



CEDER

Catch, Effort and
Discard Estimation in Real-time

Tryggvi Hjörvar



**Faculty of Industrial Engineering, Mechanical
Engineering and Computer Science
University of Iceland
2012**

CEDER
Catch, Effort and
Discard Estimation in Real-time

Tryggvi Hjörvar

60 ECTS thesis submitted in partial fulfillment of a
Magister Scientiarum degree in Industrial Engineering

Advisors
Birgir Hrafnkelsson
Guðrún Pétursdóttir
Ólafur Pétur Pálsson

Faculty Representative
Steinn Guðmundsson

Faculty of Industrial Engineering, Mechanical Engineering
and Computer Science
School of Engineering and Natural Sciences
University of Iceland
Reykjavik, October 2012

CEDER - Catch, Effort and Discard Estimation in Real-time
CEDER

60 ECTS thesis submitted in partial fulfillment of a *Magister Scientiarum* degree in
Industrial Engineering

Copyright © 2012 Tryggvi Hjörvar
All rights reserved

Faculty of Industrial Engineering, Mechanical Engineering and Computer Science
School of Engineering and Natural Sciences
University of Iceland
Hjarðarhaga 2-6
107, Reykjavík
Iceland

Telephone: 525 4000

Bibliographic information:

Tryggvi Hjörvar, 2012, *CEDER - Catch, Effort and Discard Estimation in Real-time*,
Master's thesis, Faculty of Industrial Engineering, Mechanical Engineering and Computer
Science, University of Iceland, pp. 112.

Printing: Háskólaprent, Fálkagata 2, 107 Reykjavík
Reykjavík, Iceland, October 2012

Abstract

The CEDER project aims to provide more accurate and timely information on catches, effort, landings, discards, quota and TAC (Total Allowed Catch) uptake and to assess the benefits of this information for fisheries management.

This thesis is part of a CEDER work-package dedicated to the development of classification algorithms and statistical models to identify and categorise vessel activity through analysis of positional data, estimate the fishing effort and predict the vessel catch. The data used was collected from the Atlantic Pelagic Redfish fishery on the Reykjanes-ridge.

A linear discriminant classifier on vessel speed is capable of identifying high-speed cruising activity, but no measure or alternative classifier is found to adequately differentiate between fishing and non-fishing activity at lower speeds based on the data provided. A case is made for using higher resolution data with a maximum of 15 minutes between position reports.

An estimate of fishing effort is derived from the classifier results and used in a multivariate linear regression model to predict the expected catch for individual vessels and trips within any given year. The model is found to return reasonable predictions for known years, but has a higher error rate for unfitted years.

Using a simpler effort measure of *days at sea* with the same model is shown to give at least as good or better results than the more complicated effort estimate from positional data.

A pilot-system is presented with features such as current fleet activity, total TAC uptake, alerts for suspect activity, and area fishing load.

Útdráttur

CEDER verkefnið miðar að því að þróa aðferðir til nákvæmari og hraðari miðlunar upplýsinga um afla, sókn, landanir, brottkast, kvóta og nýtingu hans og meta kosti þeirra við stjórnun og eftirlit fiskveiða.

Þessi ritgerð er hluti af CEDER vinnupakka um þróun flokkunaralgríma og tölfræðilíkana til að greina og flokka aðgerðir skipa út frá staðsetningargögnum, meta sókn og spá fyrir um afla. Gögnunum sem notuð eru var safnað á karfaveiðum á Reykjaneshrygg.

Flokkunaralgrím sem byggir á línulegum aðskilnaði eftir hraða skips getur borið kennsl á stím á mikilli ferð, en engin mælistærð eða aðrir flokkarar fundust sem gátu aðskilið fullnægjandi milli fiskveiði og annarra aðgerða á minni ferð, byggt á gögnunum sem lögð eru til grundvallar. Færð eru rök fyrir að nota gögn með hærri upplausn, eða í mesta lagi 15 mínútna milli staðsetninga.

Mat á sókn er byggt á niðurstöðum flokkunaralgrímsins og notað í margvítt línulegt aðhvarfslíkan til að spá fyrir um væntan afla fyrir ákveðin skip og veiðiferðir innan hvers árs. Líkanið skilar ásættanlegri spá fyrir þekkt ár, en er ónákvæmara fyrir ómátuð ár.

Sýnt er fram á að notkun einfaldari mælinga á *dögum til sjós* með sama líkani skilar að minnsta kosti jafn góðri eða betri spá en flóknara mat á sókn út frá staðsetningargögnum.

Sýnd er frumgerð að hugbúnaðarkerfi með möguleikum á að sýna aðgerðir flotans, nýtingu heildarkvóta, svæðisbundið veiðiálag og gefa viðvaranir um grunsamlega hegðun.

Keywords

Fishing, fisheries management, fisheries monitoring, remote monitoring, redfish, fishing effort, catch, vessel monitoring system, classification, multivariate regression models

Efnisorð

Fiskveiðar, fiskveiðistjórnun, fiskveiðieftirlit, fjareftirlit, karfi, sókn, afli, flokkun, margvíoð
aðhvarfslíkön

Quis custodiet ipsos custodes?
Who will watch the watchmen?
– Juvenal, "Satires"

Preface

This thesis is written as part of an M.Sc. degree program in industrial engineering at the University of Iceland with professor Páll Jensson, lic. techn., professor Ólafur Pétur Pálsson, Ph.D. and assistant professor Birgir Hrafnkelsson, Ph.D. acting as thesis advisors, and docent Dr. Guðrún Pétursdóttir as project leader.

The study was undertaken in collaboration with the *University of Iceland Institute for Sustainability Studies* (Stofnun Sæmundar fróða) and the *Icelandic Directorate of Fisheries* (Fiskistofa). It forms an integrated part of the CEDER¹ project, funded by the European Union's Sixth Framework Programme: Policy-Oriented Research.

Scripts, source code and printouts referenced in the thesis are available by request to the author.

Disclaimer

This study has been carried out with financial support from the Commission of the European Communities, specific RTD programme “Specific Support to Policies, sustainable management of Europe’s natural resources”. It does not necessarily reflect its views and in no way anticipates the Commission’s future policy in this area.

¹ Catch, Effort and Discard Monitoring in Real Time, <https://ceder.jrc.ec.europa.eu/>. [Last retrieved Sept 2012]

Contents

List of figures.....	xiv
List of tables.....	xvi
Glossary.....	xvii
Acknowledgements.....	xix
Chapter 1 - Introduction.....	1
1.1 Fishing for sustainability.....	1
1.2 The Tragedy of the Commons.....	1
1.3 Quota systems.....	1
1.4 The role of monitoring.....	2
1.5 CEDER.....	2
1.6 The Atlantic Pelagic (Oceanic) Redfish.....	3
1.7 NEAFC and quotas.....	4
1.8 The Pirates.....	5
1.9 Research Objectives.....	7
1.10 Structure of the thesis.....	7
Chapter 2 - Literature review.....	9
2.1 Prior work.....	9
2.2 Work of other CEDER partners.....	9
Chapter 3 - The data.....	11
3.1 Collection.....	11
3.1.1 VMS.....	11
3.1.2 Catch logbooks.....	12
3.1.3 Official catch landing reports.....	12
3.2 Preprocessing.....	12
3.2.1 Erroneous position reports.....	13
3.3 Identifying trips.....	13
3.4 Overview of the data.....	14
3.5 Visualisation of the data.....	15
3.6 Vessels.....	20
3.7 Fishing behaviour and features.....	21
3.8 Summary.....	22
Chapter 4 - Methodology and results.....	23
4.1 General approach.....	23
4.2 Classification of vessel activities.....	26
4.2.1 Supervised vs. unsupervised classification.....	26
4.2.2 Finding the “actual” activity from the catch logbook.....	26
Midpoint.....	27
4.2.3 Expert classifier based on speed.....	28
4.2.4 Calibrating the classifier.....	32
4.2.5 Adding more predictor variables.....	34
Variable transformation with Principal Components.....	35
Variable selection.....	37

4.2.6 Alternative classifiers.....	38
Fisher's Linear Discriminant.....	38
Naïve Bayes classifier.....	39
k-Nearest Neighbour classifier.....	39
CART classifier.....	40
Multilayer Perceptron.....	42
4.2.7 Cluster analysis.....	42
k-Means clustering.....	42
4.2.8 Summary.....	43
4.3 Estimating effort from classification.....	45
4.3.1 Comparison to reported effort.....	45
4.4 Estimating catch from effort.....	47
4.4.1 General approach and terms.....	47
4.4.2 The true effort, errors and bias.....	48
4.4.3 Variance of catch model.....	52
4.4.4 Least squares formulation in matrix form.....	54
4.4.5 Maximum likelihood estimation of α and σ^2_{ϵ}	55
4.4.6 Unequal variances in estimated effort U_{jk} and reported effort V_{jk}	58
4.4.7 Standardised residuals and treatment of outliers.....	60
4.4.8 Student's-t distribution.....	64
4.4.9 Curvilinear model.....	71
4.4.10 Prediction intervals on new observations with a fitted year-effect.....	76
4.4.11 Prediction intervals on new observations with an unknown year-effect.....	78
4.4.12 Usage examples for the final model.....	79
4.5 Model performance.....	81
4.5.1 Validation data.....	82
4.5.2 Comparison to Days at Sea.....	83
4.6 High-resolution GPS-data.....	86
Chapter 5 - Discussion.....	91
Chapter 6 - Conclusions and future work.....	93
Bibliography.....	95
Appendix A - Prototype system CARFI.....	A-1

List of figures

Figure 1: The Atlantic Pelagic Redfish.....	3
Figure 2: The Atlantic redfish fishing area.....	4
Figure 3: Vessels lining up on the Reykjanes-ridge.....	6
Figure 4: Trip identification algorithm results.....	14
Figure 5: Vessel VMS track.....	15
Figure 6: Vessel VMS track with satellite image	16
Figure 7: Vessel track and calculated speed.....	17
Figure 8: Vessel track, speed and leg classification	17
Figure 9: Number of legs by speed	18
Figure 10a: Catch size with 5000 kg bins.....	18
Figure 10b: Catch size with 500 kg bins.....	18
Figure 11: Reported effort.....	19
Figure 12: Reported effort vs. reported catch for each vessel trip.....	19
Figure 13: Russian trawlers on the Reykjanes-ridge.....	20
Figure 14: Scatterplot matrix for vessel attributes.....	21
Figure 15: Reported catch vs. reported effort 2001-2005.....	24
Figure 16: Strategy for building the activity classifier and catch estimation model.....	25
Figure 17: Step-by-step use of the algorithms to predict effort, activity and catch from VMS-data.....	25
Figure 18: Vessel legs and catch points.....	27
Figure 19: Number of legs by speed.....	28
Figure 20: Number of legs by speed and predicted activity.....	29
Figure 21: Simple linear discriminant classifier with 5 knots decision boundary.....	29
Figure 22: Number of legs by speed and actual activity.....	30
Figure 23: Track with predicted activity for vessel 2, year 2003.....	32
Figure 24: Speed by actual activity class.....	33
Figure 25: Scree plot of eigenvalues from the correlation coefficient matrix.....	37
Figure 26: Partial CART classification tree for all variables.....	41
Figure 27: CART classification tree for speed.....	41
Figure 28: Trip effort proportional errors.....	46
Figure 29: Trip reported effort vs. reported catch.....	46
Figure 30: log of the estimated effort vs. log of the reported effort for each trip.....	49
Figure 31: Example of a datapoint with uncertainty.....	49
Figure 32: Distance d_{jk} from datapoints to the line	50
Figure 33: Corrected log of the estimated effort vs. log of the reported effort, with bias estimator.....	51
Figure 34: Distance from datapoints to the corrected line	52
Figure 35: Matrix construction for the linear regression model.....	55
Figure 36: Estimated vessel parameters, μ_k	57
Figure 37: Estimated year parameters β_t	58
Figure 38: Normal probability plot of $\varepsilon_{jk} - e_{jk}$	61
Figure 39: Standardised residuals of $\varepsilon_{jk} - e_{jk}$	61
Figure 40: Estimated vessel parameters, μ_k without outliers.....	63
Figure 41: Estimated year parameters, β_t without outliers.....	63
Figure 42: Log-t plot of standardised residuals of y	64
Figure 43: Estimated vessel parameters μ_k without outliers using Student's-t distribution.....	67
Figure 44: Estimated year parameters β_t without outliers using Student's-t distribution.....	68
Figure 45: log-t plot of standardised residuals of y from the t-distribution.....	69
Figure 46: Empirical CDF vs. theoretical CDF of standardised residuals y from the t-distribution.....	70
Figure 47: Sorted standardised residuals of y vs. probability number for the t-distribution.....	70
Figure 48: Standardised residuals vs. log of estimated effort.....	71

Figure 49: Estimated vessel parameters, μ_k without outliers, curvilinear model.....	73
Figure 50: Estimated year parameters, β_t without outliers, curvilinear model.....	74
Figure 51: Sorted standardised residuals of y vs. probability number for the t-distribution.....	75
Figure 52: Standardised residuals vs. log of estimated effort.....	75
Figure 53: Reported catch and modelled catch with 95% prediction intervals.....	77
Figure 54: Estimation of year-effect parameters β_t	78
Figure 55: Reported catch and modelled catch with 95% prediction intervals and validation data..	79
Figure 56: Model results for vessel 5 in year 2003 – Reported catch vs. Estimated effort.....	80
Figure 57: Model results for vessel 5, all years – Reported catch vs. Estimated catch.....	80
Figure 58: Estimated vessel parameters μ_k without outliers using Student's-t distribution and a curve parameter using days at sea as effort estimate.....	84
Figure 59: Estimated year parameters β_t without outliers using Student's-t distribution and a curve parameter using days at sea as effort estimate.....	84
Figure 60: Sorted standardised residuals of y vs. probability number for the t-distribution, using days at sea as effort measure.....	85
Figure 61: Standardised residuals vs. log of estimated effort using days at sea as effort measure....	85
Figure 62: Example of VMS and high resolution GPS tracks.....	86
Figure 63: Example of missed features.....	87
Figure 64: Estimated effort at different resolutions.....	88
Figure 65: Leg speed by resolution.....	89

List of Tables

Table 1: Quota allocation and catch.....	5
Table 2: Catch by NEAFC members 2005.....	5
Table 3: Dataset overview, all data.....	14
Table 4: Dataset overview, valid data.....	15
Table 5: Classification results – Simple linear discriminant at 5,0 knots.....	31
Table 6: Classification results – Simple linear discriminant at 4,4 knots.....	34
Table 7: Principal components vectors from correlation coefficients.....	36
Table 8: Predictor variable subsets with the ten lowest error rates using Fisher's LDA.....	38
Table 9: Classification results – Fisher's Linear Discriminant.....	39
Table 10: Comparison of classifier performance.....	43
Table 11: Estimated vessel parameters μ_k	57
Table 12: Estimated year parameters β_t	58
Table 13: Estimated vessel parameters μ_k without outliers.....	62
Table 14: Estimated year parameters β_t without outliers.....	63
Table 15: Estimated vessel parameters μ_k without outliers using Student's-t distribution.....	67
Table 16: Estimated year parameters β_t without outliers using Student's-t distribution.....	68
Table 17: Estimated vessel parameters μ_k without outliers using Student's-t distribution and a curve parameter.....	73
Table 18: Estimated year parameters β_t without outliers using Student's-t distribution and a curve parameter.....	74
Table 19: Estimated vessel parameters μ_k without outliers using Student's-t distribution and a curve parameter using days at sea as effort estimate.....	83
Table 20: Estimated year parameters β_t without outliers using Student's-t distribution and a curve parameter using days at sea as effort estimate.....	84
Table 21: Estimated effort comparison between VMS and high-resolution GPS data.....	87

Glossary

An alphabetical reference of terms and abbreviations commonly used throughout this thesis.

CARFI	CEDER Atlantic Redfish Fisheries Information system, a prototype system built using the algorithms and models developed in this thesis.
CART	Classification and Regression Trees, a type of classification algorithm.
Catch	(is. afli) Usually reported catch, the total catch of the target species (redfish) during one haul or trip of a vessel.
CEDER	Catch, Effort and Discard Estimation in Real-time, the EU project under which this study was carried out.
CPUE / Catch Per Unit Effort	(is. afli á sóknareiningu) Calculated from the catch and effort, this measure gives an indication of the fishing power of the vessel and gear in question, and the stock catchability. Usually given as tonnes per hour trawling, but may also be given as tonnes per mile trawled. In this work the former is always used.
DF	Degrees of Freedom.
EEZ	Exclusive Economic Zone, a sea zone within which a state has exclusive rights to exploitation of natural resources.
Effort	(is. sókn) Measure of the work required to catch fish, usually the time spent fishing is used.
Estimated catch	The catch of a vessel estimated by the catch model, usually applies for one trip.
Estimated effort	The effort of a vessel estimated by the effort estimation algorithm, usually applies for one trip.
Gap	One or more missing VMS reports from a track.
GPS	Global Positioning System.
GRT	Gross Register Tonnes, a measure of a vessel's size.
Haul / Trawl	One fishing action, e.g. one trawl of a vessel from the time the gear is deployed until it is retrieved.
IDF	Icelandic Directorate of Fisheries.
ITQ	Individual Transferrable Quota.
ISS	University of Iceland Institute for Sustainability Studies.
IUU	Illegal, unregulated and unreported. Refers to fishing activity or vessels that are known to participate in illegal fishing in NEAFC-controlled areas.
JRC	Joint Research Centre, The European Union's scientific and technical research laboratory and an integral part of the European Commission.
k-NN	k-Nearest Neighbour, a type of classification algorithm.
Knots	Speed of a vessel, nautical miles per hour.
KW	Kilowatts, a measure of a vessel's engine power in kilowatts.
Landed catch	Sum of vessel catches for one trip, as per the official landing reports of the Icelandic Directorate of Fisheries.

Latitude	The angular distance north or south from the equator.
LDA	Linear Discriminant Analysis, a type of classification algorithm.
Leg	The vessel's movement between one VMS-position to the next, represented by a straight line.
LMSE	Log Mean Squared Error, a measure of model performance.
LOA	Length Over All, a measure of a vessel's size.
Longitude	The angular distance east or west of Greenwich, England.
MCS	Monitoring, control and surveillance, one of the three pillars of fisheries management .
MLP	Multilayer Perceptron, a type of classification algorithm.
MLR	Multivariate Linear Regression.
MSE	Mean Squared Error, a measure of model performance.
NEAFC	North East Atlantic Fisheries Commission, an international governing body that manages fisheries in the North East Atlantic.
Nml	Nautical miles, one nautical mile equals 1852 m.
PCA	Principal Components Analysis, a method of variable transformation.
Reported catch	A measurement given in a vessel's catch logbook, indicating the amount of fish in kg caught during one haul.
Reported effort	A measurement given in a vessel's catch logbook, indicating the amount of time the gear was deployed during one haul.
RSS	Residual Sum of Squares.
TAC	Total Allowed Catch.
Track	The vessel's movements during one trip, represented by VMS-points from start to finish.
Trip	The period between landings, from when the vessel leaves port until it makes a declared landing.
VMS	Vessel Monitoring System, a system onboard vessels that sends positional data, as well as catch reports while at sea.
WEKA	Waikato Environment for Knowledge Analysis, a software environment for machine learning applications.

Acknowledgements

I would like to express my sincere gratitude to all the people who have contributed to the completion of this thesis in one way or another.

To my thesis advisors; prof. Páll Jensson, lic. techn. for his guidance in the early phases of the project, to asst. prof. Birgir Hrafnkelsson, Ph.D. without whose time and insight the statistical part of this work would not have been possible, prof. Ólafur Pétur Pálsson, Ph.D. for stepping in at a short notice, and to Dr. Guðrún Pétursdóttir for her constant optimism, motivation and patience.

To all our CEDER project partners and contributors for their helpful comments, information and discussions, especially Ulrich Kröner (JRC), Erel Rosenberg (Correlation Systems), Amos Barkai (Olrac) and John Cotter (CEFAS).

To friends and family who read drafts at various stages and provided corrections, questions, and support.

Chapter 1 - Introduction

1.1 Fishing for sustainability

Fishing has been one of the mainstays of man's food supply throughout history. The waters and oceans have provided a seemingly endless supply of food and man has naturally taken advantage of it, confident that the incredible vastness of this blue planet was impervious to anything he might do to it.

Now mankind is becoming a victim of its own success. Through advances in technology and ever increasing population size, human civilisation is pushing limited natural resources towards the edge. Fleets of powerful fishing vessels are taking fish out of the sea faster than nature can replenish them, with the result that fish stocks plummet and many species are now in serious trouble.

The idea of sustainability was born as a solution to this problem; to exploit natural resources in a balance with nature, seeking wherever possible to utilise renewable resources at a rate that gives them time to replenish. In this way the resource can be sustained indefinitely, preserving both the livelihoods of humans and the biological diversity of the planet.

1.2 The Tragedy of the Commons

The problem of overfishing is in fact a manifestation of "The Tragedy of the Commons" (Hardin, 1968), the tendency for commonly held resources to be overexploited and wasted, because individuals benefit directly from using the resource while the costs are shared by all.

In ocean fisheries this appears as excessive fishing fleets and effort, overexploited (small) fish stocks, poor profitability, low personal incomes, little or no contribution to GDP, a threat to biological sustainability and a threat to economic (habitation) sustainability (Árnason, 2006).

1.3 Quota systems

Attaining a state of sustainable fishing requires some means of control over the total catch from each fish stock. This can be accomplished in many different ways, such as through biological management (area closures, seasonal closures, total allowed catch restrictions, gear restrictions), direct economic restrictions (limited fishing effort, limited vessel size or power, investment restrictions), taxation on landings (an attractive option, but nowhere used as a fisheries management method) and property rights (licenses, sole ownership, turfs, individual quotas, communal property rights) (Árnason, 2006).

The tool of choice for many countries, including Iceland, has been to introduce quasi-property rights in the form of individual transferable quotas (ITQs). Under this system vessels are granted a fishing license and tonnage quota from a specific fish stock, but their owners are free to rent or sell their quota to others, having met certain prerequisites.

1.4 The role of monitoring

Monitoring, control and surveillance (MCS) make up one of the three pillars of any proper fisheries management regime, the others being the fisheries management system (providing a regulatory framework) and the fisheries judicial system (processing violations and issuing sanctions) (Árnason, 2006). All three have to work in conjunction to achieve effective and efficient fisheries management.

The role of the MCS is twofold; firstly to gather data for use in management decisions (monitoring and surveillance), and secondly to enforce the rules of the fisheries management system (monitoring and control). However, the costs of maintaining an acceptable level of MCS has been found to be significant, between 3%-28% of the gross value of landings (Árnason, 2006). During the first eleven months of the year 2009, direct costs of in-field monitoring were estimated at 237 million ISK (Alþingi, 2010). It is obviously highly desirable to optimise the MCS to reduce costs, without sacrificing its effectiveness.

Icelandic economy relies heavily on fisheries with exports of fish and fish products representing some 57% of total exports in 2005 (Hagstofa Íslands, 2006). This has led Iceland to seek to become a leader in fisheries management, monitoring, control and surveillance.

Iceland has a unique advantage in this area. By law, every kilogramme of landed fish is to be weighed by officials at one of 62 designated landing ports around the coast. In addition, the Icelandic Coast Guard conducts on-board inspections at sea and aircraft surveillance, as well as monitoring vessel activity through the mandatory Vessel Monitoring System (VMS). The Iceland Directorate of Fisheries sends inspectors on board fishing vessels and controls each vessel's catch logbook, which is completely electronic since October 2008. Finally, the Iceland Marine Research Institute runs several research and monitoring programmes each season. Few if any nations have more comprehensive MCS activities and complete fisheries data than Iceland.

1.5 CEDER

Not all nations go to such lengths to monitor their fisheries as the Icelanders, but recognising the need for improvement in this field, the EU has shown interest in technological advances to support MCS. This is at the heart of the CEDER project. The acronym stands for "**C**atch, **E**ffort and **D**iscard **E**stimation in **R**eal-time" and has as its primary goal (CEDER consortium, 2006) to:

"provide more accurate and more timely information on catches, effort, landings, discards and quota and TAC [Total Allowed Catch] uptake and to assess the benefits of this information for fisheries management".

The project proposed to do this through utilisation of the widespread deployment of modern data gathering technologies such as the Vessel Monitoring System (VMS), electronic logbooks and landing reports, to improve the accuracy of such data available to stakeholders (e.g vessel owners, fish processing plants, authorities and scientists), increase its spatial precision and reduce the delivery time (lag).

During the project lifetime, three prototype systems for such data gathering, analysis and

dissemination were developed and tested, one by Correlation Systems Israel, one by SiriusIT Greenland, and one by the University of Iceland Institute for Sustainability Studies (ISS).

This thesis deals with one specific part of the prototype system developed by ISS, namely the task of analysing VMS, logbook and landing report data to construct a prediction model for vessel effort, catch, discard and landings.

The prototype system “CARFI” that utilises these algorithms for fisheries monitoring is presented in appendix A.

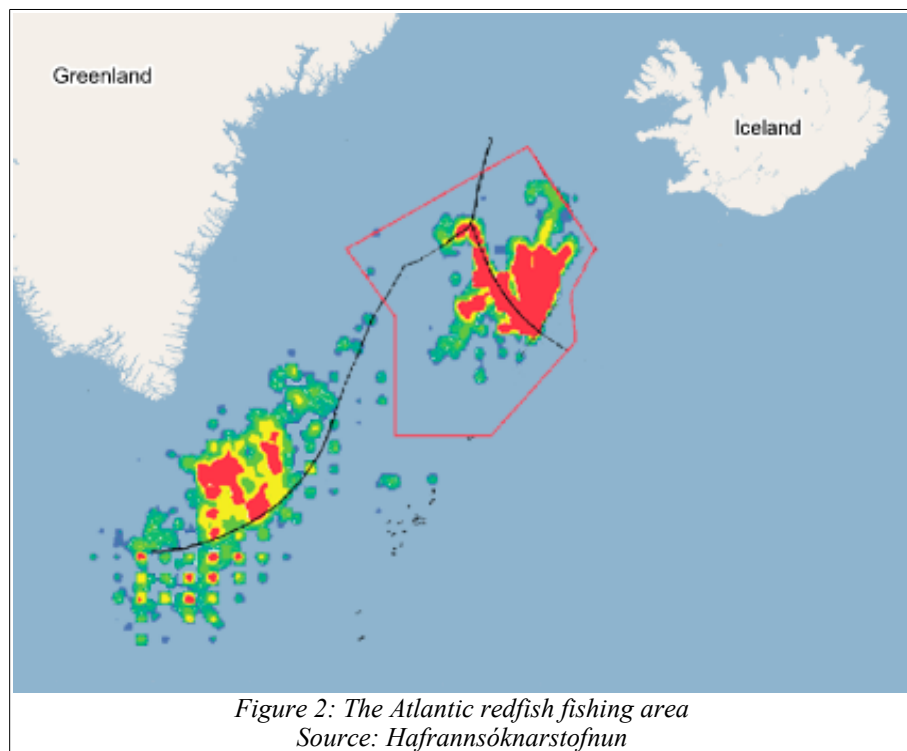
1.6 The Atlantic Pelagic (Oceanic) Redfish

One of the fisheries focused on under CEDER and the subject of this study was that of the North-Atlantic Redfish. The Atlantic pelagic redfish (*is. karfi*, *lt. Sebastes mentella*) is primarily caught on the Reykjanes ridge south-west of Iceland at a depth of 600-800m during the period of April to July. A secondary dataset is provided for the period of July to November south of the Greenland EEZ, when the redfish is caught at a depth of 200-400m.



Figure 1: The Atlantic Pelagic Redfish
Source: Jón Baldur Hlíðberg, (www.fauna.is)

The favoured spawning grounds of redfish are on the steep slopes of the Reykjanes ridge, just on the perimeter of the Icelandic EEZ. This is where most of the fishing takes place in the high season as seen in Figure 2. The figure shows the Atlantic redfish fishing areas studied, with the colour indicating the amount caught in the area (heat map). The red box marks the main fishing area during the period of April to July, and the black lines the perimeters of the Icelandic and Greenlandic EEZs respectively.



The redfish fishery is a relatively simple one, which helps the analysis and model construction. There is no secondary catch species, i.e. fish of other species than the targeted one being unavoidably caught in the net, resulting in a mixed catch. Discard is negligible, both because it is illegal by Icelandic law (with the exception of diseased fish unfit for human consumption) and perhaps more importantly, because the vessels simply do not exhaust their quotas and have no economic incentive to discard. The discard in the study period has been estimated at less than 1%, except in 2003 when it was between 5% and 7%. The discard is always due to *Sphyrion lumpi* (is. *karfaorða*) infection, a small crustacean parasite.

1.7 NEAFC and quotas

The North-East Atlantic Fisheries Commission (NEAFC) regulates the redfish fisheries, with vessels from the European Union, Faeroe Islands, Greenland, Iceland, Norway and the Russian Federation under its jurisdiction.

An important constraint is that only Icelandic vessels are allowed to catch redfish in the Icelandic EEZ. The EEZ perimeter lies directly across the most fertile fishing grounds where the redfish spawn, the schools condense and become easily catchable. Quotas are allocated separately inside and outside the EEZ; quota allocated inside the EEZ can only be caught there, while quota allocated outside can be caught either inside (applies only to Icelandic vessels) or outside the EEZ.

This results in the so-called "line-dance", where non-Icelandic vessels try to fish the most fertile grounds possible, right on the perimeter. The Icelandic coast guard generally has to keep a vessel in the area during the high season to prevent foreign vessels from fishing inside the EEZ.

There is an on-going scientific debate concerning whether the redfish in the area belong to one or two stocks (oceanic and pelagic deep-sea) (Thomson, 2002). Icelandic authorities have issued quotas for each separately, but this has not been adopted by NEAFC and remains controversial.

Both the allocated quotas and catch have declined sharply in recent years, and in fact the allocated quotas have seldom been filled. Table 1 breaks down the quota and catch for the study years, showing the total allocated (TAC) Atlantic redfish quotas for Icelandic vessels 2001-2006, actual catch and TAC uptake (total percentage of allocated quota caught).

Table 1: Quota allocation and catch

Year	Allocation (TAC)	Catch	TAC uptake
2001	45.000 tonnes	41.969 tonnes	93,26%
2002	45.000 tonnes	44.397 tonnes	98,66%
2003	55.000 tonnes	47.655 tonnes	86,65%
2004	55.000 tonnes	35.802 tonnes	65,09%
2005	34.470 tonnes	16.005 tonnes	46,43%
2006	28.610 tonnes	24.354 tonnes	85,12%

Source: Icelandic Directorate of fisheries (Fiskistofa)

For comparison, the total reported redfish catch of NEAFC member states for the year 2005 is shown in Table 2.

Table 2: Catch by NEAFC members 2005

Country	Catch	Percentage of total
European Union	1.904 tonnes	3,11%
Faeroe Islands	5.123 tonnes	8,35%
Greenland	857 tonnes	1,40%
Iceland	16.005 tonnes	26,10%
Norway	5.546 tonnes	9,04%
Russian Federation	31.885 tonnes	52,00%
Total	61.320 tonnes	100%

Source: NEAFC

1.8 The Pirates

Although the fishing is regulated by NEAFC, it is conducted in international waters which makes it difficult to take direct action against vessels that do not recognise NEAFC's jurisdiction. These vessels generally sail under the flag of states where regulation is lacking and enforcement is difficult, including Panama, Togo, Guinea and the Russian Federation. This activity is termed *illegal, unreported and unregulated* (IUU) fishing (NEAFC, 2010), but often these vessels are simply called *pirates* (Greenpeace, 2006).

NEAFC maintains a list of IUU vessels (<http://www.neafc.org/illegalfishing>) which all member states have agreed to deny access to ports and services. In 2006 the Icelandic coast

guard counted nine pirates (Landhelgisgæsla Íslands, 2006) fishing on the Reykjanes-ridge, but in the year 2007 there were none (Sjávarútvegs og landbúnaðarráðuneytið, 2007), which is attributed to more decisive actions on behalf of NEAFC based on the IUU list.

Since the vessels do not participate in NEAFC's reporting scheme by submitting data through the Vessel Monitoring System (VMS), it is difficult to monitor them or assess their impact. Both NEAFC and the European Joint Research Centre in Ispra have studied remote-sensing programmes to gather data on vessels that do not report their activities via the VMS-system. According to two such studies, the illegal catch was estimated up to 25% higher than the legal catch reported to NEAFC, or some 15.000 tonnes in 2004 (OECD, 2005).

Figure 3 shows redfish trawlers lining up like pearls on a string on the Reykjanes-ridge on May 12th 2006. The well-known pirate vessel EVA is in front.



Figure 3: Vessels lining up on the Reykjanes-ridge
Source: © Greenpeace/Martin Norman

1.9 Research Objectives

In order to facilitate monitoring and control it is desirable to construct a tool capable of delivering real-time catch estimates based on available data, as stated in the primary objectives of CEDER.

Formally stated, the aim of this study was:

- To construct an algorithm capable of identifying and categorising vessel activity through analysis of positional data (VMS). Specifically, to differentiate between fishing and non-fishing activities with the aim of estimating fishing effort.
- To construct a model capable of predicting the total catch of a vessel from its estimated effort. The model will be based on the analysis of fishing effort, catch logbooks and official landing reports.
- To verify the accuracy of the model predictions and establish a prediction interval.

The algorithms and models are based on a case study of the Atlantic Pelagic Redfish.

1.10 Structure of the thesis

After this current Chapter 1 - Introduction, Chapter 2 - Literature review outlines previous studies of interest and the work of other CEDER partners. Chapter 3 - The data introduces an analysis of the datasets used for the construction of algorithms and prediction models.

The bulk of the thesis work is then presented in Chapter 4 - Methodology and results, which starts out with the construction of a classification algorithm to identify vessel activities and estimate fishing effort, before moving on to construct a multivariate linear regression model to predict catch, based on the estimated effort. The chapter concludes with a discussion on the model performance.

In Chapter 5 - Discussion an outline of the work and the results is discussed, with Chapter 6 - Conclusions and future work presenting the main results and possible future work based on the study.

Chapter 2 - Literature review

2.1 Prior work

Extensive work has been done on predictive models for fishing in specific areas, and such models are commonly used in real-world implementations. These models however do not apply directly to the problem at hand, since they approach it from an *a-priori* standpoint, forecasting expected catch before the vessel has sailed. In contrast, this study is concerned with an *a-posteriori* scenario, estimating the vessel's catch based on where it has been, how long it has been there, and what its behaviour (VMS track) has been.

Forecasts using landings data as determined in port on the return of each vessel have been made (Czerwinski et al. 2007 and Gutierrez-Estrada et al. 2007) , but do not attempt to estimate the individual vessel effort or connect the forecasts with their geographical movement.

Estimating fishing effort with VMS data from beam trawling in the North Sea has been done in a study (Mills et al. 2007) closely related to the CEDER project, utilising 15-minute resolution data with instantaneous reported speeds (each VMS-record includes the actual recorded vessel speed at that point). The authors found that in most cases, trawls were represented by a single (2-hour resolution) VMS record, and in some cases a haul would fall entirely between VMS records, and so would not have been represented at all. The resolution of VMS position reporting was therefore found to be lower than required to adequately represent the trawl tracks. However, the algorithms could correctly identify trawling and steaming VMS points in over 95% of cases . This study did not attempt to forecast catch based on the effort estimate.

An attempt at this was made (Deng et al. 2005) using VMS data and high-frequency GPS data from Australia's northern prawn fishery, but found that VMS data with polling intervals longer than 30 min could not accurately estimate trawl tracks for the prawn trawlers.

2.2 Work of other CEDER partners

The comparable work of other CEDER partners concluded that 2 hour resolution VMS data was insufficient for accurate estimation of effort and forecasting of catch therefrom.

Specifically, a system prototype developed by Correlation Systems can classify behaviour of fishing vessels with accuracy, if it is provided with GPS positional data for each boat at least every 15 minutes. It is then able to identify fishing and cruising behaviour with less than 20% type I and II errors. The prototype can also guess landings, given recent effort, historical landings, observer figures, and historical logbook contents.

Sirius IT developed a system that is able to combine information from VMS, hail messages (logbook substitute), transcribed logbooks, and sales notes, in order to calculate a more accurate amount of catch on any given trip, but found that 2 hour VMS data was too coarse to reliably achieve the goal of effort estimation and catch forecasting from this.

The Joint Research Centre (JRC) developed a prototype that uses a set of time series models to predict quota uptake (or landings) using past quota uptake (or landings).

Chapter 3 - The data

In this chapter:

- Collection
- Preprocessing
- Identifying trips
- Overview of the data
- Visualisation of the data
- Vessels
- Fishing behaviour and features
- Summary

This chapter starts with a discussion of the collected data and what preprocessing was required to make it acceptable for analysis. Then the identification of individual trips and grouping together of VMS positions to form tracks is described, and how erroneous position reports were filtered out.

A numerical overview of the datasets is given, as well as selected illustrated examples, before the vessels included in the study are described and analysed for attributes that are important for model building.

The chapter concludes with a discussion of the fishing behaviour in the North Atlantic Redfish fisheries and how its features might be expected to manifest in the data.

3.1 Collection

The datasets used in this study were collected by the Icelandic Directorate of Fisheries (IDF) in the years 2001-2006 and stems from four sources:

- VMS tracking data
- Catch Logbook data
- Official catch landing reports
- Vessel attributes

For privacy and legal reasons the IDF encoded the names of the vessels before releasing the data, essentially making them unidentifiable. This has no bearing on the study, since the identity of specific vessels is of no consequence to the analysis or model building.

3.1.1 VMS

The tracking data from the automatic Vessel Monitoring System (VMS) includes time, latitude and longitude of the vessel with a 2 hour interval. It can be considered accurate to within ± 100 meters although instances of erroneous positions have been reported. This resolution has evolved as the industry standard simply because this is the accuracy of uncorrected GPS. The European Union has set its regulatory requirement at a slightly more relaxed ± 500 meters (Avanti Communications, 2007).

Note that because of the long reporting interval, the first and last VMS positions of each vessel trip are almost never in harbour, but on the way to or from it.

3.1.2 Catch logbooks

Catch logbook data is generally collected from hand-written logbooks aboard vessels (only a few vessels were fitted with electronic logbooks at the time of the study), recording time, latitude and longitude when the vessel starts towing, and estimated weight of the haul when on board.

Some discrepancies are present in this data, such as inaccurate or wrong position.

The estimated haul weight is either produced by the on-board processing line or by the captain visually estimating the catch on deck, in which case it is recorded only to the nearest metric ton.

Note that the catch logbook data only records the date of the catch but not the time. This requires the implementation of a rule-based method to connect the catch and positional data before analysis.

3.1.3 Official catch landing reports

All catch landed at Icelandic harbours is weighed by officials. This weighing is the basis for the quota system and can be considered accurate to within a few kilogrammes. Although fish is occasionally landed illegally, bypassing the harbour scales, it is generally assumed that Icelandic landing data is among the most accurate worldwide.

3.2 Preprocessing

Prior to analysis, the data format has to be standardised, ensuring that datatypes and formats are consistent throughout, e.g. that dates are all represented the same way, and latitude and longitude are in a *decimal format*. Also obviously erroneous data is removed, such as duplicate position reports that have no time difference, i.e. they have the same timestamp, and VMS positions that are clearly wrong (see Section 3.2.1 - Erroneous position reports).

The data preprocessing is part of the CARFI prototype system, (see script 1, `clean_positions`). Note that this script needs to run recursively, i.e. multiple times until it finds no further problems), as well as scripts for manual corrections (see script 2, `fix_positions`). This resulted in a total of 616 VMS datapoints being disqualified due to erroneous positions.

Also excluded at this stage are VMS-positions for timeperiods that do not have corresponding catch reports, and vice versa (see script 3, `exclude_messy_data`). This mainly applies to trips which start and finish in the VMS position data and catch logbooks do not coincide, e.g. where entries in the catch logbook start several days before the first VMS positions for that vessel. This constitutes the bulk of the disqualified VMS datapoints, or 45.633.

The reason for this high number of disqualified datapoints is that the datasets were delivered with both data for the fishing area on the Reykjanes ridge that is of primary

interest to this study, and data for the fishing area south of Greenland, which is of secondary interest. Most of the inconsistencies stem from this secondary area.

3.2.1 Erroneous position reports

The VMS-system is at times somewhat fickle, and can send strange reports for various reasons. Most commonly the system may send position values close to 0, indicating the equator (latitude 0°) or the prime meridian (longitude 0°). This can be caused by power surges or the system switching to a *difference* reporting format, in which case it reports the difference in degrees from the previous position, rather than the instantaneous position. Other strange readings may be observed, but are generally easily filtered with a visual inspection of the track data, or calculating the speed of the vessel and limiting it to a sensible number such as less than 50 knots.

An unfortunate feature of the VMS is that in some cases, when the system cannot send a report at the appropriate time, the report is stored and sent as soon as the system regains contact. The timestamp on these reports in the dataset is however not when the positional information was recorded, but rather when the report is actually sent. This results in a string of reports being sent within minutes or seconds, with large differences in the positional information. These reports must be excluded since otherwise the calculated mean speed lies in the hundreds of knots!

3.3 Identifying trips

The aim is to analyse and predict the catch of each vessel during one *trip*, i.e. from the time the vessel leaves harbour until it returns and lands its catch.

The catch logbook provides the date of each haul and landing. The position datapoints are grouped into trips by comparing the date and time with the catch logbook and applying a few simple rules. In the simplest terms, each trip includes all the positions between the first catch entry after a reported landing in the logbook, to the next reported landing.

This leaves out the position data while the vessel was steaming from and back to port. To include this data in the trip, the algorithm searches for *gaps* before and after in the VMS-data. Most vessels send VMS-data at 2-hour intervals while active, but generally not while in port, so missing VMS-entries indicate a change between trips.

A *gap* in VMS-reports is defined as:

- 1,5 times the previous interval (i.e. 3 hours in most cases), to account for reports being a few minutes late
- a minimum of 1,5 hours to account for extra reports sometimes being transmitted just minutes after the previous one
- more than 8 hours since a report

If no gaps are found, the algorithm applies a second rule and searches for the position *closest to shore* (actually closest to the centerpoint of Iceland), and split the trips there. This may happen e.g. when a vessel steams in to port and back out within a short period of time, transmitting VMS just before entering and/or leaving the port.

These rules effectively define the trips that are interesting, deal with the added complication of matching the catch logbook with VMS-positions only on dates and not times, and filter out any position data that was not part of an Atlantic Redfish fishing trip. An example of the resulting pairing can be seen in Figure 4. The figure shows the results from the trip identification algorithm for one vessel in the year 2001. The blue columns represent trips (a total of four trips, the height of each column corresponds to the trip number 1, 2, 3 and 4), while the gray columns represent individual trawls as reported in the vessel's catch logbook. Where the blue line is at 0 means that the algorithm found VMS position data, but determined that it did not belong to a valid trip. Where the blue line drops to -1 there are no VMS data available.

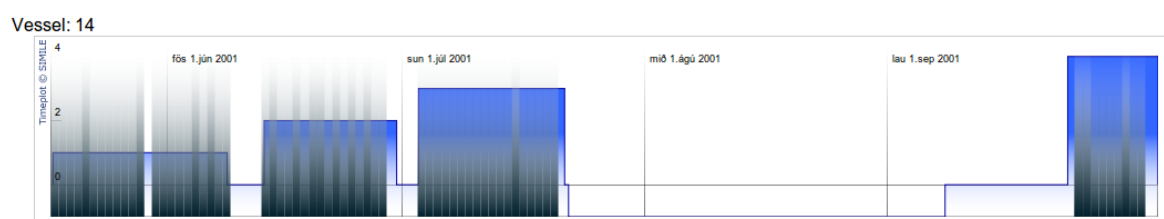


Figure 4: Trip identification algorithm results
Source: The CARFI system

In addition, the CARFI prototype system includes features to validate the trip identification algorithm, and manually apply fixes where it fails to correctly identify trips (see script 4, `fix_trips`).

3.4 Overview of the data

Having explained how the data is preprocessed and trips identified, a numerical overview of the datasets available for analysis follows. See Tables 3 and 4.

Table 3: Dataset overview, all data

Data	2001	2002	2003	2004	2005	2006	Total
Vessels in VMS	18	18	21	22	18	17	22
VMS positions	7.473	8.300	8.844	15.365	11.545	26.910	78.437
Vessels in logbooks	23	24	23	22	17	17	24
Number of Catches	2.000	1731	1639	1539	905	1030	8.844
Catches in Tonnes	41.969	44.397	47.655	35.801	16.005	24.354	210.181
Trips	111	99	103	81	51	63	508

Table 4: Dataset overview, valid data

Data	2001	2002	2003	2004	2005	2006	Total
Vessels in VMS	18	18	20	21	16	15	21
VMS positions	5.473	5.315	4.221	6.883	5.277	5.019	32.188
Vessels in logbooks	16	24	22	21	15	16	24
Number of Catches	1.061	1.624	1.441	1.009	308	595	6.038
Catches in Tonnes	22.122	41.998	42.851	23.501	4.914	13.067	148.453
Trips	57	94	93	46	20	39	349

As previously discussed, a number of datapoints were disqualified due to data quality problems, originating mostly from the secondary fishing area south of Greenland. However, a sufficient number of measurements were retained to construct the model.

3.5 Visualisation of the data

In order to better understand the datasets provided, it is useful to examine them graphically. Figure 5 shows an example of the VMS track data features discussed above. The blue circles are VMS-position reports with 2 hour resolution, and the red X-marks are positions of catch reports from the vessel logbook.

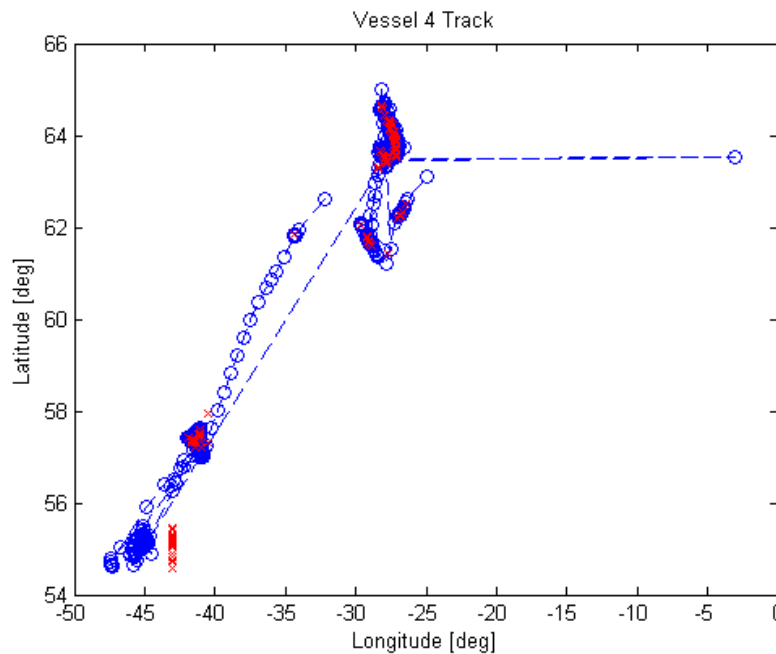


Figure 5: Vessel VMS track

As expected, the track is mostly concentrated around the reported catch positions. The long diagonal line from the upper area to the lower left is an example of where the track splits into distinct trips.

A couple of data quality problems are evident; a single position report is given far away from the main concentration (in fact close to the Greenwich meridian), and the lower left catch positions are all at the same longitude, indicating a data entry error. These problems were corrected where possible, or in cases where the data was deemed too unreliable, excluded altogether from further analysis. Mostly this is data from the secondary dataset south of the Greenland EEZ.

Figure 6 shows the same vessel track as the previous figure, overlaid on a satellite image of the area. The fishing area on the Reykjanes ridge can be clearly seen.

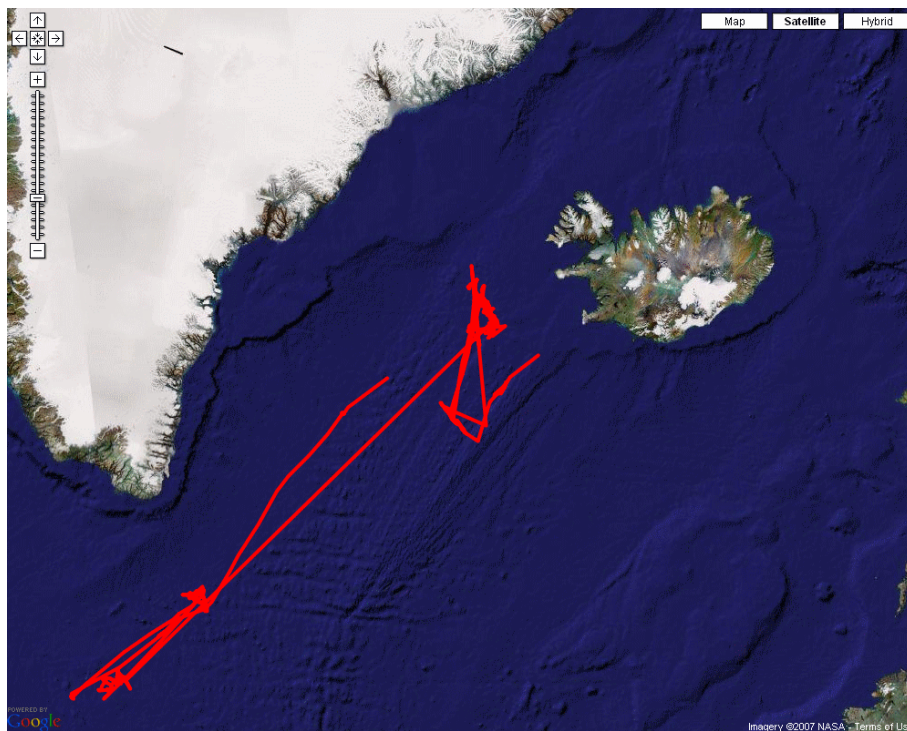


Figure 6: Vessel VMS track with satellite image

Calculating the mean speed for each leg using the reported time of the VMS-reports and Vincenty's formulae² (Vincenty, 1975) produces Figure 7. The figure shows the same vessel track as the previous figures, with the height of the columns indicating the calculated speed at each point. Red crosses represent positions of reported catches. This figure shows how the basic idea of using vessel tracks to predict catch came about, since legs with low speed are concentrated around the reported catch points.

² Vincenty's formulae are used to calculate the distance between two points on the Earth, and is accurate to within 0.5mm

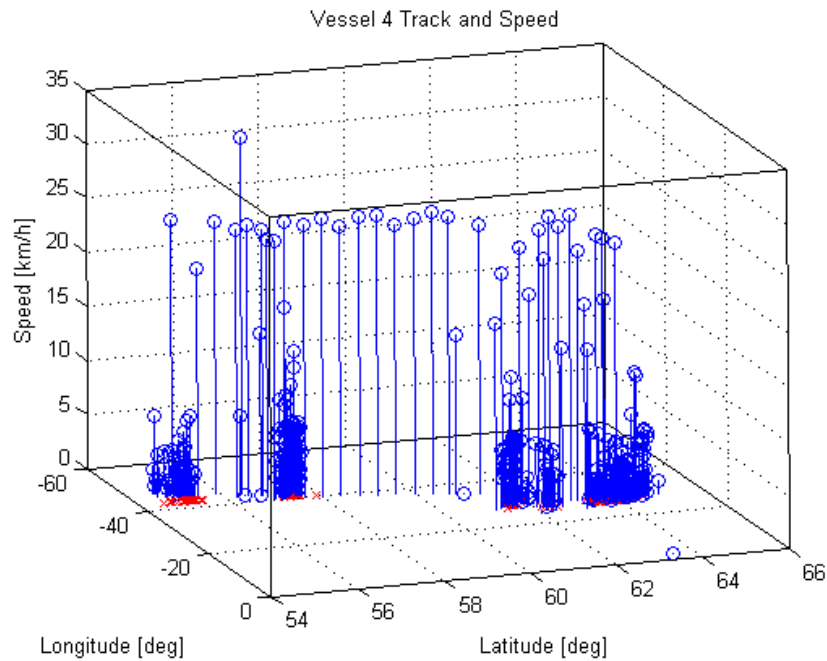


Figure 7: Vessel track and calculated speed

Jumping slightly ahead, the same data can be plotted with the added leg classification, to produce Figure 8. The figure shows the same vessel track as the previous figure, with the leg classification added. Blue columns represent cruising legs, while red columns represent trawling legs. Black crosses represent positions of reported catches. The classification criterion used here is a speed cut-off at 8 km/hr, or about 4,3 knots.

The lower speed legs are in red, and concentrate neatly around the catch points.

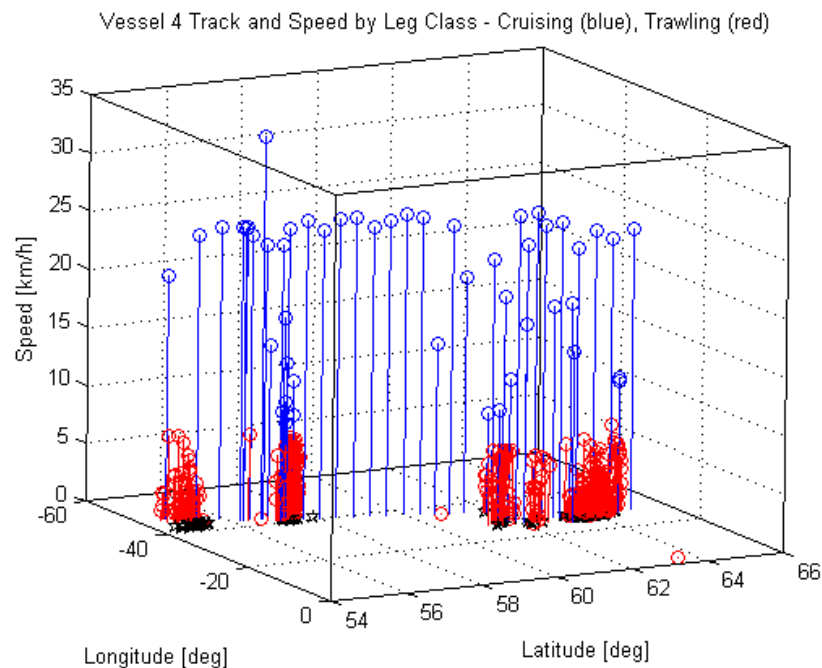


Figure 8: Vessel track, speed and leg classification

Finally for the VMS data, Figure 9 shows a histogram of the leg mean speed for all vessels, i.e. the number of legs by calculated speed at 0,5 knot intervals. It seems to indicate that there is indeed a divide at about 5 knots.

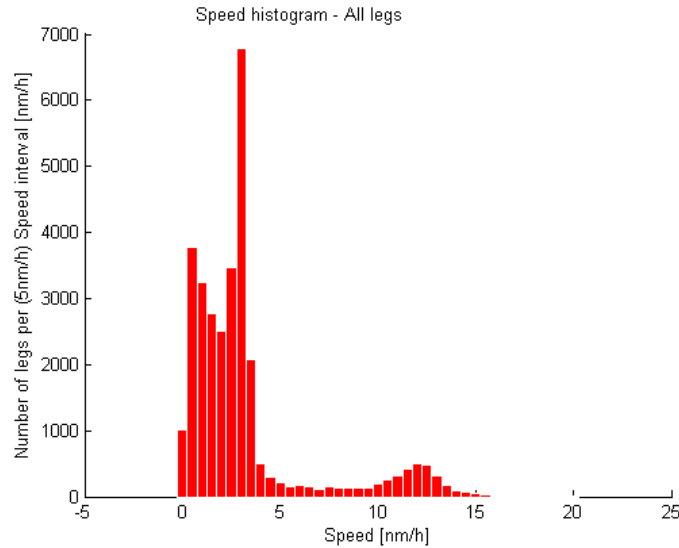


Figure 9: Number of legs by speed

Looking at the catch data, it is evident from Figure 10a and 10b that the most common catch size is about 25 tonnes, which fits nicely with information from experienced captains. The figures show a histogram of reported catch, i.e. the number of reported catches at 5 ton and 500 kg intervals respectively. Note that using the smaller bin size of 500 kg, a clear tendency to report catch in half-ton quantities is seen, leading to a *discrete effect* in the plot. This is almost certainly due to catch logbooks not being electronic and automated as discussed before.

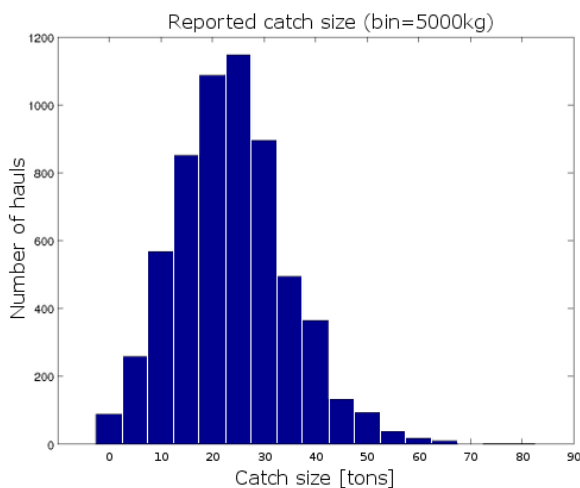


Figure 10a: Catch size with 5000 kg bins

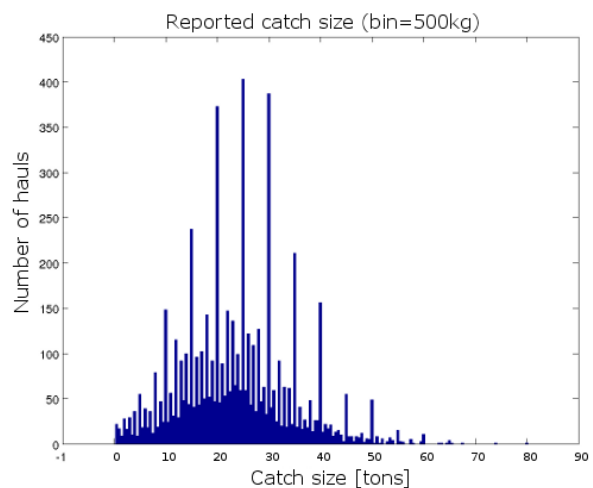


Figure 10b: Catch size with 500 kg bins

The catch logbook also gives the reported effort for each haul, and thus the distribution of

the haul length in Figure 11. The figure shows a histogram of haul length, i.e. the number of reported hauls at 60 min intervals. The most common haul duration is around 12 hours.

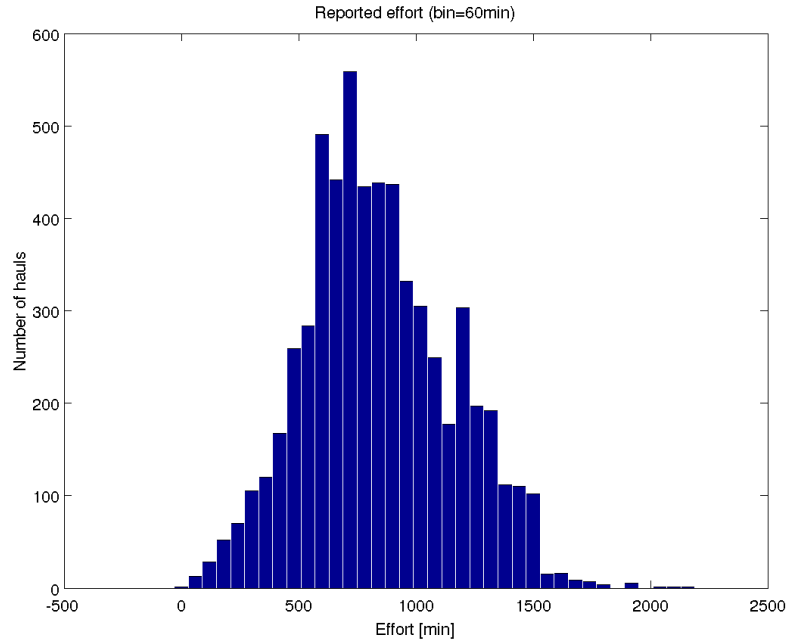


Figure 11: Reported effort

Finally, Figure 12 shows a plot of *reported effort* vs. *reported catch*. This is the relationship needed to model to predict catch from effort, and a definite suggestion of a linear relationship can be seen.

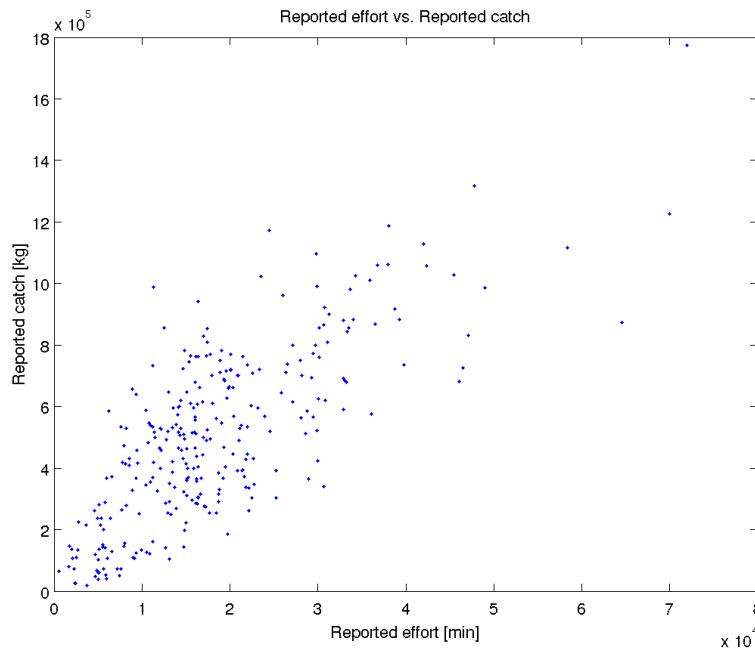


Figure 12: Reported effort vs. reported catch for each vessel trip

3.6 Vessels

At the height of the season, over 60 vessels of various nationalities are trawling for redfish on the Reykjanes-ridge, up to 10 of which are pirate vessels (Greenpeace, 2006). A total of 28 distinct Icelandic vessels have participated in the fishing at one time or another during the study period, 22 of them providing both VMS and catch logbook data of sufficient quality for analysis. These vessels are the main focus of this study and the source of the datasets provided.

Figure 13 shows two Russian examples of the vessels used in the redfish fisheries.



Figure 13: Russian trawlers on the Reykjanes-ridge
Source: Jón Páll Ásgeirsson

When analysing the catch and effort of individual vessels, the degree of variability between vessels must be assessed, and decided if the models need to take this into account.

The most important factors in this respect are vessel size (tonnage) and engine power (kilowatts), as they are the best indicators of how large a gear the vessel is capable of towing behind it.

Figure 14 is a scatterplot matrix for the three measures of *gross register tonnes (GRT)*, *length over all (LOA)* and *engine power (KW)*, and the calculated *catch per unit effort (CPUE)*, i.e. *reported catch* divided by the *reported effort* for all vessels and trips.

The plot reveals (unsurprisingly) a linear relationship between vessel size in GRT and engine power KW, and a somewhat less distinct relationship to vessel length LOA.

The column on the far left also shows that CPUE is not directly related to vessel size or power, at least not for other than the largest and most powerful vessel classes.

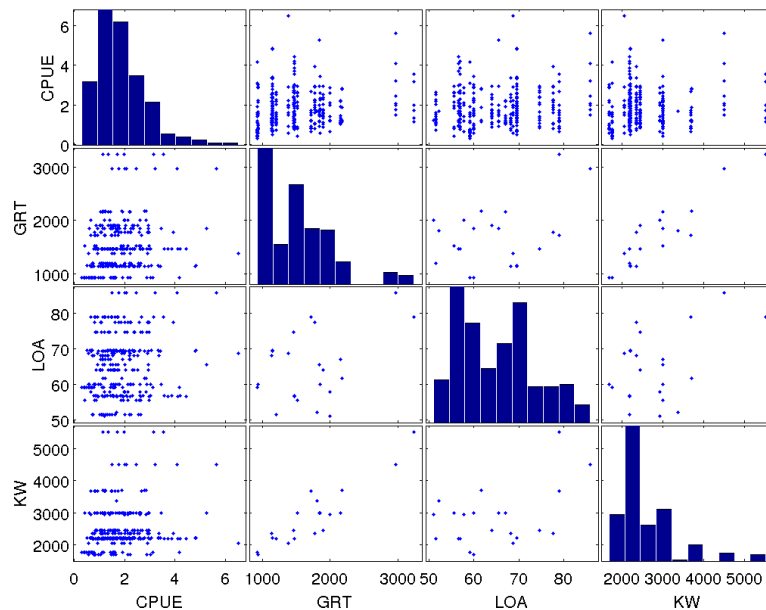


Figure 14: Scatterplot matrix for vessel attributes

The conclusion is that in terms of catching power, most of the vessels in the study are similar, and the vessel attributes are not likely to be a helpful factor in predicting catch. What little variability there is in the vessel's catching power will however be accounted for in the catch model parameters.

3.7 Fishing behaviour and features

It is useful to give a short account of the fishing activities the algorithms aim to detect, and consider what features it is possible to capture, based on the resolution of the datasets. This description is to a large extent based on interviews with one of the most experienced Icelandic captains in the area (Captain Kristinn Gestsson. Interview, 1. march 2007).

Due to the crowded conditions and tight schools of redfish, the trawlers organise themselves into two lines (see Figure 3, page 6), trawling in both directions. A vessel that reaches the end of the line will turn around and join the other line back.

The vessels keep a clearance of about 2 nml, and the gear extends up to 1 nml behind each vessel. This is to allow the redfish school to reform, since the gear of the preceding vessel will leave a fish-free gap that will settle in the meantime.

The vessels normally trawl at around 3 knots, possibly up to 4 knots, and cruise at more than 10 knots. Below a speed of 2,5 knots it becomes tricky to keep the gear open, but this also depends on depth. If the vessel is trawling at less depth, it needs more speed to keep the gear open. The currents in the area are about 0,5 knots that add or subtract from the vessel's speed in the water. There are occasional underwater peaks in the area, so the captain may need to temporarily pull the gear higher in the water to avoid them.

The duration of each trawl depends on the catch, but can easily be around 8-9 hrs (the most

common haul duration is 12 hrs, see Figure 11, page 19), and they last at least 2-3 hrs. If the vessel's factory is full and cannot receive more fish at the moment, the captain may on purpose allow the gear to go higher or lower in the water to slow the catch. Alternatively, the vessel is allowed to drift while processing catch. The optimal process is to haul 2-3 times every 24 hrs, and the optimal haul size would be around 40 tonnes. Most of the time spent in the area is used trawling.

Hauling the gear and resetting it takes about one hour, with the ship often coming to a standstill while pulling in the wire. The Icelandic vessels sometimes refuel for 4-6 hrs, but do not often transfer their catch to another vessel (transshipment), which e.g. the Russian trawlers commonly do.

There is no discard, unless the fish is seriously infected and unsuitable for consumption.

The points of most relevance here are that the hauls are relatively long, around 12 hrs, compared to the data resolution of 2 hrs. The shorter hauls however are actually at around the same length as the resolution, and this may cause problems in detection.

Similarly, picking out detailed features such as a vessel setting or hauling gear, which takes around 30 min or a quarter of the data resolution, is unlikely.

Refuelling or drifting for extended periods of time should be possible to detect.

3.8 Summary

This chapter has focused on the raw data and features of the North-Atlantic Redfish fishery, solving data-quality problems and preparing the data for analysis. The attributes of the vessels were analysed, and the fishing activities the algorithms are expected to be able to detect in the data were described.

In the next chapter, the methods applied to estimate vessel effort and build the models that use this estimate to predict the vessel catch are introduced.

Chapter 4 - Methodology and results

In this chapter:

- General approach
- Classification of vessel activities
- Estimating effort from classification
- Estimating catch from effort
- Model performance

First the general approach is outlined, i.e. how the classification of vessel activities at each point in time will give an estimate of the effort, which then is used as input to a catch estimation model. The use of the datasets as a driving factor in the algorithm development and their subsequent use is described.

The chapter then focuses on the analysis of an appropriate classification algorithm, followed by the comparatively simple estimation of effort.

Having produced the necessary effort variable, the catch estimation model is then constructed and validated, and its accuracy assessed.

The chapter concludes with a discussion of potential improvements.

4.1 General approach

Since the idea is to estimate catch, it is natural to turn to traditional fishery statistics and stock analysis for a suggestion of the initial model. The simplest and most commonly used approach is to assume a linear relationship between effort (trawling time) and catch (Stefánsson, 1997). This kind of model also makes immediate intuitive sense; the more time you spend fishing, the more you catch³.

$$\text{Catch} = \text{Catch per unit effort} \cdot \text{Effort}$$

In order to be able to estimate the catch, the effort must first be estimated. Effort in this sense is simply the total time the vessel spent trawling. When plotting the *reported effort* of the vessels vs. the *reported catch* as in Figure 15, a suggestion of a linear relationship can indeed be seen. The figure shows the reported catch vs. reported effort, based on catch logbooks of all the study vessels from 2001-2005. A simple linear regression model has been superimposed on the data.

3 Of course more sophisticated models exist that take into account a multitude of other variables, such as environmental and biological factors. For a discussion of such models developed under the CEDER project, see e.g. Cotter et al. [COT06]

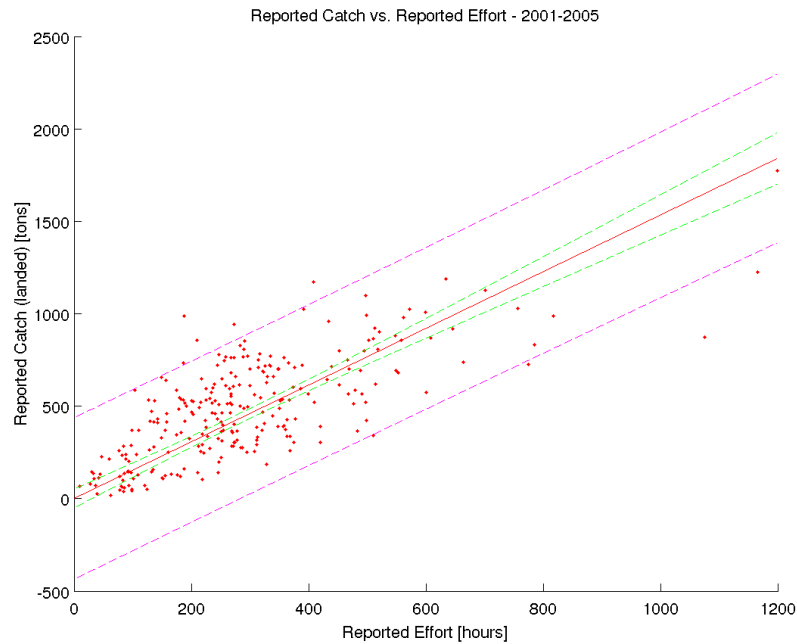


Figure 15: Reported catch vs. reported effort 2001-2005

To be able to calculate this effort, the vessel *activity* at each point in the *trip* must first be classified, i.e. what the vessel is doing during each *leg* of the vessel track must be predicted, as either *cruising* or *trawling*. The timeperiods the vessel spent trawling can then simply be added together, to produce a measure of the *estimated effort*.

The process of building the models is pictured in Figure 16, showing each step in building the proposed activity classifier and catch estimation model. To construct the classification algorithm both VMS and catch logbook data are used to produce a predicted activity and an estimated effort. To construct the catch estimation model both catch logbooks and landing reports are used in conjunction with the estimated effort to produce the final catch prediction.

Subsequently, when building the catch estimation model both the catch logbook entries and the official landing reports are used in conjunction with the *estimated effort* to arrive at suitable model parameters and finally the *estimated catch*.

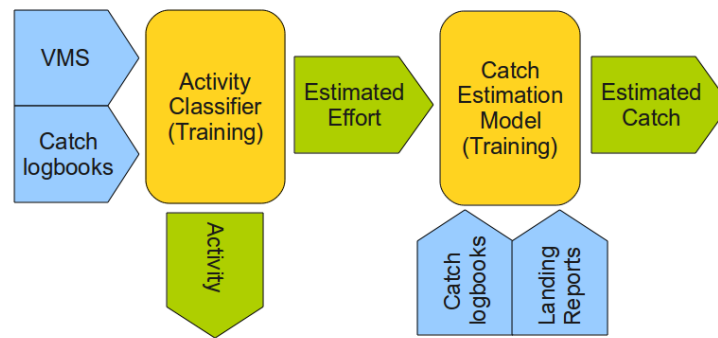


Figure 16: Strategy for building the activity classifier and catch estimation model

After calibrating the algorithms with the datasets, they should then be able, given only a vessel's VMS track, to estimate its effort and thus its catch, see Figure 17. Feeding VMS data for a vessel into the classifier gives the probable activities and estimated effort for that vessel. The activities can be compared with the catch logbook, while the estimated effort is fed into the catch estimation model, producing an estimated catch which can be compared with official landing reports. The vessel's activities can also be compared to the catch logbook, and the estimated catch with the landed catch to look for discrepancies, such as fishing activity when no catch is reported, or unusually high or low catch compared with the estimated effort.

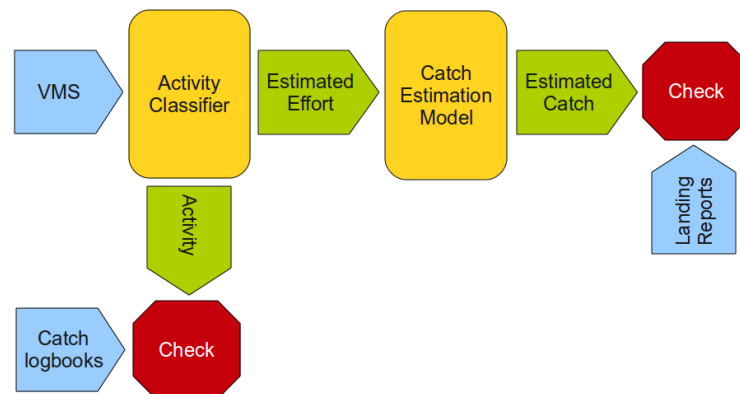


Figure 17: Step-by-step use of the algorithms to predict effort, activity and catch from VMS-data

4.2 Classification of vessel activities

In this section:

- Supervised vs. unsupervised classification
- Finding the “actual” activity from the catch logbook
- Expert classifier based on speed
- Calibrating the classifier
- Adding more predictor variables
- Alternative classifiers
- Cluster analysis

First the general approaches to classification and how to validate the activity classifier against the vessel actual activity are discussed. Then a classifier based on expert knowledge is explored, where the vessel speed is used to separate the activity classes.

Reversing this approach, the prediction of the actual activity is used to calibrate the classifier so that erroneously classified legs are minimised.

More measures are then added to the classifier and the improvement analysed, if any.

This section concludes by exploring five other classifiers, Fisher's Linear Discriminant Analysis, the Naïve Bayes classifier, k-NN classifier, CART, the Multilayered Perceptron and the unsupervised k-means clustering algorithm.

4.2.1 Supervised vs. unsupervised classification

There are essentially two approaches to classification; supervised and unsupervised. The former approach requires knowledge of the actual activity to compare the results to, whereas the latter approach tries to identify “natural groupings” in the data, without knowing if they correspond to real phenomena or not. This is also known as *cluster analysis*.

It is tempting to try supervised classification, since this approach gives a clear measure of success, an *error rate*.

4.2.2 Finding the “actual” activity from the catch logbook

The first problem is how to measure success in classifying the vessel activity and validate the classification algorithm. The obvious method is to compare the classification algorithm results with the actual activity of the vessel.

Even though the VMS is capable of sending vessel activity messages, this is not a widespread practice, and the datasets in this study do not include such reports. The catch logbooks must thus be used to match the actual vessel activity to the VMS positions. They have one entry for each haul of the vessel, with information on the gear type, position of the vessel when the trawl began, date, catch, and time trawled.

The logbooks in use for the redfish fisheries are not electronic and automated, which would increase their accuracy and reliability, but pen and paper versions. This introduces the possibility of errors in position, date and catch reporting. For example, some captains

“eyeball” the catch on deck, whereas others rely on the vessel's processing line for an accurate catch weight, and it is easy to make errors in noting the position, or misreading when the logbook data is digitised.

A problem arises here from the apparently innocuous fact that the logbook datasets do not record the precise time of the haul start and end, only the date. Including the time would make it easy to take all VMS positions during this period and declare them as trawling activity with great confidence. Since this is not the case, the reconstruction of this information requires the use of an algorithm which will be called *Midpoint*.

Midpoint

This algorithm (see script 5, `actual_activity_midpoint`) calculates the midpoint position of each leg and selects the leg with a midpoint closest to the reported catch position as the haul starting leg (within the same day as the catch entry, or last leg of previous day).

The logic behind this approach is explained in Figure 18. First, in the timeperiod between two VMS reports a vessel can actually move anywhere within an ellipse with the two known VMS positions as its vertices (Figure 18a), even though the vessel movement is represented with a straight line. There are then two possible cases; the catch point falls “outside” of the leg (Figure 18b), or the catch point falls somewhere “between” the two VMS points (Figure 18c). If the algorithm was simply to take the closest VMS position to the catch point as the start of the haul, it would not include the first leg in the trawl, during which this activity actually started.

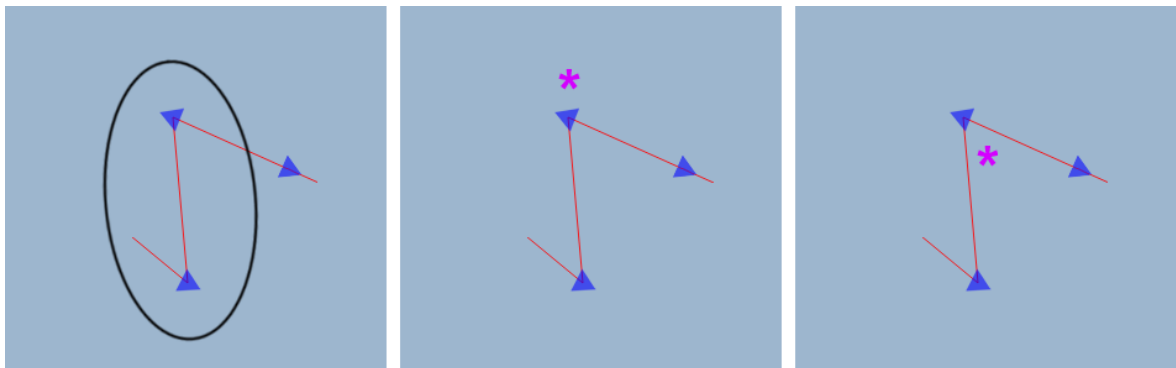


Figure 18: Vessel legs and catch points

Some uncertainty has to be accepted in this matching, since it is not always clear which of two possible legs is the actual starting leg.

It is important to note that even though the whole leg is classified as trawling activity, in fact only a portion of the first (and last) leg is actual effort. Therefore the distances from the catch point to the leg start and endpoints are calculated, the proportion of the leg that should be included as effort is estimated. This proportion is then used to pinpoint the starting time of the haul.

When the haul starting time has been reconstructed in this way, the reported haul duration from the catch logbook is added to arrive at the haul end time. In the case of many catches within the same day, the algorithm requires that the next haul never starts until the previous

one is finished.

The algorithm was implemented in the CARFI prototype system, and its source code is available from the author by request.

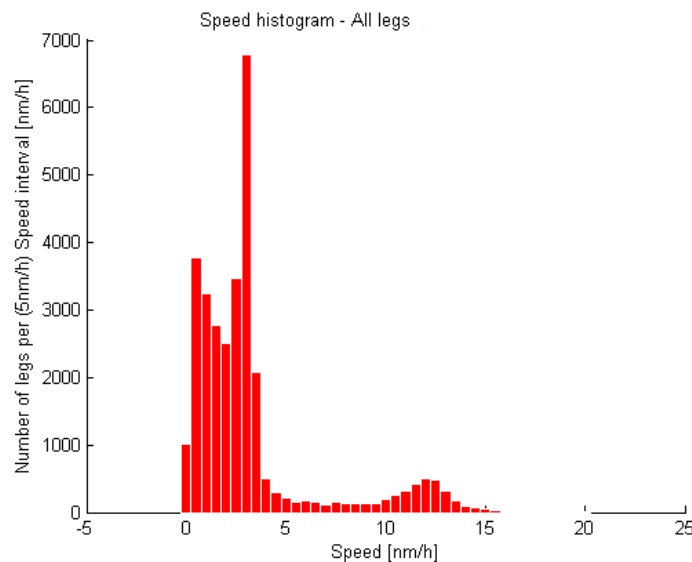


Figure 19: Number of legs by speed

The actual activity of the vessels has now been reconstructed based on the catch logbook, by connecting the legs with an activity with the midpoint algorithm. From here on, when discussing the supervised classifiers and their error rates, it is always in comparison with the actual activity as guessed by this algorithm.

To achieve a more accurate representation of actual vessel activity would require more precision in catch logbook entries, VMS activity messages, and/or higher VMS resolution.

4.2.3 Expert classifier based on speed

From the VMS positions the vessel mean speed during the leg can be calculated. As previously mentioned, VMS reporting interval is most commonly 2 hrs.

From interviews with fisheries experts, coast-guard officials, and most importantly one of the most experienced captain participating in the redfish fisheries [Captain Kristinn Gestsson. Interview, 1. march 2007], it was concluded that a vessel needs to maintain a speed of 2-4 knots while trawling. The lower bounds are determined by the need to keep a minimum speed for the gear to open fully. In addition to this relative speed, the vessel may drift at as much as 1 knot. Thus the classification rules are:

- A vessel with mean speed above or equal to 5 knots is cruising
- A vessel with mean speed below 5 knots is trawling
- A vessel with mean speed below 0,1 knot is stopped

Plotting a histogram of the leg mean speed for all vessels, i.e. the number of legs at 0,5 knots speed intervals, gives the distribution in Figure 19. This figure shows that the vessels spend most of their time below 5 knots, at trawling speeds.

Using this 5 knots limit, the resulting classification of legs is illustrated in Figure 20.

The figure shows a histogram of leg mean speed for all vessels by predicted activity, i.e. the number of legs at a specific speed classified by discriminant classifier as cruising (blue) or trawling (red). Above 5 knots all legs are classified as cruising, while below they are classified as trawling.

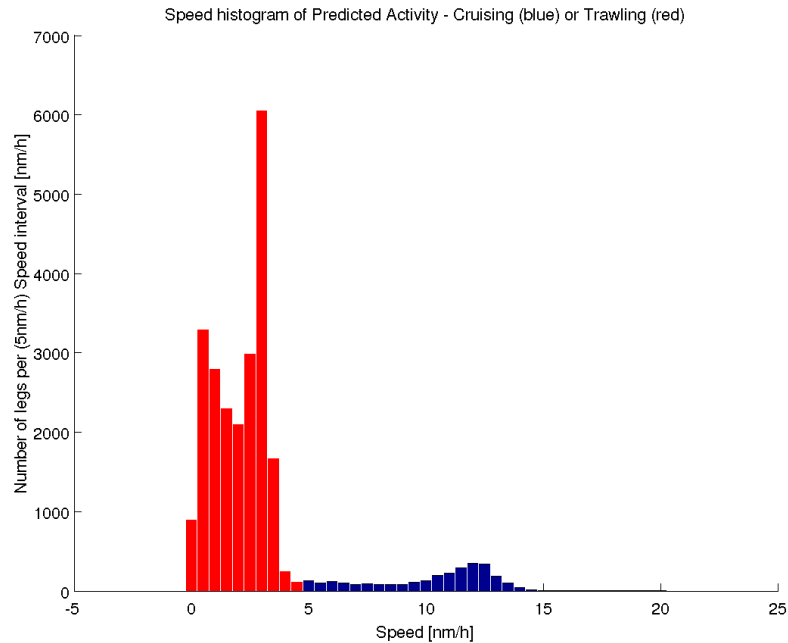


Figure 20: Number of legs by speed and predicted activity

The details of what is happening are shown in Figure 21. The figure shows the speed profile of vessel 3, year 2001. The leg mean speed is the blue line, overlaid with trawling periods from the catch logbook in gray. The red line represents the 5 knot limit. Note that the VMS data is not completely continuous (gaps when the vessel is in port, or targeting other species), hence the distinctive jumps from the end of one trip to the start of the next. Any legs with speed below the limit will be classified as trawling, and legs with speed above the limit will be classified as cruising. Ideally, the vessel should be at trawling speeds during the gray periods, and cruising speeds in between.

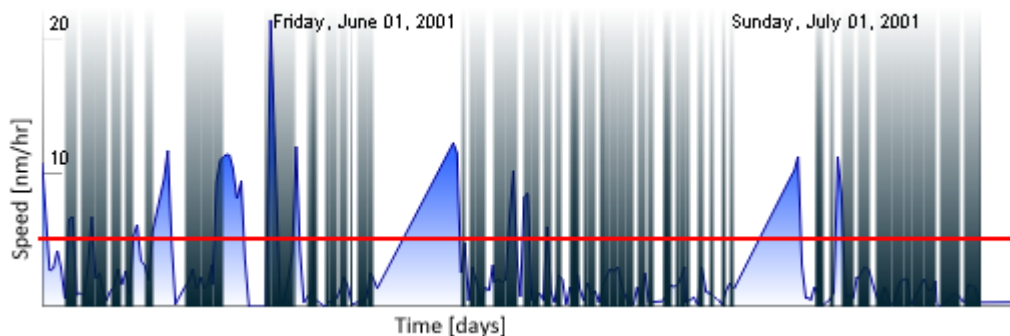


Figure 21: Simple linear discriminant classifier with 5 knots decision boundary

This type of classifier is known as a *discriminative classifier*, and the 5 knot limit as a *decision boundary*. In this case, there is only one *predictive variable*, mean leg speed, and one *class variable*, activity, which can take on three values; *stopped*, *trawling* or *cruising*.

From this, it is evidently impossible to accurately discriminate between trawling and cruising activities using only the mean leg speed, since the speed does not always cross the decision boundary during the non-trawling gaps between the gray trawling periods. Above 5 knots it is quite certain that the activity is not trawling, but below 5 knots there is actually a mixture of cruising and trawling activity.

This is reflected in the *error rate* of the classifier when compared to the actual activity, which is shown in Table 5. The table shows counts and percentages of correctly and falsely classified legs. The *Actual activity* of the leg is in the first column, followed by the *Result* of the classification. The *Total count* column shows the number of legs classified as indicated by the preceding columns, and the *Proportion of total* the percentage of the total legs. The *Class count* columns show the number of legs classified in each dataset, training and validation and the *Proportion of class* columns show the percentage of correctly or falsely classified legs within the same activity class. The error rate for each class then corresponds to the “false” result rate. The total success and error rate as well as some notable numbers mentioned in the text have been highlighted for convenience.

This table shows that 97,6% of the trawling legs are captured, and also 27,0% of cruising legs are correctly classified. However, the vessels apparently still spend about 2/3 of their cruising time below 5 knots, resulting in the high misclassification rate for that activity (see histogram in Figure 22 of leg mean speed for all vessels by actual activity, i.e. the number of legs at a specific speed classified by the midpoint algorithm as cruising (blue) or trawling (red). Above 5 knots, there are mostly cruising legs, while below there is a mixture of cruising and trawling legs.).

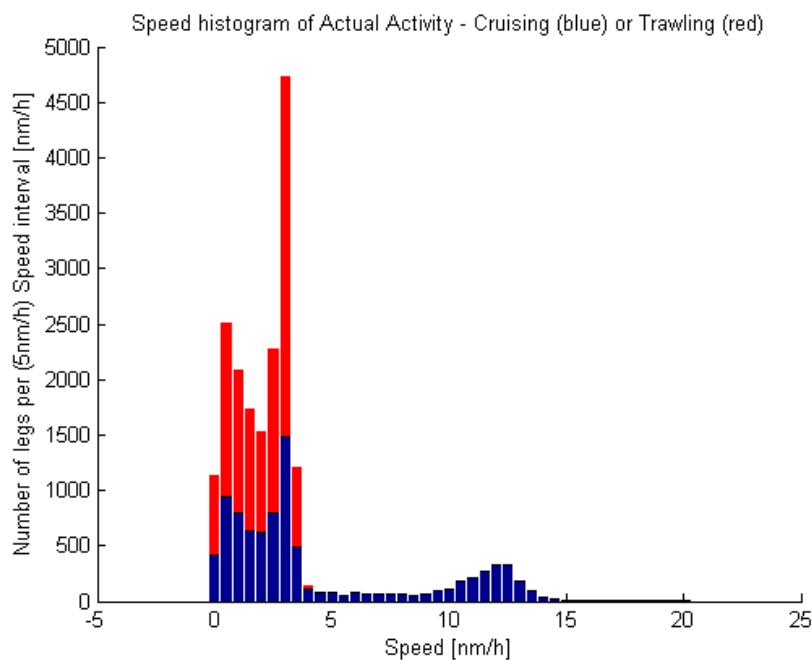


Figure 22: Number of legs by speed and actual activity

All in all, there is still a 74,6% chance of correctly identifying the leg activity, but the algorithm ends up classifying too many legs as trawling, resulting in an overestimate of the effort. The total error rate from this classifier with this speed parameter is 25,4%. It is important to note that the error rates stay consistent across both training and validation sets.

Essentially, when calculating the effort, the algorithm is adding both the gray trawling periods in Figure 21 and a majority of the gaps between them to the total effort estimate, only leaving out those legs that are clearly not trawling legs since their speed is higher than 5 knots.

Table 5: Classification results – Simple linear discriminant at 5,0 knots

Actual activity	Result	Total count	Proportion of total	Class count 2001-2005 Training set	Proportion of class	Class count 2006 Validation set	Proportion of class
Cruising	Correct	2.486	9,1%	2.001	27,0%	485	32,2%
Cruising	False	6.431	23,6%	5.410	73,0%	1.021	67,8%
Trawling	Correct	17.515	64,2%	15.061	97,6%	2.454	97,0%
Trawling	False	443	1,6%	368	2,4%	75	3,0%
Stopped	Correct	358	1,3%	329	83,3%	29	90,6%
Stopped	False	69	0,3%	66	16,7%	3	9,4%
Total	Correct	20.359	74,6%	17.391	74,9%	2.968	73,0%
Total	False	6.943	25,4%	5.844	25,2%	1.099	27,0%
Total	All	27302	100%	23235	100%	4067	100%

The conclusion from the above is that the results of this classifier can be used to identify *periods of trawling activity*, although it can not provide a highly accurate prediction of the activity during a specific leg. It should then arrive at an estimate of effort which is closer to the actual effort than e.g. using *days at sea* as a measure, since it can exclude legs where it is certain the vessel is cruising. Typically, the algorithm would be certain to exclude legs where the vessel is steaming to and from port or from one fishing grounds to the next, as can be seen in Figure 23. The figures show a track where the legs have been colour coded according to the classified activity, according to the simple mean leg speed classifier. Blue legs are cruising, red legs are trawling. The trawling legs coincide nicely with reported catch-points from the catch logbook, shown as pink stars superimposed on the right side. The classifier correctly excludes legs where the vessel is cruising to or from port and between fishing grounds, but cannot pick out individual hauls at this resolution.

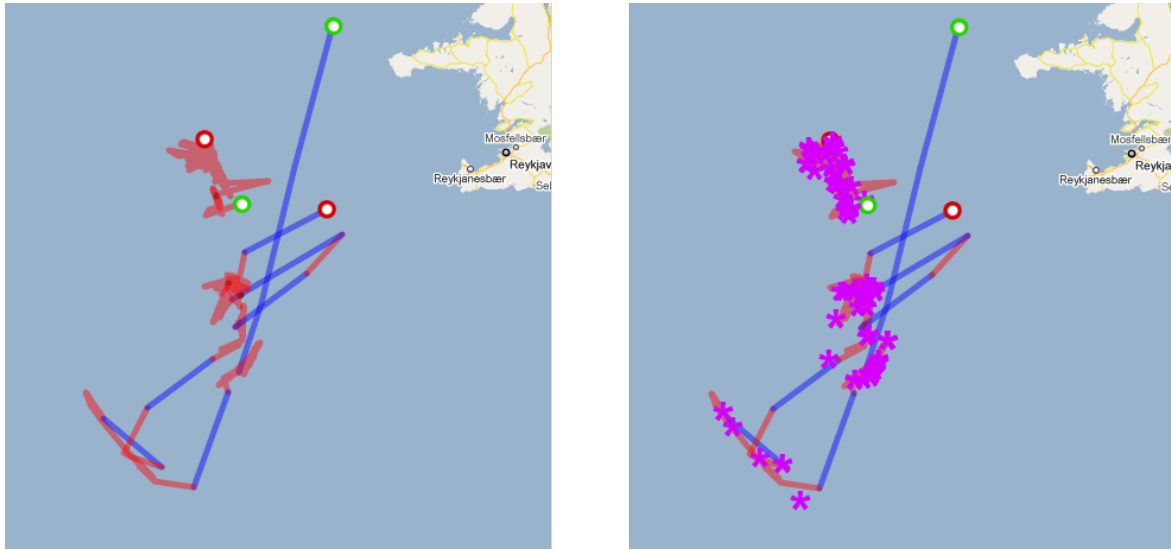


Figure 23: Track with predicted activity for vessel 2, year 2003

4.2.4 Calibrating the classifier

In the previous section expert advice was used to “guess” at the decision boundary for the classifier. But was this guess in fact accurate? Can the classifier be calibrated to improve the results?

The *error rate* was mentioned, also known as the *misclassification rate*, i.e. the proportion of incorrectly classified legs of the total. This measure of the classifier accuracy is known as the *scoring function*⁴, and can be used to evaluate the performance of the classifier. It can also be used to optimise the parameters used in the classifier, that is control where the decision boundaries lie so that the number of incorrectly classified legs is minimised.

Suppose there is a collection of N vectors of predictor variables \mathbf{x}_i (in the previous section the mean speed of each leg in the vessel track was the predictor variable) and corresponding observations y_i of the actual activity. A classifier is a function that predicts the class based on the predictor variable vector \mathbf{x}_i and a classifier parameter vector $\boldsymbol{\theta}$. If the i -th prediction is denoted as $\hat{f}(\mathbf{x}_i; \boldsymbol{\theta})$ then the scoring function can be formally written as (Hand et al. 2001)

$$S_{0/1} = \frac{1}{N} \sum_{i=1}^N I(\hat{f}(\mathbf{x}_i; \boldsymbol{\theta}), y_i)$$

where the function I is 1 if the prediction and actual observations match, and 0 if they don't.

In the simple case of one predictor variable it is easy to visualise the optimal decision boundary. Figure 24 shows the results of the *actual activity* algorithm, with the mean leg speed coloured according to actual activity class; blue is cruising, red is trawling. The decision boundary should then lie somewhere around 4 knots. Above this speed there are

⁴ For a discussion on classifiers, search and scoring functions, see e.g. *Principles of Data Mining* (Hand et al. 2001)

mostly cruising legs, while below there is a mixture of legs where the vessel is trawling or cruising at low speed. If the decision boundary speed is raised, the algorithm will start to falsely classify more cruising legs as trawling, raising the misclassification rate. If the decision boundary speed is lowered, it will falsely classify more trawling legs as cruising, but it will also include more correctly classified cruising legs. The optimal boundary lies where these two errors are balanced.

Trawling at much higher speeds than 5-6 knots would in fact be physically impossible for any normal fishing vessel. The few trawling points seen above this are an artifact of the algorithm used to reconstruct the vessel's actual activity, and reflect the definition of a trawling leg. A leg is defined as trawling if any part of it contains trawling activity, but the vessel may be starting or ending its trawl and part of it contain cruising activity at high speed, thus pushing it up on the graph.

The leg classes do not have different weights in the scoring function, i.e. there is not a greater penalty to misclassification of either cruising or trawling legs, since they are of the same timescale and would contribute equally to the final calculation of *estimated effort*.

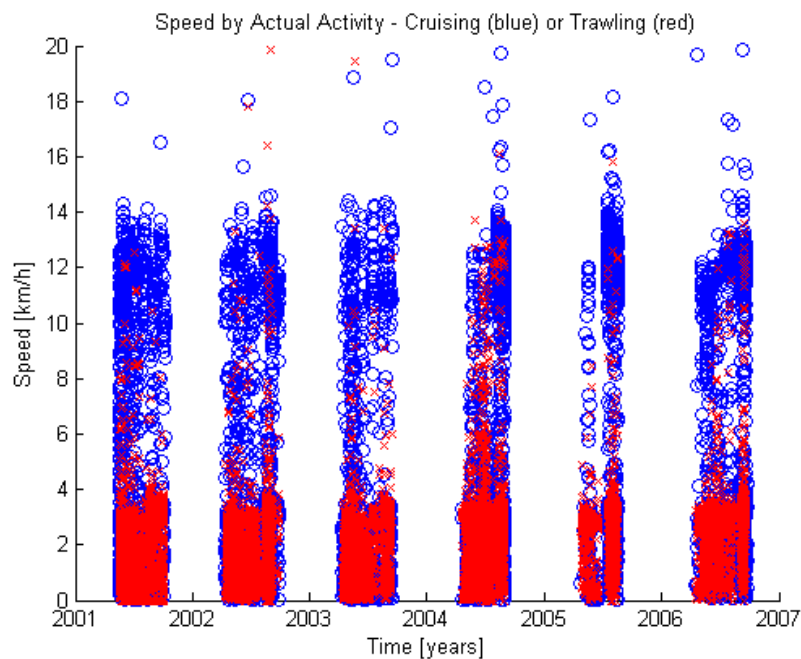


Figure 24: Speed by actual activity class

To find the optimal decision boundary requires a search of the solution space for a combination of parameters that minimise the misclassification rate. Since the solution space is small (the speed interval of about 0-8 knots) a simple linear brute-force search algorithm can be employed to search the whole space down to a precision of 0,1 knots.

With Matlab (see script 6, `calibrate_simple`), it is quickly found that the optimal speed is slightly lower than the initial assumption of 5 knots, or 4,4 knots. This lowers the total error rate given by the script only slightly, from 25,43% down to 25,36%, so the initial parameter was not very far off from the optimum.

Running the classifier again with the new parameter settings gives the more detailed results in Table 6, now with a decision boundary of mean leg speed at 4,4 knots. Comparing with the previous results in Table 5, the change in decision boundary speed slightly improves the classification of cruising legs (28,2% correct, compared to 27,0% before), but lowers the success with trawling legs (from 97,6% to 96,5%). This is as expected from the preceding discussion of the decision boundary. Also, incorrectly classifying the proportionately few cruising legs has less of an impact than incorrectly classifying the more numerous trawling legs. Keep in mind that the optimisation is moving the decision boundary through a mix of cruising and trawling datapoints.

Table 6: Classification results – Simple linear discriminant at 4,4 knots

Actual activity	Result	Total count	Proportion of total	Class count 2001-2005 Training set	Proportion of class	Class count 2006 Validation set	Proportion of class
Cruising	Correct	2.596	9,5%	2.093	28,2%	503	33,4%
Cruising	False	6.321	23,2%	5.318	71,8%	1.003	66,6%
Trawling	Correct	17.425	63,8%	14.984	97,1%	2.441	96,5%
Trawling	False	533	2,0%	445	2,9%	88	3,5%
Stopped	Correct	358	1,3%	329	83,3%	29	90,6%
Stopped	False	69	0,3%	66	16,70%	3	9,4%
Total	Correct	20.379	74,6%	17.406	74,9%	2.973	73,1%
Total	False	6.923	25,4%	5.829	25,1%	1.094	26,9%
Total	All	27.302	100%	23.235	100%	4.067	100%

4.2.5 Adding more predictor variables

It is now time to consider if the simple classifier of the preceding sections can be improved with the addition of more predictor variables. Some measures that are likely to be helpful in the classification are listed below, and the reasoning behind each is given. Including the vessel mean leg speed from the previous sections the variables are:

- *Mean leg speed*
- *Derivative of mean leg speed*
- *Course change*
- *Running average of course change*
- *Distance from last VMS position*
- *Running average of distance from last VMS position*
- *Distance to closest vessel*
- *Average distance to other vessels*

The first two measures reflect speed and changes in speed with the next two reflecting course changes. Then two measures monitor the density of the VMS positions, and lastly two measures monitor the fleet density or dispersal, for a total of 8 dimensions.

First is the *mean leg speed*, which has been dealt with extensively in the preceding sections and will not be discussed further here.

The reasoning behind trying the *derivative of mean leg speed* is that the vessel will

decrease speed before deploying its gear, and when retrieving it come to an almost complete stop before speeding up again. In other words, the suggestion is that both the current speed and the change in speed (acceleration) are predictors of the vessel behaviour. Even though this type of behaviour cannot be distinguished in detail at a resolution of 2 hrs, the effect will still be one of reducing the mean leg speed from the previous leg, thus giving a chance of detection.

When fishing, vessels tend towards much more extreme *course changes* than when cruising towards a specific goal. If the vessel has come about 180° for example, it might be reasonable to assume that it is repeatedly trawling the same area. If the vessel has changed course several times during the last few legs, it is even more likely this is the case. Therefore, the *course change* from the last leg and *running average of course change* for the last five course changes are possible predictor variables.

If a vessel were to cruise at low speed between fishing grounds, or in a relatively straight line while processing the catch between hauls, monitoring speed alone would not differentiate this behaviour from fishing activity. But when trawling, the vessel will stay in a more confined area and thus the VMS positions will be bunched closer together. So the VMS position density, or *distance from last VMS position* is examined as a predictor, as well as the *running average of distance from last VMS position* in order to catch slower trends and reduce the impact of single VMS points.

Finally, note that the vessel will seldom be alone fishing, but rather in close proximity to other vessels while trawling, but apart from the group when cruising. Thus two measures can be added to reflect the fleet dispersal, *distance to the closest vessel* and *average distance to the fleet* as a whole.

There are essentially two ways to proceed from here; *variable transformation* or *variable selection*.

The first approach seeks to transform the dataset into a more manageable number of derived variables, before letting loose with a classifier. The latter approach is simply to choose a subset of the predictor variables that gives the best classifier performance, and dropping those that yield no significant improvement. Both approaches are presented below for the sake of thoroughness.

Variable transformation with Principal Components

The Principal Components Analysis technique⁵ can best be thought of as rotating the data in all dimensions, trying to find the orientations that show *maximum variance* from that perspective, giving the best representation of the data. These vectors are called the *principal components*.

As an intermediary step before transforming the dataset into the axis of the principal components, they can be ranked by how much of the variance they explain to examine what dimensions compose each of the most important ones. When the principal components that explain e.g. 95% of the dataset variance are composed overwhelmingly of a subset of the dimensions, the dataset can actually be transformed using only these dimensions and the rest dropped. This is in effect a form of lossy compression technique,

5 For a discussion on Principal Component Analysis, see e.g. *Principles of Data Mining* (Hand et al. 2001)

since some of the information in the original dataset will be lost.

It is important to note that PCA is not necessarily well-suited to finding the best predictor variables, but rather how to best represent the data and transforming it to a lower-dimensional space (Duda et al. 2001). In some cases, it may actually give vectors that are orthogonal to the best linear discriminants. However, the dimensions chosen in this way may give an indication of what measures to use in classification, and is therefore interesting for the sake of comparison.

Note also that there is a difference between PCA results derived from the correlation matrix and the covariance matrix. For analysis of data with dimensions with different units of measure (as in this case) or very different variability sizes, the correlation matrix should form the basis of the PCA, since this is equivalent to normalising the data before the analysis.

Running Principal Components Analysis (see script 7, *classifier_analysis*) on the dataset indicates (using the correlation matrix) that at least seven principal components are needed to represent 96,7% of the variance in the data. This combination of the principal components is composed of all the dimensions (a smaller subset of the dimensions that explains the data adequately cannot be chosen), and the PCA is thus not helpful in determining which variables are more interesting than others.

Table 7 shows the principal components (eigenvectors) in order of importance, and the relative explanatory power of each principal component (percentages of variance explained by each respective eigenvalue) is illustrated in the scree⁶ plot in Figure 25. The table shows the principal components vectors from correlation coefficients representing 96,7% of the dataset variance with the vectors ranked in order of significance, that is how much of the dataset variance they explain. Note that none of the elements in the vectors are notably most influential (close to $\pm 1,0$, with the others close to 0,0).

Table 7: Principal components vectors from correlation coefficients

V1 28,6%	V2 17,4%	V3 13,7%	V4 12,6%	V5 11,1%	V6 8,2%	V7 5,1%	Dimension
0,4366	-0,0779	0,2198	-0,1386	-0,2178	0,8096	-0,09	speed
0,0282	-0,0952	0,4707	-0,7624	0,3958	-0,1709	-0,0133	speed_derivative
-0,3811	0,4553	0,3038	-0,0809	-0,2975	0,1732	0,6383	course_change
-0,4542	0,4152	0,2033	-0,0371	-0,1588	0,0547	-0,6993	course_change_running_average
0,4651	0,5004	0,0199	0,0045	-0,0454	-0,1139	-0,2384	distance_from_last
0,461	0,4845	-0,0133	-0,0019	0,0322	-0,2968	0,1858	distance_running_average
-0,0191	-0,0196	-0,5814	-0,6124	-0,5279	-0,0843	-0,0197	distance_closest_vessel
0,1667	-0,3459	0,5079	0,1286	-0,6328	-0,4173	-0,059	distance_fleet_average

6 The term “scree” means “a steep mass of detritus on the side of a mountain” and refers to the sudden leveling off of the plot, with the most important eigenvalues forming the “mountain” at the left, and the less important forming the rubble at the bottom right.

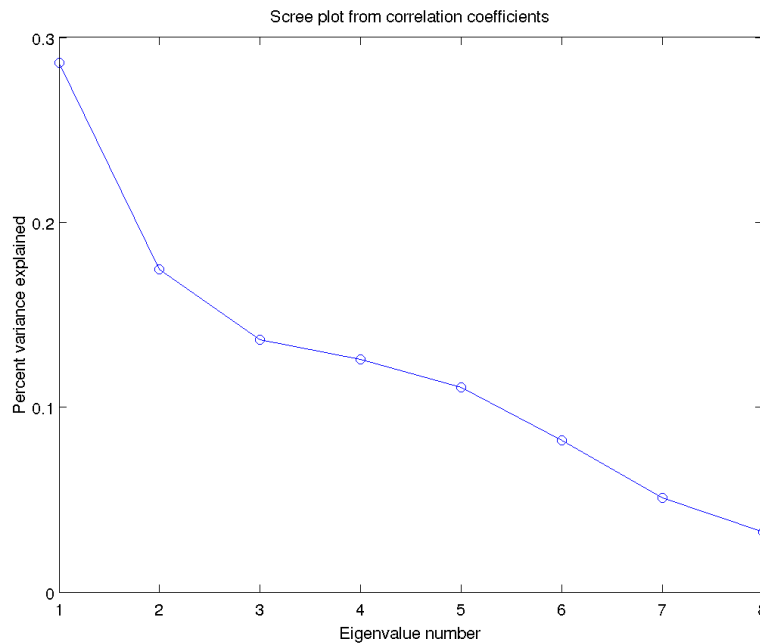


Figure 25: Scree plot of eigenvalues from the correlation coefficient matrix

Variable selection

Because of the relatively few predictor variables, it is actually preferable to take the direct approach and have the classifier try each combination in turn (for the 8 measures, there are just $2^8 - 1 = 255$ possible combinations that a classifier might use), evaluating the improvement in the classification error rate. This is simple to do using Matlab or an Open Source software package called WEKA⁷.

Using WEKA's *subset selection* algorithms with *exhaustive search* indicates that there is in fact no better combination of the predictor variables than using only vessel speed as input into the classifier (see printout 1, *weka_variable_selection*).

Similarly, using Fisher's Linear Discriminant Analysis (LDA, discussed in the next section) to test all the combinations and rank them by the error rate (see script 8, *classifier_loop_fisher_lda*), gives the same results. Using speed as the only variable is the best option and adding more variables does not increase the accuracy. The results are displayed in Table 8, showing success and error rates for the 10 best combinations of predictor variables using Fisher's LDA. Rates are calculated on the validation dataset.

⁷ <http://www.cs.waikato.ac.nz/ml/weka/> [Last retrieved september 2010]

Table 8: Predictor variable subsets with the ten lowest error rates using Fisher's LDA

Variables used	Success rate	Error rate
speed	72,4%	27,6%
speed, course_change	72,1%	27,9%
speed, course_change, course_change_running_average, distance_from_last, distance_fleet_average	72,1%	27,9%
speed, course_change, distance_from_last, distance_fleet_average	72,2%	27,8%
speed, course_change_running_average, distance_from_last, distance_fleet_average	72,2%	27,8%
speed, distance_from_last, distance_fleet_average	72,3%	27,7%
speed, course_change, distance_from_last	72,2%	27,8%
speed, course_change, course_change_running_average, distance_from_last	72,3%	27,7%
speed, course_change_running_average, distance_from_last	72,3%	27,7%
speed, distance_from_last	72,4%	27,6%

4.2.6 Alternative classifiers

To get some sense of how well the simple discriminant classifier in the preceding sections is performing, and in an attempt to improve upon the result, a comparison to some common classifiers is made:

- Fisher's Linear Discriminant
- Naïve Bayes classifier
- k-NN classifier
- CART classifier
- Multilayer Perceptron

Note that although a general description of the workings of each algorithm is given, an in-depth coverage of them is beyond the scope of this thesis. Interested readers are referred to e.g. *Principles of Pattern Classification* (Duda et al. 2001).

Fisher's Linear Discriminant

Linear Discriminant Analysis (LDA), or more specifically Fisher's Linear Discriminant⁸, is similar to Principal Components Analysis in that it rotates the data in the solution space, but seeks to find axis that are efficient for discrimination rather than representation of the data. It does this by maximising the distance between the class midpoints, and minimising the variance within each class.

Using Matlab (see script 9, `classifier_fisher_lda`), all the variants available are tried, settling on the *quadratic*, which fits multivariate normal densities to each group with

⁸ The terms *Fisher's linear discriminant* and *LDA* are often used interchangeably, although technically there are some differences.

covariance estimates for each, rather than a pooled estimate as in the *linear* variant.

Table 9 displays the results, which are just slightly worse than the simple linear discriminant. Notably, the Fisher linear discriminant fails to identify the *stopped* class altogether.

Table 9: Classification results – Fisher's Linear Discriminant

Actual activity	Result	Total count	Proportion of total	Class count 2001-2005 Training set	Proportion of class	Class count 2006 Validation set	Proportion of class
Cruising	Correct	2.593	9,5%	2.091	28,2%	502	33,3%
Cruising	False	6.324	23,2%	5.320	71,8%	1.004	66,7%
Trawling	Correct	17.428	63,8%	14.986	97,1%	2.442	96,6%
Trawling	False	530	1,9%	443	2,90%	87	3,4%
Stopped	Correct	0	0%	0	0%	0	0%
Stopped	False	427	1,6%	395	100%	32	100%
Total	Correct	20.021	73,3%	17.077	73,5%	2.944	72,4%
Total	False	7.281	26,7%	6.158	26,5%	1.123	27,6%
Total	All	27.302	100%	23.235	100%	4.067	100%

Naïve Bayes classifier

This classifier is based on Bayes' theorem of conditional probability, expressing the (posterior) probability a data point belongs to a particular class in terms of the (prior) probabilities of observing the datapoint and it belonging to the class, and the conditional probability of observing this point, given that it actually belongs to the class. The “naïveté” of the classifier stems from the strong assumption that the predictor variables are independent. In other words, the presence of a particular feature independently contributes to the probability that the object belongs to a particular class. Despite this, these classifiers can perform quite well on real-world problems, and can easily handle a high dimensionality of inputs.

Although some of the predictor variables are clearly not independent, this does not exclude the use of the Naïve Bayes classifier. One reason is that the absolute values of the posterior probabilities are not of interest, only their ranked order (to which class is the point most likely to belong). Thus some bias in the values can be tolerated.

Using WEKA, the Naïve Bayes classifier was run for all the predictor variables, as well as for speed only (see printout 2, `weka_naive_bayes`). The results are displayed in Table 10. As is evident, using all the variables results in a terrible performance (23,4% success rate), while speed alone gives comparable accuracy to, or only slightly better than the simple linear discriminant.

k-Nearest Neighbour classifier

The k-NN classifier puts each datapoint in the same class as the majority of its k nearest neighbours, where k is a user-defined constant.

Using Matlab (see script 10, `classifier_knn`) and iterating for $k=1\dots5$, this turns out to be by far the slowest classifier tried in this work.

The results are displayed in Table 10. The best performance is worse than for the other classifiers, peaking at 69,4% success for the speed variable and $k=4$.

CART classifier

CART is an acronym for *Classification And Regression Trees*, and is a binary-tree algorithm. It works by creating rules (decision boundaries) for each variable, splitting datapoints into two. At the next level of the tree the datapoints are split again, using another rule, and so on down the tree.

WEKA can easily run CART (see printout 3, `weka_cart`) and gives the resulting decision tree. As before, all variables were tried as well as speed only.

The rule-trees can be seen in Figures 26 and 27. They are interesting to compare with the linear classifiers on speed, since they could be construed as simpler versions of the CART with a more heavily pruned rule-tree.

Figure 26 shows part of a CART binary-tree pruned from all predictor variables using WEKA. The full tree is 101 nodes (rules). In this CART tree with all variables present, the first rule is indeed based on *mean leg speed*, although with slightly different speed boundaries. The next variables used are the *distance to closest vessel*, *distance from last VMS position*, and the *speed derivative*.

The CART algorithm sets the decision boundaries for cruising/trawling at 4,75 knots, and for trawling/stopped at 0,0003 knots, not too far off from the linear classifiers.

```

speed < 4.75277736034
| speed < 0.02717077489555
| | distance_closest_vessel < 22.1854433045355
| | | speed < 3.05960787782E-4: stopped(227.0/173.0)
| | | speed >= 3.05960787782E-4
| | | | speed_derivative < -0.650667025934865:
cruising(13.0/7.0)
| | | | speed_derivative >=
-0.650667025934865:trawling(24.0/14.0)
| | | | distance_closest_vessel >= 22.1854433045355:
stopped(84.0/20.0)
| | speed >= 0.02717077489555
| | | distance_from_last < 20.030060475162
| | | | course_change_running_average < 37.9
| | | | | distance_fleet_average < 160.54285064795
| | | | | speed_derivative < -1.788458335731:
cruising(25.0/15.0)
| | | | | speed_derivative >= -1.788458335731
| | | | | | speed < 3.601921508385
| | | | | | | distance_fleet_average < 4.08810454913605

```

Figure 26: Partial CART classification tree for all variables

For comparison, figure 27 shows a CART binary-tree pruned from the speed variable only using WEKA.

The results of the classifier are displayed in Table 10.

```

speed < 4.75277736034
| speed < 0.02717077489555
| | speed < 3.05960787782E-4: stopped(305.0/188.0)
| | speed >= 3.05960787782E-4: trawling(29.0/40.0)
| speed >= 0.02717077489555: trawling(17273.0/6424.0)
speed >= 4.75277736034: cruising(2533.0/510.0)

```

Figure 27: CART classification tree for speed

Multilayer Perceptron

The MLP is a simplified model of a biological neuron, with “neurons” organized in multiple layers, which are fully interconnected to the next layers on either side. Each neuron then incorporates a non-linear sigmoid activation function, which determines if it fires upon receiving input from the previous layer.

WEKA can run MLP (see printout 4, `weka_mlp`). The results are in Table 10.

4.2.7 Cluster analysis

The classification algorithms discussed so far are supervised, i.e. their results can be compared to what is actually known to be true and improvements made (such as identifying the speed limit that minimises classification error).

As previously mentioned, there exists a second type of classifiers; the unsupervised. Unsupervised classification algorithms are also called *clustering* algorithms, since their aim is to identify “clusters” in the datasets without knowing what those clusters actually represent in reality.

One reason the results of such algorithms are interesting is that they can give an idea of the features of the datasets, and thus what variables might be of most interest to use in classification. This is actually what principal components analysis attempts to do, so one unsupervised approach has in fact already been presented, in trying to do identify variables for the supervised classification.

Also, in using the supervised classification approach, an educated guess was made as to what the vessel *actual activity* at each point is, and the results of the classification based on this assumption must be confirmed.

The clustering algorithm will attempt to find natural groupings in the dataset. Since the guess of actual class is available from the *midpoint* algorithm, an error rate based on this is calculated. If the supervised algorithms used earlier are indeed delivering the same classification, they should result in comparable error rates as the unsupervised.

As one of the more popular unsupervised clustering algorithms, the *k-means* algorithm was used to compare with the previous results.

k-Means clustering

In *k-means* clustering, the datapoints are iteratively sorted into *k* number of clusters. Each datapoint is placed into the cluster that has the closest mean value. The means for each cluster are then recalculated, and the algorithm reiterated. When no datapoints change clusters between iterations, the loop is stopped.

Using WEKA (see printout 5, `weka_kmeans`), some of the most promising variable combinations were tried. The results were also compared to the guess of actual class values. The results are displayed in Table 10.

An error rate of 27,0% (or 73,0% success rate) was reached, a comparable rate to the supervised classifiers, and confirms that they are not missing some other groupings in the dataset because of the guess at the true activity.

4.2.8 Summary

Several classifiers were built and tested to identify fishing behaviour from vessel tracks. Different measures were examined and how combinations of these affect the classification success.

Comparison of the classifier performance is displayed in Table 10, showing success and error rates for the tested classifiers. Rates are calculated on the validation dataset. The best performance is very comparable for most, from 73,1% to 75,0%, with Naïve Bayes on all variables scoring worst by far (23,4% success rate).

Table 10: Comparison of classifier performance

Classifier	Variables used	Success rate
Simple Linear Discriminant	speed	73,1%
Fishers' LDA	speed	73,5%
Naïve Bayes	speed	74,1%
Naïve Bayes ⁹	all	23,4%
k-NN	speed, k=1	62,5%
k-NN	speed, k=2	68,5%
k-NN	speed, k=3	66,4%
k-NN	speed, k=4	69,4%
k-NN	speed, k=5	67,9%
k-NN	all, k=4	68,6%
CART	speed	74,4%
CART	all	75,5%
Multilayer Perceptron	speed	74,2%
Multilayer Perceptron	all	74,5%
k-means clustering	all	53,4%
	speed	
k-means clustering	speed_derivative	72,2%
	distance_closest_vessel	
	speed	
k-means clustering	speed_derivative	73,0%
	speed	
k-means clustering	speed	73,0%
	speed	
k-means clustering	distance_from_last	73,0%

The final choice of classifier should be a combination of the best performance, fewest necessary variables and simplest implementation. In the subsequent chapters the simple linear discriminant algorithm developed first will be used, since it does almost as good as the CART, MLP and Naïve Bayes, but is much simpler and easier to integrate into the prototype system code.

Also, the classification was compared to unsupervised clustering results, with no significant difference between the approaches being found, justifying the assignment of *actual activity* using the *midpoint* algorithm.

⁹ The numbers for this application of the Naïve Bayes have not been interchanged by mistake.

Four factors prevent an improvement in the classification; relatively coarse resolution of vessel tracks, lack of data on the vessel's actual activity at all times, imprecise positions of catch reports, and the fact that below the decision boundary of 4,4 knots, trawling and cruising legs tend to mix up, and no measures have been identified that can reliably differentiate between the two.

4.3 Estimating effort from classification

As discussed before, having classified the vessel activities during each of the trip legs, it should now be straight forward to estimate the effort. For each leg classified as “trawling”, its duration is added to the total effort during the trip. The accuracy is dependent on the resolution of the VMS data, i.e. the leg duration, which is most commonly 2 hrs.

Note that the objective here is not to predict the fishing effort itself, since that is a rather loosely defined concept, but rather the landed catch. The construction a full regression model for the effort is therefore not within the scope of this study. The effort measure is however used as an input into a catch prediction model, whose regression constants will take the effort variability and scale into account.

4.3.1 Comparison to reported effort

The source datasets include a measure of *reported effort* (also known as *linetime*), which is the time the vessel spent with its gear deployed (i.e. trawling), as reported in the catch logbook. The following analysis will include how well the *estimated effort* corresponds to this measure, but it is important to keep in mind that the classification algorithms are not designed to minimise this error, but rather to correctly classify each leg of the vessel VMS-track. If they were optimised to minimise the error between the *reported effort* and the *estimated effort* directly, the classification algorithms would respond by forcing more legs into a different class, even if they clearly did not belong there. The proper way of handling this error is in the effort or catch estimation models.

Examining the errors in effort estimation from this reported effort as shown in Figure 28 below, the algorithm tends to overestimate the effort using the classifier by the direct addition of intervals classified as trawling. The conclusion is that the effort estimate most likely includes time that the vessel spends on fishing-related activity that is not strictly trawling and which is not included in the reported effort. This time might be termed *turnaround*, the time spent readying the vessel for a second deployment of the gear.

Using the effort estimate in comparison with other studies must then be done with this difference in mind, and noted that it is incompatible with official statistics using the reported effort.

In Figure 28, showing $(\text{estimated effort} - \text{reported effort}) / \text{reported effort}$, not all years display the same error proportion, notably 2005 has much lower error rates than other years. This stems from differences in the reported effort vs. reported catch as seen in the following Figure 29, which shows reported effort vs. reported catch for each trip, all years. The datapoints for 2005 are plotted in green, and 2006 in red for comparison. Note the apparent difference in catch per unit effort (slope) of the two years. where reported effort for 2005 is significantly greater (the fishing in 2005 was notoriously bad, with great effort needed to catch the redfish, thus the decrease in proportional error), and problems with the source data where the reported effort and catch periods do not coincide with any of the supplied VMS-tracks.

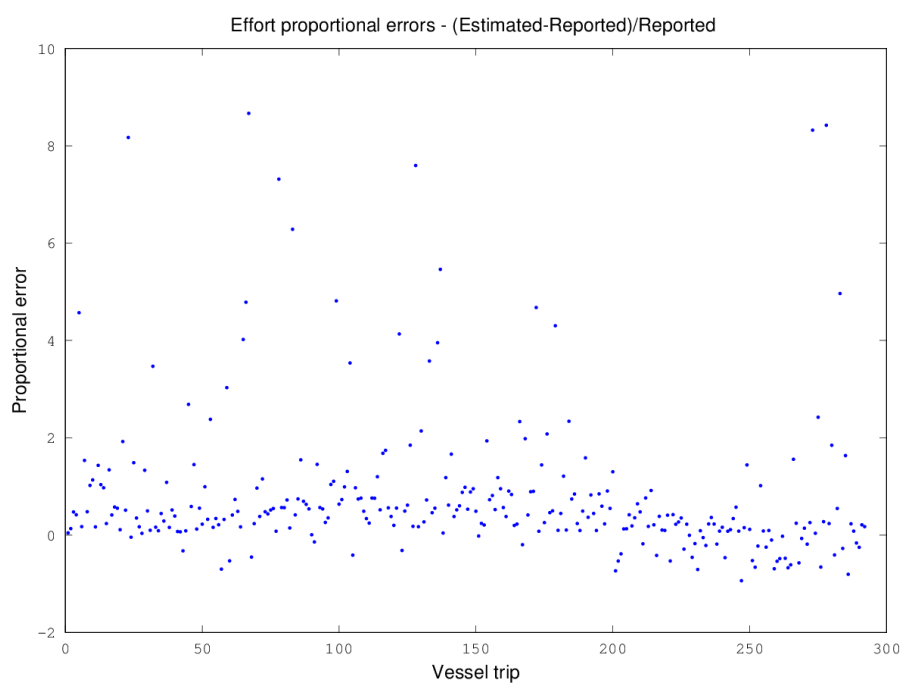


Figure 28: Trip effort proportional errors

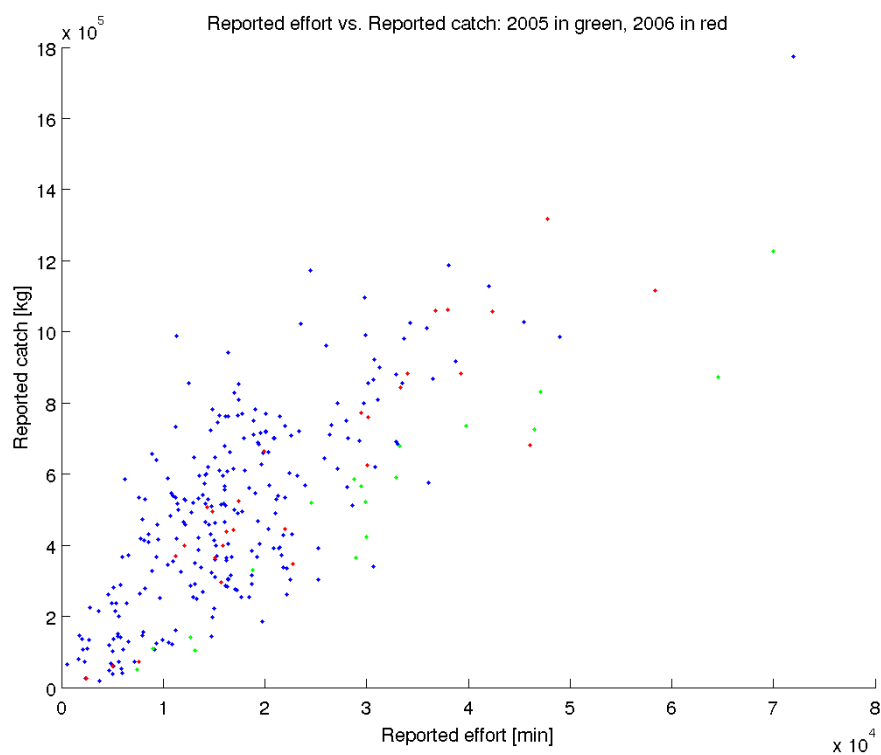


Figure 29: Trip reported effort vs. reported catch

4.4 Estimating catch from effort

In this section:

- General approach and terms
- The true effort, errors and bias
- Variance of catch model
- Least squares formulation in matrix form
- Maximum likelihood estimation of α and σ^2_ϵ
- Unequal variances in estimated effort U_{jk} and reported effort V_{jk}
- Standardised residuals and treatment of outliers
- Student's-t distribution
- Curvilinear model
- Prediction intervals on new observations with a fitted year-effect
- Prediction intervals on new observations with an unknown year-effect
- Usage examples for the final model

This section opens with definitions of some terms essential for the development of a statistical model for catch. The differences between *true effort*, *estimated effort* and *reported effort* are discussed and how the models correct for bias in these measures.

A multivariate linear regression (MLR) model is presented and its parameters estimated with the maximum-likelihood approach. The model is refined with more robust treatment of errors in estimated and reported effort, followed by an analysis of outliers.

The use of the log-Student's-t distribution rather than the log-normal distribution is explored, and then shown that an extended model that assumes a curvilinear relationship between effort and catch yields a better fit to the data.

The prediction intervals based on the model are formulated, and the estimation of parameters for the validation dataset is discussed.

Finally, simple examples on how the final model can be used on the datasets are presented.

4.4.1 General approach and terms

This section describes the second main research objective, i.e. predicting the landed catch of a vessel from each trip, given the previously calculated effort estimate from the preceding sections.

A *regression model* is constructed based on the supplied data, and as before part of the data is used for model validation.

Before starting the analysis, the definition of some key concepts and variables that will be used extensively in this section follows:

Reported catch: the total catch from one trip as reported in the landing report. This data can be considered very accurate, since it is the result of an independent weighing of the landed catch when the vessel comes to port.

Reported effort: the total effort from one trip as reported in the catch logbook. This is the recorded trawling time, and can be subject to errors. Usually it is recorded to within the hour, but can also be accurate to within 15 minutes.

Estimated catch: the estimated catch as calculated by the model, based on a supplied effort. This is essentially the prediction the statistical models deliver for some given effort.

Estimated effort: the total effort from one trip as estimated by the classifier algorithm. This effort is usually higher than the reported effort, since the classification algorithms tend to include time periods when the vessel is hauling or setting gear or performing other activities that cannot be distinguished from actual fishing activity by the algorithm.

Days at sea: the total effort from one trip as calculated from the total time spent at sea. This effort is higher than either the reported effort or the estimated effort from the classifier, since it includes time periods when the vessel is cruising or performing other non-fishing activities.

True effort: the real-world effort involved, without measurement error. This variable is a statistical entity, and can never be directly observed.

The measures are defined as:

$$\begin{aligned} U_{jk} &= \text{estimated effort for trip } j \text{ of vessel } k \\ V_{jk} &= \text{reported effort for trip } j \text{ of vessel } k \\ X_{jk} &= \text{true effort for trip } j \text{ of vessel } k \\ W_{jk} &= \text{true catch for trip } j \text{ of vessel } k \\ D_{jk} &= \text{days at sea for trip } j \text{ of vessel } k \end{aligned}$$

Referring to the very beginning of this chapter, the model will assume a linear relationship between effort and catch (see Figure 15), and as a starting point it is reasonable to assume that the *true catch*, dependent upon the *true effort* is such that the expected value and variance are

$$\begin{aligned} E[W_{jk}] &= X_{jk} \mu_{jk} \\ \text{var}(W_{jk}) &= \alpha X_{jk} \mu_{jk}^2 \end{aligned}$$

where μ_{jk} is a measure of the catching power (*catch per unit effort, CPUE*) during trip j of vessel k , and α is a variance (or scaling) parameter.

4.4.2 The true effort, errors and bias

The fact is that the measure of *true effort* in the model is unknown. However, there are two measures of the effort that both include some unknown measurement error, namely the *reported effort* and the *estimated effort* (writing the errors in exponential form here makes the subsequent notation easier to work with)

$$\begin{aligned} U_{jk} &= X_{jk} e^{\epsilon_{jk}} \\ V_{jk} &= X_{jk} e^{e_{jk}} \end{aligned}$$

or taking the natural logarithm

$$\log U_{jk} = \log X_{jk} + \varepsilon_{jk}$$

$$\log V_{jk} = \log X_{jk} + e_{jk}$$

Using these measures together can give some idea of the true effort, and plotting both on a log-log scale results in Figure 30.

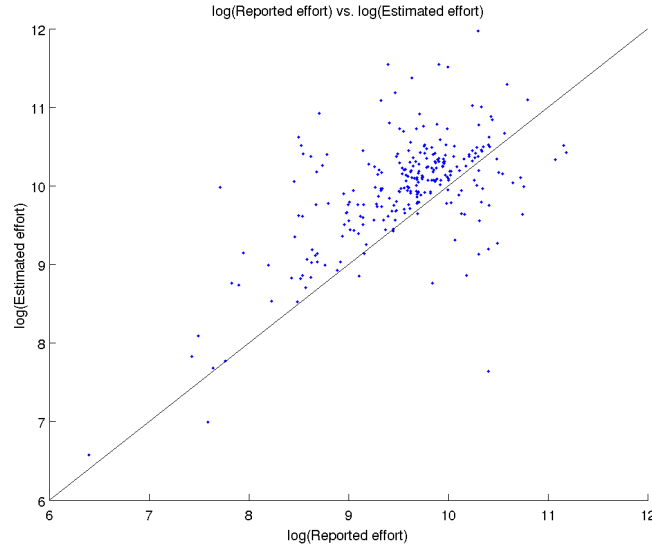


Figure 30: *log of the estimated effort vs. log of the reported effort for each trip*

Each datapoint on this plot has uncertainty in both dimensions, resulting in a circular or elliptical area where the true value is believed to be, as shown in Figure 31. The figure shows an example of a datapoint from the previous graph, plotted with errors. ε_{jk} is the error in estimated effort, $\log(U_{jk})$, e_{jk} is the error in reported effort, $\log(V_{jk})$ and d_{jk} is the shortest distance to the true effort on the straight line for trip j and vessel k .

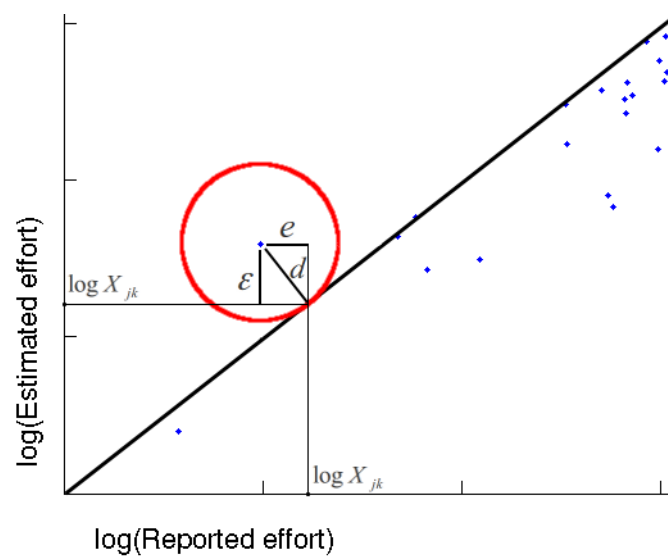


Figure 31: *Example of a datapoint with uncertainty*

In this figure the *true effort* actually lies on the straight line where $\log(V_{jk}) = \log(U_{jk})$ (if both measures had no error in them, they would in fact be measuring the true effort, and result in exactly the same value).

Assuming the errors have equal variances in both dimensions (see Section 4.4.6 for treatment of unequal errors), simple geometry gives the true effort by using the orthogonal projection of the point onto the straight line $\log(V_{jk}) = \log(U_{jk})$ (this is the shortest distance between the point and the line). If $Q1$ and $Q2$ are points on the line and P is the datapoint, the errors are then¹⁰

$$\varepsilon_{jk} = e_{jk} = \frac{|(Q2 - Q1, P - Q1)|}{\|(Q2 - Q1)\|}$$

Looking at the data in Figure 30, a slight bias towards larger values of *estimated effort* can be detected. This is to be expected, since the measurement is built from classification of track legs, and the classifier tends to incorporate cruising activity as well as trawling, as described in earlier sections. *Estimated effort* thus includes all parts of the fishing activity detected from changes in vessel speed, while the *reported effort* only measures the time which the vessel's gear is actually deployed.

This is evident by plotting the distance to the line as in Figure 32.

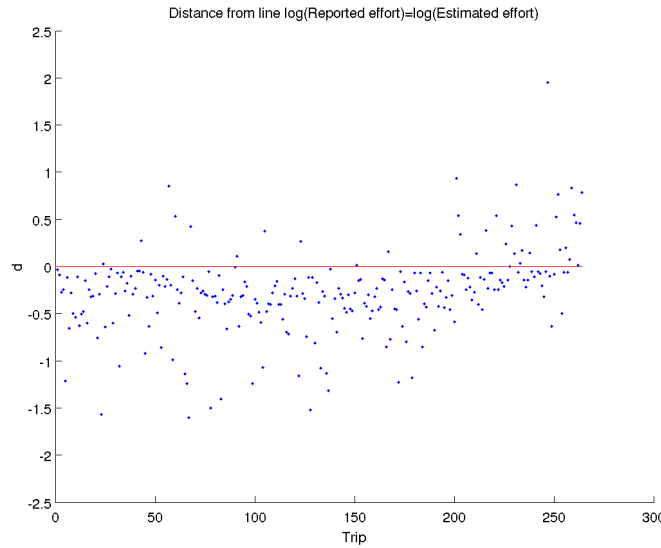


Figure 32: Distance d_{jk} from datapoints to the line $\log(V_{jk}) = \log(U_{jk})$

Based on the above arguments a *bias* term is added to the measures

$$\log U_{jk} = \log X_{jk} + \varepsilon_{jk} + \text{bias}$$

$$\log V_{jk} = \log X_{jk} + e_{jk}$$

$$\log U_{jk} - \log V_{jk} = \text{bias} + \varepsilon_{jk} - e_{jk}$$

¹⁰ Note that the vertical lines in the numerator signify the *determinant* and the double vertical lines in the denominator the *vector norm* or magnitude

A reasonable estimator of the *bias* is

$$\hat{bias} = \text{median}(\log U_{jk} - \log V_{jk})$$

Incorporating this into the calculations, Figures 30 and 32 are redrawn as Figures 33 and 34, taking the *bias* estimator into account

$$\begin{aligned} \log U_{jk} - \hat{bias} &= \log X_{jk} + \varepsilon_{jk} \\ \log V_{jk} &= \log X_{jk} + e_{jk} \end{aligned}$$

The *estimated effort* measures have now been corrected for bias, and the model can be designed based on this measure and the estimated bias.

The variances of *estimated effort* and *reported effort* are denoted by

$$\begin{aligned} \sigma_{\varepsilon}^2 &= \text{var}(\varepsilon_{jk}) \\ \sigma_e^2 &= \text{var}(e_{jk}) \end{aligned}$$

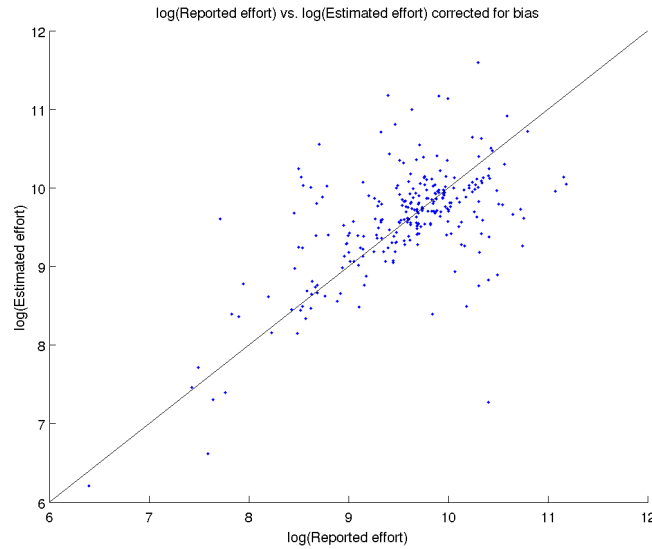


Figure 33: Corrected log of the estimated effort vs. log of the reported effort, with bias estimator

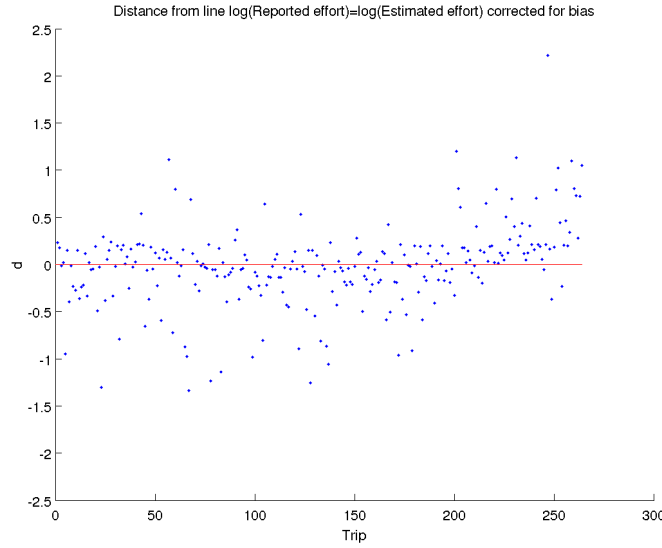


Figure 34: Distance from datapoints to the corrected line
 $\log(V_{jk}) = \log(U_{jk})$

4.4.3 Variance of catch model

The attention now turns to the variance in the catch. Based on inspection of the data in Figure 15 a variance model of the form

$$\text{var}(W_{jk}) = \alpha X_{jk} \mu_{jk}^2$$

is proposed.

Since the catch is positive and accumulates with higher effort, then a reasonable model is based on a log-normal assumption, that is

$$\log W_{jk} \sim N(\log(\mu_{jk} X_{jk}), \sigma_{00}^2)$$

The expected value is

$$E(W_{jk}) = (\mu_{jk} X_{jk}) e^{\sigma_{00}^2/2}$$

and the variance can be derived with a little algebra

$$\begin{aligned} \text{var}(W_{jk}) &= e^{2 \log(\mu_{jk} X_{jk})} [e^{2 \sigma_{00}^2} - e^{\sigma_{00}^2}] \\ &= (\mu_{jk} X_{jk})^2 e^{\sigma_{00}^2} [e^{\sigma_{00}^2} - 1] \\ &= X_{jk} \mu_{jk}^2 X_{jk} [e^{2 \sigma_{00}^2} - e^{\sigma_{00}^2}] \end{aligned}$$

To solve for σ_{00}^2 , the following substitutions are made

$$\begin{aligned} \alpha &= X_{jk} [e^{2 \sigma_{00}^2} - e^{\sigma_{00}^2}] \\ z &= e^{\sigma_{00}^2} \\ z^2 &= e^{2 \sigma_{00}^2} \end{aligned}$$

to write

$$\alpha \frac{1}{X_{jk}} = [e^{2\sigma_{00}^2} - e^{\sigma_{00}^2}]$$

$$z^2 - z - \frac{\alpha}{X_{jk}} = 0$$

and so

$$z = \frac{1}{2} \left(1 \pm \sqrt{1 + 4 \frac{\alpha}{X_{jk}}} \right)$$

or

$$\sigma_{00}^2 = \log \left(\frac{1}{2} \left(1 \pm \sqrt{1 + 4 \frac{\alpha}{X_{jk}}} \right) \right)$$

Restating the model in these terms, let $Z_{jk} \sim N(0,1)$ be a (standard) normally distributed random variable. Then

$$\begin{aligned} \log W_{jk} &= \log(\mu_{jk} X_{jk}) + \sigma_{00} Z_{jk} \\ &= \log \mu_{jk} + \log X_{jk} + \sigma_{00} Z_{jk} \end{aligned}$$

Substituting $\log X_{jk}$ with $\log U_{jk} - \hat{bias} + \varepsilon_{jk}$ yields

$$\begin{aligned} \log W_{jk} &= \log \mu_{jk} + (\log U_{jk} - \hat{bias} + \varepsilon_{jk}) + \sigma_{00} Z_{jk} \\ &= \log U_{jk} - \hat{bias} + \log \mu_{jk} + \varepsilon_{jk} + \sigma_{00} Z_{jk} \end{aligned}$$

Now the model can be presented as

$$\log W_{jk} \sim N(\log U_{jk} - \hat{bias} + \log(\mu_{jk}), \sigma_{\varepsilon}^2 + \sigma_{00}^2) \quad (1)$$

Some potential models for $\log \mu_{jk}$ are

$$\begin{aligned} \log(\mu_{jk}) &= \mu_0 && (constant) \\ \log(\mu_{jk}) &= \mu_k && k=1...K \\ \log(\mu_{jk}) &= \mu_k + \beta_t && k=1...K, t=1...T \end{aligned}$$

The first model corresponds to all the vessels having the same catching power.

In the second model, each vessel k has its own catching power independent of the others.

In the third model, the catching power is composed of the k individual vessels catching power, and a year effect for each year t . This model can accommodate the intuitive notion that in some years the fishing is simply better than in others, which is confirmed when plotting effort and catch for individual years, such as in Figure 29, Section 4.3.1.

The third model will be used, in order to take the year effect into consideration, or

$$\log W_{jk} \sim N(\log U_{jk} - \hat{bias} + \mu_k + \beta_t, \sigma_{\varepsilon}^2 + \sigma_{00}^2)$$

with the expected value of

$$E[W_{jk}] = U_{jk} e^{\hat{bias}} e^{\mu_k} e^{\beta_t} e^{(\sigma_{\varepsilon}^2 + \sigma_{00}^2)/2}$$

4.4.4 Least squares formulation in matrix form

Now that a reasonable model has been set up, its parameters can be estimated based on the data. This is done by fitting the model to the data using *least squares*.

In matrix notation, the linear regression model is

$$Y = Z\beta + \varepsilon$$

where Z is termed the design matrix, β is the parameter vector and ε are the observation errors. The elements y_i of the Y vector are calculated from Equation (1) for each vessel k and trip n in year t

$$y_{jk} = \log(W_{jk}) - \log(U_{jk}) + \hat{bias}$$

Each vessel and year combination can thus have multiple rows corresponding to the number of trips of the vessel in that year. Let N denote the total number of measurements.

The errors are assumed to be independent with unequal variance, and follow a normal distribution

$$\varepsilon \sim N(0, \Sigma_\varepsilon)$$

where Σ_ε is a diagonal matrix of the model variances

$$\Sigma_\varepsilon = \text{diag}(\hat{\sigma}_\varepsilon^2 + \sigma_{00}^2(\alpha, X_{jk}), \dots, \hat{\sigma}_\varepsilon^2 + \sigma_{00}^2(\alpha, X_{jk}))$$

Then a least squares estimator of the parameter vector is (Gelman et al. 2004, p. 374)

$$\hat{\beta} = (Z^T \Sigma_\varepsilon^{-1} Z)^{-1} Z^T \Sigma_\varepsilon^{-1} Y$$

and the covariance matrix is (Gelman et al. 2004, p. 374)

$$\text{cov}(\hat{\beta}) = (Z^T \Sigma_\varepsilon^{-1} Z)^{-1}$$

The design matrix Z is constructed so that the columns correspond to the respective vessels, followed by columns for each year, excluding one reference year (this is to avoid calculation problems in making the matrix singular. 2005 is selected as the reference year). A row then references a particular vessel and year if the value in the respective columns is 1 (otherwise 0).

The parameter vector β is similarly constructed, except that the parameters μ_k for the vessels are organised in rows from top to bottom, followed by β_t parameters for years.

A 95% confidence interval for the parameters is calculated as

$$\hat{\beta} \pm 1.96 \sqrt{(\text{diag}(Z^T \Sigma_\varepsilon^{-1} Z)^{-1})}$$

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \vdots \\ y_N \end{bmatrix}}_{\mathbf{Y}} = \underbrace{\begin{bmatrix} \text{vessel } 1 & \dots & \text{vessel } K & \text{year } 1 & \dots & \text{year } T \\ \hline 1 & 0 & & & & \\ 1 & 0 & & & & \\ 0 & 0 & \dots & & & \\ 0 & 0 & & 1 & \dots & \\ 0 & 0 & & 1 & & \\ \vdots & & & 0 & & \end{bmatrix}}_{\mathbf{Z}} \underbrace{\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_K \\ \beta_1 \\ \vdots \\ \beta_T \end{bmatrix}}_{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}$$

Figure 35: Matrix construction for the linear regression model

See Figure 35 for reference, which shows the construction of the linear regression model in matrix form. The elements of the \mathbf{Y} vector are calculated in Equation (1) for a total of N measurements, the design matrix \mathbf{Z} is a matrix of ones and zeros, divided into columns corresponding to vessels $1 \dots K$, and years $1 \dots T$. The elements of the parameter vector $\boldsymbol{\beta}$ are unknown and need to be estimated, the μ_k elements representing vessels and β_i representing years (with respect to the reference year).

4.4.5 Maximum likelihood estimation of α and σ_ε^2

The above formulation depends on two meta-parameters α and σ_ε^2 which need to be estimated in conjunction with the model parameters $\boldsymbol{\beta}$ themselves. This is achieved with *maximum-likelihood estimation* or *profile likelihood estimation*, where the likelihood function of the parameters to be maximised (α and σ_ε^2) depends on a given value of the other parameters ($\boldsymbol{\beta}$).

The normal probability density function is of the form (Gelman et al. 2004, p. 574)

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}} [\sigma^2]^{-1/2} \exp\left(-\frac{1}{2} [\sigma^2]^{-1} (x - \mu)^2\right)$$

Substituting and writing for n observations, a likelihood function for α and σ_ε^2 is constructed as

$$\begin{aligned}
L(\alpha, \sigma_\varepsilon^2) &= \prod_{i=1}^n \left\{ \frac{1}{\sqrt{2\pi}} [\sigma_\varepsilon^2 + \sigma_{00}^2]^{-1/2} \right. \\
&\quad \left. \exp \left[-\frac{1}{2} [\sigma_\varepsilon^2 + \sigma_{00}^2]^{-1} (y_i - (Z\hat{\beta})_i)^2 \right] \right\} \\
&= \prod_{i=1}^n \left\{ \frac{1}{\sqrt{2\pi}} \left[\sigma_\varepsilon^2 + \log \left(\frac{1}{2} + \frac{1}{2} \sqrt{1 + 4 \frac{\alpha}{U_{jk}}} \right) \right]^{-1/2} \right. \\
&\quad \left. \exp \left[-\frac{1}{2} \left[\sigma_\varepsilon^2 + \log \left(\frac{1}{2} + \frac{1}{2} \sqrt{1 + 4 \frac{\alpha}{U_{jk}}} \right) \right]^{-1} (y_i - (Z\hat{\beta})_i)^2 \right] \right\}
\end{aligned}$$

where the *true effort* X_{jk} has been approximated by the *estimated effort* U_{jk} .

The true effort could have been replaced by $U_{jk} e^{-bias}$ due to the relationship between X_{jk} and U_{jk} , but this is not necessary since the term e^{-bias} would be multiplied by α .

It is actually more convenient to work with the logarithm of the likelihood function

$$\begin{aligned}
l(\alpha, \sigma_\varepsilon^2) &= \log(L(\alpha, \sigma_\varepsilon^2)) \\
&= -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^n \left\{ \log \left(\sigma_\varepsilon^2 + \log(\dots) - \frac{1}{2} \sum_{i=1}^n [\sigma_\varepsilon^2 + \log(\dots)]^{-1} (y_i - (Z\hat{\beta})_i)^2 \right) \right\}
\end{aligned}$$

The function `fminsearch` in Matlab is used to find the maximum-likelihood estimation following these steps (see script 11, calculations, and script 12, `logLikelihoodCeder`):

1. Select valid initial values for α and σ_ε^2 , call them α_0 and $\sigma_{\varepsilon 0}^2$
2. Estimate β and $\Sigma_{\varepsilon, ii}$ with
$$\begin{aligned}
\hat{\beta} &= (Z^T \hat{\Sigma}_\varepsilon^{-1} Z)^{-1} Z^T \hat{\Sigma}_\varepsilon^{-1} y \\
\hat{\Sigma}_{\varepsilon, ii}^{-1} &= \left[\hat{\sigma}_{\varepsilon 0}^2 + \log \left(\frac{1}{2} + \frac{1}{2} \sqrt{1 + 4 \frac{\hat{\alpha}_0}{U_{jk}}} \right) \right]^{-1} \\
\hat{\Sigma}_{\varepsilon, ij}^{-1} &= 0, i \neq j \\
y_{jk} &= \log(W_{jk}) - \log(U_{jk}) + \hat{bias}
\end{aligned}$$
3. Run `fminsearch` to find the values of α and σ_ε^2 that maximise $l(\alpha, \sigma_\varepsilon^2)$
4. If the values for α and σ_ε^2 have changed very little from last iteration, stop here
5. Else, use the resulting values α and σ_ε^2 and repeat from step 2

The resulting parameter estimates for each vessel μ_k and year β_i (with respect to the reference year 2005, the year 2006 is used as a validation dataset and is discussed in a later section on model validation) are detailed in Tables 11 and 12 and Figures 36 and 37 below.

For a discussion of these results, see Section 4.4.7 - Standardised residuals and treatment of outliers.

The meta-parameters were estimated as $\alpha = 1411,1$ and $\sigma_\varepsilon^2 = 0,31856$

Table 11: Estimated vessel parameters μ_k

Vessel	μ_k	95% confidence interval	
1	5,10	$\pm 1,40$	[3,71;6,50]
2	3,91	$\pm 0,52$	[3,39;4,44]
3	3,39	$\pm 0,40$	[2,99;3,79]
4	3,68	$\pm 0,48$	[3,21;4,16]
5	2,51	$\pm 0,40$	[2,11;2,90]
6	3,39	$\pm 0,47$	[2,92;3,85]
7	3,27	$\pm 0,42$	[2,85;3,69]
8	3,56	$\pm 0,39$	[3,17;3,94]
9	3,43	$\pm 0,47$	[2,96;3,89]
10	3,44	$\pm 0,45$	[3,00;3,89]
11	3,82	$\pm 0,42$	[3,39;4,24]
12	3,59	$\pm 0,58$	[3,02;4,17]
13	3,16	$\pm 1,40$	[1,76;4,57]
14	3,60	$\pm 0,41$	[3,18;4,01]
15	3,64	$\pm 0,42$	[3,22;4,05]
16	3,42	$\pm 0,43$	[2,99;3,85]
17	3,49	$\pm 0,50$	[2,99;4,00]
18	3,74	$\pm 0,63$	[3,11;4,37]
19	3,58	$\pm 0,38$	[3,20;3,96]
20	3,76	$\pm 0,41$	[3,35;4,17]
21	3,60	$\pm 0,62$	[2,98;4,22]
22	3,65	$\pm 0,52$	[3,12;4,17]

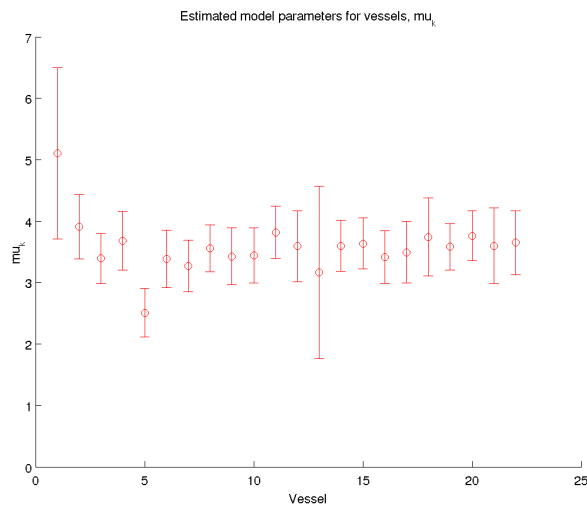
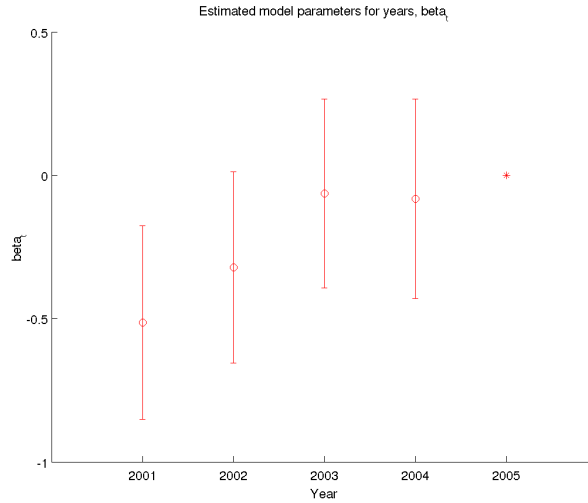


Figure 36: Estimated vessel parameters, μ_k

Table 12: Estimated year parameters β_t

Year	β_t	95% confidence interval	
2001	-0,51	0,34	[-0,85;-0,18]
2002	-0,32	0,33	[-0,66;0,01]
2003	-0,06	0,33	[-0,39;0,27]
2004	-0,08	0,35	[-0,43;0,27]
2005	0	Reference year	-

Figure 37: Estimated year parameters β_t

4.4.6 Unequal variances in estimated effort U_{jk} and reported effort V_{jk}

In the preceding section, the errors in Figure 31 were treated as equal, but this is perhaps not an accurate assumption. The above procedure is modified to take into account the possibility of unequal errors in the measurements. The same search algorithm is used to determine the maximum-likelihood estimator for the proportion between the errors.

As before,

$$\begin{aligned}
 \log U_{jk} - \hat{bias} &= \log X_{jk} + \varepsilon_{jk} \\
 \log V_{jk} &= \log X_{jk} + e_{jk} \\
 \text{var}(\varepsilon_{jk}) &= \sigma_\varepsilon^2, \quad \text{var}(e_{jk}) = \sigma_e^2
 \end{aligned}$$

Taking the difference

$$\begin{aligned}
 d_{jk} &= \log U_{jk} - \hat{bias} - \log V_{jk} \\
 &= \log X_{jk} + \varepsilon_{jk} - \log X_{jk} - e_{jk} \\
 &= \varepsilon_{jk} - e_{jk}
 \end{aligned}$$

so the joint variance is

$$\text{var}(\varepsilon_{jk} - e_{jk}) = \sigma_\varepsilon^2 + \sigma_e^2 = \sigma_{\varepsilon,e}^2$$

The proportion between the errors is defined as γ

$$\begin{aligned}\sigma_\varepsilon^2 &= \gamma \sigma_e^2, \quad \sigma_e^2 = \frac{1}{\gamma} \sigma_\varepsilon^2 \\ \text{var}(\varepsilon_{jk} - e_{jk}) &= \sigma_\varepsilon^2 + \sigma_e^2 \\ &= \sigma_\varepsilon^2 + \frac{1}{\gamma} \sigma_\varepsilon^2 = \sigma_\varepsilon^2 \left(1 + \frac{1}{\gamma}\right) \\ &= \frac{\gamma+1}{\gamma} \sigma_\varepsilon^2\end{aligned}$$

An estimator for the joint variance $(\sigma_\varepsilon^2 + \sigma_e^2)$ denoted by $\hat{\sigma}_{\varepsilon,e}^2$ is

$$\hat{\sigma}_{\varepsilon,e}^2 = \frac{1}{n-1} \sum_{i=1}^n (d_{jk} - \bar{d})^2$$

so

$$\hat{\sigma}_\varepsilon^2 = \frac{\gamma}{\gamma+1} \hat{\sigma}_{\varepsilon,e}^2$$

This estimator can be constrained somewhat. Since the joint variance is composed of the variances in *estimated effort* and *reported effort* it will always be larger than σ_ε^2 and σ_e^2 . Using a logistic function that takes values between 0,5 and 1 (assuming $\sigma_\varepsilon^2 > \sigma_e^2$), the estimator σ_ε^2 in the search algorithm is replaced with

$$\sigma_\varepsilon^2 = f(\varphi) \sigma_{\varepsilon,e}^2 = \frac{1+2e^\varphi}{2+2e^\varphi} \hat{\sigma}_{\varepsilon,e}^2$$

The Matlab search algorithm is then adapted to look for values of α and φ rather than σ_ε^2 directly. Rewriting the maximum-likelihood function gives

$$\begin{aligned}l(\alpha, \varphi) &= -\frac{n}{2} \log(2\pi) \\ &\quad - \frac{1}{2} \sum_{i=1}^n \left\{ \log \left(\frac{1+2e^\varphi}{2+2e^\varphi} \hat{\sigma}_{\varepsilon,e}^2 + \log(\dots) \right) - \frac{1}{2} \sum_{i=1}^n \left[\frac{1+2e^\varphi}{2+2e^\varphi} \hat{\sigma}_{\varepsilon,e}^2 + \log(\dots) \right]^{-1} (y_i - (Z\hat{\beta})_i)^2 \right\}\end{aligned}$$

and the calculation steps (see script 11, calculations, and script 13, logLikelihoodCeder2):

1. Select valid initial values for α , φ and σ_ε^2 , call them α_0 , φ_0 and $\sigma_{\varepsilon 0}^2$
2. Estimate β and $\Sigma_{\varepsilon,ii}$ with

$$\begin{aligned}\hat{\beta} &= (Z^T \Sigma_\varepsilon^{-1} Z)^{-1} Z^T \Sigma_\varepsilon^{-1} y \\ \hat{\Sigma}_{\varepsilon,ii}^{-1} &= \left[\hat{\sigma}_{\varepsilon 0}^2 + \log \left(\frac{1}{2} + \frac{1}{2} \sqrt{1 + 4 \frac{\hat{\sigma}_0^2}{U_{jk}}} \right) \right]^{-1} \\ \hat{\Sigma}_{\varepsilon,ij}^{-1} &= 0, i \neq j \\ y_{jk} &= \log(W_{jk}) - \log(U_{jk}) + bias \\ \hat{\sigma}_\varepsilon^2 &= \frac{1+2e^\varphi}{2+2e^\varphi} \hat{\sigma}_{\varepsilon,e}^2\end{aligned}$$

3. Run `fminsearch` to find the values of α and φ that maximise $l(\alpha, \varphi)$
4. If the values for α and φ have changed very little from last iteration, stop here
5. Else, use the resulting values α and φ and repeat from step 2

Adding this procedure to the algorithm actually does not result in any change in the estimators of the model parameters, indicating that the original assumption of equal errors was not unreasonable.

The meta-parameters were estimated as $\alpha = 1411,1$ and $\varphi = 0,83522$

4.4.7 Standardised residuals and treatment of outliers

In examining Table 11 (*estimated vessel parameters μ_k*) and the plot of the parameters in Figure 36, two parameters immediately stand out, those for vessels 1 and 5. Since all the vessels can be assumed to be similar (see Section 3.6), their parameters are not expected to differ as much as these do.

Looking at the source data, the cause for the first outlier was quickly determined to be that in the dataset there is only one trip reported for this vessel. This is not enough data to reliably construct a statistical model on, and this vessel must consequently be removed from the calculations. In fact the parameter is more than three standard deviations from the mean.

As for the second parameter, for vessel 5, no obvious inconsistencies can be found in the data and it must be concluded that this vessel simply has done worse than the others. This parameter is 2,4 standard deviations from the mean, and cannot be excluded on those grounds alone.

The standardised residuals of the dataset were calculated for search of other outliers. A *residual* of y , denoted by r_y is the difference between the measured datapoint, and the estimated function value

$$r_y = y - \mathbf{Z}\hat{\boldsymbol{\beta}}$$

Standardised residuals of y , denoted by w_y are simply scaled so that they will have a mean of zero and standard deviation of one

$$w_y = \frac{r_y - \bar{r}_y}{\text{stdev}(r_y)}$$

As a rule-of-thumb, any standardised residuals falling outside of three standard deviations should be examined as possible outliers.

Both the normal probability plot of $\varepsilon_{jk} - e_{jk}$ in Figure 38 and standardised residual plot in Figure 39 show one clear outlier marked with red. When the source data is examined for this measurement (vessel 3, trip 1 in year 2005), it is discovered that the report is highly suspect, with very low effort compared to the reported catch. This datapoint must be removed from the dataset so that the parameter estimation will not be affected by faulty data.

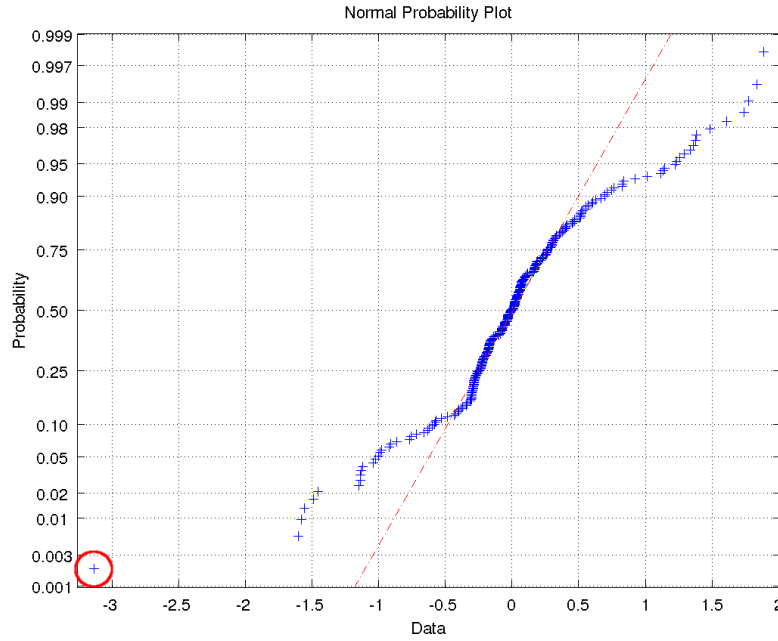


Figure 38: Normal probability plot of $\varepsilon_{jk} - e_{jk}$

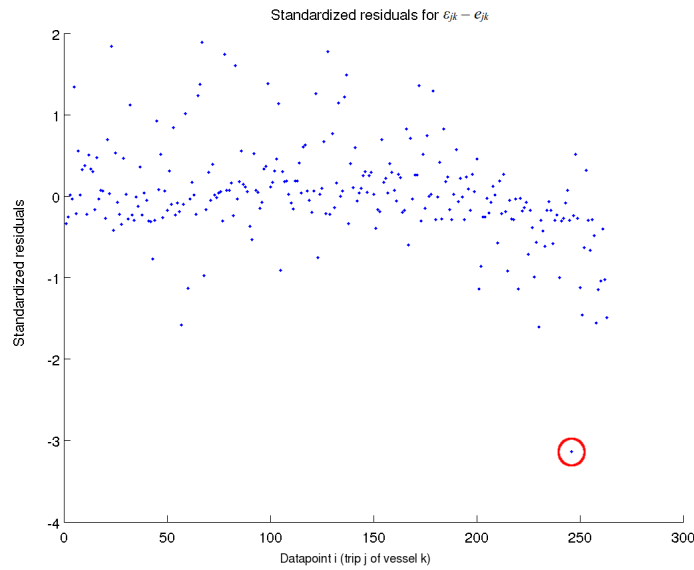


Figure 39: Standardised residuals of $\varepsilon_{jk} - e_{jk}$

The interpretation of a normal probability plot is that if the points fall on a straight line the model adequately describes the data. Some deviation is generally allowed towards the ends.

It can be seen that the normal model is a good fit in the centre, but some deviation is evident towards the ends. Possibly a Student's-t distribution would be a more accurate representation of the data.

A possible explanation is that the measurements are coming from two normal distributions rather than one. This would make the tails of the resulting combined distribution fatter, and manifest in the way seen in the plot. The source of the two distributions might be the *discrete effect* described in Section 3.5 - Visualisation of the data, where in effect the datasets would be drawing measurements from one distribution for automated logbooks and another for pen-and-paper versions.

The outliers were removed and the search algorithm run again. The parameter estimate results are in Tables 13 and 14 below, and displayed in Figures 40 and 41.

The meta-parameters were estimated as $\alpha = 998,01$, $\sigma^2_\varepsilon = 0,3252$ and $\varphi = 3,8273$

Table 13: Estimated vessel parameters μ_k without outliers

Vessel	μ_k	95% confidence interval	
1	-	-	-
2	3,83	$\pm 0,52$	[3,31;4,35]
3	3,23	$\pm 0,41$	[2,82;3,64]
4	3,60	$\pm 0,47$	[3,13;4,08]
5	2,45	$\pm 0,39$	[2,06;2,84]
6	3,31	$\pm 0,46$	[2,85;3,78]
7	3,21	$\pm 0,41$	[2,79;3,62]
8	3,49	$\pm 0,38$	[3,11;3,87]
9	3,35	$\pm 0,46$	[2,89;3,81]
10	3,36	$\pm 0,44$	[2,92;3,80]
11	3,74	$\pm 0,42$	[3,32;4,16]
12	3,52	$\pm 0,57$	[2,95;4,09]
13	3,08	$\pm 1,36$	[1,73;4,44]
14	3,52	$\pm 0,41$	[3,11;3,94]
15	3,56	$\pm 0,41$	[3,15;3,98]
16	3,35	$\pm 0,43$	[2,92;3,77]
17	3,43	$\pm 0,50$	[2,93;3,93]
18	3,65	$\pm 0,63$	[3,03;4,28]
19	3,51	$\pm 0,38$	[3,13;3,88]
20	3,69	$\pm 0,41$	[3,28;4,10]
21	3,52	$\pm 0,61$	[2,90;4,13]
22	3,58	$\pm 0,52$	[3,06;4,10]

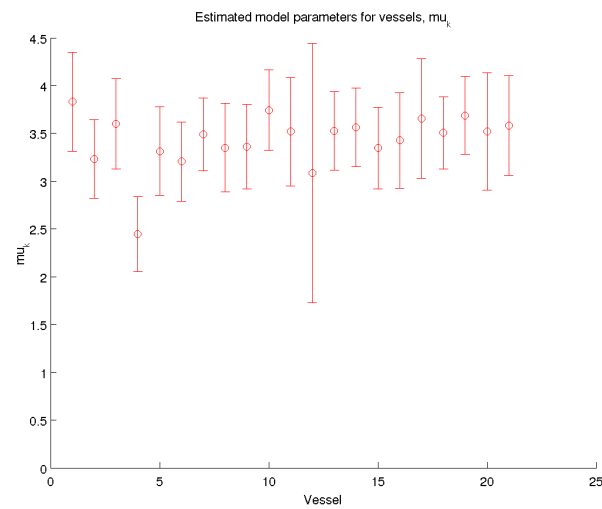


Figure 40: Estimated vessel parameters, μ_k without outliers

Table 14: Estimated year parameters β_t without outliers

Year	β_t		
2001	-0,42	$\pm 0,34$	$[-0,76; -0,09]$
2002	-0,23	$\pm 0,33$	$[-0,56; 0,10]$
2003	0,03	$\pm 0,33$	$[-0,30; 0,36]$
2004	0,00	$\pm 0,35$	$[-0,34; 0,35]$
2005	0	Reference year	-

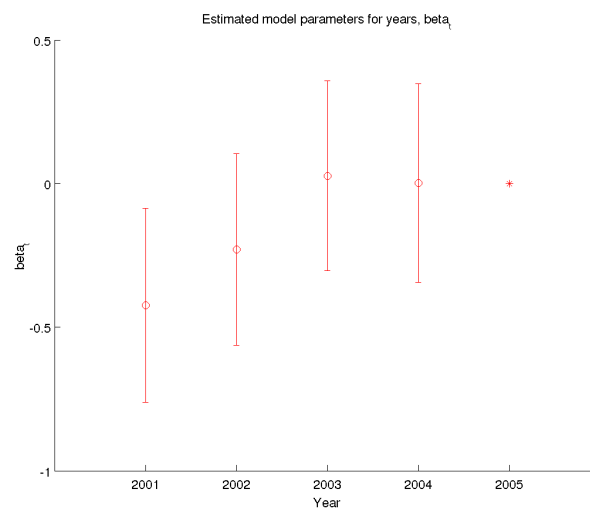


Figure 41: Estimated year parameters, β_t without outliers

4.4.8 Student's-t distribution

As indicated by Figure 38, a Student's-t distribution might describe the data more appropriately. To determine the parameters of the t-distribution which would give the best fit, the measure

$$\log t_v = \sum_{jk} \log(\text{tpdf}(\text{stdres}(y_{jk}), v))$$

is calculated for degrees of freedom v from 1 to 30 and plotted in Figure 42. Here, tpdf signifies the probability density function for the t-distribution, and $\text{stdres}(y_{jk})$ are the standardised residuals of y_{jk} .

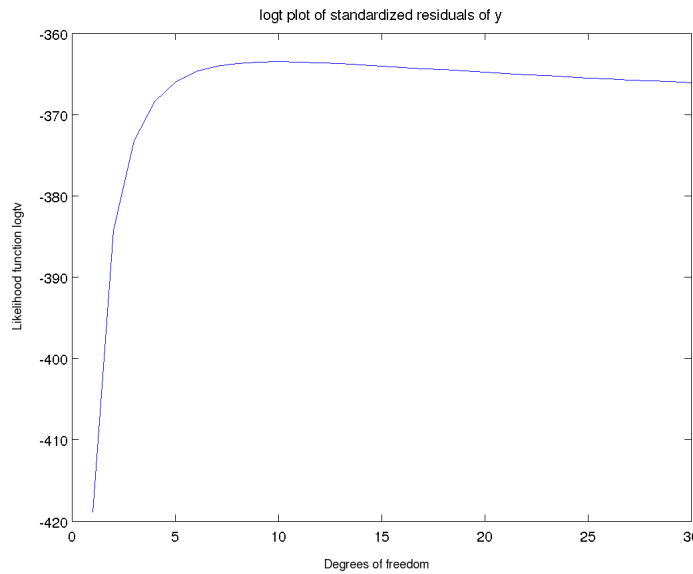


Figure 42: Log-t plot of standardised residuals of y

What is essentially being done here, is fitting t-distributions with progressively thinner tails to the data. The plot indicates that a t-distribution with $v = 10$ degrees of freedom would be appropriate. The low number for the degrees of freedom indicates that it would be worth using the t-distribution, since when the degrees of freedom approach infinity, the data approaches the normal distribution, and inversely when the degrees of freedom are low it moves further away from the normal and the t-distribution is a better fit.

It should be noted that the standardised residuals used here are calculated from a normal distribution, and are not in fact accurately estimating the degrees of freedom, but are useful in doing a quick check to see if it produces a low or high value. A more accurate treatment follows.

Student's-t probability density function with location parameter μ and scale parameter σ is of the form (Gelman et al. 2004, p. 576)

$$f(y; \mu, \sigma, \nu) = \frac{1}{\sigma} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{1}{\nu} \left(\frac{y-\mu}{\sigma}\right)^2\right)^{-\left(\frac{\nu+1}{2}\right)}$$

Substituting and writing for n observations, a likelihood function for α , σ_ε^2 and ν is

$$\begin{aligned} L(\alpha, \sigma_\varepsilon^2, \nu) &= \prod_{i=1}^n \left\{ \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} [\sigma_\varepsilon^2 + \sigma_{00}^2]^{-1/2} \right. \\ &\quad \left. \left[1 + \frac{1}{\nu} [\sigma_\varepsilon^2 + \sigma_{00}^2]^{-1} (y_i - (Z\hat{\beta})_i)^2 \right]^{-\left(\frac{\nu+1}{2}\right)} \right\} \\ &= \prod_{i=1}^n \left\{ \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left[\sigma_\varepsilon^2 + \log\left(\frac{1}{2} + \frac{1}{2} \sqrt{1 + 4 \frac{\alpha}{U_{jk}}}\right) \right]^{-1/2} \right. \\ &\quad \left. \left[1 + \frac{1}{\nu} \left[\sigma_\varepsilon^2 + \log\left(\frac{1}{2} + \frac{1}{2} \sqrt{1 + 4 \frac{\alpha}{U_{jk}}}\right) \right]^{-1} (y_i - (Z\hat{\beta})_i)^2 \right]^{-\left(\frac{\nu+1}{2}\right)} \right\} \end{aligned}$$

where, as before, the *true effort* X_{jk} has been approximated with the *estimated effort* U_{jk} .

The logarithm of the likelihood function is then

$$\begin{aligned} l(\alpha, \sigma_\varepsilon^2, \nu) &= \log(L(\alpha, \sigma_\varepsilon^2, \nu)) \\ &= n \log \Gamma\left(\frac{\nu+1}{2}\right) - \frac{n}{2} \log(\nu\pi) - n \log \Gamma\left(\frac{\nu}{2}\right) \\ &\quad - \frac{1}{2} \sum_{i=1}^n \left[\log(\sigma_\varepsilon^2 + \log(\dots)) \right] \\ &\quad - \frac{(\nu+1)}{2} \sum_{i=1}^n \log \left(\left[1 + \frac{1}{\nu} (\sigma_\varepsilon^2 + \log(\dots))^{-1} (y_i - (Z\hat{\beta})_i)^2 \right] \right) \end{aligned}$$

Recalling Equation (1) in Section 4.4.3, the model is restated in terms of the Student's-t distribution.

Let t_{jk} be a t-distributed random variable with ν degrees of freedom. Then

$$\begin{aligned} \log W_{jk} &= \log U_{jk} - \hat{bias} + \log \mu_{jk} + \varepsilon_{jk} + \sigma_{00} t_{jk, \nu} \\ &\sim t(\log U_{jk} - \hat{bias} + \log \mu_{jk}, \hat{\sigma}_\varepsilon^2 + \sigma_{00}^2, \nu) \end{aligned} \quad (2)$$

where t signifies Student's-t distribution.

Incorporating the model for vessels and years then gives

$$\log W_{jk} \sim t(\log U_{jk} - \hat{bias} + \mu_k + \beta_t, \hat{\sigma}_\varepsilon^2 + \sigma_{00}^2, \nu)$$

with the median value of

$$\text{median}[W_{jk}] = U_{jk} e^{-bias} e^{\mu_k} e^{\beta_i}$$

The function `fminsearch` in Matlab is used to find the maximum-likelihood estimation following these steps (see script 14, `calculations_t`, and script 15, `logLikelihoodCeder2_t`):

1. Select valid initial values for α , φ , ν and σ_ε^2 , call them α_0 , φ_0 , ν_0 and $\sigma_{\varepsilon 0}^2$
2. Estimate β and $\Sigma_{\varepsilon, ii}$ with

$$\begin{aligned}\hat{\beta} &= (\mathbf{Z}^T \hat{\Sigma}_\varepsilon^{-1} \mathbf{Z})^{-1} \mathbf{Z}^T \hat{\Sigma}_\varepsilon^{-1} \mathbf{y} \\ \hat{\Sigma}_{\varepsilon, ii}^{-1} &= \left[\hat{\sigma}_{\varepsilon 0}^2 + \log\left(\frac{1}{2} + \frac{1}{2} \sqrt{1 + 4 \frac{\hat{\alpha}_0}{U_{jk}}}\right) \right]^{-1} \\ \hat{\Sigma}_{\varepsilon, ij}^{-1} &= 0, i \neq j \\ \mathbf{y}_{jk} &= \log(W_{jk}) - \log(U_{jk}) + bias \\ \hat{\sigma}_\varepsilon^2 &= \frac{1 + 2e^\varphi}{2 + 2e^\varphi} \hat{\sigma}_{\varepsilon, e}^2\end{aligned}$$

3. Run `fminsearch` to find the values of α , φ and ν_0 that maximise $l(\alpha, \varphi, \nu)$
4. If the values for α , φ and ν have changed very little from last iteration, stop here
5. Else, use the resulting values α , φ and ν and repeat from step 2

Running the search algorithm gives the results in Tables 15 and 16 below, and are similar to the previous parameter results.

The meta-parameters were estimated as $\alpha = 798,04$, $\sigma_\varepsilon^2 = 0,16783$, $\varphi = -3,8337$ and $\nu = 4,3314$

Table 15: Estimated vessel parameters μ_k without outliers using Student's-t distribution

Vessel	μ_k	95% confidence interval	
1	-	-	-
2	3,83	$\pm 0,38$	[3,45;4,21]
3	3,23	$\pm 0,30$	[2,93;3,53]
4	3,60	$\pm 0,35$	[3,25;3,95]
5	2,41	$\pm 0,29$	[2,12;2,70]
6	3,31	$\pm 0,34$	[2,97;3,65]
7	3,19	$\pm 0,31$	[2,89;3,50]
8	3,48	$\pm 0,28$	[3,20;3,76]
9	3,34	$\pm 0,34$	[3,00;3,68]
10	3,36	$\pm 0,33$	[3,03;3,69]
11	3,74	$\pm 0,31$	[3,43;4,05]
12	3,49	$\pm 0,42$	[3,07;3,92]
13	3,08	$\pm 1,04$	[2,03;4,12]
14	3,52	$\pm 0,30$	[3,22;3,82]
15	3,56	$\pm 0,30$	[3,25;3,86]
16	3,33	$\pm 0,32$	[3,02;3,65]
17	3,42	$\pm 0,37$	[3,05;3,78]
18	3,66	$\pm 0,46$	[3,20;4,12]
19	3,51	$\pm 0,28$	[3,23;3,79]
20	3,68	$\pm 0,30$	[3,38;3,98]
21	3,52	$\pm 0,45$	[3,06;3,97]
22	3,57	$\pm 0,38$	[3,19;3,96]

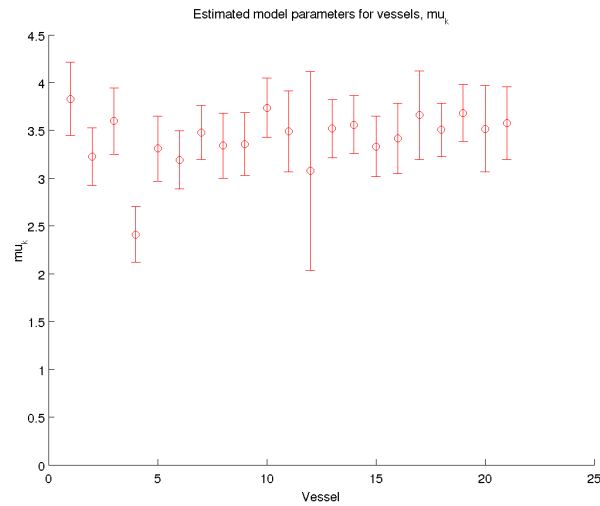
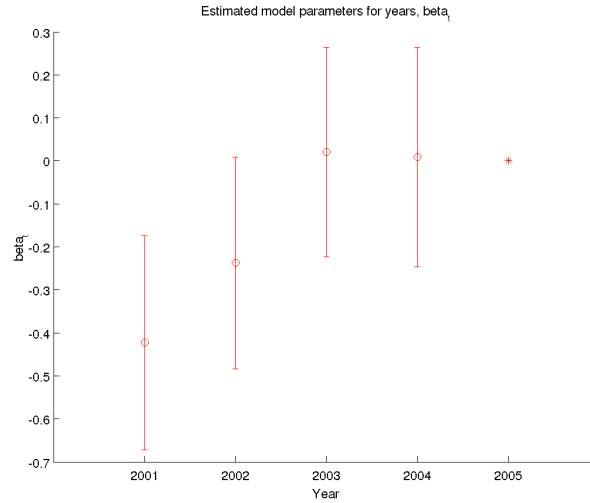
Figure 43: Estimated vessel parameters μ_k without outliers using Student's-t distribution

Table 16: Estimated year parameters β_t without outliers using Student's-t distribution

Year	β_t	95% confidence interval	
2001	-0,42	$\pm 0,25$	$[-0,67; -0,17]$
2002	-0,24	$\pm 0,25$	$[-0,48; 0,01]$
2003	0,02	$\pm 0,24$	$[-0,22; 0,26]$
2004	0,01	$\pm 0,26$	$[-0,25; 0,26]$
2005	0	Reference year	-

Figure 44: Estimated year parameters β_t without outliers using Student's-t distribution

The low value for the degrees of freedom confirms that the t-distribution is indeed a better choice than the normal distribution.

The standardised residuals for the t-distribution are similar to the ones from the normal distribution

$$w_y = \frac{r_y - \bar{r}_y}{\text{stdev}(r_y)} = \frac{r_y - \bar{r}_y}{\sqrt{\sigma_\epsilon^2 + \sigma_{00}^2}}$$

A log-t plot of the standardised residuals is displayed in Figure 45, showing the maximum value at $v = 4,3314$.

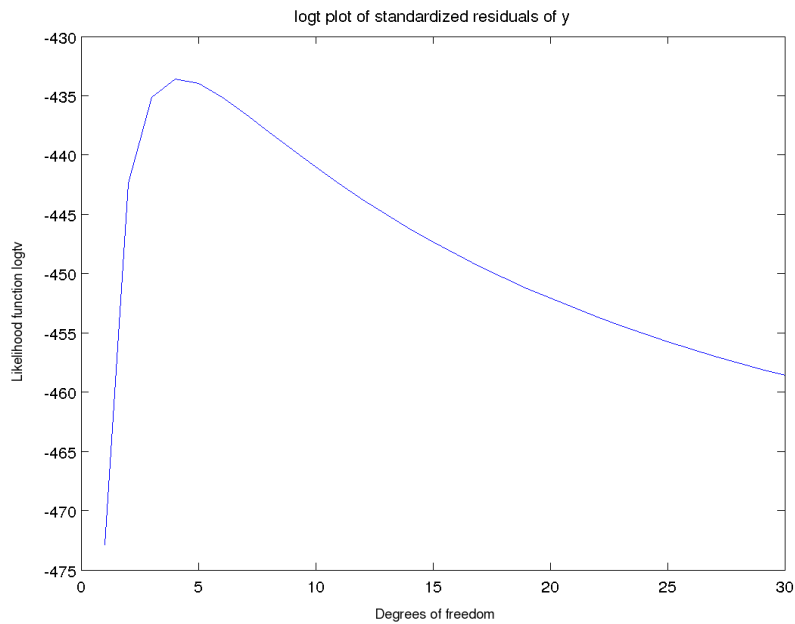


Figure 45: log-t plot of standardised residuals of y from the t -distribution

In Figure 46, plotting the empirical CDF¹¹ of the standardised residuals (in blue) vs. the theoretical CDF for the t -distribution (in magenta) shows the very close fit of the standardised residuals, as does the plot of the sorted standardised residuals vs. the probability number in Figure 47 (this figure is analogous to the normal probability plot in Figure 38, only for the t -distribution).

11 The Cumulative Distribution Function constructed from the sample data

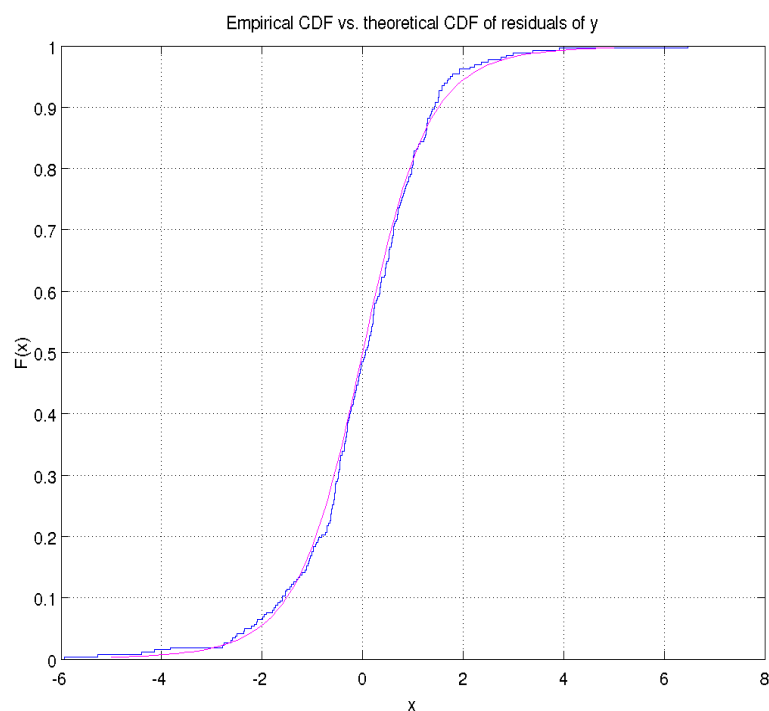


Figure 46: Empirical CDF vs. theoretical CDF of standardised residuals y from the t -distribution

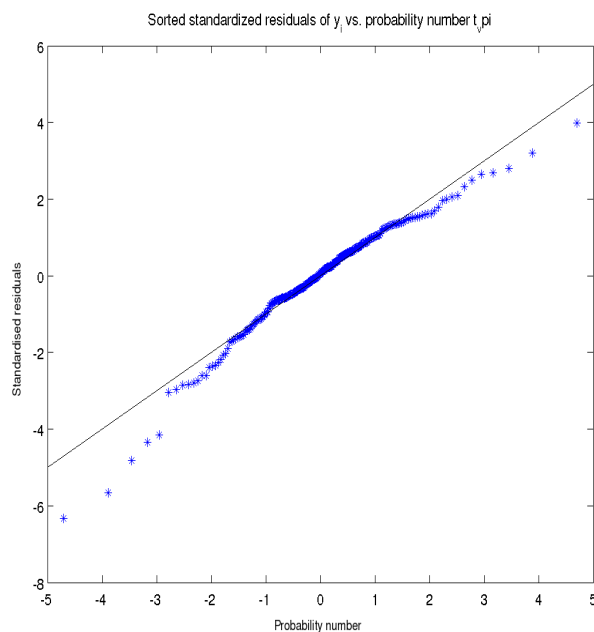


Figure 47: Sorted standardised residuals of y vs. probability number for the t -distribution

However, when looking at the residual plot vs. the log of estimated effort in Figure 48, the whole group of residuals seems slanted downwards to the right. The red lines represent 99,5% and 0,5% probability numbers from the t-distribution, while the green lines represent 97,5% and 2,5%. Most of the residuals should fall within the green limits, with only a few in the red zone.

There is really only one reason this can be explained; that the model of the catch being linearly dependent upon estimated effort is too strict.

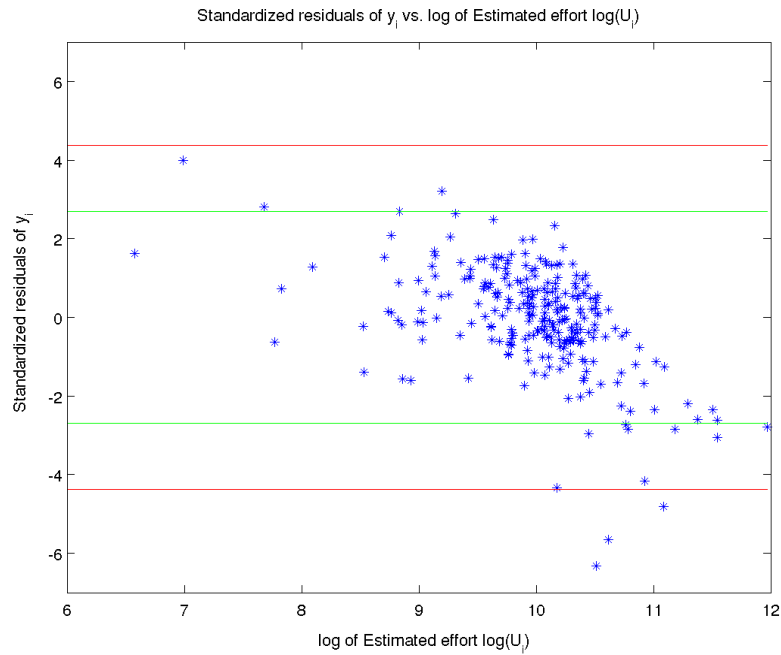


Figure 48: Standardised residuals vs. log of estimated effort

4.4.9 Curvilinear model

At the higher end of *estimated effort* in Figure 48 the residuals are all negative, that is the model is always higher than the measurements, overestimating the catch. This behaviour implies that a more realistic model would be some kind of a curve that flattens out to the right, resulting in lower catch predictions for very high estimated effort. Looking back, a suggestion of this effect can actually be seen in Figure 12, Section 3.5.

An explanation for this is that the classification and effort estimation algorithms are capturing more than pure trawling effort, especially where the vessel stays for very long periods at sea. The algorithms then overestimate the actual trawling time by including track legs that are actually cruising legs.

To give the model a curve to fit this behaviour, the parameter κ_I is added. Note that although technically the parameter also appears with the error term ε_{jk} , this term is estimated separately to simplify the calculations.

Amending the model in Equation (2) gives

$$\log W_{jk} = \kappa_1 (\log U_{jk} - \hat{bias}) + \log \mu_{jk} + \kappa_1 \varepsilon_{jk} + \sigma_{00} t_{jk} \quad (3)$$

$$\sim t(\kappa_1 (\log U_{jk} - \hat{bias}) + \log(\mu_{jk}), \kappa_1^2 \hat{\sigma}_\varepsilon^2 + \sigma_{00}^2, \nu)$$

where t signifies Student's-t distribution.

Incorporating the model for vessels and years gives

$$\log W_{jk} \sim t(\kappa_1 (\log U_{jk} - \hat{bias}) + \mu_k + \beta_t, \kappa_1^2 \hat{\sigma}_\varepsilon^2 + \sigma_{00}^2, \nu) \quad (4)$$

with the median of the catch being

$$\text{median}[W_{jk}] = (U_{jk})^{\kappa_1} e^{-\kappa_1 \hat{bias}} e^{\mu_k} e^{\beta_t} \quad (5)$$

In order for the curve to flatten out, the value of κ_l is expected to be between 0 and 1.

A column for κ_l is added to the design matrix Z , with elements $\log U_{jk} - \hat{bias}$, which also changes the y_{jk} . As before, the function `fminsearch` in Matlab is used to find the maximum-likelihood estimation of the parameters following the same steps (see script 16, `calculations_t_ext4`, and script 17, `logLikelihoodCeder2_t_ext4`):

1. Select valid initial values for α , φ , ν and σ_ε^2 , call them α_0 , φ_0 , ν_0 and $\sigma_{\varepsilon 0}^2$

2. Estimate β and $\Sigma_{\varepsilon, ii}$ with

$$\hat{\beta} = (Z^T \hat{\Sigma}_\varepsilon^{-1} Z)^{-1} Z^T \hat{\Sigma}_\varepsilon^{-1} y$$

$$\hat{\Sigma}_{\varepsilon, ii}^{-1} = \left[\kappa_0^2 \hat{\sigma}_{\varepsilon 0}^2 + \log \left(\frac{1}{2} + \frac{1}{2} \sqrt{1 + 4 \frac{\hat{\alpha}_0}{U_{jk}}} \right) \right]^{-1}$$

$$\hat{\Sigma}_{\varepsilon, ij}^{-1} = 0, i \neq j$$

$$y_{jk} = \log(W_{jk})$$

$$\hat{\sigma}_\varepsilon^2 = \frac{1 + 2e^\varphi}{2 + 2e^\varphi} \hat{\sigma}_{\varepsilon, e}^2$$

3. Run `fminsearch` to find the values of α , φ and ν_0 that maximise $l(\alpha, \varphi, \nu)$
4. If the values for α , φ and ν have changed very from last iteration, stop here
5. Else, use the resulting values α , φ and ν and repeat from step 2

Running the search algorithm gives the results in Tables 17 and 18 below, plotted in Figures 49 and 50.

The meta-parameters were estimated as $\alpha = 2680,1$, $\sigma_\varepsilon^2 = 0,3187$, $\varphi = 31,083$ and $\nu = 5,2372$ and the new curve-parameter was estimated as $\kappa_l = 0,23191$ with a 95% confidence interval of $\pm 0,09$ or $[0,15; 0,32]$.

Table 17: Estimated vessel parameters μ_k without outliers using Student's-t distribution and a curve parameter

Vessel	μ_k	95% confidence interval	
1	-	-	-
2	11,38	$\pm 0,90$	[10,49;12,28]
3	10,88	$\pm 0,89$	[9,98;11,77]
4	11,10	$\pm 0,88$	[10,21;11,98]
5	9,51	$\pm 0,84$	[8,67;10,34]
6	10,92	$\pm 0,89$	[10,03;11,81]
7	10,53	$\pm 0,87$	[9,66;11,39]
8	10,78	$\pm 0,85$	[9,92;11,63]
9	10,78	$\pm 0,88$	[9,90;11,65]
10	10,59	$\pm 0,85$	[9,74;11,44]
11	11,09	$\pm 0,86$	[10,23;11,95]
12	10,64	$\pm 0,88$	[9,76;11,51]
13	9,53	$\pm 1,32$	[8,22;10,85]
14	11,01	$\pm 0,88$	[10,14;11,89]
15	11,13	$\pm 0,88$	[10,24;12,01]
16	10,55	$\pm 0,85$	[9,70;11,41]
17	11,00	$\pm 0,91$	[10,09;11,91]
18	10,90	$\pm 0,90$	[10,00;11,79]
19	10,85	$\pm 0,85$	[10,00;11,70]
20	11,21	$\pm 0,88$	[10,34;12,09]
21	11,15	$\pm 0,92$	[10,23;12,07]
22	11,30	$\pm 0,92$	[10,39;12,22]

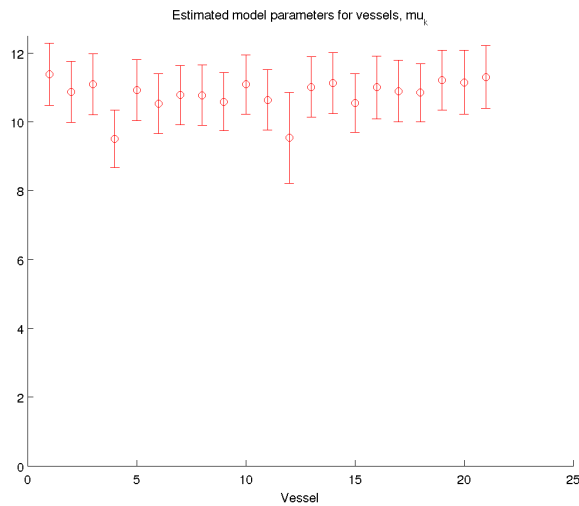


Figure 49: Estimated vessel parameters, μ_k without outliers, curvilinear model

Table 18: Estimated year parameters β_t without outliers using Student's-t distribution and a curve parameter

Year	β_t	95% confidence interval	
2001	-0,37	$\pm 0,21$	$[-0,58;-0,17]$
2002	-0,18	$\pm 0,20$	$[-0,38;0,02]$
2003	0,01	$\pm 0,20$	$[-0,19;0,21]$
2004	0,04	$\pm 0,21$	$[-0,18;0,25]$
2005	0	Reference year	-

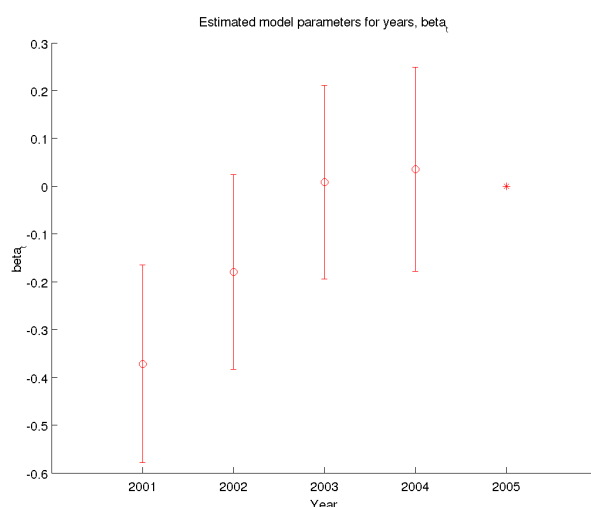


Figure 50: Estimated year parameters, β_t without outliers, curvilinear model

Even though a numerical interpretation of the vessel and year parameters is somewhat difficult, it can be useful to compare them relatively. The vessel parameter μ_k is a measure of the vessel's catching power (CPUE), and from Figure 49 it can be seen that most of the vessels are similar as expected, with the exception of vessels 5 and 13, which seem to have lower catching power than the others. This might indicate a problem with the data-quality from these vessels, or indeed warrant a closer examination of the vessel's catch and effort reports. The year parameter β_t is more difficult to relate to a physical quantity, but indicates how productive a given year was for the fleet, i.e. how much effort was needed in relation with the reference year 2005. As most of the confidence intervals include the zero, we cannot be certain the parameters are significant.

Again plotting the sorted standardised residuals and the residuals vs. estimated effort in Figure 51 and 52 shows that the introduction of the κ_l parameter has indeed improved the behaviour of the residuals.

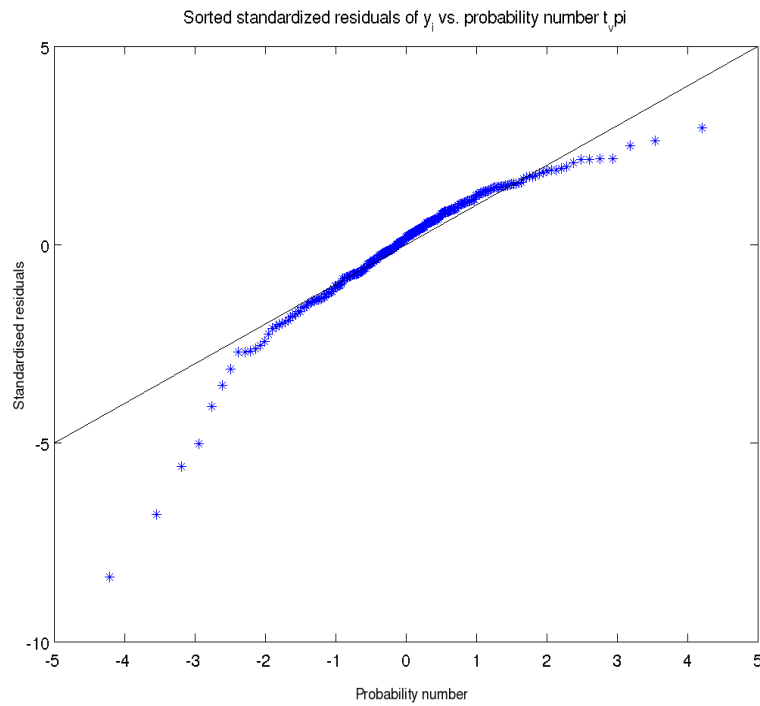


Figure 51: Sorted standardised residuals of y vs. probability number for the t -distribution

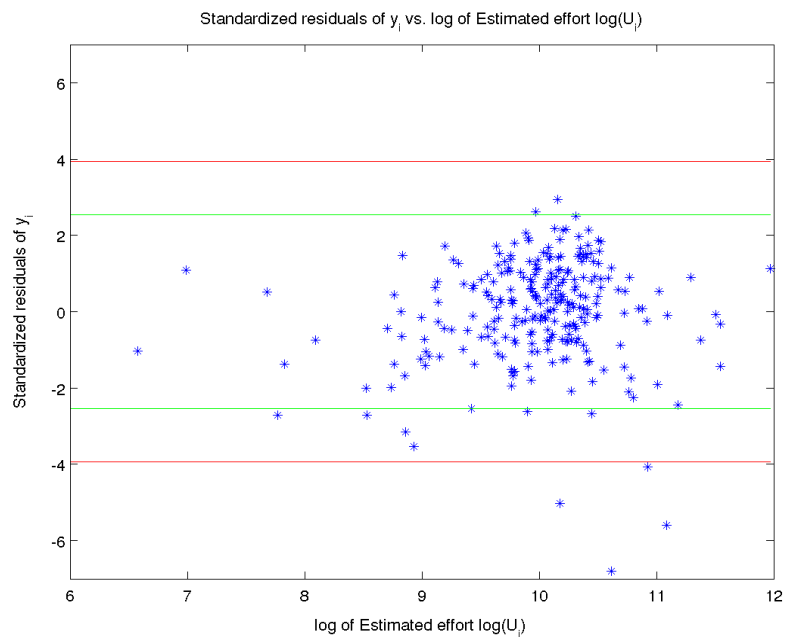


Figure 52: Standardised residuals vs. log of estimated effort

4.4.10 Prediction intervals on new observations with a fitted year-effect

The fitted values of the model can now be calculated, and subsequently prediction intervals¹² for new values. For new observations in a year which has been used for the model fit, i.e. where a new observation is made for a vessel that already has been observed within that year, the year-effect parameter β_t is already estimated by the model.

Using Equation (3) from the previous section to calculate the fitted values and prediction intervals on the composite variable y_{jk}

$$y_{jk} = \log W_{jk}$$

then the fitted values are

$$\hat{y} = \mathbf{Z}\hat{\beta}$$

where \mathbf{Z} is the design matrix (remember that κ_l is now included in β and a column with corresponding elements to κ_l has been added to the design matrix).

Since the primary interest is actually in predicting new values of catch based on an *estimated effort* measurement for a given vessel and year, the error associated with new observations must also be taken into account, along with the sampling error of the estimators for the model parameters, in order to arrive at a formulation that can be used in practice.

The estimator for the log-mean of vessel k and year t is

$$\hat{\kappa}_l(\log U_{jk} - \hat{bias}) + \hat{\mu}_k + \hat{\beta}_t$$

The variance for this estimator is denoted by

$$\hat{\sigma}_{\kappa, \mu k, \beta t}^2$$

Let

$$\begin{aligned} a^T &= [(\log U_{jk} - \hat{bias}) 0 \cdots 0 \underset{\substack{\uparrow \\ k^{th} \text{ element}}}{1} 0 \cdots 0 \underset{\substack{\uparrow \\ t^{th} \text{ element}}}{1} 0 \cdots] \\ &= \mathbf{Z}_n \end{aligned}$$

that is, the row vector in the design matrix \mathbf{Z} that corresponds to vessel k and year t .

Then

$$\begin{aligned} \hat{\sigma}_{\kappa, \mu k, \beta t}^2 &= a^T \text{cov}(\hat{\beta}) a \\ &= a^T (\mathbf{Z}^T \hat{\Sigma}_{\epsilon}^{-1} \mathbf{Z})^{-1} a \end{aligned}$$

¹² A note on terminology: a *confidence interval* is calculated on the regression parameters or the mean for particular values of the explanatory variables and indicates how well they have been estimated. Thus, calculating 95% confidence intervals on the mean from multiple samples will result in 95% of the intervals containing the true mean. A *prediction interval* indicates where to expect the next sampled datapoint, i.e. they tell us about the distribution of values, not the uncertainty in determining the population mean. Prediction intervals must account for both the uncertainty in knowing the value of the population mean, plus the variability in the data and thus, they are always wider than the confidence interval on the mean when the same values of the explanatory variables are used.

This sampling error is then added to the prediction interval.

The 95% prediction interval for a new log-value is thus

$$\hat{y} \pm t_{v,0.025} \sqrt{\hat{\kappa}_1^2 \hat{\sigma}_\varepsilon^2 + \hat{\sigma}_{00}^2 + \hat{\sigma}_{\kappa, \mu k, \beta t}^2}$$

However, the measure that is most interesting is the median catch, which according to Equation (5) is

$$\text{median}[W_{jk}] = (U_{jk})^{\hat{\kappa}_1} e^{-\hat{\kappa}_1 \hat{bias}} e^{\hat{\mu}_k} e^{\hat{\beta}_t}$$

and the 95% prediction interval on the catch is then

$$(U_{jk})^{\hat{\kappa}_1} e^{-\hat{\kappa}_1 \hat{bias}} e^{\hat{\mu}_k} e^{\hat{\beta}_t} e^{\left(\pm t_{v,0.025} \sqrt{\hat{\kappa}_1^2 \hat{\sigma}_\varepsilon^2 + \hat{\sigma}_{00}^2 + \hat{\sigma}_{\kappa, \mu k, \beta t}^2}\right)}$$

Figure 53 shows the fitted values of the modelled catch and prediction intervals along with the actual datapoints. The figure shows the estimated catch as red circles, with the corresponding prediction intervals as red bars. The reported catch is plotted as blue dots.

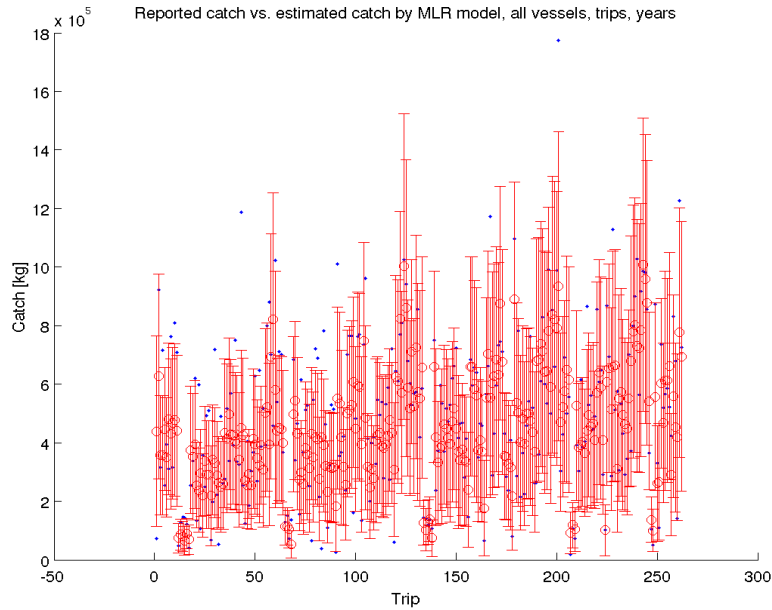


Figure 53: Reported catch and modelled catch with 95% prediction intervals

A note on the interpretation of the prediction intervals in the above plots: The true value of the measured variable lies within the given interval with 95% confidence¹³. Having some of the reported measurements fall outside of this interval is not unexpected. In fact it leads directly from the formal formulation; if the experiment was repeated one hundred times, the true value (and thus the measurement) would be expected to fall outside of the interval in about five cases. Having 262 measurements, about 13 values are then expected outside

¹³ It would *not* be correct to claim that the *probability* of the true value being within the interval is 95%. This is because in each case, the true value actually lies either within or outside the interval with a probability of 100%.

the interval. See 4.5 - Model performance for a further discussion.

Also keep in mind that there is a measurement error involved in each case, and the blue dots do not represent the *true* value of the variable. The measurements could even be plotted with their own error bars. The prediction intervals take the measurement error into account.

4.4.11 Prediction intervals on new observations with an unknown year-effect

To use the model on the validation dataset, the year effect parameter β_6 must first be estimated for the year 2006. Since in this case the assumption is that no data from that year is available to build the model, this parameter cannot be estimated along with the others.

One reasonable choice of an estimator for the year effect when there is no data for the year, would be the sample mean of the other year parameters β_t and the variance of β_t is estimated with the sample variance, that is

$$\hat{\beta}_6 = \frac{1}{5} \sum_{i=1}^5 \hat{\beta}_i$$

$$\hat{\sigma}_{\beta_6}^2 = \frac{1}{4} \sum_{i=1}^5 (\beta_i - \hat{\beta}_6)^2$$

This yields a value for $\beta_6 = -0,12689$ and $\sigma_{\beta_6}^2 = 0,030922$

Illustration 54 shows the estimated year-effect parameters. The figure shows a comparison of the estimated values of the year-effect parameters β_1 to β_6 , corresponding to years 2001-2006. The parameter for 2005 is set at 0, and the others are defined as deviations from this reference year. The year 2006 is validation data.

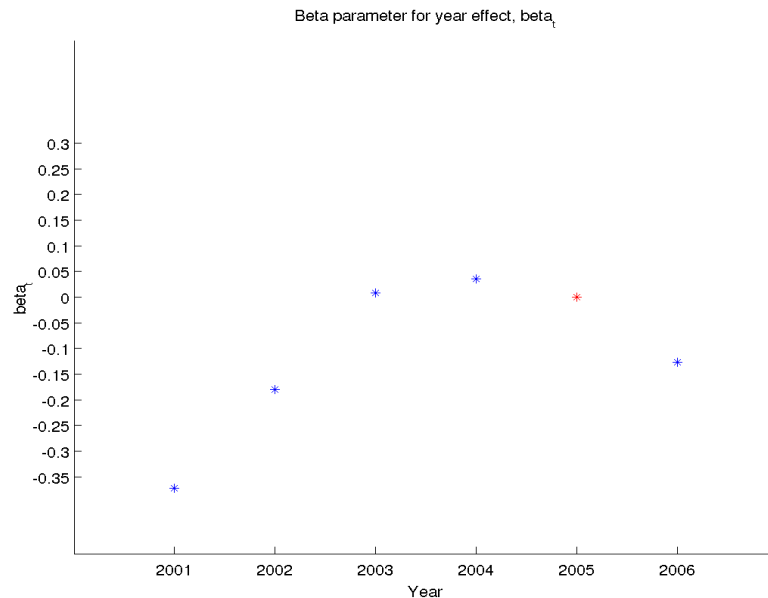


Figure 54: Estimation of year-effect parameters β_t

The prediction intervals now must take into account the uncertainty of this estimate in

addition to the uncertainty associated with the other parameters.

For y the prediction interval becomes

$$\hat{y} \pm t_{v,0.025} \sqrt{\hat{\kappa}_1^2 \hat{\sigma}_\varepsilon^2 + \hat{\sigma}_{00}^2 + \hat{\sigma}_{\beta_6}^2 + \hat{\sigma}_{\kappa, \mu k}^2}$$

and for the catch W_{jk}

$$(U_{jk})^{\hat{\kappa}_1} e^{-\hat{\kappa}_1 \hat{bias}} e^{\hat{\mu}_k} e^{\hat{\beta}_i} e^{(\pm t_{v,0.025} \sqrt{\hat{\kappa}_1^2 \hat{\sigma}_\varepsilon^2 + \hat{\sigma}_{00}^2 + \hat{\sigma}_{\beta_6}^2 + \hat{\sigma}_{\kappa, \mu k}^2})}$$

The modelled catch with prediction intervals for the training and validation datasets is depicted in Figure 55. The figure shows the estimated catch as red circles, with the corresponding prediction intervals as red bars. The reported catch is plotted as blue dots for the training dataset, and green points for the validation dataset.

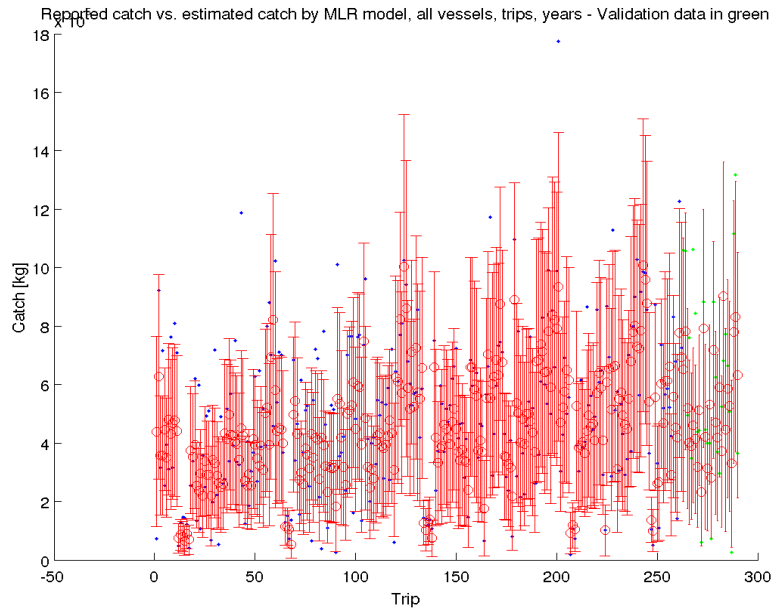


Figure 55: Reported catch and modelled catch with 95% prediction intervals and validation data

4.4.12 Usage examples for the final model

To show how the model might be used, the model results and reported data for vessel 5 in year 2003 are presented in Figure 56. The figure shows the *predicted catch* vs. *estimated effort* as a solid red line. This is the catch that the model would expect to see for the given effort. The red dots are the reported catch for the corresponding *estimated effort* of a single trip. The dotted lines are the prediction intervals.

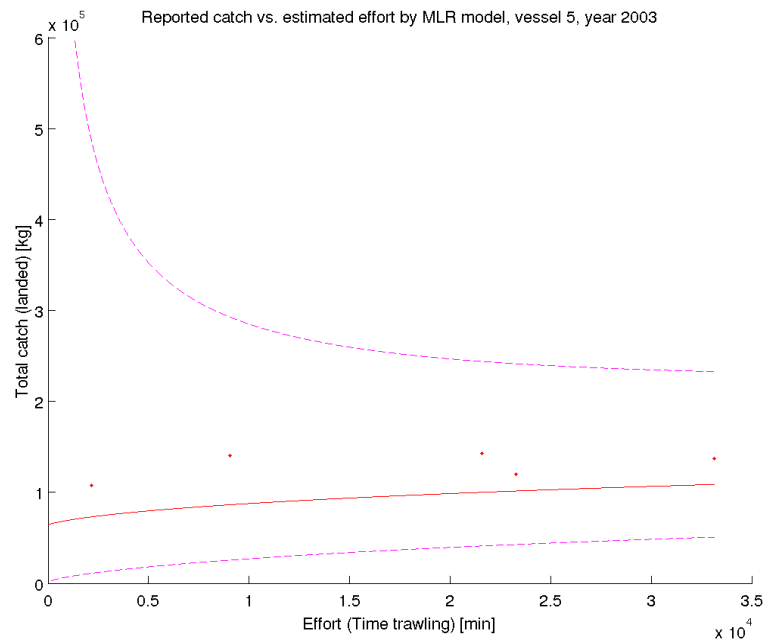


Figure 56: Model results for vessel 5 in year 2003 – Reported catch vs. Estimated effort

Figure 57 shows same vessel for all years. The figure shows the *predicted catch* from the model as red circles and the corresponding *reported catch* as blue dots. The prediction intervals on the model estimate are shown as red error-bars.

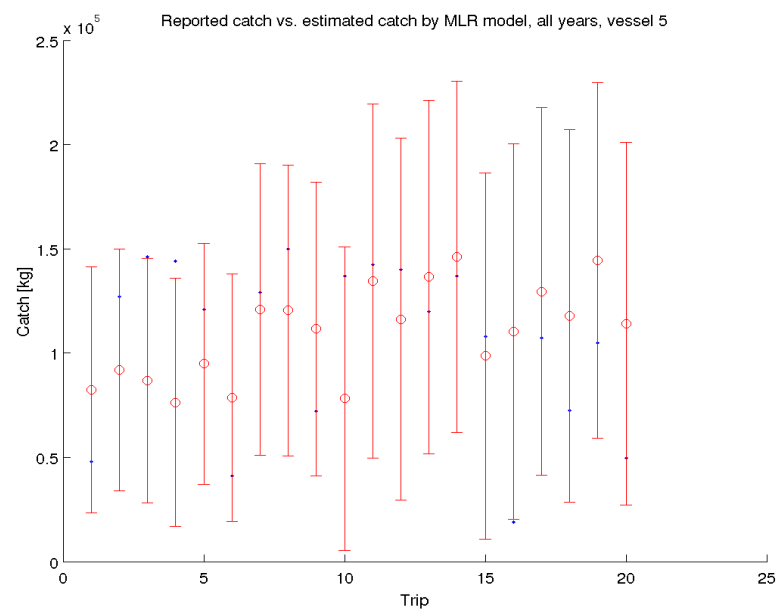


Figure 57: Model results for vessel 5, all years – Reported catch vs. Estimated catch

The two figures show how the model might be used. For example, the effort estimation algorithm might give an effort estimate for a vessel. Figure 56 can then be used to find the model prediction for the catch along with a 95% prediction interval.

The model could also be used to check if the vessel landing reports are reasonable, or if the reporting should be examined in more detail, possibly calling for extra data or plan inspections for this vessel. For example, in Figure 57 trips 4 and 16 fall outside of the prediction interval, indicating that the vessel reported higher or lower catch than the model expected from the predicted effort. This usually simply indicates a problem with reporting of either the catch or effort, but is nevertheless valuable in improving data quality.

Note that the same principles apply when predicting values for a year for which there is no data yet available, such as the validation year 2006, but with slightly expanded prediction intervals to account for the uncertainty in the model parameter estimators.

4.5 Model performance

Several approaches are available to evaluate the performance of the final model. A visual confirmation of the model fit is made by examining the standardised residual plots in Figures 51 and 52 and the reported data vs. model predictions with prediction intervals in Figures 53, 53, 56 and 55.

As can be seen in Figure 51 the t-distribution is not a perfect fit to the data, especially for low values of *estimated effort*. In Figure 52 there are 4 residuals falling outside the 99% prediction interval or 1,5% of the 262 total measurements, and 8 fall outside the 95% prediction interval (but still inside the 99% interval), or 3% of the 262 total measurements.

Also, the model errors from fitted data are examined directly, the model fit for the training and validation datasets are compared, and finally the model can be compared with alternative models using Mean Squared Error (MSE) and Log-Mean Squared Error (LMSE) measures.

A count of predicted values of catch falling inside vs. outside of the prediction intervals in Figure 53 reveals 34 values outside the 95% prediction interval, or about 13% of the predictions. Most of these are concentrated at the lower end of *estimated effort*. As previously mentioned, for a 95% prediction interval and 262 measurements, about 13 values would be expected. The reason lies in the model fit which is still a little off, especially for low effort measurements as shown in Figure 51.

Consider the variances that stem from the measurement of the catch and the fitted data, $\sigma_{00}^2 = 0,12271$ and $\kappa_1^2 \sigma_e^2 = 0,23191^2 \cdot 0,017678 = 0,0009508$ respectively.

The variance from the measurements of the catch is much greater than that of the model fit, with the latter less than 1% of the total. Assuming a 5% prediction interval, this translates into an increase from an error of 5% to 6% caused by unfitted data.

This fit of the final model, although not perfect, can be classed as “good enough” for the practical purposes of this project. The predictions of the model are to be used to highlight vessels and trips that may need closer inspection, predict landed catch for the fleet on specific days and predict the Total Allocated Quota (TAC) take-up, i.e. when the fleet will finish their allowed catch. For these goals the error is acceptable.

4.5.1 Validation data

The *mean squared error* (MSE), here defined as the *residual sum of squares* (RSS) divided by the *number of degrees of freedom* (DF) is a convenient way to compare the fit of models

$$MSE = \frac{RSS}{DF} = \frac{\sum_{i=1}^n (W_{jk} - \hat{W}_{jk})^2}{n - p}$$

The residuals are the difference between the model *estimated catch* and the reported catch, while the degrees of freedom are the difference between the number of observations n and the number of estimated population parameters in the model p . In the final model the number of μ parameters for vessels is 21, the number of β parameters for years is 4, in addition to the curvature parameter κ .

A problem with the MSE is its sensitivity to outliers, as it gives greater weight to large values compared to small values. A more reasonable comparison can be made using the *log-mean squared error* (LMSE)

$$LMSE = \frac{\sum_{i=1}^n (\log W_{jk} - \log \hat{W}_{jk})^2}{n - p}$$

One final year of data (2006) was reserved for validation of the model, and not used in the model building. As a general rule, the expected value of the MSE for a training set should be

$$\frac{n - p - 1}{n + p + 1} < 1$$

times the expected value of the MSE for the validation set, or

$$MSE_{training} = \frac{262 - 26 - 1}{262 + 26 + 1} MSE_{validation} = 0,8131 \cdot MSE_{validation}$$

Comparison of MSE and LMSE for the training and validation datasets gives an MSE of $4,3223 \cdot 10^{10}$ and LMSE of 0,2974 for the training period, and an MSE of $1,5749 \cdot 10^{12}$ and LSME of 7,7155 for the validation period.

The calculated MSE ratio is actually closer to 2,7% than 81% as expected. This essentially indicates that the validation dataset is drawn from a different population than the training set. In fact, cross-validation only yields meaningful results for stable systems, whereas features of the fishing process evolve over time with different CPUE and year parameters. Indeed, similar checks with 2-fold cross-validation (where the datasets are split into two equal parts) and repeated random sub-sampling validation (where the datasets are split randomly) which do not depend on the years used gives an MSE ratio between 30% and 70%, depending on how the datasets are split.

These results indicate that the model will do much worse when used to predict the fleet performance for a completely unknown year. One factor in this is that without any prior knowledge, the year parameter β_6 must be guessed. At least some data should be present for the year for a more reliable prediction.

4.5.2 Comparison to Days at Sea

A model that is particularly interesting for comparison is one that uses the coarser measure of *days at sea* as an estimate of effort. This will indicate if the the algorithm is doing any better by going through the procedure of classification and effort estimation, than could be done simply by taking the vessel's days at sea as a measure of effort

Recalling the final model Equations (4) and (5), a model that uses *days at sea* directly as its predicting variable is proposed. Assuming it is a continuous variable, i.e. the effort is calculated in fractions of days, this model is identical to the model given by Equation (4)

$$\log(W_{jk}) \sim t(\kappa_1(\log(D_{jk}) - \hat{bias}) + \mu_k + \beta_t, \kappa_1^2 \hat{\sigma}_\varepsilon^2 + \sigma_{00}^2, \nu) \quad (6)$$

with the median value of the catch then being

$$\text{median}[W_{jk}] = (D_{jk})^{\kappa_1} e^{-\kappa_1 \hat{bias}} e^{\mu_k} e^{\beta_t} \quad (7)$$

where D_{jk} is the effort in days at sea.

Running the search algorithm gives the results in Tables 19 and 20 below.

The meta-parameters were estimated as $\alpha = 2746,6$, $\sigma_\varepsilon^2 = 0,049076$, $\varphi = 27,317$ and $\nu = 8,6493$

The curve-parameter was estimated as $\kappa_1 = 0,61349$ with a 95% confidence interval of $\pm 0,12$ or $[0,50; 0,73]$.

Table 19: Estimated vessel parameters μ_k without outliers using Student's-t distribution and a curve parameter using days at sea as effort estimate

Vessel	μ_k	95% confidence interval	
1	-	-	-
2	7,84	$\pm 1,20$	[6,64;9,03]
3	7,35	$\pm 1,18$	[6,17;8,53]
4	7,58	$\pm 1,18$	[6,4;8,76]
5	6,35	$\pm 1,08$	[5,27;7,43]
6	7,21	$\pm 1,21$	[6,00;8,42]
7	7,19	$\pm 1,15$	[6,04;8,33]
8	7,47	$\pm 1,14$	[6,33;8,60]
9	7,41	$\pm 1,17$	[6,24;8,59]
10	7,24	$\pm 1,14$	[6,10;8,38]
11	7,62	$\pm 1,15$	[6,47;8,78]
12	7,23	$\pm 1,15$	[6,08;8,39]
13	6,58	$\pm 1,44$	[5,14;8,01]
14	7,57	$\pm 1,16$	[6,41;8,73]
15	7,63	$\pm 1,17$	[6,46;8,80]
16	7,14	$\pm 1,14$	[6,00;8,28]
17	7,48	$\pm 1,18$	[6,30;8,66]
18	7,36	$\pm 1,19$	[6,18;8,55]
19	7,31	$\pm 1,15$	[6,15;8,46]
20	7,69	$\pm 1,18$	[6,52;8,87]
21	7,55	$\pm 1,22$	[6,34;8,77]
22	7,72	$\pm 1,21$	[6,51;8,93]

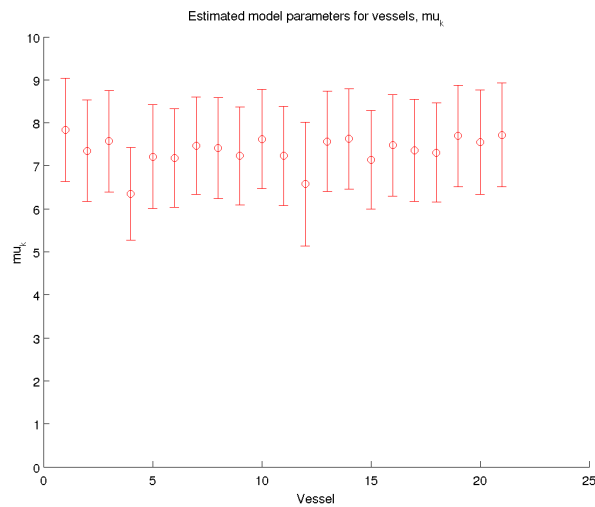


Figure 58: Estimated vessel parameters μ_k without outliers using Student's-t distribution and a curve parameter using days at sea as effort estimate

Table 20: Estimated year parameters β_t without outliers using Student's-t distribution and a curve parameter using days at sea as effort estimate

Year	β_t	95% confidence interval	
2001	-0,32	$\pm 0,19$	$[-0,51; -0,13]$
2002	0,04	$\pm 0,19$	$[-0,15; 0,24]$
2003	0,14	$\pm 0,19$	$[-0,04; 0,33]$
2004	0,00	$\pm 0,19$	$[-0,19; 0,19]$
2005	0	Reference year	-

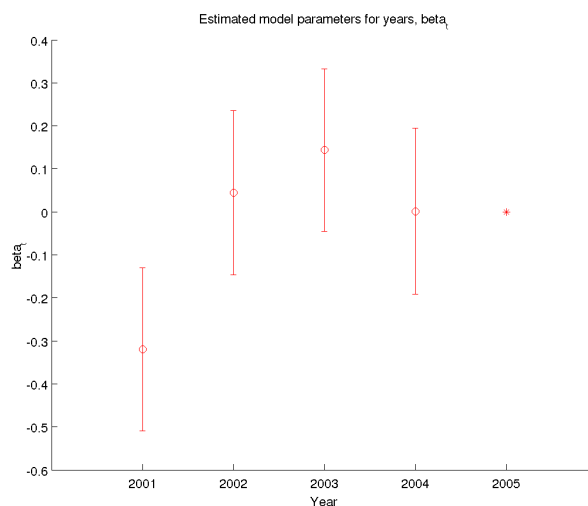


Figure 59: Estimated year parameters β_t without outliers using Student's-t distribution and a curve parameter using days at sea as effort estimate

Again plotting the sorted standardised residuals and the residuals vs. estimated effort in Figure 60 and 61 shows that the t-distribution fits the residuals reasonably well.

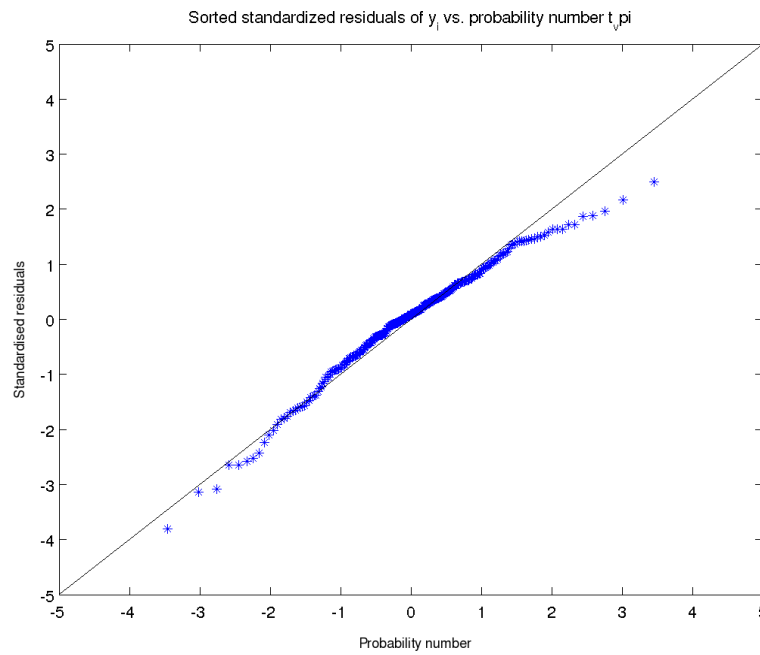


Figure 60: Sorted standardised residuals of y vs. probability number for the t -distribution, using days at sea as effort measure

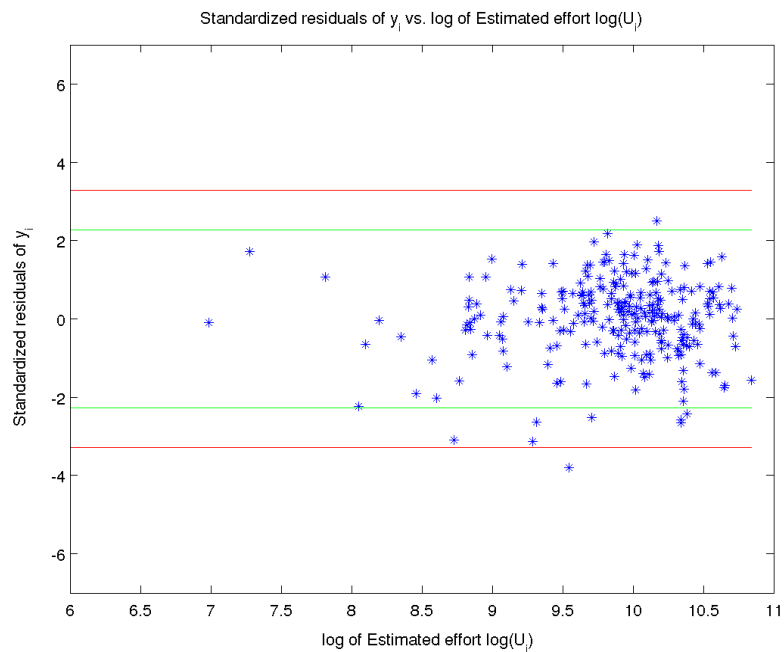


Figure 61: Standardised residuals vs. log of estimated effort using days at sea as effort measure

Using *days at sea* as an effort measure, the MSE is $6,0784 \cdot 10^{10}$ and the LMSE is 0,33286. Comparing this to the model using the effort estimate from the classifier discussed previously, the MSE and LMSE are slightly higher, but not drastically so. In Figures 60 and 61 the residuals also behave much better. The conclusion is that using *days at sea* is at least not a worse choice. The likely reason is that even though the *days at sea* measure includes a lot of time known to be non-fishing activity, the classifier is introducing extra variability to the effort measure that confuses the results.

4.6 High-resolution GPS-data

One concern for the study has been the relatively low¹⁴ resolution of the VMS data as mentioned at several points. Obviously, a resolution of one position every 2 hours cannot be expected to reliably pick up activities lasting only 30 minutes, such as deploying the gear, and with an average trawling period of about 13 hrs the uncertainty of the actual trawling time could be as much as 15%, just based on the low resolution of the VMS points. For example, registering a VMS point 30 minutes after the actual trawling activity has ended, where the vessel is now moving at cruising speed could result in the algorithm classifying that whole leg as cruising activity, essentially underestimating the effort by some 11,5%

But what is the real effect of the resolution on the classifier and effort estimation? And can an optimal or minimum resolution be recommended for reasonable accuracy?

High-resolution GPS positions for one vessel involved in the study has been graciously supplied, with as little as 3 second resolution. Selected portions of the vessel tracks are displayed in Figures 62 and 63.

Figure 62 shows the difference between a selected VMS track and a high-resolution GPS track for vessel 14, year 2003. On the left is the track from VMS data with 6 hour resolution, and on the right is the GPS track with 43 second resolution.

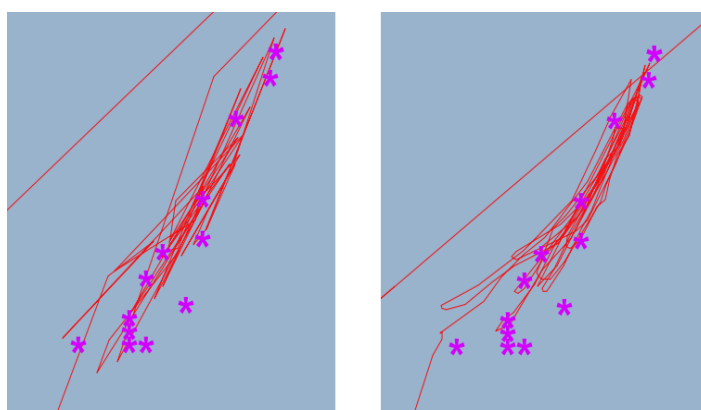


Figure 62: Example of VMS and high resolution GPS tracks

Figure 63 shows an example of a trawl that is not visible using 6 hour resolution VMS data on the left, but obvious using the higher resolution GPS data at 43 seconds on the right.

¹⁴ Note that "low resolution" means longer periods between position reports, and "high resolution" shorter periods

Track is for vessel 14, year 2003

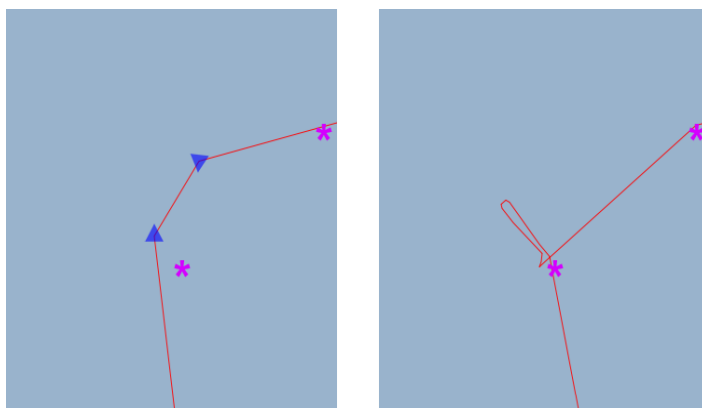


Figure 63: Example of missed features

The above figures show the more detailed features in the high-resolution data.

The classification algorithm and effort estimation were run on both the original VMS-track and the high-resolution GPS track, where the data was sampled at several different resolutions to simulate reporting periods from 60 seconds to 6,5 hours. The results can be seen in Table 19. The table shows the estimated effort at several chosen resolutions of VMS and GPS data, as well as the effort calculated by the actual activity (*midpoint*) algorithm, and the reported effort for the selected track.

A more extensive set of resolutions are plotted in Figure 64. The figure shows the estimated effort as calculated by the classification and effort estimation algorithms, when the resolution of the data is as shown on the x-axis.

Table 21: Estimated effort comparison between VMS and high-resolution GPS data

Source data and resolution	Effort [min]
Actual activity (<i>midpoint</i>)	2.339
Reported effort	2.520
Estimated effort	2.815
GPS at 60 sec resolution	2.694
GPS at 900 sec (15 min)	2.713
GPS at 1800 sec (30 min)	2.761
GPS at 7200 sec (2 hrs)	2.744
GPS at 21600 sec (6 hrs)	2.864

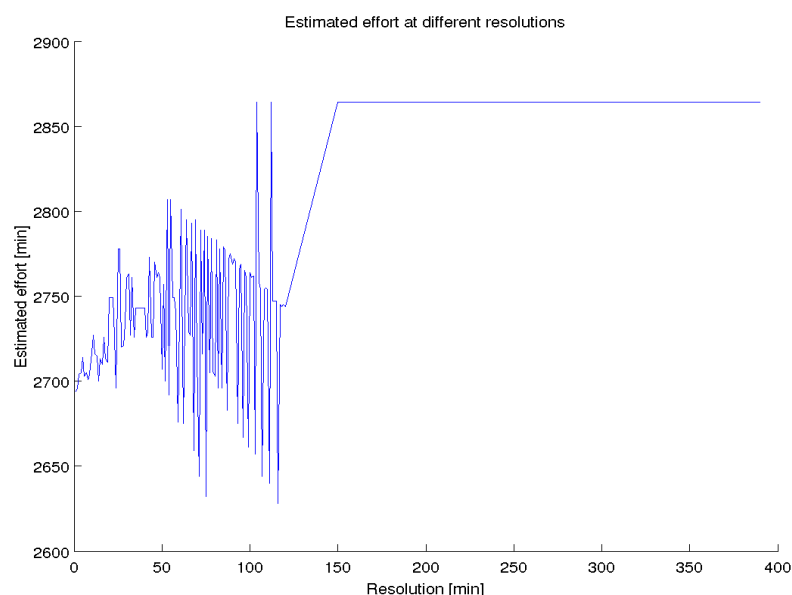


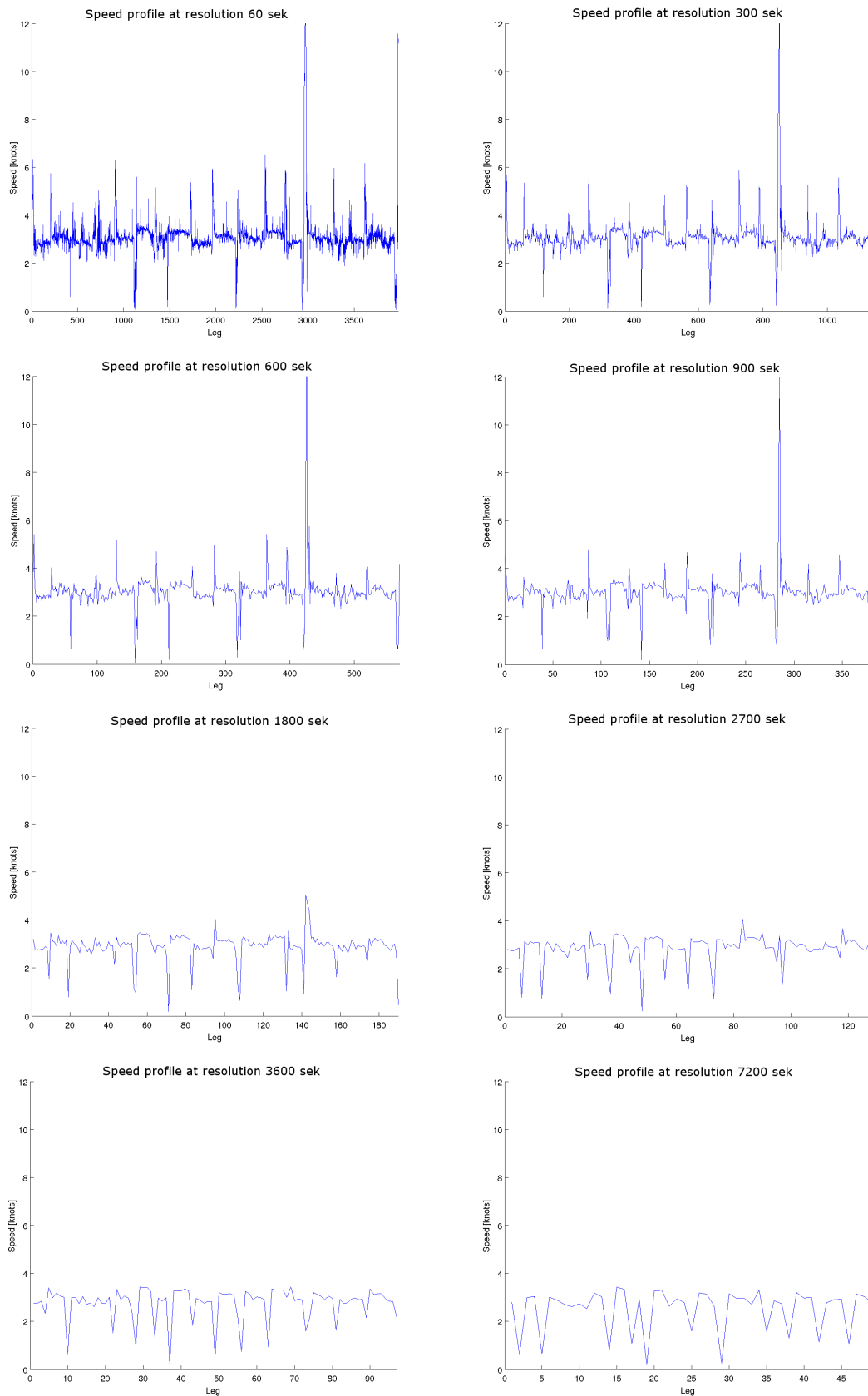
Figure 64: Estimated effort at different resolutions

It is evident that at high resolutions (1-30 min) the effort estimate steadily increases, then becomes a curiously stair-like decreasing function at resolutions between 1 and 2 hrs, and finally attains a constant value at the lowest resolutions above 2,5 hrs. There are in fact three effects observed here and these are discussed below.

First, the increasing effort estimate at the higher resolution end of the plot. The higher the resolution the closer the algorithm is to calculating the vessel's instantaneous speed. Due to a smoothing effect of averaging over the longer resolution windows, this results in more legs being classified as cruising than for lower resolutions, thus decreasing the effort estimate. This is perhaps best shown by looking at the speed plots in Figure 65. The figure shows vessel mean speed during legs calculated at progressively coarser resolutions, from 60 sec to 2 hrs. Ideally the 14 reported hauls should be detectable during this track.

The classifier is in fact dependent on the resolution it was optimised for, and in this case the classifier's decision boundary of 4,4 knots is optimised for a resolution of 2 hrs. Legs with speed over 4,4 knots are thus classified as cruising. At lower resolutions, as the timeperiod between position reports increases, the calculation of the mean speed tends to smooth out any spikes and fewer legs reach this speed.

This is also reflected in the fact that both the *reported effort* and the guess at the *actual effort* from the *midpoint* algorithm are lower than that estimated at the highest resolution (see Table 19). To use the classifier at the higher resolutions it would be necessary to optimise it again and find a new decision boundary. A quick check reveals that lowering the decision boundary to 3,4 knots would result in an estimated effort of 2,513 minutes, very close to the reported effort. Looking at the speed plots for the higher resolutions in Figure 65, it can also be seen that this would indeed seem to be a reasonable decision boundary to capture the trawling activity.

*Figure 65: Leg speed by resolution*

The second feature is the curious stair-like effect, caused by an offset effect of the resolution windows. As the windows grow, they may from time to time suddenly start or end at a position report where the vessel was speeding up considerably. This may cause the leg to suddenly be classified as cruising, where it was previously classified as trawling. Coupled with the fact that as the windows get larger, each is more influential in the effort estimate, this results in a downwards trend with increasing steps. This indicates that at resolutions above about 1 hrs, the classification algorithms become vulnerable not only to the large timeperiod between position reports, but also to the relative time of the position reports with respect to the start of a haul. Depending on when the leg start and endtimes are reported, the leg may be classified differently, even if it is of the same length of time.

Finally, at the very low resolutions above 2,5 hrs, the algorithm finally stops detecting any difference between the legs, and simply classifies them all as trawling activity.

A final note on the speed plots in Figure 65. In order to get an idea of the minimum resolution required to have any hope of detecting trawling activity, it is sufficient to follow the signal degradation from one resolution to the next. Ideally it should be possible to count the 14 reported hauls for this period, and it is no problem at the highest resolution of 60 seconds. As the resolution lowers, this gets progressively more difficult, although some beneficial reduction in noise is also seen. An ideal compromise seems to be at a resolution of 300 sec (5 minutes), while it is starting to get more difficult to identify hauls at 600 sec (10 minutes) and 900 sec (15 minutes), and the signal seem to be definitely lost by 1800 sec (30 minutes).

From these examinations it can be concluded that, above a data resolution of about 15 minutes, any algorithm will have significant trouble detecting trawling activity with any accuracy, and using a resolution of 5 minutes would be recommended. At the 2 hrs resolution provided by the VMS system trawling activity cannot be reliably detected. Also note that the classifier dependence on the resolution should be taken into account when using data with widely differing resolutions.

Chapter 5 - Discussion

In this work, an algorithm to identify and categorise vessel activity through analysis of positional data (VMS) is developed. The algorithm is capable of identifying high-speed cruising track legs, but no measure is found to adequately differentiate between fishing and non-fishing vessel activity at lower speeds, based on the data provided.

The data resolution of one VMS position every 2 hrs is found to be insufficient to detect the necessary features in vessel activity, and a case is made for a minimum resolution 15 minutes to be able to differentiate between fishing and non-fishing activity based on speed or other measures.

A statistical model to predict the total catch of a vessel from its estimated effort is presented, with prediction intervals. The model does reasonably well at prediction when partial data for the year is available, but has more inaccuracy when the year parameters are completely unknown a-priori.

Using a simpler effort measure of *days at sea* with the same model is shown to give at least as good or better results than the more complicated and calculations-intense effort estimate from VMS tracks.

In addition, *Appendix A* presents a pilot-system based on the algorithms, with several interesting features such as current fleet activity, total TAC uptake, alerts for effort and reported catch that do not match (falling outside of the 95% prediction intervals) and a heat-map showing a more realistic area fishing load using vessel tracks and activity, than using only single reported catch-points.

Chapter 6 - Conclusions and future work

Reviewing the research objectives, the conclusions of this thesis can be summarised as follows:

- A classification algorithm (linear discriminant) based on calculated speed from 2 hour VMS position reports reaches 75% classification accuracy. The algorithm cannot sufficiently differentiate between cruising and fishing at low speeds, but still returns a reasonable estimate of effort.
- A multivariate linear regression model based on the t-distribution with an additional curve parameter gives an adequate prediction of catch, given an estimated effort measurement.
- 95% prediction intervals were established on the model.

Other conclusions were:

- Using a simpler effort measure of *days at sea* with the same prediction model is shown to give at least as good or better results.
- VMS reports with a resolution of 15 minutes between positions is likely the minimum required resolution to adequately identify and classify vessel activity.

As suggestions for future work, the main points of interests would be

- using high-resolution GPS tracks and fully electronic catch logbooks with precise start and end times for hauls to classify activity.
- allowing vessel parameters to change between years to give more flexibility in fitting the data.
- using measures of distance to the nearest vessel to explore an effect on catch, i.e. if a vessel trawling in line after another experiences lower catch than expected.
- using measures to detect the pattern of vessels, such as when three or more vessels line up, indicating fishing activity.
- using more sophisticated stock models to predict catch, rather than assuming a linear relationship. An example would be “catch-at-length” curves.
- using catch logbooks of nearby vessels as input into the catch prediction for a vessel, or as comparison with the predicted catch.
- using high-resolution position data to detect and count number of hauls, and then using this measure as input into the prediction model

Bibliography

- Alþingi [The Icelandic Parliament]. Svar sjávarútvegs- og landbúnaðarráðherra við fyrirspurn Jóns Gunnarssonar um kostnað við fiskveiðieftirlit hjá Fiskistofu, þskj. 1012, 322. mál. *Vefútgáfa Alþingistiðinda*, 2010.
<<http://www.althingi.is/altext/138/s/1012.html>> [Last retrieved 29. september 2012].
- Avanti Communications. Analysis of Information Flow in Fisheries, CEDER work package deliverable D1.3. European Commission Joint Research Center, 2007.
- Árnason R. Operations Research and Fisheries Management, *EURO XXI, 21st European Conference on Operations Research*, Reykjavík. EURO XXI, 2006.
- CEDER consortium. *CEDER – Catch, Effort and Discard Estimates in Realtime*, European Commission Joint Research Center, 2006 <<https://ceder.jrc.ec.europa.eu>> [Last retrieved 29. september 2012].
- Cotter A.J.R., B. Mesnil, G.J. Piet. Estimating stock parameters from trawl cpue-at-age series using year-class curves, *ICES Journal of Marine Science* (1054-3139), Oxford university press, 2007-03 , Vol. 64 , N. 2 , P. 234-247, 2007
- Czerwinski I.A., J.C. Gutierrez-Estrada, J.A. Hernando-Casal, Short-term forecasting of halibut CPUE: Linear and non-linear univariate approaches. *Fisheries Research* 86 (2007) 120–128. 2007.
- Deng R., C. Dichmont, D. Milton, M. Haywood. Can vessel monitoring system data also be used to study trawling intensity and population depletion? The example of Australia's northern prawn fishery, *Canadian Journal of Fisheries and Aquatic Sciences*, Vol. 62, Iss. 3, March 2005, pp. 611-623(12), 2005
- Duda R.O., P.E. Hart, D.G. Stork. *Pattern Classification*. New York: Wiley Interscience, 2001.
- Gelman A., J.B. Carlin, H.S. Stern, D.B. Rubin. *Bayesian Data Analysis, 2nd edition*. Boca Raton: Chapman & Hall/CRC, 2004.
- Greenpeace. *Greenpeace Case Study on IUU fishing # 3 - Caught, RED-handed: Daylight Robbery on the High Seas*. Greenpeace, 2006.
<http://www.greenpeace.de/fileadmin/gpd/user_upload/themen/meere/IUU-3.pdf> [Last retrieved 29. september 2012].
- Gutierrez-Estrada J.C., C. Silva, E. Yáñez, N. Rodríguez, I. Pulido-Calvo. Monthly catch forecasting of anchovy *Engraulis ringens* in the north area of Chile: Non-linear univariate approach, *Fisheries Research* 86 (2007) 188–200, 2007.
- Hagstofa Íslands [Statistics Iceland]. *Hagtiðindi – Utanríkisverslun [Statistical Series – Foreign trade]* 91, no. 28, 2006:1. <<https://hagstofa.is/lisalib/getfiletrack.aspx?ItemID=4266>> [Last retrieved 29. september 2012].
- Hand D., H. Mannila, P. Smyth. *Principles of Data Mining*. Cambridge: MIT-press, 2001.

Hardin G. The tragedy of the commons. *Science* 162:1243-7, 1968.

Landhelgisgæsla Íslands [Icelandic Coast Guard]. Léleg veiði á úthafskarfaslóð á Reykjaneshrygg - níu sjóræningjaskip á miðunum [Low catch of Atlantic Redfish on the Reykjanes ridge – nine pirate vessels on the fishing grounds]. *Landhelgisgæsla Íslands*, 2006 <<http://www.lhg.is/frettirogutgafa/frettir/nr/180>> [Last retrieved 29. september 2012].

Mills C.M., S.E. Townsend, S. Jennings, P.D. Eastwood, C.A. Houghton . Estimating high resolution trawl fishing effort from satellite-based vessel monitoring system data, *ICES Journal of Marine Science*, 64: 248 – 255 , 2007.

NEAFC. *Illegal Fishing*. North East Atlantic Fisheries Commission, NEAFC, 2010 <<http://neaftc.org/illegalfishing>> [Last retrieved 29. september 2012]

OECD. *Draft Synthesis Report on IUU Fishing Activities*. OECD AGR/FI (2004), Paris, 2005.

Sjávarútvegs- og landbúnaðarráðuneytið. 26. ársfundur Norðaustur-Atlantshafs fiskveiðinefndarinnar, NEAFC. *Sjávarútvegs- og landbúnaðarráðuneytið*, 2007 <<http://www.sjavarutvegsraduneyti.is/frettir/sjreldra/nr/7522>> [Last retrieved 29. september 2012].

Stefánsson G. *Stærðfræðileg fiskifræði [Mathematical Fisheries Analysis]*. Reykjavík: Hafrannsóknastofnunin [Icelandic Marine Research Institute], 1997.

Thomson A. The management of Redfish (*Sebastes Mentella*) in the North Atlantic Ocean - A Stock in Movement. *FAO Fisheries Report*, no. 695, Supplement (Papers presented at the Norway-FAO expert consultation on the management of shared fish stocks, Bergen, Norway, 7-10 October 2002). Rome, 2002.

Vincenty T. Vincenty's formulae, *Wikipedia*, 2009. <http://en.wikipedia.org/wiki/Vincenty%27s_formulae> [Last retrieved 29. september 2012].

Appendix

Appendix A - Prototype system CARFI

In this appendix a prototype software system developed in conjunction with the classification algorithms and prediction models is presented. The aim of the system is to demonstrate the feasibility of bringing real-time information on fisheries to stakeholders. It illustrates the use of the algorithms developed in a real-world environment.

A general description of the system is given, with special emphasis on illustrating practical use cases for the developed algorithms. The particulars of the system design are however outside of the scope of this work, interested readers are referred to the workpackage deliverables of the CEDER project.

Table of Contents

A.1 - System overview.....	A-3
A.2 - Data collection	A-4
A.3 - Processing	A-4
A.3.1 - Clean positions.....	A-4
A.3.2 - Run trip identifier.....	A-4
A.3.3 - Run activity classification.....	A-4
A.3.4 - Run actual activity and catch/leg connection using midpoint.....	A-5
A.3.5 - Run actual activity using 30 nml proximity.....	A-5
A.3.6 - Run alarms check.....	A-5
A.3.7 - Calculate Additional Predictor Variables.....	A-5
A.4 - Analysis.....	A-6
A.4.1 - Alarms.....	A-6
A.5 - Vessel tracks.....	A-7
A.5.1 - Vessel speed.....	A-8
A.5.2 - Vessel predicted activity.....	A-8
A.5.3 - Vessel actual activity.....	A-10
A.5.4 - Fleet positions.....	A-10
A.5.5 - Area load.....	A-11
A.5.6 - Vessel Catch.....	A-12
A.5.7 - Vessel TAC uptake.....	A-14
A.6 - Other features.....	A-15
A.6.1 - Dataset overview.....	A-15
A.6.2 - Validate trip identifier.....	A-15
A.6.3 - Activity classification accuracy.....	A-15
A.6.4 - Leg activity classification comparison.....	A-15
A.6.5 - High resolution GPS tracks.....	A-15
A.6.6 - Create database.....	A-15
A.6.7 - Import dataset.....	A-15
A.6.8 - Export dataset.....	A-15

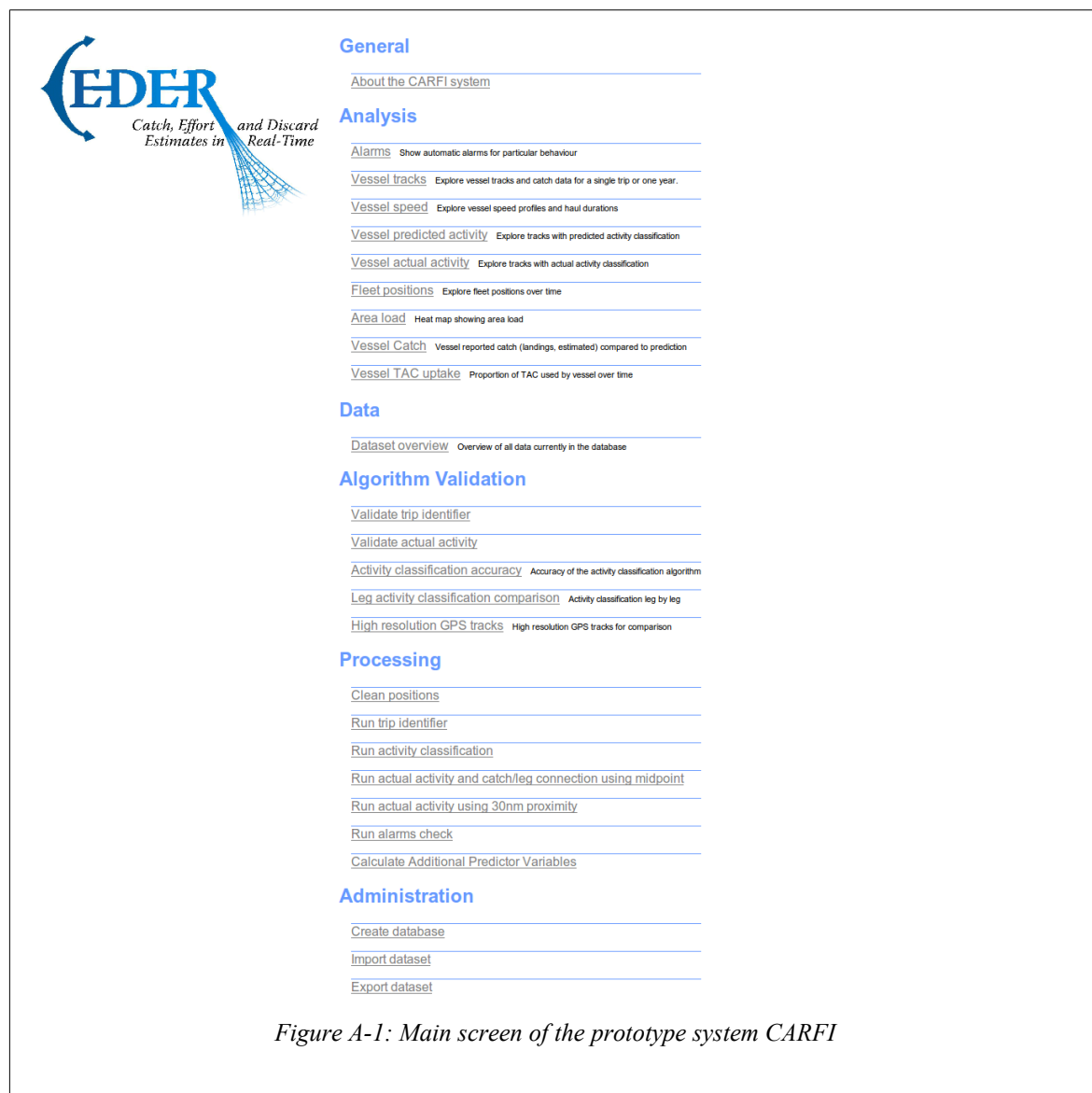


Figure A-1: Main screen of the prototype system CARFI

A.1 - System overview

The name of the system is "CARFI" (Ceder Atlantic Redfish Fisheries Information system) and is comprised of three parts:

- Data collection module
- Processing module
- Analysis module

Figure A-2 shows how data in the form of VMS-reports, catch-reports (logbooks) and landing reports are collected under the data collection module and stored in a database, where they are available to the processing and analysis modules.

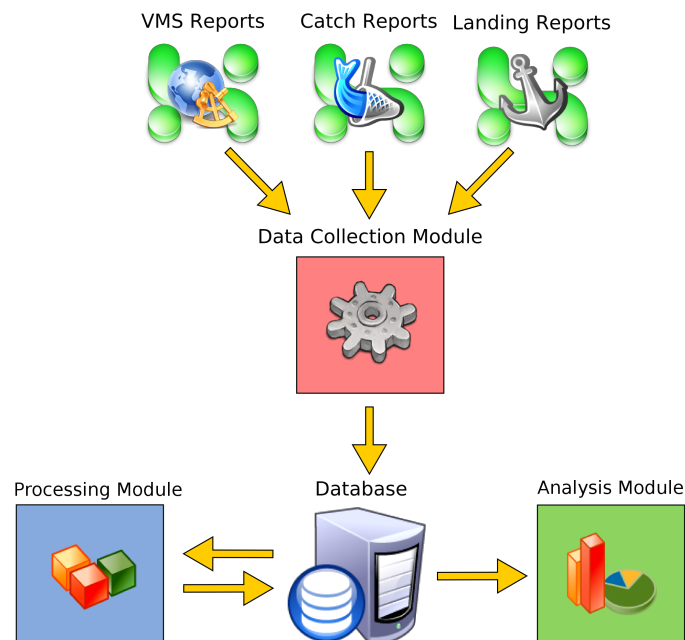


Figure A-2: System components of the CARFI prototype system.

A.2 - Data collection

The system collects data delivered from the already implemented systems of the Icelandic Oceanographic Institute and the Icelandic Directorate of Fisheries and stores in a database for analysis.

The data collected includes

- VMS data
 - Date and time
 - Position (longitude, latitude)
 - Target Species
 - Gear type
- Logbook / eLogbook
 - Date and time of gear deployment
 - Position of gear deployment
 - Catch weight (green weight)
- Landing reports
 - Trip dates and times
 - Landed catch

A.3 - Processing

The system includes a processing module whose purpose is mainly to run classification algorithms on the VMS-data to determine when a vessel is trawling (gear deployed) and when it is cruising. The result is an estimate of effort. A secondary purpose of this module is to provide various functionality required during the development of the models in this thesis.

A.3.1 - Clean positions

Purpose: To do initial cleaning and sanity checks on incoming positional data.

Description: Using this page will show positions with impossible speeds, short time intervals between VMS reports or large distances between points. It will automatically invalidate one of any duplicate positions where the time difference is zero and the coordinates are unchanged. Other positions can be cleaned manually.

A.3.2 - Run trip identifier

Purpose: To group VMS-positions into trips.

Description: Using this page will run an algorithm to identify fishing trips in VMS-position data, based on landings from catch logbooks. The trips are labelled numerically within each year.

A.3.3 - Run activity classification

Purpose: To classify legs by vessel activity.

Description: Using this page will run an algorithm to classify legs in VMS-position data, based on speed.

A.3.4 - Run actual activity and catch/leg connection using midpoint

Purpose: To determine the actual activity of vessels and connect legs to catch-points, using the midpoint algorithm.

Description: Using this page will run an algorithm to connect each catch entry to all legs during the haul. The algorithm assigns actual_activity to each leg based on this connection. This algorithm calculates the midpoint of each leg and selects the midpoint closest to the catch as the haul starting leg (within same day or last position of previous day). Also, maximum distance from first position is 30 nml .

A.3.5 - Run actual activity using 30 nml proximity

Purpose: To determine the actual activity of vessels and connect legs to catch-points, using the 30 nml proximity algorithm.

Description: Using this page will run an algorithm to find all VMS-positions within 30 nml of any catch point. The algorithm assigns actual_activity to each leg based on this distance, and connects the leg to the catch-point. This algorithm calculates the proximity of each leg to any catch and assigns it to that catch if the distance is less than 30 nml.

A.3.6 - Run alarms check

Purpose: Check if any predefined alarms have been triggered

Description: Using this page will run an algorithm to check predefined alarms and store results in the database. Some alarms allow a sensitivity setting.

The currently defined alarms are:

- **EFFORT_HIGH:** The estimated effort is very high compared to the vessel's reported catch, i.e. it falls outside of the upper prediction interval.
- **EFFORT_LOW:** The estimated effort is very low compared to the vessel's reported catch, i.e. it falls outside of the lower prediction interval.
- **TRAWLING_NO_CATCH:** The activity classifier has assigned a leg activity as "trawling", but there was no reported catch-point found within the sensitivity limits. Default sensitivity limits are within 18 hrs and 20 nml.
- **CATCH_NO_TRAWLING:** The activity classifier has not found any legs classified as "trawling" connected to a reported catch-point within the sensitivity limits. Default sensitivity limits are within 18 hrs and 20 nml.

A.3.7 - Calculate Additional Predictor Variables

Purpose: Used during development to calculate additional predictor variables for analysis

Description: Using this page will run an algorithm to calculate additional predictor variables for classifier analysis. The variables are:

- Derivative of mean leg speed
- Course change
- Running average of course change
- VMS position density
- Fleet dispersal

A.4 - Analysis

The system includes an analysis module to display the results of the prediction models and classification algorithms. This includes vessel tracks, activity classification, fleet positions, area load, catch and TAC uptake. In the next section we will show some examples of these.

To predict individual vessel and total fleet catch based on effort, the analysis module includes the SLR and MLR models developed in the preceding chapters. The design of the module is such that it is easy to add further models for this purpose.

A.4.1 - Alarms

Purpose: Show automatic alarms for particular behaviour

Description: This page displays results of automated alarms.

The active alarms are:

- **EFFORT_HIGH** - Vessel reports catch ABOVE the effort model prediction interval, indicating unusually high catch for the estimated effort
- **EFFORT_LOW** - Vessel reports catch BELOW the effort model prediction interval, indicating unusually low catch for the estimated effort
- **TRAWLING_NO_CATCH** - Vessel shows trawling behaviour, but fails to report any catch in the area or within a reasonable timeframe
- **CATCH_NO_TRAWLING** - Vessel reports catch where no trawling behaviour in the area or within a reasonable timeframe

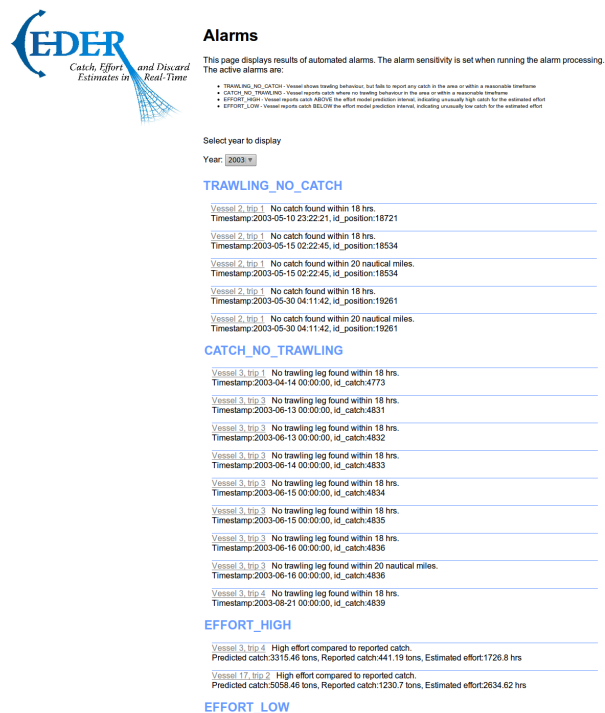


Figure A-3: Alarms page

Looking at the list of alarms shown in Figure A-3, the first group labelled

TRAWLING_NO_CATCH lists instances where a vessel has been detected trawling, but reported no catch. The second group lists instances where a vessel has not been detected trawling, but nevertheless reported catch. Both of these groups are likely to be indicative of data-quality issues, rather than actual infractions, but such activities might include fishing outside of quota or transshipment between vessels (not common in the Icelandic North-Atlantic fisheries fleet, but very common among other national fleets and pirate vessels).

The groups labelled EFFORT_HIGH and EFFORT_LOW are particularly interesting, since there we see an example of the catch prediction models in action. The alarms are triggered when the vessel reported catch falls outside of the prediction intervals. E.g. here the predicted catch is 3315.46 tons, but the reported catch is 441.19 tons, for an estimated effort of 1726.8 hrs during the trip.

Each alarm can be clicked to examine the vessel track for the area and timeperiod using the *Vessel predicted activity* page described in a following section.

A.5 - Vessel tracks

Purpose: Explore vessel tracks and catch data for a single trip or one year.

Description: This page displays the vessel tracks as delivered from the data sources. The page uses Google Maps to plot vessel VMS-points in tracks, grouped by trip, and Timeplot from MIT's SIMILE project to show vessel speed and trawls.

As illustrated in Figure A-4 and A-5, the user can select a vessel and year, zoom in and click specific points for further information. They can also select the option of displaying the reported catch-points from the vessel's catch logbook as pink stars, seen in Figure A5.

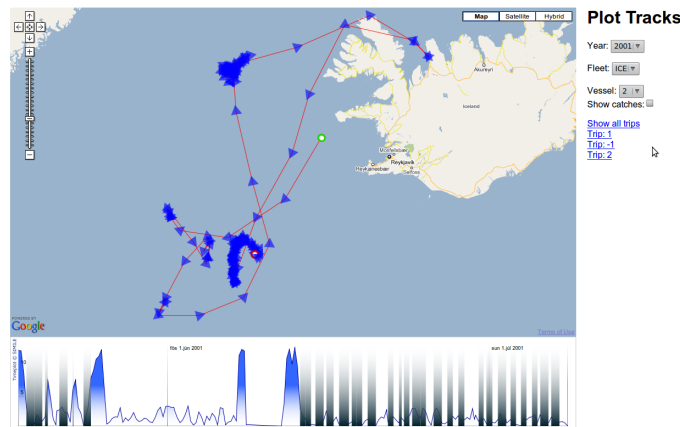


Figure A-4: Vessel Tracks page

Below the map of the vessel track is the timeplot, showing the vessel speed in blue overlayed with gray columns representing reported trawling. In figure A-5 a VMS-point towards the end of the vessel's trip has been selected, and highlighted the corresponding time in the timeplot. It is obvious that indeed, as the vessel approaches the end of its trip and heads to port, its speed increases.

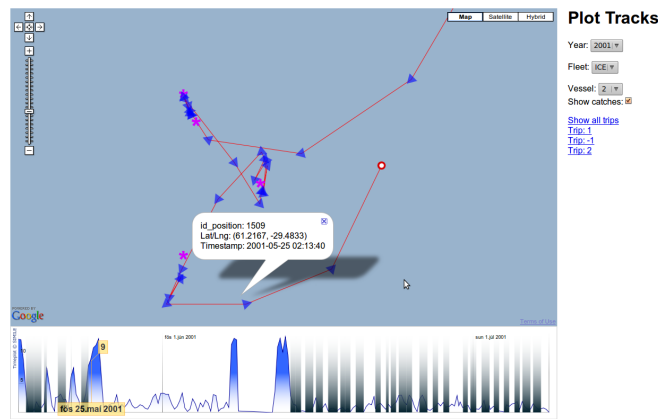


Figure A-5: Vessel Tracks page detail

Google maps gives us the interesting option of showing the oceanographic features of the sea floor, where we can pick out the Reykjanes-ridge in Figure A-6.

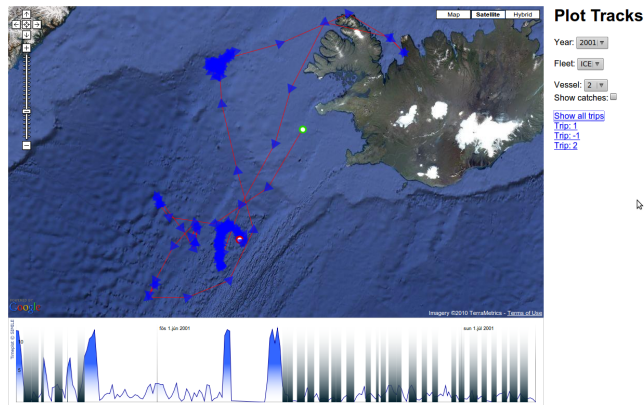


Figure A-6: Vessel Tracks with oceanographic features

A.5.1 - Vessel speed

Purpose: Explore vessel speed profiles and haul durations

Description: This page displays only the speed profiles for each vessel, and is the same as the bottom plot from the *Vessel tracks* page described in the preceding section.

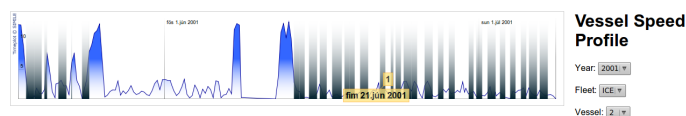


Figure A-7: Vessel speed page

A.5.2 - Vessel predicted activity

Purpose: Explore tracks with predicted activity classification

Description: This page displays the vessel tracks similarly to the *Vessel tracks* page, but with the added classification of track legs. Blue legs are classified as *cruising*, while red legs are classified as *trawling*.

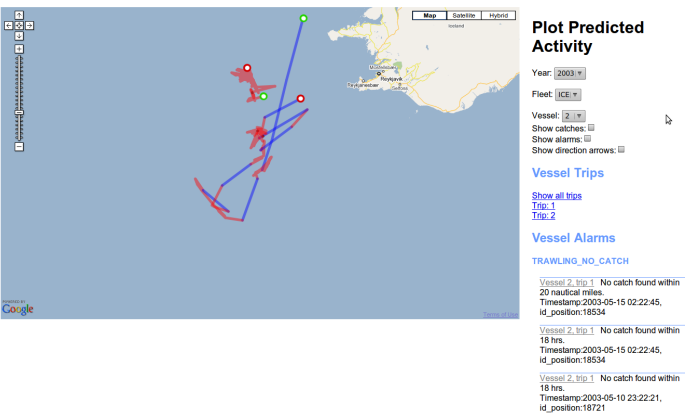


Figure A-8: Vessel predicted activity page

As before the user can select the option of displaying the reported catch-points. In addition, they can select the option of displaying any alarms raised for this vessel, which then show up as red triangles seen in Figure A-9. Clicking the alarm sign brings up further details, such as in the illustrated example, where the vessel has reported higher catch than the estimated effort would seem to allow. All alarms for the vessel are listed to the right of the screen.

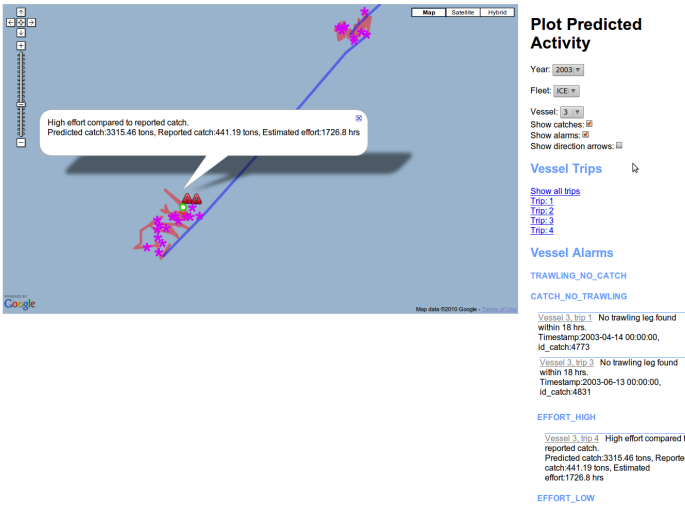


Figure A-9: Vessel predicted activity page detail

A.5.3 - Vessel actual activity

Purpose: Explore tracks with actual activity classification

Description: This page is essentially the same as the preceding *Vessel predicted activity* page, with the important difference that here we are examining the *actual* activity of the vessel, as determined by the system algorithms and shown in Figure A-10.

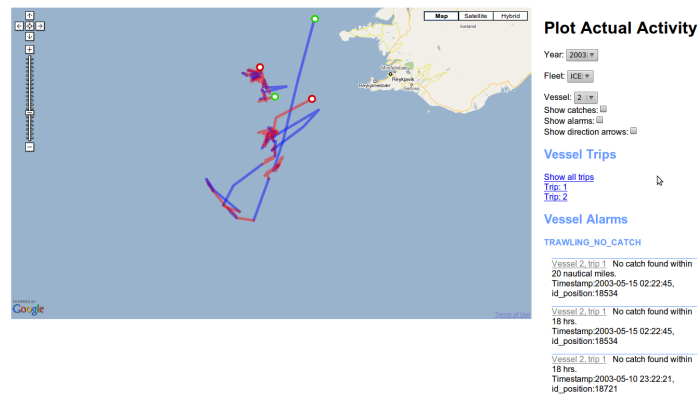


Figure A-10: Vessel actual activity page

A.5.4 - Fleet positions

Purpose: Explore fleet positions over time

Description: This page shows the fleet positions at any given time. Each vessel in Figure A-11 is depicted on the map in blue if it has been determined to be cruising, or red if it is trawling. This information comes from the classification algorithms. The user can select the option of displaying the Icelandic EEZ, and vessel wake (last leg) to better identify vessel movements.

Note that the slider at the bottom of the map can be moved to examine the vessel positions in time, creating an interesting time-lapse graphic.

A previously mentioned feature of the Atlantic Redfish fishery can be seen in Figure A-11, where the (red) vessels line up in a row to trawl.

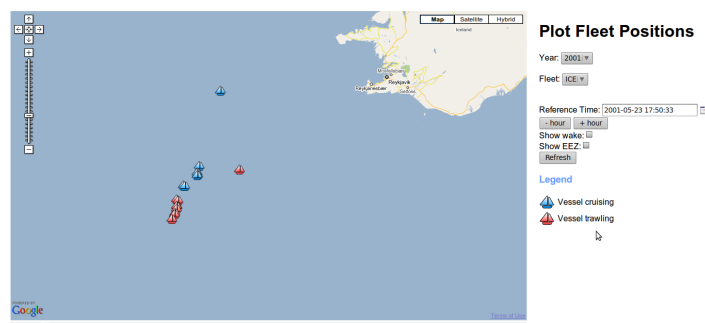


Figure A-11: Fleet positions page

Another feature can be seen in Figure A-12, where some vessels are engaged in what is known as the "line-dance", where they are required to stay out of the Icelandic EEZ (red line), but try to trawl as close as possible.

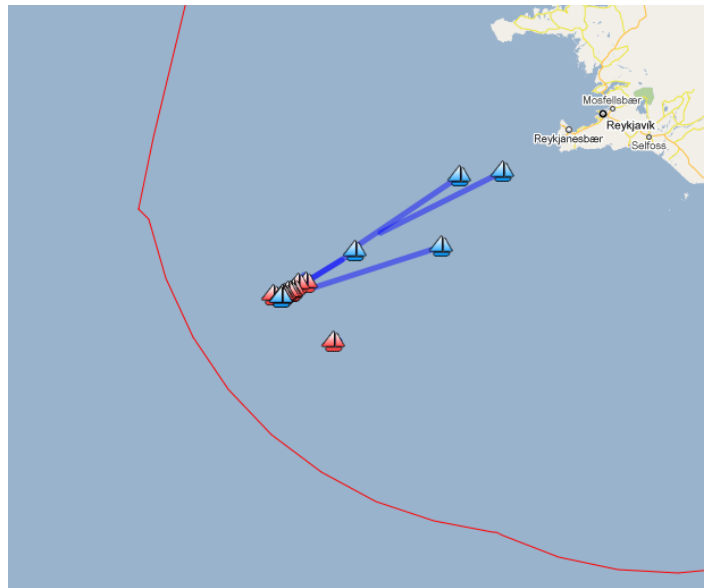


Figure A-13: Vessels on the move with wake marks

Finally, wake marks can be turned on as seen in Figure A-13 to show the vessel last movements.

A.5.5 - Area load

Purpose: Heat map showing area load

Description: This page shows a heat-map of the area load on a cell grid of 0.1 degrees square (this precision can be adjusted in the source code). The user can select either catch reports or estimated effort as the basis for the map.

Usually, catch-points have been used to generate this type of estimate of area load, but with the effort estimation algorithms we believe we can approach something more reasonable. With catch reports as the basis for the map as in Figure A-14 each trawl only applies to one map cell, while with the effort estimation as in Figure A-15, each cell containing a trawling leg is affected resulting in a wider distribution of load.

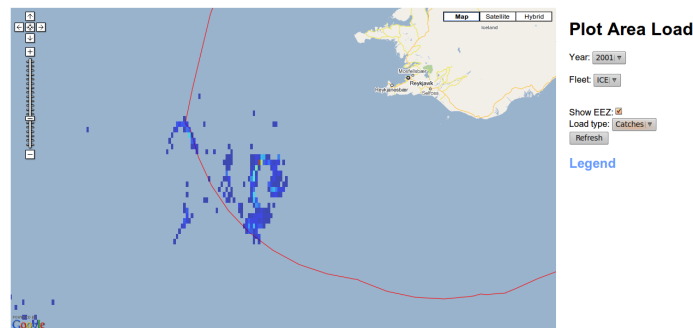


Figure A-14: Area load page from catch-reports

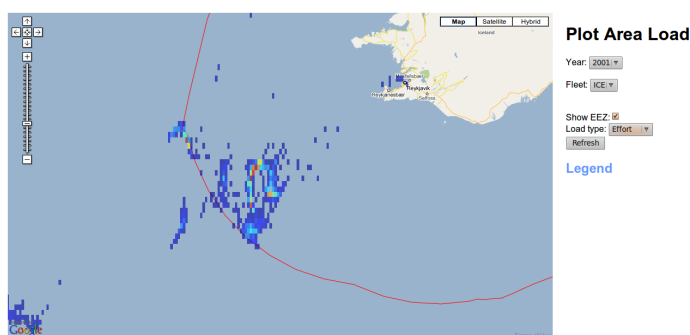


Figure A-15: Area load page from estimated effort

A.5.6 - Vessel Catch

Purpose: Vessel reported catch (landings, estimated) compared to prediction

Description: This page displays reported catch and predicted catch for each vessel and trip, according to the selected prediction model. The graph also shows the prediction interval for each prediction point.

In Figure A-16 a simple linear regression model (SLR) has been selected for all years and all vessels. The figure shows the simple linear regression model (SLR) for all years and all vessels. The blue line indicates the model prediction, the red lines represent confidence intervals on the regression parameter and the green lines are prediction intervals. The red dots represent reported catch.

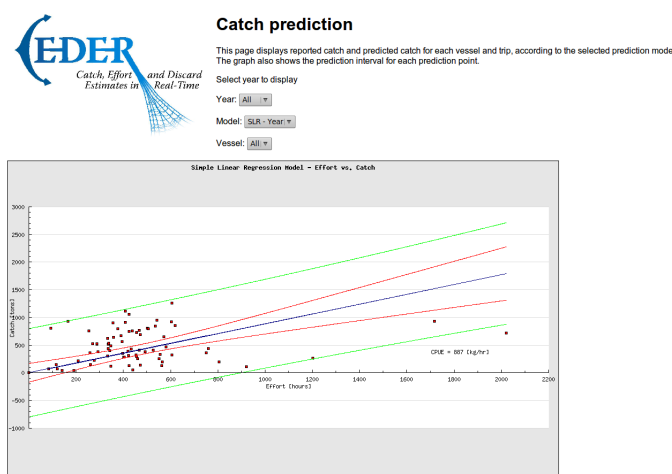


Figure A-16: Vessel Catch page – SLR.

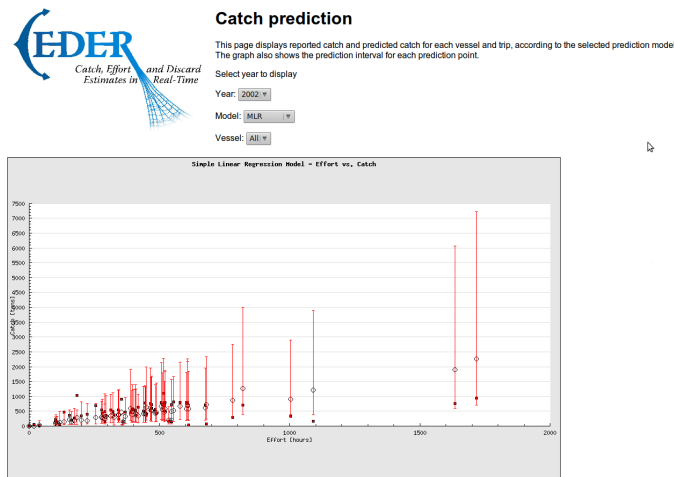
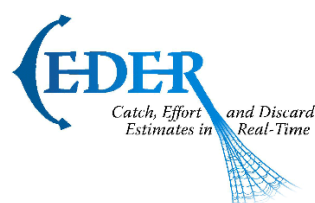


Figure A-18: Vessel Catch page – MLR, all vessels, year 2002.

In Figure A-17 we select the multivariate linear regression model (MLR) for one year and all vessels. The figure shows the multivariate linear regression model (MLR) for the year 2002 and all vessels. The black circles indicate the model prediction, with the red error bars representing the prediction interval for each. The red dots are the reported catch.

For this combination we cannot draw a linear plot, since the model parameters are dependent on the year and vessel.

In Figure A-19 we have further constrained the model to one vessel, and this means we can plot the predicted linear relationship between estimated effort and catch. The figure shows the multivariate linear regression model (MLR) for the year 2002, further constrained on vessels 8. The blue line indicates the model prediction, with the red lines representing the prediction interval. The red dots are the reported catch.



Catch prediction

This page displays reported catch and predicted catch for each vessel and trip, according to the selected prediction model. The graph also shows the prediction interval for each prediction point.

Select year to display

Year: 2002

Model: MLR

Vessel: 8

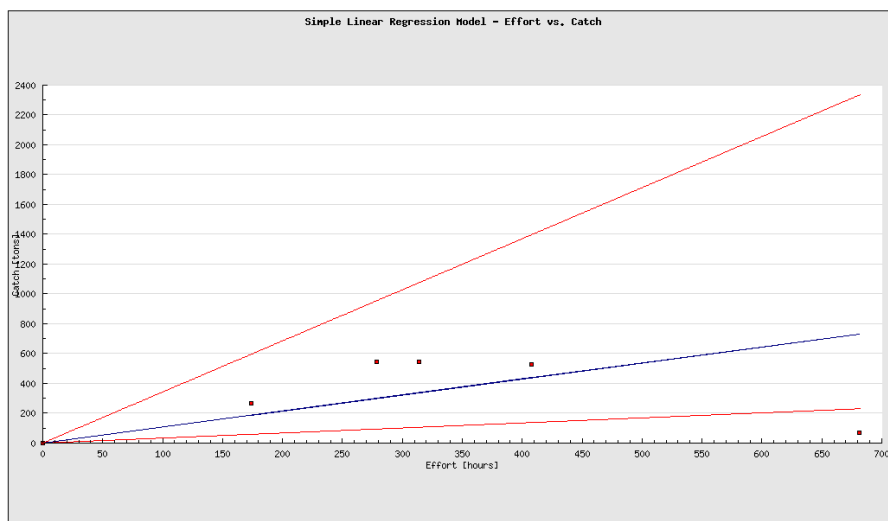


Figure A-19: Vessel Catch page – MLR, vessel 8, year 2002.

A.5.7 - Vessel TAC uptake

Purpose: Proportion of TAC used by vessel over time

Description: This page displays reported catch and predicted catch for the fleet over time, according to the selected prediction model. This enables the user to predict when a certain TAC (Total Allowed Catch) level is likely to be reached.



TAC Uptake

This page displays reported catch and predicted catch for each vessel over time, according to the selected prediction model. The graph also shows the prediction interval for each prediction point.

Select year to display

Year: 2003

Model: MLR

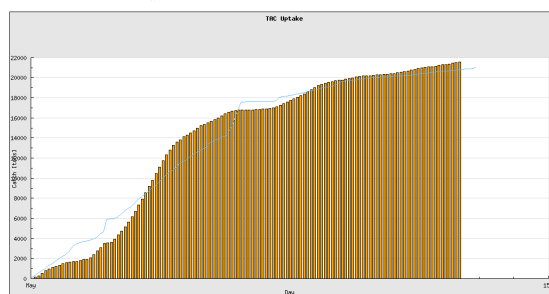


Figure A-20: TAC uptake. The figure shows the cumulative reported (landed) catch each day and the MLR model prediction.

A.6 - Other features

In addition, the system includes some administrative and secondary features.

A.6.1 - Dataset overview

Purpose: Overview of all data currently in the database

Description: This page lists basic statistics of the data in the database, number of vessels, total catch, catch-reports, landing reports, and VMS-points for each year.

A.6.2 - Validate trip identifier

Purpose: Overview of all data currently in the database

Description: This page shows in a visual way the results of the trip identifier algorithm and validation, so as to confirm that the algorithm is working properly and enable the user to make corrections.

A.6.3 - Activity classification accuracy

Purpose: Accuracy of the activity classification algorithm

Description: This page gives an overview of the accuracy of the activity classification algorithm.

A.6.4- Leg activity classification comparison

Purpose: Activity classification leg by leg

Description: This page lists each leg and shows it's actual activity and the predicted activity from the classification algorithm.

A.6.5 - High resolution GPS tracks

Purpose: High resolution GPS tracks for comparison

Description: This page shows high-resolution GPS tracks provided for one vessel for comparison.

A.6.6 - Create database

Purpose: Create the database and it's tables from scratch

Description: This page enables the user to set up the system database initially.

A.6.7 - Import dataset

Purpose: Import data

Description: This page enables the user to import catch logbooks, landing reports and VMS data files in .csv format to the database.

A.6.8 - Export dataset

Purpose: Export data

Description: This page enables the user to export various datasets for use in the development of the models.