference must be due to the fact that the parameter μ had a high standard error (0.885).
- Mean of technical efficiency estimates: 0.739.
- Conclusions: The inefficiency error term could not be ignored and the Half Normal distribution should be the one considered, since the μ was statistically null.

Model 5 - Panel data - Half Normal - CBSF 2009 - Tab. 3.5

Once found the most convenient model for the LBSF data (Model 3), the same model was fitted to the data from the logbooks - CBSF.

- Variables selected: XCOMP, CTRIPS and NLAND (using 0.1 as significance level).
- Inefficiency parameters: γ and σ^2 were statistical different from zero (i.e. significant).

Table 3.5: Panel data and $u_i \frown N^+(0, \sigma_u^2)$

	Estimate	Std. Error	z value	p-value
Intercept	3.629	1.315	2.760	**0.006**
XCOMP	1.068	0.497	2.150	**0.032**
NLAND	0.485	0.267	1.816	**0.069**
CTRIPS	0.934	0.117	8.000	**1e-15**
σ^2	0.264	0.094	2.806	**0.005**
γ	0.825	0.072	11.38	$<$ **2e-16**

- Mean of technical efficiency estimates: 0.722.
- Conclusions: The model shows that the inefficiency error term should be present and the results were quite similar to those of model 3.

Model 6 - EEF 2010 - Tab. 3.6

Before running the EEF Model, it was tried for 2010 the time invariant model 3, which proved to be the best for 2009. Unfortunately, it was not possible to run such a model for 2010, since the algorithm did not converge. In fact, the OLS estimates given in the first step (without inefficiency) gave a better fit than the estimates given on the second step (with inefficiency), showing that this type of error should not to be considered in the model. For the EEF model the convergence turned out to be reached.

Table 3.6: EEF (2010) and $u_i \frown N(m_{it}, \sigma_u^2)$

	Estimate	Std. Error	z value	p-value
(Intercept)	-8.716	0.860	-10.13	$<$ **2e-16**
XCOMP	1.593	0.225	7.082	**1e-12**
HOOKS	1.461	0.165	8.864	$<$ **2e-16**
MAT-2	-0.343	0.068	-5.037	**4e-07**
MAT-3	0.681	0.103	6.633	**3e-11**
SCHOLAR	-0.130	0.036	-3.634	**0.0003**
NSEA	-0.549	0.126	-4.363	**1e-05**
LTRIPS	1.009	0.110	9.136	$<$ **2e-16**
PERCCYOGUQ	1.868	0.866	2.157	**0.031**
σ^2	0.106	0.029	3.696	**0.0002**
γ	0.9996	0.003	362.20	$<$ **2e-16**

- Variables selected: XCOMP, LTRIPS, HOOKS, MAT, NSEA and SCHOLAR.
- Z variables: PERCCYOGUQ was significant, i.e. statistically different from zero.
- Inefficiency parameters: γ and σ^2 were statistical different from zero.

- Conclusions: The inefficiency must be taken into consideration and based on the parameter estimate, the PERCCYOGUQ had a positive impact on inefficiency.

In Table 3.7 are presented the technical efficiencies per quarter, i.e. the mean of technical efficiencies of all 15 vessels for each time period. As can be seen there was some tendency that the efficiency gradually increased throughout the year achieving 0.85. As result, the mean efficiency for the four periods was 0.731.

Table 3.7: Mean of technical efficiency for each time period (model 6)

Quarter	Efficiency
Quarter 1	0.62
Quarter 2	0.69
Quarter 3	0.77
Quarter 4	0.85

Model 7 - EEF 2009 - Tab. 3.8

In order to compare the results between 2009 e 2010, model 6 procedure was ran for 2009.

Table 3.8: EEF (2009) and $u_i \frown N(m_{it}, \sigma_u^2)$

	Estimate	Std. Error	z value	p-value
Intercept	-1.881	0.858	-2.194	**0.028**
XCOMP	1.036	0.205	5.046	**4e-07**
HOOKS	0.621	0.138	4.487	**7e-06**
MAT-2	0.135	0.074	1.827	**0.068**
MAT-3	0.104	0.139	0.748	0.455
LTRIPS	1.282	0.137	9.386	$<$ **2e-16**
PERCCYOGUQ	3.353	0.399	8.412	$<$ **2e-16**
σ^2	0.126	0.033	3.884	**1e-04**
γ	1.000	0.000	81219.6	$<$ **2e-16**

- Variables selected: XCOMP, LTRIPS, HOOKS and MAT (using 0.1 as significance level).
- Z variables: PERCCYOGUQ was highly significant.
- Inefficiency parameters: γ and σ^2 were statistical different from zero.
- Conclusions: The results pointed at the same direction as model 6. The inefficiency need to be considered and the PERCCYOGUQ seemed to have even a larger impact on inefficiency.

Table 3.9 presents the technical efficiencies per quarter of 15 vessels for each time period and there was no positive trend throughout the year 2009. The mean of technical efficiency ranged between 0.67 and 0.75 reaching 0.67 as average of all observations.

Table 3.9: Mean of technical efficiency for each time period (model 7)

Quarter	Efficiency
Quarter 1	0.66
Quarter 2	0.63
Quarter 3	0.64
Quarter 4	0.75

Model 8 - EEF 2009/2010 - Tab. 3.10

This model comprises the two years (2009 and 2010) in the same data set, thus it was considered not four but eight time periods, four for each year. Thus it was possible to evaluate in a single model the trend and the seasonality of technical efficiency along the years.

Table 3.10: EEF (2009) and $u_i \frown N(m_{it}, \sigma_u^2)$

	Estimate	Std. Error	z value	p-value
Intercept	-5.603	1.298	-4.318	**2e-05**
XCOMP	0.901	0.212	4.259	**2e-05**
HOOKS	1.147	0.181	6.345	**2e-10**
MAT-2	-0.065	0.083	-0.785	0.4323
MAT-3	0.535	0.103	5.204	**2e-07**
LTRIPS	1.094	0.071	15.383	$<$ **2e-16**
PERCCYOGUQ	3.089	0.399	7.747	**9e-15**
σ^2	0.094	0.018	5.157	**3e-07**
γ	0.823	0.126	6.558	**5e-11**

- Variables selected: XCOMP, LTRIPS, HOOKS and MAT.
- Z variables: PERCCYOGUQ was highly significant.
- Inefficiency parameters: The null hypothesis of γ and σ^2 were statistical null was rejected for all usual significance levels.
- Conclusions: The inefficiency should be present in the model and once again the PERCCYOGUQ had a positive impact on inefficiency.

3.4 Discussion

The results concerning the technical efficiency of the black scabbarfish fishery fleet during 2009, considering only the the two time invariant approaches, revealed no significative differences between the two assumed distributions, Half and Truncated Normal. In fact, in both cases (cross sectional or panel-time invariant model), the μ estimate was not statistically significant for all the usual significance levels, and the estimates of the mean efficiency as well as of the coefficients of the variables selected do not greatly differ. Note that for the inefficiency parameters, (λ and σ^2) analogous comparisons could not be performed, because their estimates depend on the distribution assumed.

However, depending on the type of data, the producer (i.e. vessel) technical efficiency estimates for both distributions differ more or less (upper plots of Fig. 3.1 and Tab. 3.11). For cross sectional data the differences were higher than for the panel data where they were almost null. The same conclusions can be drawn from the upper plots of Fig. 3.2 where the ranking of the 16 producers technical efficiency estimates are compared between distributions. In the panel case (right plot) there were no differences between the rankings and in the cross sectional data (left plot) they were small. Despite these differences, the Wilcoxon test null hypothesis of identical individual technical efficiencies for the two distributions, was not rejected for this type of data.

Table 3.11: Summary of technical efficiencies estimates for 2009 with LBSF.

	Cross - Half	Cross - Truncated	Panel - Half	Panel - Truncated
Vessel 1	0.402	0.392	0.558	0.557
Vessel 2	0.941	0.910	0.915	0.915
Vessel 3	0.799	0.830	0.716	0.714
Vessel 4	0.636	0.612	0.601	0.601
Vessel 5	0.917	0.920	0.892	0.891
Vessel 6	0.752	0.783	0.711	0.710
Vessel 7	0.944	0.911	0.903	0.903
Vessel 8	0.586	0.602	0.753	0.750
Vessel 9	0.746	0.722	0.773	0.771
Vessel 10	0.492	0.490	0.459	0.458
Vessel 11	0.664	0.636	0.625	0.624
Vessel 12	0.710	0.697	0.898	0.897
Vessel 13	0.975	0.999	0.921	0.920
Vessel 14	0.401	0.389	0.462	0.462
Vessel 15	0.841	0.829	0.814	0.813
Vessel 16	0.892	0.918	0.843	0.841

Analyzing now the differences between the two invariant cases, for each of the distributions, one can conclude that, despite the non rejection of the Wilcoxon test null

hypothesis, the differences between the rankings of individual technical efficiency estimates are much more apparent than before for either distributions (lower plots of Fig. 3.2). Similar conclusions can be drawn about the magnitude of the differences between the individual technical efficiency estimates coming from the two data types (lower plots of Fig. 3.1).

In fact, one would expect such differences to occur, since in the panel data set there were four times more observations than in the cross sectional data set and the set of the model selected variables included one more variable in the panel data case. Finally, one should note also that the differences in magnitude as well as in ranking are more apparent with the Truncated Normal distribution hypothesis.

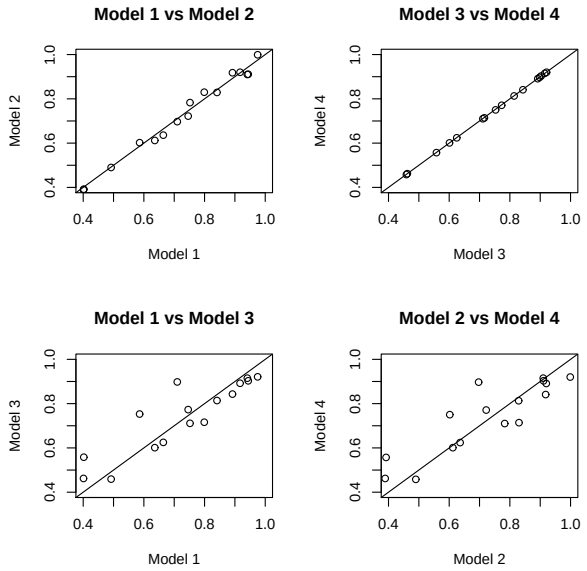

Figure 3.1: Technical efficiencies estimates of the four Time-Invariant models for 2009.

The results obtained regarding the use of the Half or Truncated Normal distributions showed that the distribution choice did not influence the outcome. This way, the Half Normal is preferable, since no extra parameter needs to be estimated. Concerning the analysis of the type of data to be used, cross sectional or panel data, the results were not so clear. In spite of that, the second type of data should be used since for obtaining similar consistency the cross sectional data requires a large set of producers to be observed during one period of time, while the panel data requires a smaller set of producers to be observed during several periods of time. The number of producers being limited, it is more feasible to observe them longer.

Bearing in mind these conclusions, model 3 was selected to evaluate the TE of the black scabbardfish fishery fleet, especially since the estimates for the inefficiency parameters γ and σ^2 were significant.

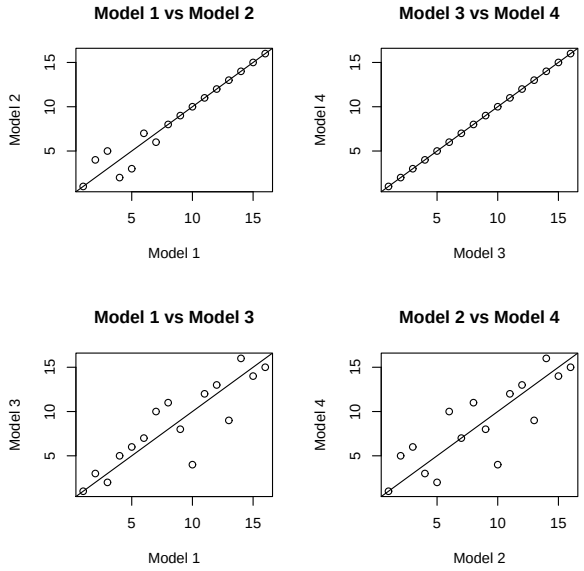

Figure 3.2: Ranking vessels in relation to technical efficiencies estimates of the four first models.

Under model 3 there was a strong positive correlation (Pearson's correlation coefficient 0.75) between the response variable (LBSF) and the technical efficiency estimates (Fig. 3.3). Also under this model no strong correlation was observed between technical efficiency estimates and any of the explanatory variables, since the higher estimate obtained was around 0.25 with HOOKS (Fig. 3.4). This fulfilled the independence assumptions that enable the estimation procedure to be applied.

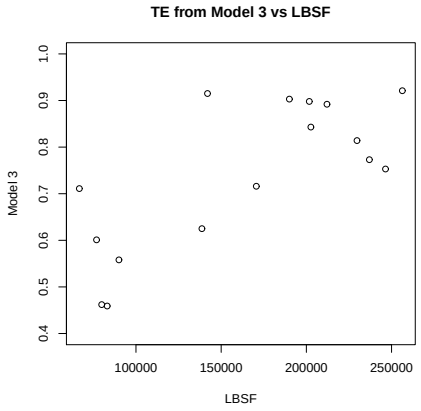

Figure 3.3: Technical efficiencies per vessel vs LSBF (only for Model 3).

Comparing the estimates of technical efficiency from each vessel with the overall mean efficiency (Fig. 3.5), two vessels (numbers 10 and 14) have a much smaller efficiency than the others vessels. In the same figure, it can be seen that four vessels (numbers 3, 6, 8

and 9) presented a technical efficiency similar to the overall mean.

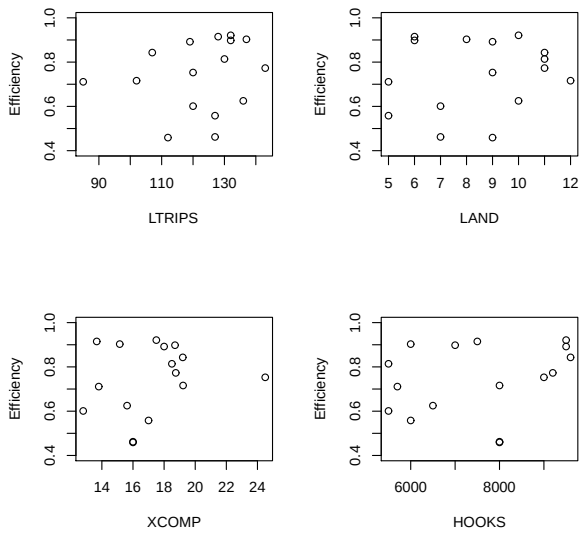

Figure 3.4: Technical efficiencies per vessel vs different inputs (only for Model 3).

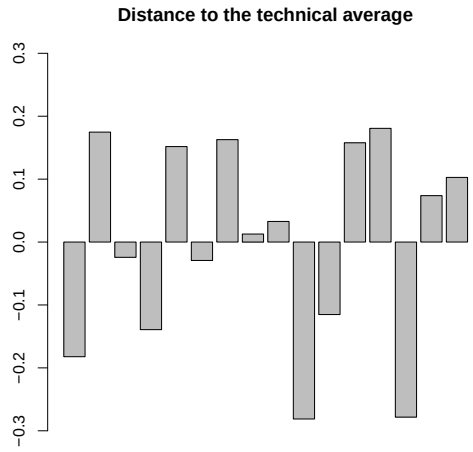

Figure 3.5: Distance of technical efficiencies estimates per vessel to the overall mean (only for Model 3).

The model 5 was fitted using a different response variable CBSF, but adopting the same model with the same explanatory variables as in the model 3. No differences were observed between the estimates of any quantities nor between the vessels rankings (Fig. 3.6). In addition, both the inefficiency parameters and the explanatory variables were statistically significant. Thus it was concluded that the two databases lead to the same

results. Such consistency may come from the fact that only two variables were collected from these two databases (the output and the number of trips).

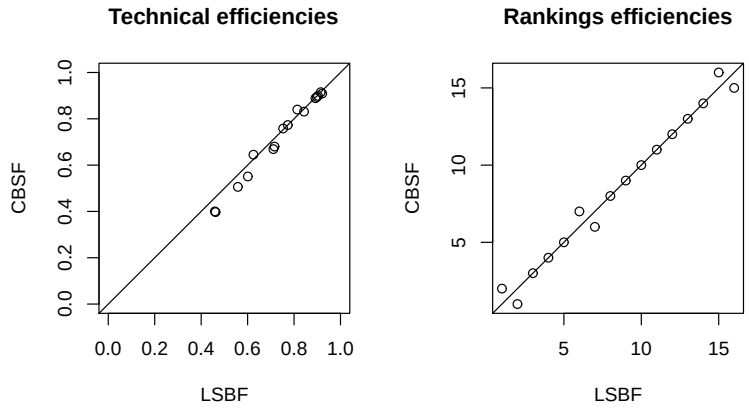

Figure 3.6: Technical efficiency estimates of Model 3 and Model 5 (left) and respective assigned rankings of Model 3 and Model 5 (right).

The adjustment of the EEF models gave support to the hypothesis that PERCCYOGUQ is related to inefficiency. The capture of deep-water sharks has a negative impact on technical efficiency, which is in agreement with the feedback given by the stakeholders and with our suspicions, since the presence of deep-water sharks decreases the catches of black scabbardfish and therefore the efficiency.

The adjustment of EEF for 2009 and for 2010 data separately showed that the variable PERCCYOGUQ was significant in both cases. The results obtained for the two years put also in evidence:

1. differences in the variables selected;
2. differences on technical efficiencies estimates per vessel (Fig. 3.7);
3. major differences on rankings of 8 vessels in 15 (Fig. 3.7);
4. difference on the trend of technical efficiency along the quarters of the year;
5. the overall mean of technical efficiency estimate in 2009 (0.67) was lower than in 2010 (0.73).

Despite the first two items, the Wilcoxon test does not reject the null hypothesis, but with a very low *p-value* (0.11). In what regards the fifth item, differences in overall mean, they are related to the PERCCYOGUQ coefficient, since in 2009 the coefficient estimate was almost the double of what was obtained in 2010, making 2009 less efficient than 2010.

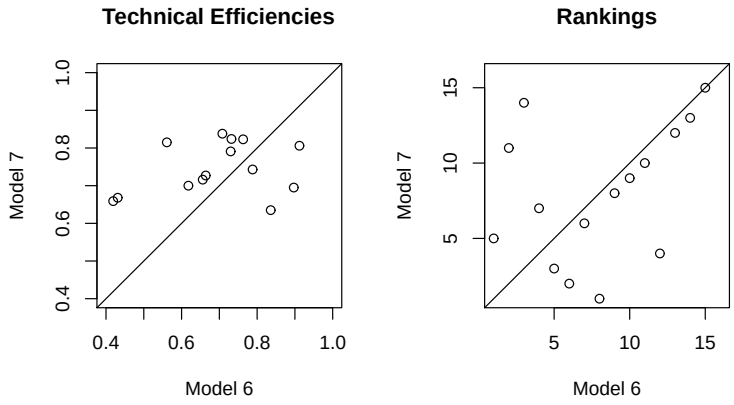

Figure 3.7: Model 6 vs Model 7.

The model 8 encompasses the two years and the 15 vessels, considering at the total 8 time periods; the first four time periods corresponds to 2009 while the last four corresponds to 2010. In both years, the estimates of the technical efficiency average by quarter were higher in the fourth quarter, so this model showed some seasonality. However, while in the 2009 no trend was detected on the technical efficiency, in 2010 there was a slight positive trend along the year (Tables 3.12 and 3.13). The overall mean of technical efficiency was 0.7, which is the arithmetic mean of the estimates obtained for the two years separately (0.67 for 2009 and 0.73 for 2010).

Table 3.12: Summary of technical efficiencies estimates for Model 8 (EEF 09/10).

Time Period	1	2	3	4	5	6	7	8
Vessel 1	0.603	0.56	0.452	0.449	0.587	0.51	0.692	0.902
Vessel 2	0.876	0.716	0.623	0.644	0.86	0.582	0.778	0.821
Vessel 3	0.629	0.716	0.837	0.9	0.874	0.802	0.721	0.858
Vessel 4	0.844	0.648	0.584	0.676	0.666	0.626	0.561	0.752
Vessel 5	0.628	0.623	0.645	0.872	0.487	0.74	0.769	0.794
Vessel 6	0.927	0.929	0.888	0.896	0.672	0.563	0.653	0.56
Vessel 7	0.522	0.62	0.656	0.878	0.454	0.633	0.841	0.846
Vessel 8	0.605	0.645	0.614	0.784	0.625	0.756	0.674	0.83
Vessel 9	0.244	0.307	0.533	0.724	0.764	0.575	0.77	0.783
Vessel 10	0.83	0.774	0.655	0.644	0.709	0.734	0.705	0.849
Vessel 11	0.724	0.829	0.807	0.839	0.715	0.895	0.932	0.916
Vessel 12	0.745	0.78	0.756	0.856	0.656	0.632	0.717	0.612
Vessel 13	0.484	0.331	0.308	0.376	0.439	0.469	0.446	0.55
Vessel 14	0.668	0.768	0.798	0.808	0.489	0.695	0.877	0.93
Vessel 15	0.675	0.668	0.801	0.928	0.562	0.84	0.852	0.888

Table 3.13: Mean of technical efficiency for each time period (model 8)

Quarter	Efficiency
Quarter 1	0.67
Quarter 2	0.66
Quarter 3	0.66
Quarter 4	0.75
Quarter 5	0.64
Quarter 6	0.67
Quarter 7	0.73
Quarter 8	0.79

Regarding to the methodology, it has the positive side of getting the process and producers efficiencies through a simple, quickly and accessible procedure. The negative side is that the implementation of the estimation methodology uses iterative processes that may present some convergence problems. Note also that the estimation method used in the package frontier does not provide asymptotically consistent estimates for technical efficiency in the case of cross sectional data. For this type of data, to obtain more reliable results a large sample of producers is required, preventing this approach to be used in applications where only a small number of producers exists.

For the two other studies included in SFA, Cost and Profit Efficiency, their application is extremely difficult in this area since collecting information about prices, which fluctuate throughout the years, is presently an impossible mission in Portugal. In fact, the collection of data, allowing a reasonable work exclusively for the technical efficiency estimation, proved to be very difficult. The first cause is that the process is complex in itself and the second cause is the resistance shown by some authorities to allow the full access to data.

In general, the black scabbardfish fishery in Sesimbra can be considered efficient since over the selected models the values of technical efficiency did not differ greatly and were quite high, around 0.70. Thought the significant variables changed with from model to model, yet there were variables that had a constant presence: XCOMP and LTRIPS. These two variables have a positive impact in the black scabbardfish catches which are positively related to the efficiency.

The variable AGE and XP were never selected. While for the first variable (vessel age in years), as the vessels are being repaired and renovated the variable loses its impact on efficiency, for the second one (years of experience as a skipper) the experience of the fisherman before being a skipper may somehow also remove the significance of this variable. For future work it would be interesting to consider these two variables that measure the fisherman experience and analyze their relative weight.

There are other ideas for future work that involve approaches that unfortunately could not be applied this time due to the lack of data. Those require the collection of other

variables that are less dependent on memory and were recorded in some data base, as well as the use of economic variables, like expenses on fuel, hooks, bait and salaries. For this purpose the information contained on the balance sheets should be thereof collected and stored in a data base, whose access would allow a more accurate analysis of the different types of efficiency. The last suggestion for future work would be the implementation of the Gamma distribution for the efficiency type errors in a software package.

Chapter 4

Final Remarks

After a long work and having reached several results, it is important to emphasize the principal conclusions and look for the future. The logbooks are one of the most used source of data and this way their importance is huge. Thus, errors in this data source may have a considerable impact or assume a vital role in conclusions and in advances made on this field. In this study, the identified discrepancies between the logbooks and other sources of data were significant, and therefore it is of the utmost importance to instruct and alert the fishing community about the necessity of a correct filling of the logbooks. In reality, only combining the work of scientists with the work of the fishing community will the sustainability of the sea life and of the artisanal fishing be achieved.

The sustainability of the fishing together with the resource, is the ultimate goal of the fishing community, and the knowledge and management of the fishery activity efficiency are vital for that purpose. In the present case, the black scabbardfish, our studies concluded that this fishery can be considered efficient. However, there is a lot to do in future works, such as collecting economic variables that would enable a more accurate analysis of balance sheets and the use of other approaches as Cost and Profit Efficiency.

Bibliography

Bibliography

Aigner, D. J., L. C. A. K., Schmidt, P., 1977. Formulation and estimation of stochastic frontier production function models. Journal of Econometrics 6(1), 21–37.

Battese, G. E., Coelli, T. J., 1988. Prediction of firm-Level technical efficiencies with generalized frontier function and panel data. Journal of Econometrics 38, 387–399.

Bishop, J., 2006. Standardizing fishery-dependent catch and effort data in complex fisheries with technology change. Fish Biology and Fisheries 16, 21–38.

Coelli, T., 1996. A guide to frontier version 4.1: A computer program for stochastic frontier production and cost function estimation. Centre for Efficiency and Productivity Analysis Working Papers 7.

Coelli, T., Battese, G., 1993. A stochastic frontier production function incorporating a model for technical inefficiency effects. Working Papers in Econometrics and Applied Statistics 69.

Figueiredo, I., B.-M. P., Gordo, L., 2005. Deep-water sharks fisheries off Portuguese continental cost. J. Northw. Atl. Fish. Sci. 35, 291–298.

Figueiredo, I., Bordalo-Machado, P., 2007. The fishery for black scabbardfish (*Aphanopus carbo* Lowe, 1839) in the Portuguese continental slope. Reviews in Fish Biology and Fisheries 19, 49–67.

Flores-Lagunes, A., H. W., Schnier, K., 1999. Technical efficiency of the longline fishery in Hawaii: an aplication of a stochastic production frontier. Marine Resources Economics 13, 259–274.

Flores-Lagunes, A., H. W., Schnier, K., 2007. Identifying technically efficient fishing vessels: a non-empty, minimal subset approach. Journal of Applied Econometrics 22, 729–745.

Kumbhakar, S., Lovell, C., 2000. Stochastic Frontier Analysis. Cambridge University Press.

Lee, L.-F., Schmidt, P., 1978. The stochastic frontier production function and average efficiency. Journal of Econometrics 7(3), 385–389.

Maunder, M., Punt, A., 2004. Standardizing catch and effort data: a review of recent approaches. Fisheries Research 70, 141–159.

McCullagh, P., Nelder, J., 1989. Generalized Linear Models, 2nd Edition. Chapman & Hall.

Meeusen, W., van den Broeck, J., 1977. Efficiency estimation from Cobb-Douglas production functions with composed error. International Economic Review 18, 435–444.

Pascoe, S., A. J., de Wilde, J., 2001. The impact of management regulation on the technical efficiency of vessels in the Dutch beam trawl fishery. Eur. Rev. Agric. Econ. 28(2), 187–206.

Schmidt, P., Sickles, R. C., 1984. Production frontier and panel data. Journal of Business and Economic Statistics 2(4), 367–374.

Squires, D., Kirkley, J., 1999. Skipper skill and panel data in fishing industries. Can. J. Fish. Aquat. Sci. 56(11), 2011–2018.

Tingley, D., P. S., Coglan, L., 2005. Factors affecting technical efficiency in fisheries: stochastic production frontier versus data envelopment analysis approaches. Fisheries Research 73, 363–376.

Turkman, M., Silva, G., 2000. Modelos Lineares Generalizados, da teoria à prática.

Williams, D. A., 1987. Generalized linear model diagnostics using deviance and single case deletions. Applied Statistics 36, 181–191.

Zar, J. H., 1996. Biostatistical analysis. Prentice-Hall, Englewood Cliffs.

ANNEX

Annex 1 - Models

Models used in Chapter 2 (Generalized Linear Models)

1^{st} *data set*

Model 1:

Step (glm (BSF $\sim$ *as.factor* (YEAR) $\times$ *as.factor*(QUARTER) + *as.factor* (ERECTAN)+ HOOKS + PERCCYOGUQ + *as.factor* (CLUSTER-ALL)))

Model 2:

Step (glm (BSF $\sim$ *as.factor* (YEAR) $\times$ *as.factor* (QUARTER) + *as.factor* (ERECTAN)+ HOOKS + PERCCYOGUQ + *as.factor* (CLUSTER-XCOMP)))

Model 3:

Step (glm (BSF $\sim$ *as.factor* (YEAR) $\times$ *as.factor* (QUARTER) + *as.factor* (ERECTAN)+ HOOKS + PERCCYOGUQ + *as.factor* (CLUSTER-XTAB)))

Model 4:

Step (glm (BSF $\sim$ *as.factor* (YEAR) $\times$ *as.factor* (QUARTER) + *as.factor* (ERECTAN)+ HOOKS + PERCCYOGUQ + *as.factor* (CLUSTER-XPOW)))

Model 5:

Step (glm (BSF $\sim$ *as.factor* (YEAR) $\times$ *as.factor* (QUARTER) + *as.factor* (ERECTAN) + PERCCYOGUQ + *as.factor* (CLUSTER-XCOMP)))

Model 6:

Step (glm (log (BSF) $\sim$ *as.factor* (YEAR) $\times$ *as.factor* (QUARTER) + *as.factor* (ERECTAN) + HOOKS + PERCCYOGUQ + *as.factor* (CLUSTER-XCOMP)))

2^{nd} *data set*

Model 1:

Step (glm (CBSF $\sim$ *as.factor* (YEAR) + *as.factor* (QUARTER) + *as.factor* (ERECTAN) + PERCCYOGUQ + *as.factor* (CLUSTER-XCOMP)))

Model 2:
Step (glm (log(CBSF) ∼ *as.factor* (YEAR) + *as.factor* (QUARTER) + *as.factor* (ERECTAN) + PERCCYOGUQ + *as.factor* (CLUSTER-XCOMP)))

Models used in Chapter 3 (Stochastic Frontier Analysis)

Model 1:
sfa (log(LBSF) ∼ log(XCOMP) + log(LTRIPS), ineffDecrease = TRUE, truncNorm = FALSE)

Model 2:
sfa (log(LBSF) ∼ log(XCOMP) + log(LTRIPS), ineffDecrease = TRUE, truncNorm = TRUE)

Model 3:
sfa (log(LBSF) ∼ log(XCOMP) + log(LTRIPS) + log(LAND), ineffDecrease = TRUE, truncNorm = FALSE)

Model 4:
sfa (log(LBSF) ∼ log(XCOMP) + log(LTRIPS) + log(LAND), ineffDecrease = TRUE, truncNorm = TRUE)

Model 5:
sfa (log(CBSF) ∼ log(XCOMP) + log(CTRIPS) + log(LAND), ineffDecrease = TRUE, truncNorm = TRUE)

Model 6:
frontier (y = LBSF, x = c(XCOMP, HOOKS, MAT, SCHOLAR, SEA ,LTRIPS), z = c("PERCCYOGUQ"))

Model 7:
frontier (y = LBSF, x = c(XCOMP, HOOKS, MAT, LTRIPS), z = c("PERCCYOGUQ"))

Model 8:
frontier (y = LBSF, x = c(XCOMP, HOOKS, MAT, LTRIPS), z = c("PERCCYOGUQ"))

Annex 2 - Demonstrations of expressions used in SFA

Cross-Sectional data and assuming Half-Normal

Demonstration of the $f(\varepsilon)$:

$$
\begin{aligned}
f(\varepsilon) &= \int_0^{+\infty} f(u,\varepsilon)du \\
&= \int_0^{+\infty} \tfrac{2}{2\pi\sigma_u\sigma_v} \exp\left\{-\tfrac{u^2}{2\sigma_u^2} - \tfrac{(\varepsilon+u)^2}{2\sigma_v^2}\right\} du \\
&= \tfrac{2}{\sqrt{2\pi}} \int_0^{+\infty} \tfrac{1}{\sqrt{2\pi}\sigma_u\sigma_v} \exp\left\{-\tfrac{u^2}{2\sigma_u^2} - \tfrac{\varepsilon^2+u^2+2\varepsilon u}{2\sigma_v^2}\right\} du \\
&= \tfrac{2}{\sqrt{2\pi}} \int_0^{+\infty} \tfrac{1}{\sqrt{2\pi}\sigma_u\sigma_v} \exp\left\{\tfrac{-u^2\sigma_v^2-\varepsilon^2\sigma_u^2-u^2\sigma_u^2-2u\varepsilon\sigma_u^2}{2\sigma_u^2\sigma_v^2}\right\} du \\
&= \tfrac{2}{\sqrt{2\pi}} \exp\left\{-\tfrac{\varepsilon^2\sigma_u^2}{2\sigma_v^2\sigma_u^2}\right\} \int_0^{+\infty} \tfrac{1}{\sqrt{2\pi}\sigma_u\sigma_v} \exp\left\{\tfrac{-u^2(\sigma_u^2+\sigma_v^2)-2u\varepsilon\sigma_u^2}{2\sigma_u^2\sigma_v^2}\right\} du \\
&= \tfrac{2}{\sqrt{2\pi}} \exp\left\{-\tfrac{\varepsilon^2}{2\sigma_v^2}\right\} \int_0^{+\infty} \tfrac{1}{\sqrt{2\pi}\sigma_u\sigma_v} \exp\left\{\tfrac{-u^2(\sigma^2)-2u\varepsilon\sigma_u^2}{2\sigma_u^2\sigma_v^2}\right\} du \\
&= \tfrac{2}{\sqrt{2\pi}} \exp\left\{-\tfrac{\varepsilon^2}{2\sigma_v^2}\right\} \int_0^{+\infty} \tfrac{1}{\sqrt{2\pi}\sigma_u\sigma_v} \exp\left\{\tfrac{-(u\sigma+\varepsilon\sigma_u^2/\sigma)^2+\varepsilon^2\sigma_u^4/\sigma^2}{2\sigma_u^2\sigma_v^2}\right\} du \\
&= \tfrac{2}{\sqrt{2\pi}} \exp\left\{-\tfrac{\varepsilon^2}{2\sigma_v^2} + \tfrac{\varepsilon^2\sigma_u^4/\sigma^2}{2\sigma_u^2\sigma_v^2}\right\} \int_0^{+\infty} \tfrac{1}{\sqrt{2\pi}\sigma_u\sigma_v} \exp\left\{-\tfrac{\sigma^2(u+\varepsilon\sigma_u^2/\sigma^2)^2}{2\sigma_u^2\sigma_v^2}\right\} du \qquad (4.1) \\
&= \tfrac{2}{\sqrt{2\pi}} \exp\left\{-\tfrac{\varepsilon^2}{2\sigma_v^2} + \tfrac{\varepsilon^2\sigma_u^2}{2\sigma^2\sigma_v^2}\right\} \int_0^{+\infty} \tfrac{1}{\sqrt{2\pi}\sigma_u\sigma_v} \exp\left\{-\tfrac{\sigma^2(u+\varepsilon\sigma_u^2/\sigma^2)^2}{2\sigma_u^2\sigma_v^2}\right\} du \\
&= \tfrac{2}{\sqrt{2\pi}} \exp\left\{-\tfrac{\varepsilon^2}{2\sigma^2}\left(\tfrac{\sigma^2}{\sigma_v^2} - \tfrac{\sigma_u^2}{\sigma_v^2}\right)\right\} \int_0^{+\infty} \tfrac{1}{\sqrt{2\pi}\sigma_u\sigma_v} \exp\left\{-\tfrac{(u+\varepsilon\sigma_u^2/\sigma^2)^2}{2\frac{\sigma_u^2\sigma_v^2}{\sigma^2}}\right\} du \\
&= \tfrac{2}{\sqrt{2\pi}} \exp\left\{-\tfrac{\varepsilon^2}{2\sigma^2}\left(\tfrac{\sigma_u^2+\sigma_v^2}{\sigma_v^2} - \tfrac{\sigma_u^2}{\sigma_v^2}\right)\right\} \cdot \tfrac{1}{\sigma} \int_0^{+\infty} \tfrac{1}{\sqrt{2\pi}\frac{\sigma_u\sigma_v}{\sigma}} \exp\left\{-\tfrac{(u+\varepsilon\sigma_u^2/\sigma^2)^2}{2\frac{\sigma_u^2\sigma_v^2}{\sigma^2}}\right\} du \\
&= \tfrac{2}{\sqrt{2\pi}\sigma} exp\left\{-\tfrac{\varepsilon^2}{2\sigma^2}\right\} \cdot \left[1 - \Phi\left(\tfrac{\varepsilon\sigma_u^2/\sigma^2}{\sigma_u\sigma_v/\sigma}\right)\right] \\
&= \tfrac{2}{\sigma}\tfrac{1}{\sqrt{2\pi}} exp\left\{-\tfrac{\varepsilon^2}{2\sigma^2}\right\} \cdot \left[1 - \Phi\left(\tfrac{\varepsilon\sigma_u}{\sigma\sigma_v}\right)\right] = \tfrac{2}{\sigma}\phi\left(\tfrac{\varepsilon}{\sigma}\right)\Phi\left(-\tfrac{\varepsilon\lambda}{\sigma}\right)
\end{aligned}
$$

$Note:\ \lambda = \frac{\sigma_u}{\sigma_v}$

Demonstration of the $f(u|\varepsilon)$:

$$
\begin{aligned}
f(u|\varepsilon) &= \tfrac{f(u,\varepsilon)}{f(\varepsilon)} \\
&= \frac{\frac{2}{2\pi\sigma_u\sigma_v}\exp\left\{-\frac{u^2}{2\sigma_u^2}-\frac{(\varepsilon+u)^2}{2\sigma_v^2}\right\}}{\frac{2}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{\varepsilon^2}{2\sigma^2}\right\}\left[1-\Phi\left(\frac{\varepsilon\lambda}{\sigma}\right)\right]} \\
&= \frac{\exp\left\{-\frac{u^2}{2\sigma_u^2}-\frac{(\varepsilon+u)^2}{2\sigma_v^2}+\frac{\varepsilon^2}{2\sigma^2}\right\}}{\sqrt{2\pi}\frac{\sigma_u\sigma_v}{\sigma}\left[1-\Phi\left(\frac{\varepsilon\lambda}{\sigma}\right)\right]} \\
&= \frac{\exp\left\{-\frac{u^2}{2\sigma_u^2}-\frac{\varepsilon^2+u^2+2u\varepsilon}{2\sigma_v^2}+\frac{\varepsilon^2}{2\sigma^2}\right\}}{\sqrt{2\pi}\sigma_*\left[1-\Phi\left(\frac{\varepsilon\sigma_u}{\sigma\sigma_v}\right)\right]} \\
&= \frac{\exp\left\{-\frac{u^2}{2\sigma_u^2}-\frac{u^2}{2\sigma_v^2}-\frac{2u\varepsilon}{2\sigma_v^2}-\frac{\varepsilon^2}{2\sigma_v^2}+\frac{\varepsilon^2}{2\sigma^2}\right\}}{\sqrt{2\pi}\sigma_*\left[1-\Phi\left(\frac{\frac{\varepsilon\sigma_u^2}{\sigma^2}}{\frac{\sigma_u\sigma_v}{\sigma}}\right)\right]} \\
&= \frac{\exp\left\{-\frac{u^2\sigma_v^2+u^2\sigma_u^2}{2\sigma_v^2\sigma_u^2}+\frac{2u\left(-\varepsilon\frac{\sigma_u^2}{\sigma^2}\right)}{2\frac{\sigma_v^2\sigma_u^2}{\sigma^2}}+\frac{-\varepsilon^2\sigma^2+\varepsilon^2\sigma_v^2}{2\sigma_v^2\sigma^2}\right\}}{\sqrt{2\pi}\sigma_*\left[1-\Phi\left(-\frac{\mu_*}{\sigma_*}\right)\right]} \\
&= \frac{\exp\left\{-\frac{u^2\sigma^2}{2\sigma_v^2\sigma_u^2}+\frac{2u\mu_*}{2\sigma_*^2}-\frac{\varepsilon^2(\sigma^2-\sigma_v^2)}{2\sigma_v^2\sigma^2}\right\}}{\sqrt{2\pi}\sigma_*\left[1-\Phi\left(-\frac{\mu_*}{\sigma_*}\right)\right]} \\
&= \frac{\exp\left\{-\frac{u^2}{2\sigma_*^2}+\frac{2u\mu_*}{2\sigma_*^2}-\frac{\varepsilon^2\sigma_u^2}{2\sigma_v^2\sigma^2}\right\}}{\sqrt{2\pi}\sigma_*\left[1-\Phi\left(-\frac{\mu_*}{\sigma_*}\right)\right]} \\
&= \frac{\exp\left\{-\frac{u^2}{2\sigma_*^2}+\frac{2u\mu_*}{2\sigma_*^2}-\frac{\varepsilon^2\sigma_u^4}{2\sigma_u^2\sigma_v^2\sigma^2}\right\}}{\sqrt{2\pi}\sigma_*\left[1-\Phi\left(-\frac{\mu_*}{\sigma_*}\right)\right]} \\
&= \frac{\exp\left\{-\frac{u^2}{2\sigma_*^2}+\frac{2u\mu_*}{2\sigma_*^2}-\frac{\frac{\varepsilon^2\sigma_u^4}{\sigma^4}}{\frac{2\sigma_u^2\sigma_v^2}{\sigma^2}}\right\}}{\sqrt{2\pi}\sigma_*\left[1-\Phi\left(-\frac{\mu_*}{\sigma_*}\right)\right]} \\
&= \frac{\exp\left\{-\frac{u^2}{2\sigma_*^2}+\frac{2u\mu_*}{2\sigma_*^2}-\frac{\mu_*^2}{2\sigma_*^2}\right\}}{\sqrt{2\pi}\sigma_*\left[1-\Phi\left(-\frac{\mu_*}{\sigma_*}\right)\right]} = \frac{\exp\left\{-\frac{(u-\mu_*)^2}{2\sigma_*^2}\right\}}{\sqrt{2\pi}\sigma_*\left[1-\Phi\left(-\frac{\mu_*}{\sigma_*}\right)\right]}
\end{aligned}
\tag{4.2}
$$

$$
Note: \quad \mu_* = -\tfrac{\varepsilon\sigma_u^2}{\sigma^2} \text{ and } \sigma_*^2 = \tfrac{\sigma_u^2\sigma_v^2}{\sigma^2}
$$

Demonstration of the $M(u|\varepsilon)$:

$$\begin{aligned}
&\frac{f(u|\varepsilon)}{du} = \left(\frac{\frac{1}{\sqrt{2\pi}\sigma_*} \exp\left\{ -\frac{(u-\mu_*)^2}{2\sigma_*^2} \right\}}{1-\Phi(-\mu_*/\sigma_*)} \right)' = 0 \Leftrightarrow \\
&\Leftrightarrow \frac{\frac{1}{\sqrt{2\pi}\sigma_*}}{1-\Phi(-\mu_*/\sigma_*)} \frac{2(u-\mu_*)}{2\sigma_*^2} \exp\left\{ -\frac{(u-\mu_*)^2}{2\sigma_*^2} \right\} = 0 \Leftrightarrow \\
&\Leftrightarrow \frac{2(u-\mu_*)}{2\sigma_*^2} = 0 \Leftrightarrow \\
&\Leftrightarrow u = \mu_* = -\frac{\varepsilon_i \sigma_u^2}{\sigma^2}
\end{aligned} \tag{4.3}$$

$$u_i \geq 0 \Leftrightarrow -\frac{\varepsilon_i \sigma_u^2}{\sigma^2} \geq 0 \Leftrightarrow -\varepsilon_i \geq 0 \Leftrightarrow \varepsilon_i \leq 0 \tag{4.4}$$

Demonstration of the $E(\exp(-u)|\varepsilon)$:

$$\begin{aligned}
E(\exp(-u)|\varepsilon) &= \int_0^{+\infty} \exp(-u) f(u|\varepsilon) du \\
&= \int_0^{+\infty} \exp(-u) \frac{\exp\left\{ -\frac{(u-\mu_*)^2}{2\sigma_*^2} \right\}}{\sqrt{2\pi}\sigma_* \left[1-\Phi\left(-\frac{\mu_*}{\sigma_*}\right)\right]} du \\
&= \frac{1}{\left[1-\Phi\left(-\frac{\mu_*}{\sigma_*}\right)\right]} \int_0^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_*} \cdot \exp\left\{ -\frac{(u-\mu_*)^2}{2\sigma_*^2} - u \right\} du \\
&= \frac{1}{\left[1-\Phi\left(-\frac{\mu_*}{\sigma_*}\right)\right]} \int_0^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_*} \cdot \exp\left\{ -\frac{u^2-2u\mu_*+\mu_*^2+2u\sigma_*^2}{2\sigma_*^2} \right\} du \\
&= \frac{1}{\left[1-\Phi\left(-\frac{\mu_*}{\sigma_*}\right)\right]} \int_0^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_*} \cdot \exp\left\{ -\frac{u^2-2u(\mu_*-\sigma_*^2)+(\mu_*-\sigma_*^2)^2-(\mu_*-\sigma_*^2)^2+\mu_*^2}{2\sigma_*^2} \right\} du \\
&= \frac{\exp\left\{ -\frac{-(\mu_*-\sigma_*^2)^2+\mu_*^2}{2\sigma_*^2} \right\}}{\left[1-\Phi\left(-\frac{\mu_*}{\sigma_*}\right)\right]} \int_0^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_*} \cdot \exp\left\{ -\frac{(u-(\mu_*-\sigma_*^2))^2}{2\sigma_*^2} \right\} du \\
&= \frac{\exp\left\{ \frac{\mu_*^2-2\mu_*\sigma_*^2+\sigma_*^4-\mu_*^2}{2\sigma_*^2} \right\}}{\left[1-\Phi\left(-\frac{\mu_*}{\sigma_*}\right)\right]} \left[1 - \Phi\left(-\frac{(\mu_*-\sigma_*^2)}{\sigma_*} \right) \right] \\
&= \frac{\left[1-\Phi\left(\sigma_*-\frac{\mu_*}{\sigma_*}\right)\right]}{\left[1-\Phi\left(-\frac{\mu_*}{\sigma_*}\right)\right]} \cdot \exp\left\{ -\mu_* + \sigma_*^2/2 \right\}
\end{aligned} \tag{4.5}$$

Panel data and assuming Half-Normal

Demonstration of the $f(u, \boldsymbol{\varepsilon})$:

$$
\begin{aligned}
f(u,\varepsilon) &= \tfrac{2}{(2\pi)^{(T+1)/2}\sigma_u\sigma_v^T}\cdot\exp\left\{-\tfrac{(u-\mu_*)^2}{2\sigma_*^2}-\tfrac{\varepsilon'\varepsilon}{2\sigma_v^2}+\tfrac{\mu_*^2}{2\sigma_*^2}\right\}\\
&= \tfrac{2}{(2\pi)^{(T+1)/2}\sigma_u\sigma_v^T}\cdot\exp\left\{\tfrac{-u^2-\mu_*^2+2u\mu_*+\mu_*^2}{2\sigma_*^2}-\tfrac{\varepsilon'\varepsilon}{2\sigma_v^2}\right\}\\
&= \tfrac{2}{(2\pi)^{(T+1)/2}\sigma_u\sigma_v^T}\cdot\exp\left\{-\tfrac{u^2}{2\frac{\sigma_u^2\sigma_v^2}{\sigma_v^2+T\sigma_u^2}}-\tfrac{2u\frac{T\sigma_u^2\bar{\varepsilon}}{\sigma_v^2+T\sigma_u^2}}{2\frac{\sigma_u^2\sigma_v^2}{\sigma_v^2+T\sigma_u^2}}-\tfrac{\varepsilon'\varepsilon}{2\sigma_v^2}\right\}\\
f(u,v) &= \tfrac{2}{(2\pi)^{(T+1)/2}\sigma_u\sigma_v^T}\cdot\exp\left\{-\tfrac{u^2(\sigma_v^2+T\sigma_u^2)}{2\sigma_u^2\sigma_v^2}-\tfrac{2uT\sigma_u^2\bar{\varepsilon}}{2\sigma_u^2\sigma_v^2}-\tfrac{(v_1-u,\ldots,v_T-u)'(v_1-u,\ldots,v_T-u)}{2\sigma_v^2}\right\}\\
&= \ldots\exp\left\{-\tfrac{u^2}{2\sigma_u^2}-\tfrac{Tu^2}{2\sigma_v^2}-\tfrac{2uT\bar{\varepsilon}}{2\sigma_v^2}-\tfrac{\sum_t(v_t-u)^2}{2\sigma_v^2}\right\}\\
&= \ldots\exp\left\{-\tfrac{u^2}{2\sigma_u^2}+\tfrac{-Tu^2-2u\sum_t\varepsilon_t-\sum_t v_t^2-\sum_t u^2+\sum_t 2uv_t}{2\sigma_v^2}\right\}\\
&= \ldots\exp\left\{-\tfrac{u^2}{2\sigma_u^2}+\tfrac{-Tu^2-2u\sum_t(v_t-u)-v'v-Tu^2+2u\sum_t v_t}{2\sigma_v^2}\right\}\\
&= \ldots\exp\left\{-\tfrac{u^2}{2\sigma_u^2}+\tfrac{-2Tu^2+2u\sum_t u-2u\sum_t v_t+2u\sum_t v_t-v'v}{2\sigma_v^2}\right\}\\
&= \ldots\exp\left\{-\tfrac{u^2}{2\sigma_u^2}+\tfrac{-2Tu^2+2Tu^2-v'v}{2\sigma_v^2}\right\}\\
&= \tfrac{2}{(2\pi)^{(T+1)/2}\sigma_u\sigma_v^T}\cdot\exp\left\{-\tfrac{u^2}{2\sigma_u^2}-\tfrac{v'v}{2\sigma_v^2}\right\}
\end{aligned}
\tag{4.6}
$$

Demonstration of the $f(\boldsymbol{\varepsilon})$:

$$
\begin{aligned}
f(\varepsilon) &= \int_0^{+\infty} f(u,\varepsilon)du \\
&= \int_0^{+\infty} \tfrac{2}{(2\pi)^{(T+1)/2}\sigma_u\sigma_v^T} \cdot \exp\left\{-\tfrac{(u-\mu_*)^2}{2\sigma_*^2} - \tfrac{\varepsilon'\varepsilon}{2\sigma_v^2} + \tfrac{\mu_*^2}{2\sigma_*^2}\right\} du \\
&= \tfrac{2}{(2\pi)^{T/2}\sigma_v^{T-1}} \cdot \exp\left\{-\tfrac{\varepsilon'\varepsilon}{2\sigma_v^2} + \tfrac{\mu_*^2}{2\sigma_*^2}\right\} \int_0^{+\infty} \tfrac{1}{\sigma_u\sigma_v(2\pi)^{1/2}} \cdot \exp\left\{-\tfrac{(u-\mu_*)^2}{2\sigma_*^2}\right\} du \\
&= \tfrac{2}{(2\pi)^{T/2}\sigma_v^{T-1}} \cdot \exp\left\{-\tfrac{\varepsilon'\varepsilon}{2\sigma_v^2} + \tfrac{\mu_*^2}{2\sigma_*^2}\right\} \int_0^{+\infty} \tfrac{1}{(2\pi)^{1/2}} \tfrac{1}{\frac{\sigma_u\sigma_v}{(\sigma_v^2+T\sigma_u^2)^{1/2}}} \tfrac{1}{(\sigma_v^2+T\sigma_u^2)^{1/2}} \cdot \exp\left\{-\tfrac{(u-\mu_*)^2}{2\sigma_*^2}\right\} du \\
&= \tfrac{2}{(2\pi)^{T/2}\sigma_v^{T-1}} \cdot \exp\left\{-\tfrac{\varepsilon'\varepsilon}{2\sigma_v^2} + \tfrac{\mu_*^2}{2\sigma_*^2}\right\} \tfrac{1}{(\sigma_v^2+T\sigma_u^2)^{1/2}} \int_0^{+\infty} \tfrac{1}{(2\pi)^{1/2}\sigma_*} \cdot \exp\left\{-\tfrac{(u-\mu_*)^2}{2\sigma_*^2}\right\} du \\
&= \tfrac{2}{(2\pi)^{T/2}\sigma_v^{T-1}(\sigma_v^2+T\sigma_u^2)^{1/2}} \cdot \exp\left\{-\tfrac{\varepsilon'\varepsilon}{2\sigma_v^2} + \tfrac{\mu_*^2}{2\sigma_*^2}\right\} \cdot \left[1 - \Phi\left(-\tfrac{\mu_*}{\sigma_*}\right)\right]
\end{aligned}
\tag{4.7}
$$

Demonstration of the $f(u|\boldsymbol{\varepsilon})$:

$$
\begin{aligned}
f(u|\varepsilon) &= \tfrac{f(u,\varepsilon)}{f(\varepsilon)} \\
&= \frac{\frac{2}{(2\pi)^{(T+1)/2}\sigma_v^T\sigma_u} \cdot \exp\left\{-\frac{(u-\mu_*)^2}{2\sigma_*^2} - \frac{\boldsymbol{\varepsilon'\varepsilon}}{2\sigma_v^2} + \frac{\mu_*^2}{2\sigma_*^2}\right\}}{\frac{2}{(2\pi)^{T/2}\sigma_v^{T-1}(\sigma_v^2+T\sigma_u^2)^{1/2}} \cdot \exp\left\{-\frac{\boldsymbol{\varepsilon'\varepsilon}}{2\sigma_v^2} + \frac{\mu_*^2}{2\sigma_*^2}\right\} \cdot \left[1-\Phi\left(-\frac{\mu_*}{\sigma_*}\right)\right]} \\
&= \tfrac{(\sigma_v^2+T\sigma_u^2)^{1/2}}{\sigma_u\sigma_v\sqrt{2\pi}} \cdot \frac{\exp\left\{-\frac{(u-\mu_*)^2}{2\sigma_*^2}\right\}}{1-\Phi\left(-\frac{\mu_*}{\sigma_*}\right)} = \tfrac{1}{\sigma_*\sqrt{2\pi}} \cdot \frac{\exp\left\{-\frac{(u-\mu_*)^2}{2\sigma_*^2}\right\}}{1-\Phi\left(-\frac{\mu_*}{\sigma_*}\right)}
\end{aligned}
\tag{4.8}
$$

Demonstration of the $M(u|\boldsymbol{\varepsilon})$:

$$
\begin{aligned}
&\tfrac{f(u|\varepsilon)}{du} = \left(\frac{1}{\sqrt{2\pi}\sigma_*} \frac{\exp\left\{-\frac{(u-\mu_*)^2}{2\sigma_*^2}\right\}}{1-\Phi(-\mu_*/\sigma_*)}\right)' = 0 \Leftrightarrow \\
&\Leftrightarrow \tfrac{\frac{1}{\sqrt{2\pi}\sigma_*}}{1-\Phi(-\mu_*/\sigma_*)} \tfrac{2(u-\mu_*)}{2\sigma_*^2} \exp\left\{-\tfrac{(u-\mu_*)^2}{2\sigma_*^2}\right\} = 0 \Leftrightarrow \\
&\Leftrightarrow \tfrac{2(u-\mu_*)}{2\sigma_*^2} = 0 \Leftrightarrow u = \mu_* = -\tfrac{T\bar{\varepsilon}\sigma_u^2}{\sigma_v^2+T\sigma_u^2}
\end{aligned}
\tag{4.9}
$$

Efficiencies Effects Frontier and assuming $N(m_{it}, \sigma^2)$:

Demonstration of the $f(u, \varepsilon)$:

$$
\begin{aligned}
f(u,\varepsilon) &= \frac{\exp\left(-\frac{1}{2}\left[\frac{(\varepsilon+u)^2}{\sigma_v^2}+\frac{(u-z\delta)^2}{\sigma_u^2}\right]\right)}{2\pi\sigma_u\sigma_v\Phi(z\delta/\sigma_u)} = \frac{\exp\left(-\frac{1}{2}\left[\frac{\varepsilon^2+u^2+2u\varepsilon}{\sigma_v^2}+\frac{u^2+(z\delta)^2-2uz\delta}{\sigma_u^2}\right]\right)}{2\pi\sigma_u\sigma_v\Phi(z\delta/\sigma_u)} \\
&= \frac{\exp\left(-\frac{1}{2}\left[\frac{\varepsilon^2}{\sigma_v^2}+\frac{(z\delta)^2}{\sigma_u^2}+\frac{u^2+2u\varepsilon}{\sigma_v^2}+\frac{u^2-2uz\delta}{\sigma_u^2}\right]\right)}{2\pi\sigma_u\sigma_v\Phi(z\delta/\sigma_u)} = \frac{\exp\left(-\frac{1}{2}\left[\frac{\varepsilon^2}{\sigma_v^2}+\frac{(z\delta)^2}{\sigma_u^2}+\frac{u^2(\sigma_u^2+\sigma_u^2)}{\sigma_u^2\sigma_v^2}+\frac{2u\varepsilon}{\sigma_v^2}-\frac{2uz\delta}{\sigma_u^2}\right]\right)}{2\pi\sigma_u\sigma_v\Phi(z\delta/\sigma_u)} \\
&= \frac{\exp\left(-\frac{1}{2}\left[\frac{\varepsilon^2}{\sigma_v^2}+\frac{(z\delta)^2}{\sigma_u^2}+\frac{u^2(\sigma_v^2+\sigma_u^2)}{\sigma_u^2\sigma_v^2}-\frac{2u(z\delta\sigma_v^2-\varepsilon\sigma_u^2)}{\sigma_v^2\sigma_u^2}\right]\right)}{2\pi\sigma_u\sigma_v\Phi(z\delta/\sigma_u)} \\
&= \frac{\exp\left(-\frac{1}{2}\left[\frac{\varepsilon^2}{\sigma_v^2}+\frac{(z\delta)^2}{\sigma_u^2}+\frac{u^2}{\frac{\sigma_u^2\sigma_v^2}{\sigma_v^2+\sigma_u^2}}-\frac{2u\frac{(z\delta\sigma_v^2-\varepsilon\sigma_u^2)}{\sigma_v^2+\sigma_u^2}}{\frac{\sigma_v^2\sigma_u^2}{\sigma_v^2+\sigma_u^2}}\right]\right)}{2\pi\sigma_u\sigma_v\Phi(z\delta/\sigma_u)} \\
&= \frac{\exp\left(-\frac{1}{2}\left[\frac{\varepsilon^2}{\sigma_v^2}+\frac{(z\delta)^2}{\sigma_u^2}+\frac{u^2}{\sigma_*^2}-\frac{2u\mu_*}{\sigma_*^2}\right]\right)}{2\pi\sigma_u\sigma_v\Phi(z\delta/\sigma_u)} = \frac{\exp\left(-\frac{1}{2}\left[\frac{\varepsilon^2}{\sigma_v^2}+\frac{(z\delta)^2}{\sigma_u^2}+\frac{(u-\mu_*)^2}{\sigma_*^2}-\frac{\mu_*^2}{\sigma_*^2}\right]\right)}{2\pi\sigma_u\sigma_v\Phi(z\delta/\sigma_u)}
\end{aligned}
\tag{4.10}
$$

Demonstration of the $f(\varepsilon)$:

$$
\begin{aligned}
f(\varepsilon) &= \int_0^{+\infty} f(u,\varepsilon)du \\
&= \int_0^{+\infty} \frac{\exp\left(-\frac{1}{2}\left[\frac{\varepsilon^2}{\sigma_v^2}+\frac{(z\delta)^2}{\sigma_u^2}+\frac{(u-\mu_*)^2}{\sigma_*^2}-\frac{\mu_*^2}{\sigma_*^2}\right]\right)}{2\pi\sigma_u\sigma_v\Phi(z\delta/\sigma_u)}du \\
&= \frac{\exp\left(-\frac{1}{2}\left[\frac{\varepsilon^2}{\sigma_v^2}+\frac{(z\delta)^2}{\sigma_u^2}-\frac{\mu_*^2}{\sigma_*^2}\right]\right)}{\sqrt{2\pi}\sigma_u\sigma_v\Phi(z\delta/\sigma_u)}\int_0^{+\infty}\frac{\exp\left(-\frac{1}{2}\left[\frac{(u-\mu_*)^2}{\sigma_*^2}\right]\right)}{\sqrt{2\pi}}du \\
&= \frac{\exp\left(-\frac{1}{2}\left[\frac{\varepsilon^2}{\sigma_v^2}+\frac{(z\delta)^2}{\sigma_u^2}-\frac{\mu_*^2}{\sigma_*^2}\right]\right)\cdot\sigma_*}{\sqrt{2\pi}\sigma_u\sigma_v\Phi(z\delta/\sigma_u)}\int_0^{+\infty}\frac{\exp\left(-\frac{1}{2}\left[\frac{(u-\mu_*)^2}{\sigma_*^2}\right]\right)}{\sqrt{2\pi}\cdot\sigma_*}du \\
&= \frac{\exp\left(-\frac{1}{2}\left[\frac{\varepsilon^2}{\sigma_v^2}+\frac{(z\delta)^2}{\sigma_u^2}-\frac{\mu_*^2}{\sigma_*^2}\right]\right)\cdot(\sigma_u\sigma_v)}{\sqrt{2\pi}\Phi(z\delta/\sigma)\cdot\sigma_u\sigma_v\sqrt{\sigma_u^2+\sigma_v^2}}\int_0^{+\infty}\frac{\exp\left(-\frac{1}{2}\left[\frac{(u-\mu_*)^2}{\sigma_*^2}\right]\right)}{\sqrt{2\pi}\cdot\sigma_*}du \\
&= \frac{\exp\left(-\frac{1}{2}\left[\frac{\varepsilon^2}{\sigma_v^2}+\frac{(z\delta)^2}{\sigma_u^2}-\frac{\mu_*^2}{\sigma_*^2}\right]\right)}{\sqrt{2\pi}\Phi(z\delta/\sigma_u)(\sigma_u^2+\sigma_v^2)^{1/2}}\left[1-\Phi\left(-\frac{\mu_*}{\sigma_*}\right)\right] \\
&= \frac{\exp\left(-\frac{1}{2}\left[\frac{\varepsilon^2}{\sigma_v^2}+\frac{(z\delta)^2}{\sigma_u^2}-\frac{\mu_*^2}{\sigma_*^2}\right]\right)\Phi\left(\frac{\mu_*}{\sigma_*}\right)}{\sqrt{2\pi(\sigma_u^2+\sigma_v^2)}\Phi\left(\frac{z\delta}{\sigma_u}\right)}
\end{aligned}
\tag{4.11}
$$

Demonstration of the $f(u|\varepsilon)$:

$$\begin{aligned} f(u|\varepsilon) &= \frac{\frac{\exp\left(-\frac{1}{2}\left[\frac{\varepsilon^2}{\sigma_v^2}+\frac{(z\delta)^2}{\sigma_u^2}+\frac{(u-\mu_*)^2}{\sigma_*^2}-\frac{\mu_*^2}{\sigma_*^2}\right]\right)}{2\pi\sigma_u\sigma_v\Phi(z\delta/\sigma_u)}}{\frac{\exp\left(-\frac{1}{2}\left[\frac{\varepsilon^2}{\sigma_v^2}+\frac{(z\delta)^2}{\sigma_u^2}-\frac{\mu_*^2}{\sigma_*^2}\right]\right)}{\sqrt{2\pi(\sigma_u^2+\sigma_v^2)}}\frac{\Phi\left(\frac{\mu_*}{\sigma_*}\right)}{\Phi\left(\frac{z\delta}{\sigma_u}\right)}} \\ &= \frac{\frac{\exp\left(-\frac{1}{2}\left[\frac{(u-\mu_*)^2}{\sigma_*^2}\right]\right)}{\sqrt{2\pi}\sigma_u\sigma_v}}{\frac{\Phi\left(\frac{\mu_*}{\sigma_*}\right)}{\sqrt{\sigma_v^2+\sigma_u^2}}} = \frac{\exp\left(-\frac{1}{2}\left[\frac{(u-\mu_*)^2}{\sigma_*^2}\right]\right)}{\sqrt{2\pi}\sigma_*\Phi\left(\frac{\mu_*}{\sigma_*}\right)} \end{aligned} \tag{4.12}$$

Demonstration of the $E(\exp(-u)|\varepsilon)$:

$$\begin{aligned} E(\exp(-u)|\varepsilon) &= \int_0^{+\infty}\exp(-u)f(u|\varepsilon)du \\ &= \int_0^{+\infty}\exp(-u)\frac{\exp\left(-\frac{1}{2}\left[\frac{(u-\mu_*)^2}{\sigma_*^2}\right]\right)}{\sqrt{2\pi}\sigma_*\Phi\left(\frac{\mu_*}{\sigma_*}\right)}du \\ &= \frac{1}{\Phi\left(\frac{\mu_*}{\sigma_*}\right)}\int_0^{+\infty}\frac{1}{\sqrt{2\pi}\sigma_*}\exp\left(-\frac{1}{2}\left[\frac{u^2+\mu_*^2-2u\mu_*}{\sigma_*^2}\right]-u\right)du \\ &= \frac{1}{\Phi\left(\frac{\mu_*}{\sigma_*}\right)}\int_0^{+\infty}\frac{1}{\sqrt{2\pi}\sigma_*}\exp\left(-\frac{u^2-2u\mu_*+2u\sigma_*^2+\mu_*^2}{2\sigma_*^2}\right)du \\ &= \frac{1}{\Phi\left(\frac{\mu_*}{\sigma_*}\right)}\int_0^{+\infty}\frac{1}{\sqrt{2\pi}\sigma_*}\exp\left(-\frac{u^2-2u(\mu_*-\sigma_*^2)+(\mu_*-\sigma_*^2)^2-(\mu_*-\sigma_*^2)^2+\mu_*^2}{2\sigma_*^2}\right)du \\ &= \frac{1}{\Phi\left(\frac{\mu_*}{\sigma_*}\right)}\int_0^{+\infty}\frac{1}{\sqrt{2\pi}\sigma_*}\exp\left(-\frac{(u-(\mu_*-\sigma_*^2))^2-(\mu_*-\sigma_*^2)^2+\mu_*^2}{2\sigma_*^2}\right)du \\ &= \frac{\exp\left(-\frac{\mu_*^2-(\mu_*-\sigma_*^2)^2}{2\sigma_*^2}\right)}{\Phi\left(\frac{\mu_*}{\sigma_*}\right)}\int_0^{+\infty}\frac{1}{\sqrt{2\pi}\sigma_*}\exp\left(-\frac{1}{2}\left[\frac{u-(\mu_*-\sigma_*^2)}{\sigma_*}\right]^2\right)du \\ &= \frac{\exp\left(-\frac{\mu_*^2-(\mu_*^2+\sigma_*^4-2\mu_*\sigma_*^2)}{2\sigma_*^2}\right)}{\Phi\left(\frac{\mu_*}{\sigma_*}\right)}\left[1-\Phi\left(-\frac{\mu_*-\sigma_*^2}{\sigma_*}\right)\right] \\ &= \frac{\exp\left(\frac{\sigma_*^4-2\mu_*\sigma_*^2)}{2\sigma_*^2}\right)}{\Phi\left(\frac{\mu_*}{\sigma_*}\right)}\left[\Phi\left(\frac{\mu_*-\sigma_*^2}{\sigma_*}\right)\right] \\ &= \exp\left(\frac{\sigma_*^2}{2}-\mu_*\right)\frac{\Phi\left(\frac{\mu_*}{\sigma_*}-\sigma_*\right)}{\Phi\left(\frac{\mu_*}{\sigma_*}\right)} \end{aligned} \tag{4.13}$$