NTNU

Norwegian University of
Science and Technology

Forecasting the Atlantic Salmon Spot Price Using the Autoregressive Distributed Lag Model

Magnus Lien

Industrial Economics and Technology Management
Submission date: June 2018
Supervisor: Sjur Westgaard, IØT

Norwegian University of Science and Technology
Department of Industrial Economics and Technology Management

Problem description

The price of Atlantic salmon is characterized by high volatility both in terms of frequency and magnitude, which imposes uncertainty and costs on the entire value chain of salmon farming. The salmon industry has experienced a rapid growth over the past decades, and the production, processing and marketing of salmon has become a multibillion dollar industry. In addition, the industry has an increased presence in capital markets. As the market of salmon farming is becoming more globalized and competitive, the companies' ability to limit unnecessary costs is an important requirement to maintain market positions. It is useful for the participants of the salmon market to have reliable estimates of future salmon prices. A satisfying price model is an important tool in order to establish operational efficiency, and can benefit the entire supply chain for salmon farming. Financially, a price model can help improve investment decisions, valuation of bonds and stocks, and risk management.

In this study, I propose a new forecast model for the prediction of the Atlantic salmon spot price, represented by the NQSALMON index. The methodology builds on a multivariate ordinary least square regression model presented by Sandaker et al. (2016). The approach presented in this paper is twofold:

1. Methodologically, I build a regression framework based on a combination of the autoregressive distributed lag model and the partial least square regression, in order to predict the 3-, 6- and 9-months ahead log transformed salmon spot price. I use a genetic algorithm for variable selection, which selects sets of covariates that have a unique cointegrated relationship with the salmon spot price, which enables the use of non stationary data in the regression framework.

2. Qualitatively, I assess if the avoidance of stationary transformation of data, and the removal of regression problems related to intercorrelated predictors, can help improve the forecast accuracy of salmon prediction models. This is done by comparing the results of the proposed methodology with the results of a multivariate ordinary least square regression model implemented on stationary transformed data. In addition, I assess the value of using exogenous variables in salmon prediction, relative to using only past salmon prices.

II

Preface

This thesis is motivated by the operations of Norwegian salmon producers and previous masters' theses written on the same subject. It is written as a cooperation between Applied Economics and Operations Management and Financial Engineering, at the Department of Industrial Economics and Technology Management, Norwegian University of Science and Technology (NTNU). It is original and independent work performed by Magnus August Lien, during the spring of 2018.

I would like to thank my supervisor, Professor Sjur Westgaard at the Department of Industrial Economics and Technology Management (NTNU), for helpful guidance and advice.

Abstract

Salmon farming is the largest growing food supply sector in the world. Alongside the growth, the industry is becoming more competitive and is strengthening its position in the capital markets. However, the salmon price is characterized by high volatility, which imposes uncertainty and additional costs on the value chain of salmon production. The participants of the salmon market can benefit from a reliable price model, as it can be used to improve decision making regarding the operational and financial aspects of the industry that are subject to the price uncertainty. This includes the timing of salmon harvest, required machine capacity, investment decisions and stock valuation.

This study aims to provide a framework for predicting the spot price of the Atlantic salmon, represented by the NQSALMON index. I present a new model, called the ARDL-PLS model, which is used to predict the 3-,6- and 9-month ahead spot price. The ARDL-PLS model is a combination of the autoregressive distributed lag model and the partial least square regression. Each of the three months to be predicted are appointed a designated submodel, which are independent on the other submodels. The covariates used in each submodel are selected from a database of 61 candidate predictors, using a genetic algorithm (GA) for feature selection. The GA-search is constrained to only select subsets of predictors that have a unique cointegrated relationship with the salmon price, which enables the use of non-stationary data in the ARDL-PLS model. The out-of-sample results of the ARDL-PLS model is compared to the results of an ordinary least square regression (OLR) model, which is implemented on stationary transformed data

There are several encouraging results of this study. The genetic algorithm functions well as a feature selection tool, as it is relatively quickly able to select subsets of predictors that cointegrate with the salmon price, and that results in a favourable goodness-of-fit. Generally the ARDL-PLS model is able to explain a large degree of the variance in the salmon price, and the predictive accuracy of the ARDL-PLS model is better than the OLR model on all forecast horizons. The two models perform relatively similar in the 3 month-ahead prediction, but the ARDL-PLS model excel in the longer horizons. I attribute the performance difference between the OLR model and the ARDL-PLS model with two factor: (1) In contrast to the ARDL-PLS model, the OLR model requires the data to be

stationary transformed, which removes the long term information in the data. (2) Unlike the OLR model, the use of intercorrelated predictors in the ARDL-PLS model does not affect the stability of the estimated regression coefficients, as the ARDL-PLS model transforms the covariates into orthogonal uncorrelated factors. Finally, the results indicate that the use of exogenous variables in regression based salmon price models, significantly increases the prediction accuracy.

Sammendrag

Lakseindustrien er i dag den raskest voksende matforsyningssektoren i verden. I tillegg til den høye veksten, har lakseindustrien blitt mer kompetitiv og fått økt tilstedeværelse i kapitalmarkedene. Samtidig er lakseprisen svært volatil, noe som resulterer i økt usikkerhet og økte kostnader i hele verdikjeden for lakseproduksjon. Markedsdeltagere vil kunne dra nytte av en pålitelig modell for prediksjon av lakseprisen, da den kan brukes til å forbedre beslutningsgrunnlaget knyttet til operasjonelle og finansielle aspekter ved industrien som er påvirket av prisusikkerheten. Eksempler på slike aspekter er slaktetidspunkt, påkrevd maskinkapasitet, investeringsbeslutninger og verdivurdering av aksjer.

Målet ved dette studiet er å utvikle et rammeverk for prediksjon av laksespotprisen, som er representert ved NASDAQ Salmon Index (NQSALMON). Jeg presenterer en ny modell, kalt ARDL-PLS, som blir brukt til å predikere prisen 3, 6 og 9 måneder frem i tid. ARDL-PLS modellen er en kombinasjon av en autoregressive distributed lag (ARDL) modell og en partial least square (PLS) regresjonsmodell. Hver av de tre månedene som predikeres, tilegnes en uavhengig delmodell. Kovariatene som blir brukt i hver delmodell er utvalgt fra et datasett bestående av 61 potensielle forklaringsvariabler, ved hjelp av en genetisk algoritme. Den genetiske algoritmen er begrenset til å kun velge variabelsett som har et unikt kointegrert forhold med lakseprisen, noe som muliggjør bruken av ikke-stasjonære tidsserier i ARDL-PLS modellen. Out-of-sample resultatene til ARDL-PLS modellen sammenliknes opp mot resultatene til en ordinary least square regression (OLR) modell, hvor den sistnevnte modellen er implementert på stasjonær transformert data.

Det er flere lovende resultater fra dette studiet. Den genetiske algoritmen fungerer godt som et verktøy for variabelseleksjon, da den relativt raskt finner kointegrerte sett av variabler med tilfredsstillende forklaringsevne. Generelt klarer ARDL-PLS modellen å beskrive en stor andel av variansen i lakseprisen, og leverer mer nøyaktige prediksjoner enn OLR modellen på alle de predikerte månedene. Jeg forklarer ytelsesforskjellen mellom OLR-modellen og ARDL-PLS-modellen med følgende to faktorer: (1) til forskjell fra ARD-PLS modellen, så krever OLR modellen at dataen blir stasjonærtransformert, noe som fjerner den langsiktige forklaringskraften i tidsseriene. (2) Til forskjell fra OLR modellen, så blir ikke regresjonskoeffisientene i ARDL-PLS modellen svekket av

at forklaringsvariablene er korrelerte, da den transformerer forklaringsvariablene til ortogonale ukorrelerte komponenter som blir brukt i regresjonen. Avslutningsvis så indikerer resultatene at prediksjonsmodeller for laksepriser blir drastisk forbedret ved bruk av eksogene variabler.

Content

List of Figures

List of Tables

1 Introduction

The spot price of Atlantic salmon is characterized by high volatility, which imposes uncertainty and additional costs on the value chain of salmon farming. The participants of the salmon market can benefit from a reliable price model, as it can be used to improve decision making regarding the operational and financial aspects of the industry that are subject to the price uncertainty. Consequently, this study aims to provide a framework for predicting the spot price of the Atlantic salmon, represented by the NQSALMON index. I present a new model, called the ARDL-PLS model, which is used to predict the 3-, 6- and 9-month ahead salmon spot price.

In 2014, a significant milestone was reached. For the first time in World history, aquaculture's contribution to global fish supply for human consumption overtook that of wild-caught fish (Seafish 2017). In 2014 total world aquaculture production increased to 101 million tonnes, with a value of 165.8 billion USD. Aquaculture is the fastest growing food supply sector in the world, and for global fish availability to meet projected demand, it has been estimated that aquaculture production will need to more than double by mid-century. In Norway, the aquaculture industry is responsible for 22,700 jobs, delivers fish to 100 vastly different countries, and has developed into a second industrial fairy-tale in the shadow of the petroleum industry. The industry has a long history at local level, and form the cornerstone of many Norwegian coastal communities. In addition, aquaculture can be considered as one of Norway's most important responses to one important challenge facing the World today. Namely, to produce sufficient healthy food for a rapidly growing population (Sjømat Norge 2011).

Today, the salmon farming industry is characterized by high price volatility and seasonal patterns (Oglend & Sikveland 2008). The seasonal effects can among others be explained by seasonally changing sea water temperatures (Oglend 2013)) and a higher demand around Christmas (Misund & Asche 2016). Furthermore, several variables such as the price of other commodities and exchange rates, has impact on the salmon price (Misund & Asche 2016). The price volatility imposes uncertainty on the value chain of salmon production resulting in additional costs. As the market of salmon farming is becoming more globalized and competitive, companies' ability to limit these unnecessary costs is an important factor for maintaining market positions. A satisfying price forecast model is an

important tool in order to establish functional risk management and operational efficiency, and can benefit the entire supply chain for salmon. For instance, reliable estimates of future salmon prices would serve as vital decision support in determining the optimal timing of salmon harvest, which terms producers should engage in forward contracts, or what should be the available machine capacity. In addition, reasonable estimates of future salmon prices are useful from a financial standpoint, as the salmon industry is increasing its presence in the capital markets. Many companies in the salmon industry are listed on stock exchanges and there have been vast amounts of corporate bond offerings. In addition, Fish Pool have been established, as a response to the increasing demand for financial hedging instruments. Consequently, the salmon farming industry receives increased attention from analysts and investors. Thus, a reliable price model can benefit to the financial aspect of the industry, as it can be used to improve risk management and investments decisions, and contribute to better valuation of bonds and stocks.

Although there exists a clear motive for the development of a reliable price model, there are only a few articles on the explicit modelling and prediction of the salmon spot price. To the authors knowledge, the last published study on modeling point forecasts of the salmon spot price is carried out by Sandaker et al. (2016), who uses exogenous variables in an ordinary least square regression (OLR) model. This is the only attempt of using exogenous variables as predictors in salmon price models, although there exists a number of successful applications of exogenous variables in the prediction of other commodities (Coleman 2012). Although the model developed by Sandaker et al. (2016) results in satisfying short term predictions, literature on econometric and forecasting indicate that there might exists several inefficient aspects of the methodology, including the removal of long term information in the variables (Bentzen & Engsted 2001) and unstable regression coefficients due to intercorrelated predictors (Wold et al. 1984). The results of Sandaker et al. (2016) motivates for the investigation of methods that can build upon the current framework, while coping with the problem regarding intercorrelated explanatory variables, and that enables the retainment of the long term information in the data. Research indicate that a partial least square (PLS) regression gives a solution to the multiple regression problem which is stabilized in comparison with a ordinary least square (OLR) solution (Wold et al. 1984). In addition, an Auto regressive distributed lag (ARDL) framework enables consistently estimation of both the short-run and long-run relationships, and can easily be solved by PLS regression (Bentzen & Engsted 2001).

There are two key purposes of this study: Firstly to present a tool that market participants can use to predict the salmon price. Secondly, to investigate if the methodology presented by Sandaker et al. (2016) is enhanced when substituting the OLR model with a combined model consisting of an ARDL framework and a PLS regression. The combined model is referred to as the *ARDL-PLS* model. The NASDAQ Salmon Index (NQSALMON), in USD, is used to represent the spot price of salmon. The explanatory variables used in the model are selected from a database of 17 time series. The time series included in the database and the corresponding lag structure, are based on discussions with industry experts (Sandaker et al. 2017). In addition, seasonal dummy variables and a trend variable are included in the dataset. In order to decide which combination of variables, at what lags, from the dataset to use in the model, I deploy a genetic algorithm (GA) for variable selection, which have been shown to be an effective feature selection tool in the field of finance (Kuhn & Johnson 2013). In contrast to Sandaker et al. (2016) who uses *log return* transformed time series, I use *log* transformed time series, as this is shown to maintain the long term information in the data. Consequently. some of the time series included in the database are non-stationary, and in order to avoid spurious regression, the set of covariates used in the model has to satisfy certain cointegration requirements (Bentzen & Engsted 2001). These requirements are formulated as optimization constraints in the GA-search. Thus all the sets of predictors selected by the GA-search are cointegrated with the salmon price. I use the variables selected by the GA-search as covariates in the ARDL-PLS model to predict the 3-, 6- and 9-month ahead value of the NQSALMON. To obtain each forecast, I develop a *submodel*. Each submodel use a set of maximum 6 covariates selected by the GA. Furthermore, in order to verify the performance of the ARDL-PLS model, I implement an OLR model in accordance with the methodology presented by Sandaker et al. (2016), as reference. The covariates used in the OLR model are selected by a GA from a database consisting of the same 17 time series, and corresponding lag structures, as used by the ARDL-PLS model. However, in the case of the OLR model, the time series are *log return* transformed, which removes the need for cointegration constraints in the GA-search used by the OLR model. Lastly, I implement an ARIMA model and a Naive model which are used for benchmarking purposes.

As aforementioned, the existing literature on point prediction of the salmon spot price is scarce, and to my knowledge this is the second attempt at using exogenous variables in salmon price modelling. Also, this is the first attempt at

combining an ARDL framework for predicting non-stationary time series, with a PLS regression. Lastly, this paper is the first to formulate the cointegration requirements of the ARDL model as optimization constraints in a feature selection algorithm, in order to enable the use of large databases of non-stationary time series in regressions models. Hence, this paper represents a significant contribution to research, and should provide industry participants with an important tool to improve risk management and investments decisions, in addition to help increase operational efficiency.

There are several encouraging results from this study. The use of a genetic algorithm as a feature selection tool looks promising, as it is relatively quickly able to find subset of variables that result in favourable goodness-of-fit, and that satisfy the cointegration requirements. Generally the ARDL-PLS model is able to explain a large degree of the variance of the salmon price. The respective R^2 values from the out-of-sample predictions for the 3-, 6- and 9-months ahead forecasts are 58%, 41% and 44%, which is relatively high given the fact that economic time series are considered as low signal-to-noise environments. The 9-month ahead predictions from the ARDL-PLS model are more accurate and less volatile than the 6-months ahead predictions, which is somewhat surprising. This is likely to result from the fact that the availability of candidate predictors are larger for the shorter horizons, due to the aforementioned lag structure, which makes the genetic algorithm more prone to overfit for the shorter horizons. Nonetheless, the predictive accuracy of the ARDL-PLS model is better than the OLR model for all horizons. The 3-month ahead prediction accuracy of the two models are relatively similar, but the ARDL-PLS model excel in the longer horizons. I attribute the performance difference between the OLR model and the ARDL-PLS model with two factor: (1) By not return transforming the data used by the ARDL-PLS models, the long term information is maintained, which increases the models ability to forecast the longer horizons. (2) Intercorrelations among the explanatory variables does not affect the stability of the regression coefficients in the ARDL-PLS model, as it, in contrast to the OLR model, transforms the predictors into uncorrelated orthogonal components. The out-of-sample performance of the ARDL-PLS model is superior to that of the Naive model, yielding a forecast error of 56% of the one produced by the Naive model, for the 9-months ahead forecast. Lastly, the OLR model performs far better than the ARIMA model. As the two models only differ from the fact that the ORL model utilize exogenous variables, the result indicate that the use of exogenous variables in salmon prediction greatly enhances performance. In sum, I have been able to create a parsimonious forecast model that is easy to

implement, and that seem to deliver forecasts with favourable precision compared to the OLR model following the methodology of Sandaker et al. (2016). However, there are some misbehaviour, including unexpected regression coefficient signs and residual autocorrelation, that can be explored in future research.

This paper consist of 7 chapters. Following this introduction, Chapter 2 gives an overview of relevant literature on the topic of salmon price modelling. Chapter 3 contains an overview of the dataset used in the ARDL-PLS model, in addition to some necessary preprosessing and descriptive statistics. Chapter 4 gives an introduction to the ARDL-PLS model and the cointegration requirements, and the genetic algorithm for variable selection. Chapter 5 describes the whole implementation process of the ARDL-PLS model and the OLR model, in addition to presenting the most important validation techniques. Chapter 6 proceeds with the result, presenting the performance of the variable selection procedure, the model tuning, the residual diagnostics and the out-of-sample results. Finally, Chapter 7 presents the conclusion.

2 | Relevant literature

The aim of this chapter is to give the reader an overview of the most relevant research related to the field of salmon price prediction. Firstly, I presents papers that address the explicit modelling and forecasting of salmon prices. Next, I present papers that consider other aspects of the salmon price, such as volatility. Lastly, I give an overview of how this study is positioned in context of the existing literature.

2.1 Literature on predicting salmon spot prices

There exists only a few articles that consider the explicit modelling and prediction of the salmon spot price. To the authors knowledge, the last published study modelling the point forecast of the salmon price is carried out by Sandaker et al. (2016). Prior to that, Guttormsen published a report in 1999. On the other hand, studies conducting implicit modeling of the salmon price, such as volatility and price elasticity, are more frequently represented in literature.

2.1.1 Explicit modeling and forecasting of the salmon price

Vukina & Anderson (1994) compare four state-space models for predicting non-stationary time series, in order to conduct a short term forecast of five different salmon products on the Tokyo wholesale market. The state-space models introduced includes an error correction model, a cointegration model, an impulse response model and a model that combines a structural model with an innovation model. The results indicate that all time series used have pronounced cyclic behaviour, and the out of sample prediction accuracy of the models are satisfying which encourage further research.

Gu & Anderson (1995) build a model that combines seasonality removal with a multivariate state space framework, in order to predict the US salmon market, and concluded that models applying seasonal components in modelling have substantial predictive power. Out-of-sample predictions for 3-, 6- and 12-months ahead are generated in order to test the forecast precision. The study show that adjusting the time series for seasonality before modeling improves the forecast performance, and thus indicates that salmon prices exhibit seasonal movement. However, as mentioned by Guttormsen (1999), the methodology may be im-

practical to use for market participants, as it is quite complicated and likely requires the user to have moderate knowledge within the field of statistics and econometric.

Guttormsen (1999) presents simpler and more intuitive methods for forecasting weekly producer prices for the Norwegian salmon market. The study uses six easily applicable procedures, including classical additive decomposition, Holt-Winters exponential smoothing, autoregressive moving average, vector autoregression, and two different naive models. The models are applied to predict 4-, 6-, 8- and 12- week ahead prices. The out-of-sample results were promising, but the author did not find evidence of a superior model.

Sandaker et al. (2016) develop a multivariate autoregressive model for predicting the log return transformation of the Norwegian salmon spot price (represented by the Fish Pool Index), using ordinary least square regression (OLR) and a forward selection algorithm for feature selection. The forward selection algorithm is used to determine the best in-sample subset of lags of 37 explanatory variables, for 1-, 2-, 3- and 4-weeks ahead forecasts. These subsets are used as predictors in the out-of-sample test. The out-of-sample error for the log return spot price is half of the naive forecast, and the results indicate that variables with a fundamental relationship with the spot price have strong predictive power. In addition, the results show that deploying feature selection techniques to shrink the predictor space, substantially decreases the degree of over-fit, and increases the predictive power of the model. The methodology presented by Sandaker et al. (2016) is interesting, as it is relatively simple to implement while giving satisfying results. However, many of the predictors used in the model are intercorrelated, and Sandaker et al. (2016) points out that this makes it nearly impossible to distinguish the individual contributions of the predictors. In addition, Wold et al. (1984) notes that in multiple linear regression, collinearities among the independent variables can cause the estimated coefficients to be very unstable and thereby far from their target values, which affects the models ability to consistently deliver good predictions. Lastly, it can be argued that the methodology is only suitable for short term forecasting given the loss of long term information due to stationarity transformation of the data (Nkoro et al. 2016).

Sandaker et al. (2017) build a database of 25 explanatory variables, and use this to develop a 1- to 12-month ahead quantile regression prediction model for the salmon spot price distribution. Each of the twelve month to be predicted

were appointed a designated submodel, using 8 explanatory variables picked by a genetic algorithm from the database. In addition, the paper presents recommended lag structures of each of the explanatory variables in the database, based on industry assumptions. The results indicate that the genetic algorithm quickly finds submodels with satisfying goodness-of-fit. In addition, the study show that exogenous variables such as standing biomass, feed consumption and prices of alternative protein, have strong predictive power over the salmon spot price. Since the study does not include an out-of-sample test, it can be argued that the actual performance of the applied methodology is not fully evaluated. However, the results and the database presented in the study creates a basis for further modeling of the salmon price.

2.1.2 Modeling and forecasting the salmon price implicitly

Although this paper focuses on the explicit modelling of salmon spot prices, it is worth mentioning relevant research regarding implicit modelling aspects. Oglend (2013) and Oglend & Sikveland (2008) investigated methods for modeling the volatility of the salmon price. These studies indicate that the volatility is correlated with the spot price, thus resulting in higher volatility when the spot price is high. The studies of Oglend & Sikveland (2008) indicates that the 1- and 5- week lags of the salmon spot price log returns have predictive power. In addition, they argue that the assumptions regarding independent and identical distributed error terms is not necessarily a valid assumption when modelling the salmon price. Oglend & Sikveland (2008) states that the salmon price is characterized by seasonal patterns. The seasonal effects can among others be explained by seasonally changing sea water temperatures and a higher demand around Christmas. In addition, Oglend (2013) found that the salmon price has followed an upwards trend since the early part of the last decade, due to an overall global growth in demand for protein sources, which is likely to remain in the future.

Asche et al. (2016) analyses the Fish Pool salmon forward contract. The authors examine whether the forward market provides a price discovery function, and how well the market performs in terms of the forward price being an unbiased estimator of the spot price. The authors find that the salmon forward market has not reached a stage where forward prices are able to predict future spot prices, indicating that it is still immature. For our report this is relevant because it suggests that the forward price does not provide a good forecast of the future

spot price, and should therefor not be used as an explanatory variable. Bloznelis (2016) employ an ARMA GARCH model and a dynamic conditional correlation (DCC) model on weekly data in order to examine the behaviour of weight-class-specific prices. The results indicate that there are two periods of different volatility regimes in the development of the salmon price, as the volatility and the correlation increased from 1996-2005 and 2007-2013.

2.2 This study in the context of exciting literature

This paper utilize an autoregressive distributed lag (ARDL) model solved by a partial least square (PLS) regression, in order to forecast the 3-, 6- and 9-month ahead log salmon spot price. The combination of the ARDL model and the PLS regression is noted as the *ARDL-PLS* model. In addition, a genetic algorithm (GA) for feature selection is used in order to select appropriate subsets of predictors from a dataset containing 61 explanatory variables. Most of the time series used in the model are nonstationary, and in order to avoid spurious regression, the set of covariates used in the model has to satisfy certain cointegration requirements. These requirements are formulated as optimization constraints in the GA-search for the ARDL-PLS model. In addition, the data used in the model is based on the database and recommended lag structure presented by Sandaker et al. (2017).

The methodology used in this paper builds on Sandaker et al. (2016), as their approach is promising. This is due to the following aspects:

- The methodology uses a parsimonious linear regression model. The user of a salmon price prediction model is likely to benefit on easy interpretability, and linear regression models conform this property. In addition, decades of professional experience suggest that simple parsimonious models tends to be best for out-of-sample forecasting in business, finance and economics. (Diebold 2014)

- The regression model utilize exogenous variables. As the salmon price is characterized by cyclical movements, the use of exogenous variables with similar cyclical movement can improve the prediction accuracy. This because the cyclicality of these variables possibly will correlate with the cyclical movement of the salmon price.

- The methodology utilize a feature selection algorithm. A model with less predictors may be more interpretable. In addition, some models are negatively affected by non-informative predictors. Regression models estimates parameters for every term in the model, thus non-informative parameters can add uncertainty to the prediction and reduce the overall effectiveness of the model (Kuhn & Johnson 2013). Thus, the use of feature selection algorithms is likely to improve salmon prediction model.

However, this paper aims to improve some of the issues related to the model presented by Sandaker et al. (2016), which are:

1. The correlation between the predictors used in ordinary least square regression can result in unstable regression coefficients and difficulties regarding the interpretation of the predictor contribution (Wold et al. 1984).

2. The stationary transformation can remove the long term predictive power of the time series (Nkoro et al. 2016). Therefor it can be argued that the model is likely to perform poorly when applied to long term forecasting. Since some applications such as salmon production planning, may require a prediction horizon of several months ahead, the application of the model is limited.

In an attempt to handle issue (1), the model in this paper is solved by a PLS regression, which transforms the original predictors into uncorrelated factors which are used in the regression. The results from Wold et al. (1984) indicates that a PLS method gives a solution to the multiple regression problem which is stabilized in comparison with a OLR model presented by Sandaker et al. (2016). Issue (2) is handled by utilizing non stationary data in an autoregressive distributed lag (ARDL) framework. The results from Bentzen & Engsted (2001) show that the ARDL framework enables both the short term and the long term predictive relationships to be estimated consistently, so that the model can be used to perform both short term and long term forecasts.

Except for the methodology presented by Sandaker et al. (2016), the modelling approaches applied in literature for forecasting the salmon price is of limited value to this study, as:

- The econometric models applied in literature differ from the one applied in this paper, as there are few, if any, regression based methods, and non that are specifically using PLS regression.

- Most approaches in literature which are considering the point wise modelling of the salmon spot price, only consider short term forecasting.
- No approaches which model the salmon spot price, apply exogenous variables or variable selection algorithms. ´

Salmon market participants are in need of a prediction model that is relatively simple, and that can be used for predicting the salmon spot price over both short and long horizons. These two aspects are both covered by my study. The methodology in this paper is the first attempt at using PLS regression to forecast the salmon spot price. Also, this is the first attempt at combining a ARDL framework for predicting non stationary time series, with a PLS regression. Lastly, this paper is the first to formulate the cointegration requirements of the ARDL model as optimization constraints in a feature selection algorithm, in order to enable the use of large databases of non stationary time series in regressions models.

Although past research on the salmon market is generally of limited value to this study (excluding Sandaker et al. (2016) and Sandaker et al. (2017)), there are a few key results that should be accounted for in the model formulation:

1. Several studies, including Gu & Anderson (1995) and Oglend & Sikveland (2008), state that the salmon price is characterized by seasonal behaviour due to seasonal changing sea water temperature and seasonal changes i demand.
2. Oglend (2013) found that the salmon price follows an upwards trend, due to a strong growth in demand for protein sources.

In order to account for these characteristics in the salmon price, trend and seasonality are explicitly modelled as part of the ARDL framework presented in chapter 4. There are however, other useful findings in literature concerning explanatory variables, which could be utilized when desiding on the composition of the dataset used in the modelling. For instance, Oglend & Sikveland (2008) states that the 1- and 5- week lags of the salmon spot log returns have predictive power. However, Sandaker et al. (2017) have already performed a thorough investigation of which explanatory variables, and corresponding lag structures, to be used in prediction models for the salmon spot price. In addition, the primary aim of this report is to investigate the usefulness of an ARDL-PLS model applied to salmon prediction, rather than increasing the knowledge of which drivers that affect the price of salmon the most. Consequently, I have chosen to base the

dataset of explanatory variables used in this paper on the dataset proposed by Sandaker et al. (2017), rather than conducting an independent analysis.

3 | Data analysis

This chapter contains a brief overview of the time series and corresponding lag structures to be used in the prediction model, in addition to data preprosessing and descriptive statistics.

3.1 The NASDAQ salmon index

In this paper I develop a regression model for the point forecast of the log salmon spot price, represented by the NASDAQ Salmon index (NQSALMON) in USD on a monthly resolution. The NQSALMON is reported in NOK, and reflects the weekly market spot price for Fresh Atlantic Superior Salmon, Head on Gutted (HOG). The index is widely accepted as the best assessment of the salmon spot prices in the market, and the index value is a volume weighted average price across the following weight classes: 1-2 kg, 2-3 kg, 3-4 kg, 4-5 kg, 5-6 kg, 6-7 kg, 7-8 kg and 9+ kg (The Nasdaq Group Inc. 2016a). The prices are converted to USD using the average NOK/USD exchange rate for Monday through Thursday weighted 60 %, plus the Friday prior weighted 40%. This weighting is based on the method used by NASDAQ Clearing when converting Atlantic Superior Salmon transaction in terms of foreign currencies (The Nasdaq Group Inc. 2016b).

3.2 The dataset

The dataset of explanatory variables used in this paper is based on the analysis of Sandaker et al. (2017), which presents an overview of 24 exogenous time series with recommended lag structures, resulting from discussions with industry experts. As the dataset used in this paper is directly derived from existing literature, the reader is referred to Sandaker et al. (2017) for the reasoning behind the composition of variables and the recommended lag structure.

Table 3.1 give an overview of the explanatory variables in the database used in this paper, and the corresponding sources. Although the database used in this paper is based on the recommendations of Sandaker et al. (2017), I have chosen to exclude some of the time series due to many missing values. This

include the time series for shrimp prices, consumption of Atlantic Salmon in EU, US, Russia, Japan and emerging markets, and the global harvest volume of Salmon. Based on the results of Sandaker et al. (2017) these time series are not significantly important for the model. In addition, the maximum allowed lag-length for the explanatory variables used in this paper, is set to 12 months, as the size of the dataset is relatively small (only 132 samples). Therefor, the time series for smolt release used in Sandaker et al. (2017) is excluded, as the recommended lag structure for this variable is 15-17 months. The database used in this paper consists of 17 time series, including the NQSALMON. An overview of the applied lag structure and the expected impact of each time series (illustrated by expected coefficient sign), is shown in table 3.2. The expected impacts of the time series are retrieved from the analysis of Sandaker et al. (2017).

3.3 Data preprocessing

The NQSALMON and all the exogenous variables are assessed on a monthly time resolution. For time series with weekly and daily observations, the last observation of the month is selected, and the rest is discarded. Some of the time series contain missing values (e.g N/As, zeroes and blanks). These are replaced by an interpolation between the prior value and the next value.

The time series are log transformed, in order to stabilize the variance, which grow over time. Figure 3.1 and figure 3.2 shows the spot price of salmon and the log spot price of salmon, respectively. The variance of the log spot price is more stable, and it's therefor the series for which we'll build the forecast model. The log transformation is convenient when including trends in regression models as economic trends often are exponential in levels, and thus linear in logs. This is further explained in the section 4.1.

As an alternative to predicting the log spot price, econometric researchers often turn to differencing in order to transform time series into a stationary process. However, this method tend to remove the long term predictive relationship between the explanatory variables and the response variable (Nkoro et al. 2016). As an alternative, I will turn to the use of an autoregressive distributed lag (ARDL) model on non stationary data, in order to include both the short term and the long term predictive information. An important requirement for using this model is that all explanatory variables are at most integrated of an order

1, i.e all time series are either I(0) or I(1) (Bentzen & Engsted 2001). This is tested by performing a unit root test, using a Augmented Dickey–Fuller test, on the first differences of the variables in the data set. The test rejects the null hypothesis of a unit root on a 1% confidence for the first difference of all the time series. Consequently, all time series satisfy either I(0) or I(1).

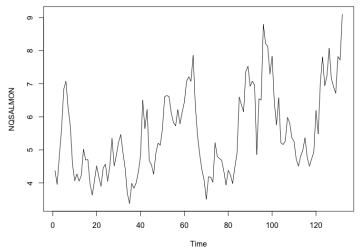

Figure 3.1: NQSALMON

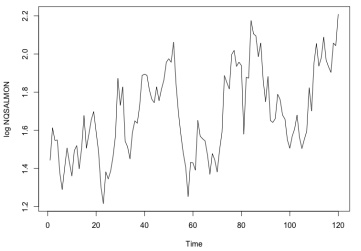

Figure 3.2: Log NQSALMON

3.4 Descriptive statistics

Table 3.3 display a selection of descriptive statistics for the monthly log of the NQSALMON and the other 16 time series used in the modelling. In addition, it includes two stationarity tests (ADF and KPSS), an autocorrelation test (Box-Pierce) and a normality test (JB). It can be seen that all times series yield a significant Box-Pierce statistics indicating the presence of significant autocorrelation. The Jarque-Bera (JB) test rejects the null-hypothesis of normality for 11 of the 17 time series (including NQSALMON) on a 10% significance level. Moreover, the KPSS test rejects the hypothesis of stationarity for 14 of the 17 times series (including the NQSALMON), while the Augmented Dickey-Fuller (ADF) test confirms the null hypothesis of non stationarity for 9 of the time series, on a 10% significance. Thus we see that the results from the two stationarity tests differ for some of the time series. It should be noted that the ADF test is sensitive to the number of lags included which can influence the results. It can be seen that most of the time series are significantly correlated with the NQSALMON. However, there are some exceptions, including the feed consumption of salmon and the Poultry index. It should be noted that the low correlation does not necessarily imply that these time series are useless in the

model. As will be seen in chapter 6, these low-correlated time series have lags that exhibit high correlation with the NQSALMON.

Table 3.4 shows the correlation matrix of the explanatory variables. As can be seen, many pair of the candidate predictors exhibit significant correlation, including the biomass of salmon and the harvest volume of salmon (0.91), the poultry index and the biomass of salmon (0.81), the meat price and the poultry index (0.92). The high correlation in the data indicates that the use of PLS can be beneficial relative to OLR, as the PLS model bypasses regression problems related to intercorrelated predictors by using orthogonal uncorrelated factors.

Time series, Unit	Description	Data source	Resolution
NQSALMON, USD/kg	Volume-weighted average of weekly reported sales prices of fresh Atlantic Superior Salmon, head-on-gutted	The NASDAQ group	Weekly
Standing biomass, #Individuals (Norway)	Sum of the biomass of large salmon in all cages of Norwegian waters	Directorate of Fisheries	Monthly
Standing biomass, Tonnes (Norway)	Sum of the biomass of large salmon in all cages of Norwegian waters	Directorate of Fisheries	Monthly
Feed consumption, Tonnes (Norway)	The amount of feed consumed by all the salmon in cages in Norwegian waters	Directorate of Fisheries	Monthly
Harvest volume, Tonnes (Norway)	The amount of salmon taken out from the cages and prepared for sales	Directorate of Fisheries	Monthly
Standing biomass of trout, #Individuals (Norway)	Sum of the biomass of trout in all cages in Norwegian waters	Directorate of Fisheries	Monthly
Standing biomass of trout, Tonnes (Norway)	Sum of the biomass of trout in all cages in Norwegian waters	Directorate of Fisheries	Monthly
Harvest volume of trout, Tonnes (Norway)	Trout taken out from the cages and prepared for sale	Directorate of Fisheries	Monthly
Sea lice occurrence, #Lice/fish (Norway)	Avg. number of sea lice per salmon	Lusedata	Weekly
Sea lice treatments, % of fish being treated (Norway)	The share of all salmon being treated for sea lice	Lusedata	Weekly
Sea temperature, Degrees Celsius (Norway)	Avg. sea temperature in Norwegian waters	Lusedata	Weekly
Meat price index, Index	Index composed of four other indices: poultry, pig, bovine, and ovine	FAO	Monthly
Poultry index, Index	Index based on export values for broiler cuts (US, export price) and chicken (Brazil, export price)	FAO	Monthly
Beef price, US cents/pound	Beef, Australian and New Zealand 85% lean fores (US import price)	Quandl	Monthly
Currency pair, USD/EUR	-	Oanda	Daily
Trout price, NOK/kg (Norway)	Price of trout farmed in Norway	Seafood.no	Weekly
Average harvest weight, kg (Norway)	Harvest volume divided by the number of salmon harvested	Directorate of Fisheries	Monthly

Table 3.1: Explanatory time series

Time series, Unit	**Lags(s)**[1]	**Impact**[1]
NQSALMON, USD/kg	1-2	+
Standing biomass, #Individuals (Norway)	3,6,9	-
Standing biomass, Tonnes (Norway)	3,6,9	-
Feed consumption, Tonnes (Norway)	2-4	-
Harvest volume, Tonnes (Norway)	1	-
Standing biomass of trout, #Individuals (Norway)	6,9,12	-
Standing biomass of trout, Tonnes (Norway)	6,9,12	-
Harvest volume of trout, Tonnes (Norway)	3,6,9,12	-
Sea lice occurrence, #Lice/fish (Norway)	3,12	+
Sea lice treatments, % of fish being treated (Norway)	3,12	+
Sea temperature, Degrees Celsius (Norway)	3,6	-
Meat price index, Index	3,6,9,12	+
Poultry index, Index	3,6,9,12	+
Beef price, US cents/pound	3,6,9,12	+
Currency pair, USD/EUR	6,9,12	-
Trout price, NOK/kg (Norway)	3,6,9,12	+
Average harvest weight, kg (Norway)	1	-

Table 3.2: Lag structure and expected regression coefficient signs of the explanatory variables

General information[1]		**Descriptive statistics**					**Tests, p-value**			
Time series	N#	Mean	Std. dev	Min	Max	NQSALMON corr.	KPSS	ADF	Box-Pierce $p=6$	JB
ln[NQSALMON, USD/kg]	132	1.69	0.23	1.22	2.21	1	0.01	0.36	0	0.08
ln[Standing biomass, #Individuals (Norway)]	132	12.72	0.17	12.24	12.96	0.31	0.01	0.07	0	0.00
ln[Standing biomass, Tonnes (Norway)]	132	13.28	0.19	12.85	13.54	0.32	0.01	0.61	0	0.01
ln[Feed consumption, Tonnes (Norway)]	132	11.53	0.42	10.73	12.19	-0.01	0.01	0.01	0	0.04
ln[Harvest volume, Tonnes (Norway)]	132	11.34	0.22	10.83	11.72	0.18	0.01	0.01	0	0.05
ln[Standing biomass of trout, #Individuals (Norway)]	132	10.09	0.14	9.83	10.48	-0.59	0.62	0.71	0	0.46
ln[Standing biomass of trout, Tonnes (Norway)]	132	10.59	0.17	10.17	10.96	-0.56	0.51	0.29	0	0.83
ln[Harvest volume of trout, Tonnes (Norway)]	132	8.68	0.29	7.75	9.45	-0.32	0.1	0.04	0	0.13
ln[Sea lice occurrence, #Lice/fish (Norway)]	132	-1.68	0.63	-3.48	-0.30	-0.28	0.1	0.01	0	0.41
ln[Sea lice treatments, % of fish being treated (Norway)]	132	-2.35	0.57	-3.71	-0.87	-0.25	0.01	0.23	0	0.68
ln[Sea temperature, Degrees Celsius (Norway)]	132	2.09	0.38	1.27	2.69	-0.19	0.1	0.01	0	0.03
ln[Meat price index, Index]	132	5.11	0.13	4.78	5.36	0.27	0.01	0.69	0	0.07
ln[Poultry index, Index]	132	5.19	0.13	4.87	5.41	0.11	0.01	0.69	0	0.05
ln[Beef price, US cents/pound]	132	5.09	0.23	4.64	5.61	0.38	0.01	0.62	0	0.09
ln[Currency pair, USD/EUR]	132	-0.27	0.09	-0.46	-0.05	0.22	0.01	0.06	0	0.05
ln[Trout price, NOK/kg (Norway)]	132	3.58	0.28	2.90	4.31	0.82	0.01	0.34	0	0.66
ln[Average harvest weight, kg (Norway)]	132	1.62	0.08	1.45	1.89	-0.39	0.01	0.01	0	0.00

[1] Note the choice of the time period, namely Jan. 2006 to Des. 2016, corresponding to N = 132 observations present in the table. I, however, note that the modelling dataset is of 120 observations.

Table 3.3: Descriptive statistics of the log transformed time series, time period: Jan. 2006 – Des. 2016

Correlation matrix

	NQSALMON	*Standing biomass, #Individuals*	*Standing biomass, Tonnes*	*Feed consumption*	*Harvest volume*	*Standing biomass of trout, #Individuals*	*Standing biomass of trout, Tonnes,*	*Harvest volume of trout*	*Sea lice occurrence*	*Sea lice treatment*	*Sea temperature*	*Meat price Index*	*Poultry Index*	*Beef price*	*Currency pair, USD/EUR*	*Trout price*	*Average harvest weight*
NQSALMON	1.00	0.29	0.26	0.03	0.15	-0.61	-0.54	-0.31	-0.28	-0.26	-0.16	0.26	0.12	0.36	0.19	0.79	-0.39
Standing biomass, #Individuals	0.29	1.00	0.90	0.77	0.91	-0.29	0.01	0.23	0.07	-0.22	0.31	0.81	0.67	0.80	0.30	0.45	-0.26
Standing biomass, Tonnes	0.26	0.90	1.00	0.67	0.89	-0.25	0.05	0.25	0.27	-0.13	0.09	0.73	0.58	0.81	0.35	0.48	-0.18
Feed consumption	0.03	0.77	0.67	1.00	0.70	-0.07	0.30	0.41	0.41	0.05	0.79	0.55	0.41	0.50	0.20	0.19	-0.30
Harvest volume	0.15	0.91	0.89	0.70	1.00	-0.14	0.10	0.37	0.13	-0.20	0.22	0.75	0.61	0.77	0.31	0.36	-0.10
Standing biomass of trout, #Individuals	-0.61	-0.29	-0.25	-0.07	-0.14	1.00	0.74	0.38	0.11	0.17	0.12	-0.13	-0.02	-0.27	-0.39	-0.72	0.51
Standing biomass of trout, Tonnes,	-0.54	0.01	0.05	0.30	0.10	0.74	1.00	0.58	0.26	0.09	0.37	-0.04	-0.04	-0.10	-0.06	-0.35	0.22
Harvest volume of trout	-0.31	0.23	0.25	0.41	0.37	0.38	0.58	1.00	0.36	0.13	0.36	0.03	-0.00	-0.01	0.05	-0.27	0.09
Sea lice occurrence	-0.28	0.07	0.27	0.41	0.13	0.11	0.26	1.00	1.00	0.65	0.28	-0.45	-0.03	-0.01	-0.11	-0.11	0.20
Sea lice treatment	-0.26	-0.22	-0.13	0.05	-0.20	0.17	0.09	0.13	0.65	1.00	0.17	-0.31	-0.22	-0.42	-0.25	-0.18	0.31
Sea temperature	-0.16	0.31	0.09	0.79	0.22	0.12	0.37	0.36	0.28	0.17	1.00	0.13	0.05	0.03	0.04	-0.10	-0.27
Meat price Index	0.26	0.81	0.73	0.55	0.75	-0.13	-0.04	0.03	-0.03	-0.31	0.13	1.00	0.92	0.83	-0.05	0.27	-0.00
Poultry Index	0.12	0.67	0.58	0.41	0.61	-0.02	-0.04	-0.00	-0.03	-0.22	0.05	0.92	1.00	0.62	-0.32	0.04	0.17
Beef price	0.36	0.80	0.81	0.50	0.77	-0.27	-0.10	-0.01	-0.10	-0.42	0.03	0.83	0.62	1.00	0.42	0.53	-0.24
Currency pair, USD/EUR	0.19	0.30	0.35	0.20	0.31	-0.39	-0.06	0.05	-0.11	-0.25	0.04	-0.05	-0.32	0.42	1.00	0.39	-0.45
Trout price	0.79	0.45	0.48	0.19	0.36	-0.72	-0.35	-0.27	-0.11	-0.18	-0.10	0.27	0.04	0.53	0.39	1.00	-0.47
Average harvest weight	-0.39	-0.26	-0.18	-0.30	-0.10	0.51	0.22	0.09	0.20	0.31	-0.27	-0.00	0.17	-0.24	-0.45	-0.47	1.00

Table 3.4: Correlation matrix for the explanatory time series

4 | Methodology

As aforementioned, the aim of this paper is to develop a combined model consisting of an autoregressive distributed lag (ARDL) model and a Partial Least Square (PLS) regression, in order to predict the h-step ahead log transformation of the NQSALMON. More precisely, for each time step t, the aim is to construct a model that predicts the log spot price at $t + h$, where

$$h \in \{3, 6, 9\} = H \tag{4.1}$$

Consequently, we say that the model *horizon* is up to 9 months. The aim of this paper is to make a model for predicting the h-step ahead log spot price

$$y_t^{(h)} | F_t, \text{ where } y_t^{(h)} = ln(p_{t+h}), \ \forall h \in H \tag{4.2}$$

p_{t+h} is the value of NQSALMON (in USD) at times $t + h$. F_t represents the information available at time t, which is the data available in a subset of covariates up to t, i.e $F_t \subset \{x_1, x_2..., x_t\}$.

The combination of ARDL and PLS is noted "the model" or "the ARDL-PLS model". Although it might seem complicated that the model combines both ARDL and PLS, it is rather straight forward: the ARDL model is a framework that defines which terms to include in the regression equation, and the PLS model comprises the algorithm that carries out the regressing and computes the forecasts. In addition, a genetic algorithm for feature selection is applied in order to pick a suitable subset of predictors to use in the ARDL-PLS model.

This chapter starts with introducing the ARDL framework, followed by a description of the cointegration requirements, which are necessary in order to avoid spurious regression when using non stationary data. Next, the PLS regression is presented, followed by a discussion on why this model is found suitable for the application in this paper, relative to other linear and nonlinear methods. Lastly, the genetic algorithm (GA) is presented. As aforementioned, the cointegration requirements of the ARDL-PLS model are formulated as optimization constraint in the GA-search, which restricts the GA-search to only select subset of variables that are cointegrated with the salmon price. This modification of the GA-search is explained in the end of the chapter.

4.1 The ARDL model

The auto regressive distributed lag (ARDL) model is widely used in time-series contexts. In this model, the response variable is explained by trend, seasonality, lags of the response variable and lags of the explanatory variables. The ARDL model is linear and can be solved by regression.

4.1.1 The trend component of ARDL

As aforementioned, Oglend (2013) found that the salmon price follow an upwards trend, due to a strong growth in demand for protein sources. Trends can be linear or nonlinear, depending on the data to be analyzed. In order to decide the type of trend that is most appropriate for the model, we can fit various trend models to the relevant time series, and analyze the resulting goodness-of-fit (Diebold 2014). Figure 4.1 shows the fit of different trend models, and table 4.1 shows the corresponding root mean square error (RMSE), which quantifies the accuracy of the fit. An intercept is included in each trend model.

y =	$\beta ln(t)$	βt	βt^2
RMSE	0.200	0.196	0.199

Table 4.1: Quality of fit for various trend models

Table 4.1 shows that it is most appropriate to use the linear trend $y = \beta t$. Although this model only consists of linear terms, it describes an exponential correlation with the salmon price since the variables are log transformed. This is illustrated below.

$$T_t = \beta_0 e^{\beta_1 t} \tag{4.3}$$

$$ln(T_t) = ln(\beta_0) + \beta_1 t \tag{4.4}$$

Exponential trends are very common in business, finance and economics, since economic variables often display roughly constant growth rates (Diebold 2014).

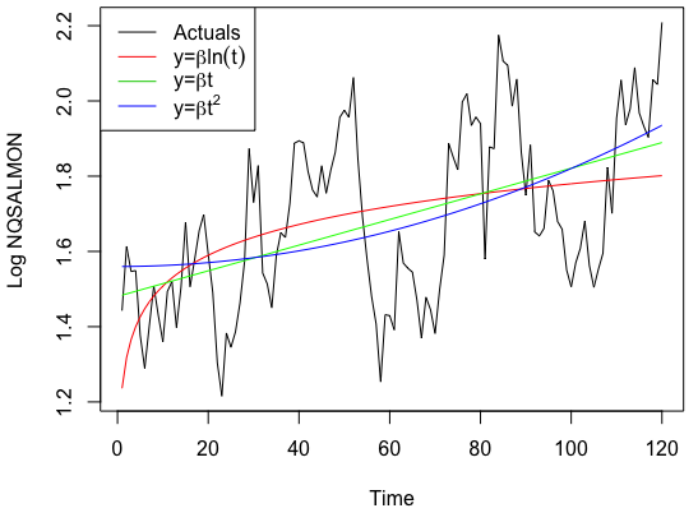

Figure 4.1: Plot of various trend models

4.1.2 The seasonal component of ARDL

As mentioned in chapter 2, several studies state that the salmon price is characterized by seasonal behaviour due to seasonal changing sea water temperature and seasonal changes in demand. Seasonality can be modeled by advanced techniques, but it is often sufficient to regress seasonal dummy variables. The number of dummy variables should equal the number of observation per year of the time series (Diebold 2014). Consequently, we use 12 dummy variables as we are dealing with monthly resolutions. The dummy variables D_s have the value 1 if and only if we are in season s, 0 otherwise. Thus $D_1 = 1$ indicates that we are in the first season, $D_2 = 1$ indicates that we are in the second season etc. Since we can only be in one season at a time, only one dummy variable is non-zero at any time t. By multiplying each dummy variable D_{it} with a seasonal factor γ_i,

we can model the seasonal component of the time series as:

$$S_t = \sum_{i=1}^{12} \gamma_i D_{it} \tag{4.5}$$

4.1.3 The complete ARDL model and the cointegration requirements

By combining the expressions for trend and seasonality with the lagged values of the response variables and the exogenous variables, we can formulate the ARDL model. y_t is the log NQSALMON at time t, $x_{i,t-p}$ is the p-lag of the explanatory variable i at time t, q is the maximum lag length, and ϵ_t is the error term. The general ARDL model can be written as:

$$y_t = \beta_0 + \beta_1 t + \sum_{i=1}^{n} \gamma_i D_{it} + \sum_{t=1}^{q} \phi_p y_{t-p} + \sum_{i=1}^{n} \sum_{t=1}^{q} \theta_{i,p} x_{i,t-p} + \epsilon_t \tag{4.6}$$

Findings in economic literature concerning unit root and non-stationarity, has led to an implicit dismissal of the use of ARDL models on non-stationary data. Instead, researchers have turned to cointegration models and error correction models, or utilized a stationarity transformation of the data (e.g differencing) in order to avoid spurious regressions. However, as mentioned by Bentzen & Engsted (2001), the use of the ARDL model on non-stationary data is valid if there exist a unique cointegrated relationship between the variables. In that case, both the short-run and long-run relationships can be consistently estimated by regression. To be specific, the following requirements must hold in order to use the ARDL model on non-stationary data:

1. There exists a unique cointegrated relationship between the salmon price and the explanatory variables
2. The explanatory variables are not cointegrated among themselves
3. The explanatory variables are exogenous
4. All variables are at most integrated at order 1

It is important to confirm that these criteria hold because the contrary can result in spurious regressions and invalid standard statistical inference (e.g t- and F-test of linear restrictions). Requirement (4) was confirmed satisfied by the

ADF test in section 3.3 for all time series in the dataset. Requirement (1)-(3) are formulated as optimization constraints in the genetic algorithm, so the GA-search is forced to only select subset that satisfy the requirements. Now follows an explanation of how the optimization constraints are formulated.

Variables are cointegrated if a linear combination of the variables is integrated of order 0, i.e is stationary. The Johansen analysis (Johansen 1991) can be used to test for the number of cointegrated relationships among a set of non-stationary variables. Let $r(y, X)$ be the number of cointegrated relationships among the joint set of the salmon price y, and a subset of explanatory variables X. Let $r(x)$ be the number of cointegrated relationships among only the subset X. Than the corresponding cointegration constraints for requirement (1) and (2) can be formulated as:

$$r(y, X) = 1 \tag{4.7}$$

$$r(X) = 0 \tag{4.8}$$

Now equation 4.7 ensures that there are exactly one cointegrated relationship between the salmon price and the subset of explanatory variables, and equation 4.8 ensures that the subset of explanatory variables are not cointegrated among themselves. In order to satisfy requirement (3), we must ensure that the subset of explanatory variables are exogenous. The variables used in this paper are assumed exogenous by Sandaker et al. (2017). However, as an extra precaution, we utilize the Johansen α-test (Johansen 1991). In short the null hypothesis of the test is that the variables are exogenous, and we reject this if the p-value is below a threshold K. Let p(X) be the p-value from the α-test when conducted on subset X of explanatory variables. Thus, to fully satisfy requirement (3) we add the following constraint to the feature selection algorithm:

$$p(X) > K \tag{4.9}$$

Consequently, by implementing the constraints in equation 4.7, 4.8 and 4.9, we ensure that the cointegration requirements are satisfies for all subsets selected by the GA-search. The Johansen analysis and the α-test are implemented in R using the "urca" package.

4.2 The partial least square regression

The ARDL model illustrated by equation 4.6 can be solved by any linear regression model, and in this thesis I apply partial least square (PLS) regression.

Consequently, the model is noted as a combination of ARDL and PLS. This section contains an introduction to PLS regression, as it is a novel approach within the field of salmon prediction, and a discussion on why this model seem to be more suitable for the application of this paper, relative to other linear and nonlinear models.

4.2.1 Introduction to PLS

Partial least square regression (PLS) originated with Herman Wold, who presented a nonlinear iterative partial least squares (NIPALS) algorithm in 1966, that linearized models that were nonlinear in the parameters (Wold 1966). Later he adapted NIPALS to a regression setting that involved correlated predictors, and named the method "PLS". The NIPALS algorithm aims to identify underlying relationships among the predictors, in the form of factors, which are highly correlated with the response variable (Kuhn & Johnson 2013).

PLS regression seek to minimize the sum of squared errors (SSE), which is showed in equation 4.10. Each iteration of the algorithm investigates the relationship between the predictors $\mathbf{X}$ and the response $\mathbf{Y}$, and summarizes the relationship in a vector of weights $\mathbf{W}$. The predictor data is orthogonally projected onto the vector of weights, which creates scores $\mathbf{t}$. The scores are used to generate loadings $\mathbf{p}$, which measure the correlation of the score vector and the original predictors. To simplify, one can imagine the loadings as vectors that describe how the latent variable space is connected to the original predictor space, while the scores describe the position of each sample in the latent variable space. These quantities, in addition to the weights, are stored sequentially in matrices $\mathbf{W}$, $\mathbf{T}$ and $\mathbf{P}$, and are needed for predicting new samples. In the end of each iteration, the predictors and the response are "deflated" by subtracting the information explained by the current estimated structure, thus the next generated loadings and scores only seek to model the remaining unexplained variance between the predictor and the response. The PLS regression has one tuning parameter, which is the number of components to use. In order to tune the model we use cross validation, which is explained in the chapter 5, as this is the most common tuning technique for PLS-models (Kuhn and Johnson, 2016). Now follows a discussion on why I consider PLS to be the most suitable model for salmon price prediction.

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{4.10}$$

4.2.2 Alternatives to using PLS regression

The "family" of linear regression models include ordinary linear regression (OLR) and factor models like partial least square regression (PLS) and principal component regression (PCR). The most common linear model, and the one used by Sandaker et al. (2016), is OLR. Like PLS, the objective of the OLR is to find a plane that minimizes the sum of squared errors between the observed and the predicted response. Let $\mathbf{X}$ be the matrix of predictors. It can be shown that the optimal regression coefficients are:

$$\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y \tag{4.11}$$

It is common in time series analysis that predictors are correlated and contain similar predictive information. If the correlations are high, the OLR solution for multiple linear regression will have high variance and become unstable. Common solutions to this problem include pre-processing the predictors by either (1) removing highly correlated predictors or (2) conducting principal component analysis (PCA) on the predictors. The removal of highly correlated predictors ensures that the pairwise correlations among all predictors are below a pre-specified threshold, but does not ensure that a linear combination of predictors are uncorrelated with other predictors. If this is the case, the OLR-solution will still be unstable. In addition, we lose our ability to meaningfully interpret the coefficients, since the coefficients used in the prediction are not unique (they depend on which predictors that are removed). PCA transforms the predictors into orthogonal predictors (factors) which are uncorrelated. The method of pre-processing predictors via PCA prior to performing regression, is known as principal component regression (PCR). The PCA creates factors on the aim of maximizing the explained variance among the predictors. However, if the variability in the predictor space is not related to the response variable, the PCR can have problems with identifying a predictive relationship (Kuhn & Johnson 2013). Like PCR, PLS creates linear combinations of the predictors. However, while PCR creates components only on the basis of maximizes the explained variance of the predictors, PLS also requires that the components have maximum correlation with the response which makes the components from PLS more relevant for predicting the response variable. Since PLS handles the issue of colinearity among predictors found in OLR, in addition to creating more promising factors than PCR, it is the method applied in this paper.

Although the PLS model seem to be the most promising linear model, one could be tempted to apply one of the more complex nonlinear models instead. Kuhn

& Johnson (2013) presents several nonlinear regression models, including neural networks, k-nearest neighbours and random forest. When including nonlinearity in a linear model, additional predictors that are functions of the original predictors must be added directly in an attempt to capture curvilinear relationships, while in nonlinear models the exact form of the nonlinearity does not need to be known explicitly or specified prior to model training. However, although nonlinear models are theoretically able to capture a much wider span of relationships between the predictors and the response compared to linear models, decades of professional experience suggest that simple parsimonious models tends to be best for out-of-sample forecasting in business, finance and economics (Diebold 2014). A distinct advantage of linear models is that they are highly interpretable. This makes it simple to analyze each predictors contribution to the response variable, as it can be done by evaluating the regression coefficients. Another consequent of the parsimony, is that it is easy to get a precise estimate of each parameter in the model. Lastly, enforcing simplicity reduces the scope of "data mining", i.e enforcing the model to maximize the fit to historic data. Data mining tends to tailor models in part to the idiosyncrasies of historical data, which have no structural relationship to unrealized future data, which in turn results in miserable out-of-sample forecasts (Diebold 2014). Due to the above-mentioned arguments, I have chosen to use a parsimonious PLS regression instead of a complex nonlinear model.

4.3 Feature selection

A model with less predictors may be more interpretable. In addition, some models are negatively affected by non-informative predictors. Regression models estimates parameters for every term in the model, thus non-informative parameters can add uncertainty to the prediction and reduce the overall effectiveness of the model. Kuhn & Johnson (2013) conducted an empirical study on the consequences of using non informative predictors in several different models, and the results indicated that PLS models are highly sensitive to over-fit when the predictor space becomes large, which makes it beneficial to use feature selection prior to model building. Another important benefit of using feature selection for the application in this paper, is that we can instruct the algorithm to only select subsets of explanatory variables that satisfy the cointegration requirements. This enables the use of large datasets containing non stationary time series in regression modelling, without the need to manually discover and

remove subsets that violates the cointegration requirements.

As aforementioned, a genetic algorithm (GA) is applied as feature selection tool for the application of this paper. The GA-search selects a subset of explanatory variables to be used as covariates in each ARDL-PLS submodel, and the selected subsets are constrained to satisfy the cointegration requirements presented in section 4.1.3. This section contains a general presentation of the GA, and a description of how it is modified in terms of parameter tuning and the implementation of the cointegration constraints. Lastly, I include a discussion on why I consider the GA-search to be superior for the application of this paper, relative to other variable selection methods.

4.3.1 The genetic algorithm

Feature selection can be viewed as a complex optimization problem, where we seek the combination of features that provides optimal prediction of the response variable. Genetic algorithms (GA) are heuristics, which imitate the evolutionary process by allowing a population (a set of solutions) to reproduce, in order to create children (new solutions) that compete to survive. The most fit children are allowed to reproduce, which creates the next generation. Each solution is represented as a chromosome, which is a string of genes. To create the next generation of children, two chromosomes reproduce through crossover and mutations. When applied to a feature selection setting, each chromosomes is a binary vector, where each gene represents the presents or absence of a particular predictor. The fitness of each chromosome is determined by the model using the predictors indicated by the binary vector. The most fit chromosome is the subset of predictors that result in the most accurate predictions (Kuhn & Johnson 2013).

GAs are initiated with a random population of chromosomes. The fitness of each chromosome is evaluated, which determines the chromosome's probability of participating in reproduction. Pairs of chromosomes from the population are picked to reproduce. The cross over face consists of splitting the parents chromosomes at a random position, before combing the head of one parent with the tail of the other parent and vica versa. This results in two children per couple of parents. After crossover, the individual entries of the new chromosomes can be randomly selected for mutation in which the current binary value is changed to the other value. The crossover phases drives the subsequent generations towards the subspace defined by the most fit chromosomes, which can make the

algorithm getting trapped in a local optimum. In order to prevent this from happening, the mutation phase is included which randomly perturbs the genes (Kuhn & Johnson 2013). A pseudo-code for the feature selection GA used in this paper is shown below. The code is based on the algorithm used the R-package "gaselect".

Genetic algorithm

1. Define the number of generations ($numGenerations$), size of each population ($populationSize$), and probability of mutation ($mutationProbability$)
2. Generate an initial random set of $populationsSize$ binary chromosomes, each of length p
3. **for** $i = 1...numGenerations$ do
   (a) **for** each chromosome do
      i. Compute each chromosome's fitness
   (b) **end**
   (c) **for** $k = 1...populationsSize/2$ do
      i. Select two chromosomes based on the fitness criterion
      ii. Crossover: Randomly select a loci and exchange each chromozome's genes beyond the loci
      iii. Mutation: Randomly change binary values of each gene in each new child with probability $mutationProbability$
   (d) **end**
4. **end**

GA input parameters

The parameters settings used in the GA are shown in table 4.2, and are based on a combination of initial testing and recommendations from Obitko (2014). The number of iterations and the population size are set sufficiently high in order to increase the probability of finding the global optimum. Note that the $mutationProbability$, which normally is set to ≤ 1, is set to 2. This is done to increase the probability of the GA finding valid solutions, as the number of valid solutions are drastically reduces by the cointegration constraints.

Parameter	Value	Description
Population size	populationSize= 400	The number of chromosomes in each population. A small population creates a small search space, while a large population slows down the convergence time
Generations/Iterations	numGenerations =400	The number of new populations created, i.e the number of iterations of the algorithm.
Cross over	crossover = "single"	Defines the way parent chromosomes are combined in order to create new offsprings.
Elitism	elitism = 9	Copies the best chromosomes to the next generation, in order to prevent loosing the best found solution. It is found that elitism rapidly increases the performance of the GA.
Mutation probability	mutationProbability=2	Randomly changes the genes of new offsprings in order to prevent the GA from falling into local optimums.

Table 4.2: Parameter settings for the GA-search used by the ARDL-PLS model

Lastly, the GA requires a specification of the maximum number of variables to be selected for each solution. Multiple studies have shown that regression models are more likely to be reliable if the number of predictors p is less than $T/10$ or $T/20$ (Harrell 2015). The intuition is that the more degrees of freedom used by the predictors, the easier it is for the model to adjust these variables to random patterns in the data, thus creating overfit. This is especially relevant for the application in this paper, as financial time series are low signal-to-noise environments. Taking the risk of overfit seriously, I choose to use $p = T/20$, which results in a maximum of 6 predictors per submodel.

GA fitness evaluation

The tasks of the feature selection algorithm is to find a subset of predictors that maximize the performance of each PLS-ARDL submodel. However, the time series used in the subset must satisfy the cointegration requirements presented in section 4.1.3, in order to avoid spuriosity. In order to ensure the fulfillment of the cointegration requirements, the fitness function of the GA *backtracks* which of the original 16 exogenous time series are used in the subset, and checks if these time series together satisfy the requirements. If this is the case, the PLS model is trained on the chosen subset of lagged variables, and the corresponding training SIC value is calculated and used as the fitness value of the solution in the GA-search. If the subset of variables found by the genetic variable does not satisfy the cointegration requirements, the fitness value is set to 0. The SIC criteria is a metric that punishes the in-sample RMSE to reflect the degrees of freedom used. This is in order to compute a subset of predictors that results in a satisfying in-sample performance while maintain the parsimony of the model. SIC is the recommended model selection metric of Diebold (2014), and can be expressed as:

$$SIC = T^{\frac{K}{T}} \frac{\sum_{t=1}^{T} e_t^2}{T} \tag{4.12}$$

Where T is the sample size of the train set, K is the number of parameters estimated by the model, and e_t are the residuals from the model training.

The cointegration requirements mentioned in section 4.1.3 are:

1. The number of cointegrated relationships among the joint set of the salmon price y and the subset X of explanatory variable, $r(y, X)$, must equal to 1.
2. The number of cointegrated relationships among the subset X of explanatory variable, $r(X)$ must equal to 0
3. The explanatory variables must be exogenous.

Requirements (1) and (2) are checked by conducting a Johansen test, and requirement (3) is checked by conducting an α-test. The testes are run within the fitness function of the genetic algorithm and conducted on a 5% significance. An overview of the fitness evaluation within the GA is shown below, where S is a subset of explanatory variables, and O_S is the corresponding backtracked original exogenous time series. $r(y, O_S)$, $r(O_S)$ and $p(O_S)$ are the constraint values of the ARDL restrictions.

Finess evaluation of subset S

1. Backtrack the set O_S of original time series used in the subset S
2. **if** $r(y, O_S) = 1$, & $r(O_S) = 0$, & $p(O_S) \geq K$

   Fitness = $\frac{1}{SIC}$, computed on subset S
3. **else**

   Fitness = 0
4. **return** Fitness

Note that the term $\frac{1}{SIC}$ is used because the genetic algorithm seek to maximize the fitness function, while we want to minimize the SIC-value.

4.3.2 Alternative methods for feature selection

There are in general two types of feature selection methods, filters and wrappers. Wrappers are algorithms that add and remove predictors to the forecast model in order to find the subset of predictors that result in the best model fit. Filter methods evaluate each predictor on the basis of a criteria before the model is solved, thus the actual affect on the predictive accuracy is not included in the algorithm. A typical filter criteria is that the correlation between the predictor and the response must be above a defined threshold in order for the predictor to be included in the model (Kuhn & Johnson 2013). Filter methods are usually more computational efficient than wrapper methods, but the selection criteria is not directly connected to the model performance. As we aim to find a parsimonious set of important covariates to include in the model, wrappers seems most appropriate.

Although there exists many wrappers, I have limited this discussion to forward selection, backwards selection and genetic algorithms. Forward selection is considered to be one of the most common form of wrappers, and is applied by Sandaker et al. (2016). A classical forward selection evaluates the predictors, by using a statistical hypothesis test, in order to see if each newly added predictor is statistically significant. If at least one predictor has a p-value below the threshold, the predictor associated with the smallest value is added to the model and the process starts again. The reader is referred to Kuhn & Johnson (2013), for an elaboration. This method has several issues, including (1) the procedure is greedy meaning it does not reevaluate past solutions, (2) the use of repeated hypothesis tests in this manner invalidates many of their statistically properties as the same data is being evaluated many times, and (3) the optimization criteria is not directly linked to model performance. Backwards selection is a common modification of the forward selection, in order to reduce the problems regarding greediness. The algorithm starts with all predictors, and then iteratively removes predictors in order to determine which are not significantly contributing to the model. The algorithm starts with sorting the predictors with respect to an importance measure. A threshold is applied in order to eliminate the least important variables. Then the model is fitted to the remaining predictors, and the performance is measured. The procedure is repeated until maximum model performance is achieved. Although backwards selection makes the search procedure less greedy than the forward selection, it exacerbates the problem of repeated hypothesis testing (Kuhn & Johnson 2013).

Repeated hypothesis testing is not an issue in genetic algorithms, as it uses fitness scores to evaluate solutions, which is obtained directly by calculating the corresponding goodness-of-fit. In addition, Ga's search parallel from a population of points. Therefore, it has the ability to avoid being trapped in local optimal solution, in contrast to traditional methods which search from a single point. Lastly, genetic algorithms have been shown to be effective feature selection tools in the fields of chemometrics, image analysis and finance (Kuhn & Johnson 2013). As the research on applying genetic algorithms for feature selection looks promising, and less complex wrappers have already been applied in the context of salmon spot price prediction, I use a genetic algorithm in this paper.

5 | Model implementation and validation techniques

This chapter describes the whole implementation process of the ARDL-PLS model, covering everything from data transformation to performance evaluation. As mentioned earlier, the ARDL-PLS model is compared to an OLR model, where the latter model is based on the methodology by Sandaker et al. (2016). Therefor this chapter also contains a description of the implementation process of the OLR model. Lastly, the most important performance metrics and validation techniques are presented.

5.1 Model implementation

The model implementation is computationally demanding, as the GA-search has to be run for each submodel of both the OLR model and the ARDL-PLS model. Consequently, only the $h \in \{3, 6, 9\}$-step ahead forecast are computed. It would be optimal to compute the 1- to 12-step ahead forecasts, but due to time restrictions I had to limit the model testing. However, the horizons $h \in \{3, 4, 5\}$ have the same available candidate predictors resulting from the lag constraints. The same yields for $h \in \{6, 7, 8\}$ and $h \in \{9, 10, 11\}$. Thus, despite the fact that the model testing only covers $h \in \{3, 6, 9\}$, it can be argued that the results give a suitable indicator of the models general ability to forecast both short term and long term.

As explained in chapter 4, the GA-search is not guaranteed to converge, which can make the optimal subset selected by the GA-search to vary between runs. Consequently, in an attempt to reduce randomness in the model performance evaluation, the implementation process is performed 5 times for each submodel in order to test the consistency of the models. This is done for both the ARDL-PLS model and the OLR model.

5.1.1 Implementation of the ARDL-PLS model

In short, the modelling procedures involves the creation of candidate predictors, followed by the selection of a subset of predictors for each horizon h, followed by the training and tuning of the corresponding submodel for each horizon h, and

finally the out-of-sample testing. This procedure is run 5 times. The procedure for each run r can be summarized in the following steps:

1. 17 time series (including the salmon price) are log transformed and lagged according to table 3.2 in chapter 3. In addition 1 trend variable and 12 seasonal dummy variables are created. These variables comprise a set of 61 candidate predictors and make up the relevant information set F_t.

2. The total set of sample points is divided into S_I and S_O, representing the respective in-sample set and out-of-sample set, for the prediction of the h-step ahead log NQSALMON. S_I consist of 80% of the total samples, and S_O consist of 20% of the total samples.

3. An ARDL-PLS model is applied to create a forecast of the h-step ahead log NQSALMON. In order to determine the U most relevant predictors for each horizon h, a variable selection algorithm is utilized on the training set S_I. That is, for each $h \in H$:

   (a) A genetic algorithm for feature selection is used to determine which subset of predictors that results in the lowest in-sample prediction error. The size of the subset can vary between 4 and 6 variables, thus $U \in \{4, 5, 6\}$. The performance of each subset is evaluated by fitting a PLS model and computing the resulting SIC value, in accordance with the GA's fitness function illustrated in section 4.3.1. The cointegration constraints in the fitness function ensures that the selected subset satisfy the cointegration requirements of the ARDL model.

   (b) The results from the GA algorithm is a subset of predictors that will be applied as covariates for the submodel for horizon h.

4. Now $|H| = 3$ ARDL-PLS submodels are trained and tuned, using the training set S_I and the selected covariates from the GA-search. The model tuning consist of selecting the optimal number of PLS-components for each submodel, based on cross-validation.

5. The 3 submodels are used to predict the test set S_O, and the resulting out-of-sample predictions are evaluated and compared to the OLR model.

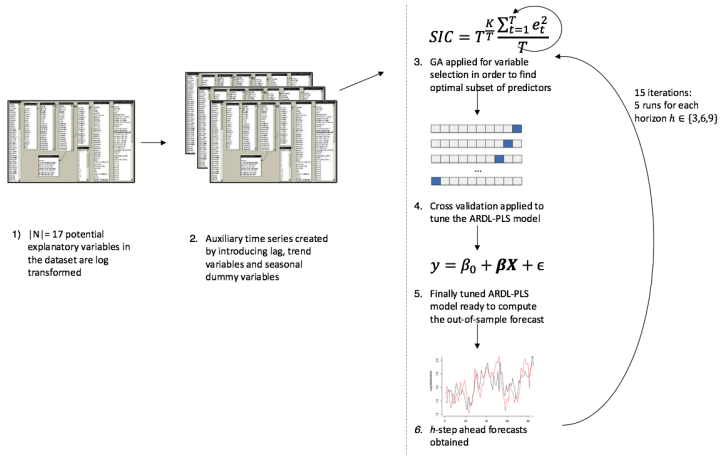

Figure 5.1: Overview of the h-step ahead log ARDL-PLS prediction model

5.1.2 Implementation of the OLR model

Now follows a description of the implementation process of the OLR model. As mentioned in chapter 3, all the time series in the dataset are at most integrated at order 1. Consequently, since the OLR model applies log return transformed data, non stationarity and spurious regression is not an issue.

The OLR model consider the prediction of the h-step ahead log return transformation of the NQSALMON. More precisely, for each time step t, the aim is to construct a model that predicts the log return spot price at $t + h$, where

$$h \in \{3, 6, 9\} = H \tag{5.1}$$

The aim is to make a model for predicting the h-step ahead log return spot price

$$y_t^{(h)} | F_t, \text{where } y_t^{(h)} = ln(p_{t+h}) - ln(p_t), \ \forall h \in H \tag{5.2}$$

p_t is the value of NQSALMON (in USD) at time t. F_t represents the information available at time t, which is the data available in a subset of covariates up to t, i.e $F_t \subset \{x_1, x_2..., x_t\}$.

The implementation of the OLR model is very similar to the ARDL-PLS model. The main differences are that the applied time series are *log return* transformed,

and that there is no model tuning required as the OLR model has no tuning parameter. In addition, the fitness function of the genetic algorithm applied for the OLR model does not contain cointegration constraints. An overview of the fitness function for the genetic algorithm used in the OLR model can be found in appendix A.2. In short, the modelling procedures involves the creation of candidate predictors, followed by the selection of a subset of predictors for each horizon h, followed by the training of the corresponding submodel for each horizon h, and finally the out-of-sample testing. This procedure is run 5 times, and each run can be summarized in the following steps:

1. 17 time series (including NQSALMON) are log return transformed and lagged according to table 3.2 in chapter 3. These variables comprise a set of 48 candidate predictors and make up the relevant information available in F_t.

2. The total set of sample points are divided into S_I and S_O, representing the respective in-sample set and out-of-sample set for the prediction of the h-step ahead log return NQSALMON. S_I consist of 80% of the total samples, and S_O consists of 20% of the total samples.

3. A multivariate ordinary linear regression (OLR) model is applied to create a forecast of the h-step ahead log return NQSALMON. In order to determine the U most relevant predictors for each horizon h, we utilize a variable selection algorithm on the training set S_I. That is, for each $h \in H$:

   (a) A genetic algorithm for feature selection is used to determine which subset of predictors that results in the lowest prediction error. The subsets are restricted to a maximum of 6 predictors. The performance of each subset is evaluated by computing the corresponding in-sample SIC value of the model.

   (b) The result from the GA-search is a subset of predictors that will be applied as covariates in the OLR model for horizon h.

4. Now $|H| = 3$ OLR submodels are trained, using the training set S_I and the selected predictors from the genetic algorithm.

5. The 3 submodels are used to predict the test set S_O, and the results are transformed back to log-values and compared to the results of the ARDL-PLS model

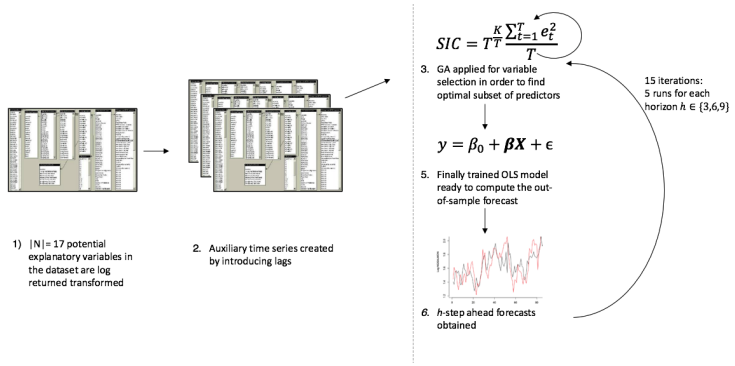

Figure 5.2: Overview of the h-step ahead log return OLR prediction model

5.2 Model evaluation

Now follows a presentation of the metrics and methods used to evaluate the performance of the models.

5.2.1 Measures of forecast accuracy

In this thesis, the *root mean square error* (RMSE) and the *coefficient of determination* (R^2) are the applied measures of forecast accuracy. In the context of point forecast models, RMSE is the most common method for evaluating a model's predictive capabilities. The value can be interpreted as the average distance between the observed values and the models predictions (Kuhn & Johnson 2013). R^2 describes the proportion of the variation in the dataset that is explained by the model. The simplest way of calculating the R^2 is to square the correlation between the observed and predicted values (Kuhn & Johnson 2013). Let y_t denote the tth observation of the target variable, and let $\hat{y_t}$ denote a forecast of y_t.

$$RMSE = \frac{1}{T}\sum_{t=1}^{T}[y_t - \hat{y_t}] \tag{5.3}$$

$$R^2 = 1 - \frac{\sum_{t=1}^{T}[y_t - \hat{y_t}]}{\sum_{t=1}^{T}[y_t - mean(y)]} \tag{5.4}$$

The ARDL-PLS model will be compared to the performance of the naive model, which is an estimating technique using the present actuals as the next periods forecast. The naive h-step ahead forecast can be expressed as:

$$\hat{y}_{t+h} = y_t \tag{5.5}$$

In addition, an ARIMA(1,1,0) model will be used as benchmark. The tuning of the ARIMA model is based on Hyndman-Khandakar algorithm for automatic ARIMA modelling presented by Hyndman & Athanasopoulos (2014a). The reader is referred to appendix A.3 for an elaboration on the ARIMA model.

5.2.2 Residual diagnostics

Hyndman & Athanasopoulos (2014a) mentions two essential characteristics of a good forecast: the residuals need to be uncorrelated and have zero mean. If the residuals have a non zero mean, the forecast is biased. If there are correlation between the residuals, there is information left that can be used for improving the forecast. Hyndman & Athanasopoulos (2014a) also mentions that constant variance and normal distributed residuals are useful properties. However, these properties are only required when computing prediction intervals, which is outside the scope of this report. The residual mean is assessed visually by analyzing the residual plot. Autocorrelation is detected by analyzing the ACF plots, and the residual plots. As mentioned in chapter 4, it is essential that the residuals of the ARDL-PLS model are stationary in order to confirm that the explanatory variables cointegrate with the response variable. Although a Johansen test is included in the genetic algorithm in order to obtain cointegration among the selected variables, unit root tests are performed on the residuals as a second safety measure. The tests used are ADF and KPSS. The unit root tests are only performed on the residuals from the cross validation of the training set, as the size of the test set is very limited.

5.2.3 Validation techniques

A model's predictive abilities can only be determined by considering how well the model predicts data that was not used during the fitting-process. Two of the most widely used procedures for evaluating regressions models are: K-fold cross validation (CV) and out-of-sample (OOS) evaluation. As mentioned in section 5.1.1, CV is used to tune the ARDL-PLS model, and OOS evaluation is used to test the models predictive performance. Now follows a description of

the two validation techniques.

OOS evaluation consists of using a portion of the available data for fitting the model, and using the rest for testing the model. Consequently, the predictive performance on the test data indicates the model's ability to forecast new data. It is important to consider the appropriate share of samples that should be withheld for model testing. By enlarging the train set, the model fit can improve but the empirical basis of the model testing decreases, and vice versa. Hyndman & Athanasopoulos (2014a) generally recommends a test set size that amounts to 20% of the total data. Consequently, I use 20% of the total samples for the model testing in this paper. Cross validation can be vied as a more sophisticated version of the OOS evaluation. In K-fold cross validation, the samples are randomly partitioned in K equal sized subset. A model is fit using all samples except the first subsets. The held-out samples are predicted by the model, and the performance is evaluated. This procedure is repeated for all subsets, and the K performance measures are averaged. The choice of K is usually 5 or 10, but there is no formal rule. I have chosen to use 20 folds as results from initial testing indicate that this gives stable performance throughout different runs of the model. It should be noted that the results of the cross validation should not be used as basis when evaluating the ARDL-PLS model's ability to predict new data. This is because the cross validation is performed on the same data as used by the GA search, which can result in an overly optimistic result (Kuhn & Johnson 2013). As some researchers are critical to the use of cross validation on time series, I have included a discussion on this topic in appendix A.4.

6 | Results and discussion

This chapter presents the results from the implementation of the ARDL-PLS model, including the GA-search, the model tuning and the out-of-sample testing. The results from the out-of-sample testing of the ARDL-PLS model are compared to the results from the OLR model, the naive model and the ARIMA model. The reader is referred to appendix A.2 for details about the OLR model regarding selected variables from the genetic algorithm and the corresponding regression coefficients. Lastly, an overview of the ARIMA model coefficients can be found in appendix A.3.

It is important to specify that the results presented in this chapter are not made on a complete empirical basis, as the implementation of the ARDL-PLS model and the OLR model are computed for only 5 runs of the genetic algorithm. Thus, it can not be denied that the prediction accuracy can vary when the algorithm is rerun, as other subset of predictors can be selected which results in different out-of-sample performances. Optimally we would run the GA-search many times for each forecast horizon in order to get an overview of the consistency of the results of each model. However, the genetic algorithm is computationally demanding and it needs to be run for both the ARDL-PLS model and the OLR model, which is time consuming. Due to computational difficulties and limited available time, the scope of the model testing is restricted.

Lastly, we need to distinguish between submodels corresponding to different runs *and* different horizons, as the GA-search is run 5 times for each forecast horizon. Thus, we let the notation $r = k, h = l$ refer to the submodel that results from run k of the GA-search, and that conducts the l-step ahead forecast. In addition, the term *intra-horizon* variance refers to the variance between submodels that forecast the same horizon.

6.1 Variable selection for the ARDL-PLS model

This section contains the results of the GA-search for the ARDL-PLS model. The aim of the GA-search is to evaluate the set of $p = 61$ candidate predictors and select subsets of $U \leq 6$ variables to be used as covariates for the submodels across the three horizons $H = \{3, 6, 9\}$. The GA-search, which is tuned according to the settings displayed in section 4.3.1, is run 5 times for each horizon $h \in \{3, 6, 9\}$. Figure 6.1, presents a visualization of the development of the per-generation performance of the GA-search for the first run of each horizon. More specifically, the figure shows the best fitness value (i.e smallest SIC-value) for each of the 400 generation of the GA-search. As evident from the plot, the improvement in fitness value (i.e reduction in the SIC value) is very small after 200 generations for $r = 1, h = 3$ and $r = 1, h = 6$, and after 100 generations for $r = 1, h = 9$, indicating that the algorithm is close to convergence. The submodel $r = 1, h = 9$ requires the fewest number of generations before converging, which is likely to result from the fact that this submodel has a smaller set of candidate predictors to select from, due to the aforementioned lag constraints. The fact that the algorithm is close to convergence, does not necessarily imply that it is close to the global optimal solution, as it may very well be trapped in a local optimum. However, by experimentally varying the parameter settings and running the GA several times, I have found that the solutions are generally close to the global optimum. Lastly, it can be seen that the earliest generations of the GA-search have a fitness score of 0. This is because the best solutions corresponding to these generations do not satisfy the aforementioned cointegration constraints.

The selected variables to be used as covariates in the submodels corresponding to run $r \in \{1, 2, .., 5\}$ and horizon $h \in \{3, 6, 9\}$, are presented in table 6.1. Empty cells for a particular submodel (for some horizon h) imply that the use of the variable is not allowed due to lag constraints. In total, 7 of the original 17 explanatory times series have been utilized as covariates in at least one sub model. In addition, the trend variable is used. We see that the intra-horizon variation in selected predictors is relatively small. For instance, submodel $r = 2, h = 9$ and $r = 5, h = 9$ use the same predictors. It can be seen from the overview of regressions coefficients of the final tuned submodels in appendix A.1 that these models are actually identical, i.e they have identical regression coefficients. This can result from the fact that the cointegration constraints in the genetic algorithm drastically reduces the number of valid subsets, which increases the possibility

Time series, Unit	Lags[1]														
	h=3					h=6					h=9				
	r=1	*r=2*	*r=3*	*r=4*	*r=5*	*r=1*	*r=2*	*r=3*	*r=4*	*r=5*	*r=1*	*r=2*	*r=3*	*r=4*	*r=5*
Trend-variable[2]	-	-	-	-	-	-	-	1	1	-	1	1	1	1	1
Seasonal dummy variables	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
NQSALMON, USD/kg															
Standing biomass, #Individuals (Norway)	6	6	-	-	6	6	6	-	-	9	-	-	-	-	-
Standing biomass, Tonnes (Norway)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Feed consumption, Tonnes (Norway)	-	-	-	-	-										
Harvest volume, Tonnes (Norway)															
Standing biomass of trout, #Individuals (Norway)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Standing biomass of trout, Tonnes (Norway)	6	6	6,9	6,9	-	12	6	9,12	12	12	9	9	9	9	9
Harvest volume of trout, Tonnes (Norway)	-	-	-	-	-	-	-	-	-	-	-	9	12	9	9
Sea lice occurrence, #Lice/fish (Norway)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Sea lice treatments, % of fish being treated (Norway)	12	3	3	12	12	12	12	12	12	12	12	12	12	12	12
Sea temperature, Degrees Celsius (Norway)	-	-	-	-	-	-	-	-	-	-					
Meat price index, Index	6	6	6	6	6	6	6	6	6	6	9	9	9	12	9
Poultry index, Index	6	6	6	6	6	6	6	6	6	6	9	9	9	9	9
Beef price, US cents/pound	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Currency pair, USD/EUR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Trout price, NOK/kg (Norway)	3	3	3	3	3	6	6	-	6	6	-	-	-	-	-
Average harvest weight, kg (Norway)	-	-	-	-	-	-	-	-	-	-					

[1] All lags are denoted relative to the horizon being forecasted.
To get the number of lags relative to the forecasters point of view, subtract by h months for any horizon h
[2] 1 indicates that the trend variable is included

Table 6.1: Overview of the ARDL-PLS covariates, with corresponding lags, employed for each submodel

of the same predictors being selected in multiple runs. It can also be seen that the GA tends to select the same predictors for different forecast horizons. For instance, the trend variable is the only covariate that is utilized in the submodels for h=6, but is not used in the submodels for h=3. Generally, the GA's tendency to select the same variables for different runs and different horizons, indicate that the algorithm is relatively stable and close to convergence.

It is somewhat surprising that the seasonal dummy variables are not used as covariates in any of the submodels, as former research indicate that the inclusion of seasonality increases the forecast accuracy of salmon price models. However, some of the exogenous variables used in the model, such as the trout price, are likely to have the same seasonal properties as the salmon price, which can correlate with the salmon price is a manner that makes the seasonal dummy variables excessive.

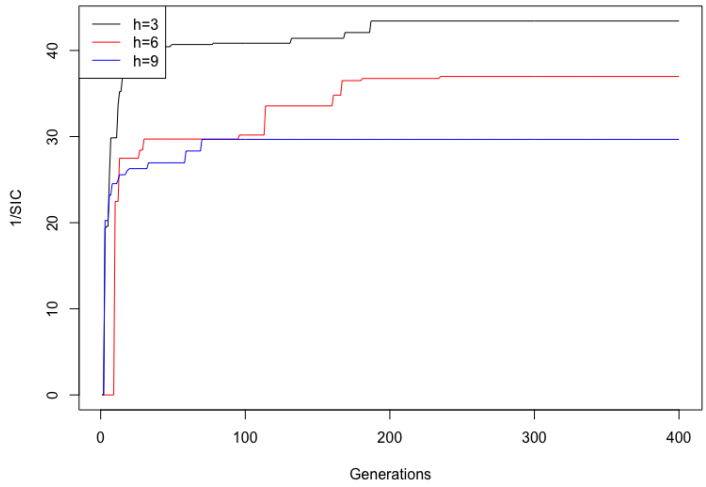

Figure 6.1: Plot of GA fitness evaluation scores for the first run of the ARDL-PLS submodels

6.2 Tuning the ARDL-PLS model

This section contains the results from the cross validation of the ARDL-PLS model, and the corresponding model tuning. As aforementioned, the ARDL-PLS submodels are trained and cross validated using the covariates selected by the GA-search. The aim of the cross validation is to decide the tuning parameter, i.e the number $L \leq 6$ PLS-components to use in each submodel. The results from the cross validation of the sub-models are shown in table 6.2. Empty cell indicate that the relevant number of PLS-components is constrained by the number of predictors selected by the genetic algorithm.

RMSE is used as the performance measure for choosing the number of PLS-components to use in the submodels. More specifically, the submodels are tuned according to the number of PLS-components that result in the lowest RMSE-value. An alternative would be to use R^2. However, the R^2 is a measure of correlation not accuracy. In addition the R^2 has a tendency to not properly

Submodels		PLS-components						Optimal[1]
		1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	
h=3	*r=1*	0.1912	0.1733	0.1615	0.1518	0.1460	0.1419	6
	r=2	0.2019	0.1757	0.1559	0.1531	0.1466	0.1417	6
	r=3	0.2053	0.1825	0.1658	0.1633	0.1608	0.1514	6
	r=4	0.1921	0.1753	0.1689	0.1620	0.1597	0.1528	6
	r=5	0.1947	0.1759	0.1668	0.1629	0.1591		5
h=6	*r=1*	0.1975	0.1881	0.1782	0.1652	0.1536	0.1519	6
	r=2	0.1974	0.1860	0.1718	0.1609	0.1509	0.1511	5
	r=3	0.1963	0.1887	0.1846	0.1836	0.1723	0.1628	6
	r=4	0.1970	0.1916	0.1920	0.1931	0.1824	0.1722	6
	r=5	0.1980	0.1920	0.1916	0.1885	0.1859	0.1789	6
h=9	*r=1*	0.1966	0.1885	0.1857	0.1794	0.1725		5
	r=2	0.1972	0.1922	0.1938	0.1877	0.1853	0.1786	6
	r=3	0.1992	0.1919	0.1948	0.1905	0.1844	0.1767	6
	r=4	0.1959	0.1894	0.1919	0.1876	0.1889	0.1881	4
	r=5	0.1975	0.1899	0.1906	0.1872	0.1815	0.1742	6

[1] Number of PLS-components used in the tuned ARDL-PLS submodels

Table 6.2: Overview of the cross validated RMSE-values computed for 1 to 6 PLS-components, and the optimal number of PLS-components, employed for each submodel

punish models that overpredict low values and underpredict high values (Kuhn & Johnson 2013). Consequently RMSE is utilized. As mentioned before, the results of the cross validation is not a good indicator of the model's ability to predict new data, as it is based on the same data as used in the variable selection. Thus, only the out-of-sample results, which are presented in section 6.5, should be used to evaluate the model performance.

It can be seen from table 6.2 that it is optimal in most cases to use all available PLS-components. The fact that most of the submodels utilize all available PLS-components, give a general indication that the covariates selected by the GA contain strong predictive power. The exceptions are the cases for $r = 2, h = 6$ and $r = 4, h = 9$, where it is optimal to leave out 1 component and 2 components respectively. For these cases, a share of information in the covariates are left unused as the results from the cross validation indicates that the inclusion of this information decreases the signal-to-noise ratio of the data. Generally, the RMSE-values are relatively similar for submodels corresponding to the same

horizon, which indicates that there is low intra-horizon variance in the optimal fitness function of the GA-search. It can be seen that the identical submodels $r = 2, h = 9$ and $r = 5, h = 9$ do not have the same RMSE-values. This is because the cross validation algorithm randomly partitions the dataset into folds. If the folds are not identical, the RMSE-values corresponding to these models can differ although they have identical regression coefficients.

6.3 The regression coefficients of the ARDL-PLS model

The regression coefficients of the tuned ARDL-PLS submodels can be found in appendix A.1. Following is an analysis of the regression coefficients for the submodels corresponding to the first run of the GA-search, i.e for $r = 1, h \in \{3, 6, 9\}$. Table 6.3 shows the regression coefficients of each predictor, the t-ratio and the correlation between the predictor and the response variable (i.e the log NQSALMON). The t-ratio is used as the measure for the predictor significance. In this paper, a predictor is considered significant when having an absolute t-ratio of greater or equal to 1.440, corresponding to a 10% significance. Although the hypothesis of the relationship between the response and the explanatory time series are summarized in chapter 3, these relationships might differ for the lags of the respective time series. Therefor the coefficient signs are compared to the *correlation* between the predictors and the response(i.e the log NQSALMON).

The variation in the set of covariates used in the submodels are surprisingly small. When disregarding the lag structure, the submodels $r = 1, h = 3$ and $r = 1, h = 6$ deploy the same set of explanatory time series as predictors. In addition, 4 out of 5 time series used in $r = 1, h = 9$ are used for $r = 1, h = 3$ and $r = 1, h = 6$. The applied lags for $r = 1, h = 3$ and $r = 1, h = 6$ are also very similar, and 4 out of 6 predictors used by these submodels are identical. They only differ for the selected lags of the trout price and the biomass of trout. The fact that the GA-search tends to select the same covariates inter-horizon, indicates that the selected covariates have strong predictive power.

All predictors selected by the GA-search are considered significant for at least one forecast horizon. Most of the regression coefficients signs equal to the correlation signs. For $r = 1, h = 9$, 5 out of 6 coefficient signs match the correlation

signs. For $r = 1, h = 3$ and $r = 1, h = 6$, 4 out of 6 coefficient signs match the correlation signs. Although the regression coefficient signs resulting from the ARDL-PLS model generally equal to the correlation signs, there are some exceptions. For instance, the meat price index has a negative regression coefficient, but has a positive correlation with the salmon price. Kennedy (2005) mentions that the omission of key explanatory variables, is a common reason for unexpected regression coefficient signs. In the case of the meat price index, there could be an excluded variable that has a positive correlation with the salmon price, and negative correlation with the meat price index. When excluding this "hidden" variable, it is possible that the meat price index ends up with a negative regression coefficient, although it has a standalone positive correlation with the salmon price. The phenomenon is known as omitted variable bias, and might introduce spuriosity to the model.

A more intuitive explanation for the "wrong" coefficient signs, is based on the fact that there exists only an indirect connection between the regression coefficients and the response, as they are both directly linked to the underlying latent structure of the PLS model. If there are several variables that describe the same variation in the salmon price, one of these variables can make up the majority of the PLS-component describing this variance. This leaves the remaining variables to contributing to other PLS-components, which describes other aspects of the salmon price. For instance, it can be shown that the correlation between the poultry index and the meat price index is 0.95, indicating that these variables are likely to describe similar aspects of the the salmon price variance. If the latent structure of the PLS-model is comprised in a way that leaves the majority of this common variance to be explained by the poultry index variable, the meat price index will be used to explain other aspects of the variation in the salmon price. Thus, if the share of the meat price index utilized by the latent structure has a negative correlation with the salmon price, the regression coefficient sign will be negative, although the meat price index has a standalone positive correlation with the salmon price. In contrast to an OLR model, these multicollinearity effects does not affect the predictive abilities of the PLS model, as the model regresses uncorrelated PLS-components. One can conduct a thorough analysis of the loading plots and score plots resulting from the PLS-model, in order to get a full overview of how the latent structure is comprised. However, the aim of this report is merely to propose a satisfying prediction model, rather than conducting a thorough investigation of the variable contribution. Consequently, this analysis is left for future researchers.

Time series, Unit	h=3				h=6				h=9			
	Lag	*t-ratio*	β	*Corr*[1]	*Lag*	*t-ratio*	β	*Corr*[1]	*Lag*	*t-ratio*	β	*Corr*[1]
Trend-variable									-	3.74	0.01	0.51
Standing biomass, #Individuals (Norway)	6	4.66	0.82	0.55	6	7.87	1.19	0.55				
Standing biomass of trout, Tonnes (Norway)	6	2.26	0.27	-0.15	12	1.29	0.12	-0.24	9	-3.46	-0.26	-0.23
Harvest volume of trout, Tonnes (Norway)												
Sea lice treatments, % of fish being treated (Norway)	12	-1.04	-0.04	-0.43	12	-3.59	-0.12	-0.43	12	-5.41	-0.15	-0.43
Meat price index, Index	6	-4.66	-1.81	0.28	6	-7.51	-2.28	0.28	9	-4.45	-2.05	0.35
Poultry index, Index	6	2.78	1.07	0.29	6	3.77	1.19	0.29	9	4.90	1.32	0.37
Trout price, NOK/kg (Norway)	3	5.82	0.58	0.65	6	5.36	0.35	0.42				

[1] Correlation between the respective covariate and the log NQSALMON

Table 6.3: Overview of regression coefficient, t-ratio and correlation, employed for the ARDL-PLS submodels corresponding to the first run of the GA-search

6.4 Residual diagnostics of the tuned ARDL-PLS model

Now follows an analysis of the residuals from the cross validation of the tuned ARDL-PLS submodels, considering stationarity and autocorrelation.

Residual unit root test

As explained in chapter 4, stationary residuals implicate that the explanatory variables cointegrate with the salmon price. In order to confirm this, an ADF test and a KPSS test is performed on the residuals resulting from the cross validation of the ARDL-PLS submodels. For illustrative purposes, the test-results for the submodels $r = 1, h = \{3, 6, 9\}$ are shown in table 6.4. Stationarity was confirmed by both tests for all submodels, which implicate that the variables cointegrate. Plots of the residuals are shown in figure 6.2-6.4.

Horizon	KPSS		ADF	
	test stat	p-value	test stat	p-value
h=3	0.1136	>0.1	-3.8266	0.0207
h=6	0.0461	>0.1	-3.5213	0.0441
h=9	0.08325	>0.1	-4.4618	0.0100

Table 6.4: Stationarity tests on model residuals for the ARDL-PLS submodels $r = 1, h = \{3, 6, 9\}$

Residual autocorrelation

Figure 6.2-6.4 shows the residual plots and the ACF plot for the submodels $r = 1, h = \{3, 6, 9\}$. Generally, all submodels exhibit residual autocorrelation, which is not surprising as the residuals correspond to multi-step ahead forecasts. The requirement of i.i.d error terms only applied to one-step ahead forecast models (Hyndman & Athanasopoulos 2014a). The ACF plots shows that the autocorrelation is generally only significant for lags that are prohibited by the model, due to the lag constraints. E.g for submodel $r = 1, h = 3$, the ACF values are high for the lags $l \in \{1, 2, 13, 14, 15\}$. Autocorrelation in the lags $l > 12$ can be reduced by including these lags in the model, thus relaxing the lag constraints. Consequently lag 13,14 and 15 of the explanatory variables could be included as candidate predictors for all submodels in an attempt to reduce the residual autocorrelation. However, this would increase the set of candidate predictors used in the GA-search, which can increase the probability of overfit.

The autocorrelation in the early lags (i.e lags $l < h$), can not be reduced in this model, as the PLS model only can utilize information in lags by including them explicitly in the model equation. This is due to the fact that there is no mechanism in the PLS model that dynamically adjust the forecasts based on former forecast errors, in contrast to e.g ARIMA models. An alternative model, that is perhaps better at coping with residual autocorrelation, is the regression model with ARIMA errors presented by Hyndman & Athanasopoulos (2014b). When forecasting with ARIMA errors, we forecast the regression part of the model and the ARIMA part of the model separately, before combing the results. However, this procedure requires stationary transformed data, which can limit the long term predictive ability of the model, in contrast to the ARDL-PLS model.

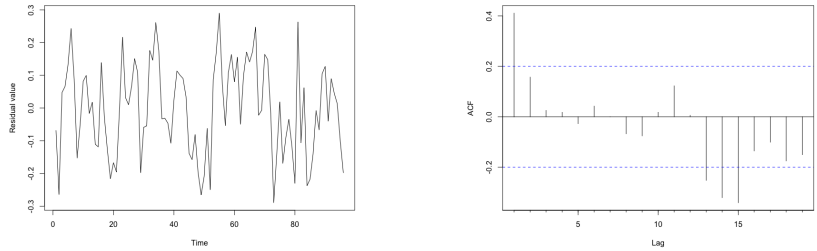

Figure 6.2: Residual plot and ACF plot for the ARDL-PLS submodel $r = 1, h = 3$

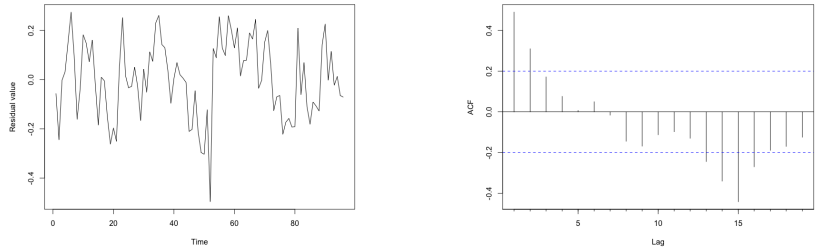

Figure 6.3: Residual plot and ACF plot for the ARDL-PLS submodel $r = 1, h = 6$

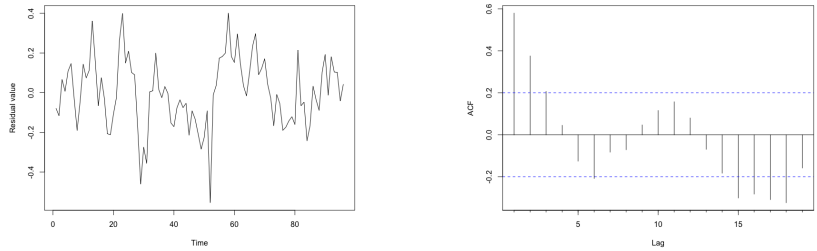

Figure 6.4: Residual plot and ACF plot for the ARDL-PLS submodel $r = 1, h = 9$

6.5 Out of sample results

Table 6.5 shows the out-of-sample performance (RMSE and R2) of the ARDL-PLS model, the OLR model, the Naive model and the ARIMA(1,1,0) model. The best average RMSE and R2 values for each horizon is marked in bold. Note that the results of the Naive model and the ARIMA model is only shown for one run, as these models don't utilize the genetic algorithm.

Generally, the ARDL-PLS model is able to explain a large degree of the variance in the salmon price. The average R^2 values of the out-of-sample predictions are 0.58, 0.41 and 0.44 for the respective 3-, 6- and 9-months ahead forecasts. The average performance of the ARDL-PLS model is superior to the OLR model in all horizons. For $h = 3$, the ARDL-PLS model and the OLR model perform relatively similarly, with respective average RMSE values of 0.1386 and 0.1428. For $h = 6$ and $h = 9$ the ARDL-PLS model is significantly better performing than the OLR model. In addition, the ARDL-PLS model greatly outperforms the ARIMA model and the Naive model on all horizons. The average RMSE values of the ARDL-PLS model is 17%, 26% and 43% less than those produced by the the Naive forecast for the respective 3-, 6- and 9-months ahead forecasts.

As aforementioned, the prediction accuracy of the OLR model and the ARDL-PLS model only differ substantially for the longer horizons (i.e $h = 6$ and $h = 9$). A likely explanation for the difference in the models ability to perform long term forecasts, is that the OLR model does not use orthogonal components in the regression, which makes the model vulnerable to multicollinearity affects and unstable regression coefficients. As the signal-to-noise ratio is likely to be higher for the longer horizons, the quality of the regression coefficients estimates are more crucial in long term forecasting. In addition, the use of "undifferenced" data allows for the retainment of long term predictive information, which is likely to increase the long term prediction performance of the ARDL-PLS model, relative to the OLR model.

For $h = 6$, the ARDL-PLS model has a relatively large intra-horizon variance in performance, compared to the other horizons, with RMSE values ranging between 0.1175 and 0.2094. In addition, the ARDL-PLS model performs better on $h = 9$ than on $h = 6$, which is somewhat surprising. This is likely to result from the fact that the submodels for $h = 6$ has access to a larger set of candidate predictors due to the lag constraint. Thus, the variable selection has a

Submodels		OOS-performance[1]							
		ARDL-PLS		OLR		Naive		ARIMA	
		RMSE	R^2	*RMSE*	R^2	*RMSE*	R^2	*RMSE*	R^2
h=3	*r=1*	0.1295	0.6395	0.1372	0.6073				
	r=2	0.1438	0.5552	0.1498	0.5314				
	r=3	0.1372	0.5952	0.1476	0.5354				
	r=4	0.1398	0.5797	0.1335	0.6281				
	r=5	0.1429	0.5610	0.1457	0.5573				
	Mean	**0.1386**	**0.5861**	0.1428	0.5719	0.1679	0.3935	0.1671	0.3999
h=6	*r=1*	0.1207	0.6866	0.2037	0.1343				
	r=2	0.1175	0.7034	0.2114	0.0675				
	r=3	0.1784	0.3156	0.2081	0.0964				
	r=4	0.1827	0.2826	0.1746	0.3636				
	r=5	0.2094	0.0577	0.1857	0.2804				
	Mean	**0.1617**	**0.4092**	0.1967	0.1884	0.2202	-0.0005	0.2148	0.0079
h=9	*r=1*	0.1514	0.5074	0.2142	0.0425				
	r=2	0.1480	0.5292	0.1431	0.5729				
	r=3	0.1461	0.5409	0.2082	0.0962				
	r=4	0.2041	0.1042	0.2359	-0.1615				
	r=5	0.1480	0.5292	0.2062	0.1119				
	Mean	**0.1595**	**0.4421**	0.2015	0.1324	0.2845	-0.6699	0.2789	-0.6725

[1] Bold figures indicates the best average scores among the models, intra-horizon.

Table 6.5: Comparison of the out-of-sample performance of the ARDL-PLS model, the OLR model, the Naive model and the ARIMA(1,1,0) model

larger search space and more combinations to assess. Due to the large amount of noise in the data, the increase in candidate predictors is likely to increase the probability of the GA-search finding patterns among the predictors that are not reproduced in the future, thus resulting in an overfit. The intra-horizon performance variance is larger for the ARDL-PLS model than for the OLR model, indicating that the ARDL-PLS model is more sensitive to the choice of subset selected by the GA-search. This is likely to be caused by the fact that the ARDL-PLS predicts in logs, in contrast to log returns, which makes the choice of covariates somewhat more crucial for the ARDL-PLS model, compared to the OLR model. That being said, the worst run of the ARDL-PLS model is still better than the worst run of the OLR model, for all horizons.

It is somewhat surprising that the OLR model greatly outperforms the ARIMA model in all horizons. As both models are linear regressions (note that the MA term in the ARIMA model is unused) and utilize log return transformed lagged time series, they are mathematically very similar. The OLR model only differ

in terms of applied input data, as the model utilize exogenous variables. In addition, the ARIMA model is only slightly better performing than the Naive model. In sum, this shows that past salmon price observation are of limited use to salmon price prediction models, relative to exogenous variables. Thus, in context of salmon price prediction, this study is a proof of concept of the usage of exogenous variables as covariates.

Figure 6.6-6.8 shows the out-of-sample residuals and the corresponding predictions for the ARDL-PLS submodels $r = 1, h \in \{3, 6, 9\}$. Generally, the submodels overpredict the first half of the test set, and underpredict the second half of the test set. This is likely to be explained by examining the plot of the log salmon price, shown in figure 6.5. At the beginning of the test set (*t=96*), the salmon price has just dropped after a relatively long peak period. As the ARDL-PLS model only utilize lagged time series as predictors, it can not fully follow the sudden changes in the price, which results in an optimistic estimate of the following periods. The same mechanism results in a pessimistic estimate of the second half of the test set, as the salmon price experience a major increase after *t=110*. Note that the submodel for $r = 1, h = 3$ actually manage to follow the sudden increase in the price at around *t=110*, and slightly overpredicts the following periods. This can result from the fact that the submodel conducting the 3-month ahead forecast is more "updated", relative to the submodels conducting the 6- and 9-month ahead forecasts. However, it can not be denied that the models have a general tendency to overpredict. Yet, given the fact that the test set is relatively small and exhibits a relatively high volatility, we can not draw any general conclusion upon whether the models is biased or not. By examining the ACF plots in figure 6.9, it can be seen that there are more autocorrelation in the residuals corresponding to the out-of-sample predictions, compared to the residuals corresponding to the cross validation predictions, where the latter is shown in figure 6.2-6.4. This is likely to result from the fact that the cross validation use the same samples as the GA-search, which can increase the goodness-of-fit in the cross validation predictions, relative to the out-of-sample predictions. Lastly, it can be seen that the out-of-sample residuals exhibit largest autocorrelation in the lags that are left out due to the aforementioned lag constraints

Table 6.5 shows that the R^2 values of the Naive model and the ARIMA model are negative in several of the horizons. This can be explained by looking at the formula for R^2, which is presented in section 5.2. If the denominator in the fraction is larger than the numerator, the R^2 becomes negative. This happens if

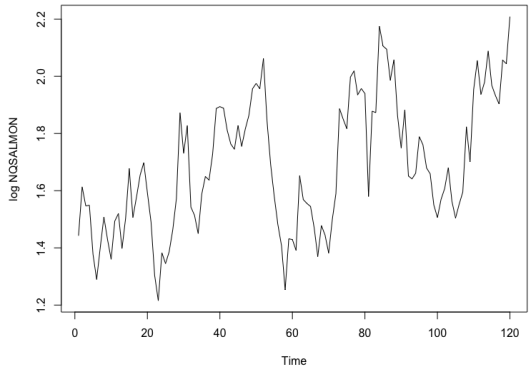

Figure 6.5: Plot of the log NQSALMON

the average forecast error is larger than the average deviation between the true value and the mean of the true value. Note that the difference in performance between the OLR model and the Naive model are much less than the values presented by Sandaker et al. (2016), who's models produce forecast errors of approximately 50% of the forecasts produced by their naive model. One main reason for why the results in this paper deviates strongly from the results of Sandaker et al. (2016) is because the naive model used in this report is based on naive forecasting in *log* values, while the naive model in Sandaker et al. (2016) is based on naive forecasting in *log return* values. The naive forecast performs much better in log values, which reduces the difference in performance between the OLR and the Naive model. To illustrate this effect, I have used the naive forecasting approach to predict the log return transformed salmon price, transformed the resulting predictions back to log values and calculated the corresponding RMSE-value. Then I used the naive forecast to predict the log transformed salmon price and calculated the corresponding RMSE-value. This is done for $h \in \{1, 2\}$. The result of both approaches are shown in table 6.6. Evidently, the naive forecast used in this paper is superior to the one used by Sandaker et al. (2016), which makes it harder to outperform.

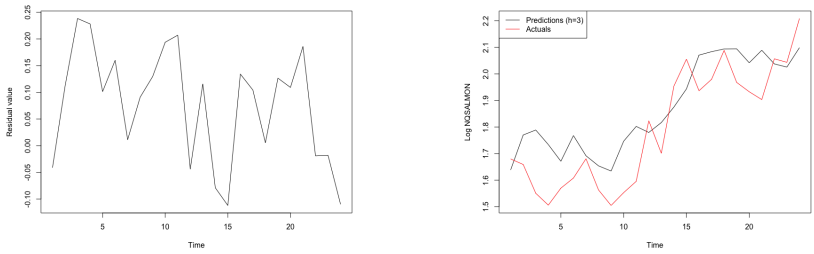

Figure 6.6: Out-of-sample residual plot and prediction plot for the ARDL-PLS submodel $r = 1, h = 3$

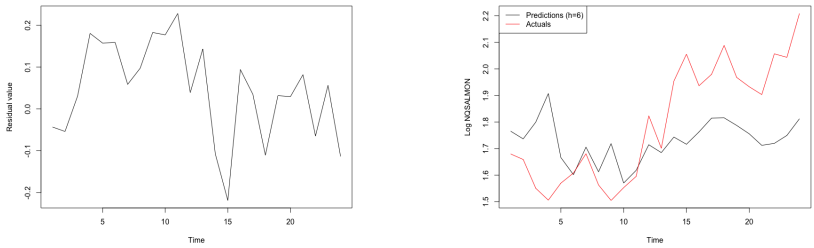

Figure 6.7: Out-of-sample residual plot and prediction plot for the ARDL-PLS submodel $r = 1, h = 6$

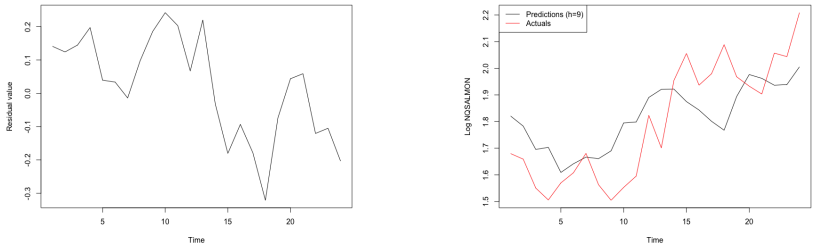

Figure 6.8: Out-of-sample residual plot and prediction plot for the ARDL-PLS submodel $r = 1, h = 9$

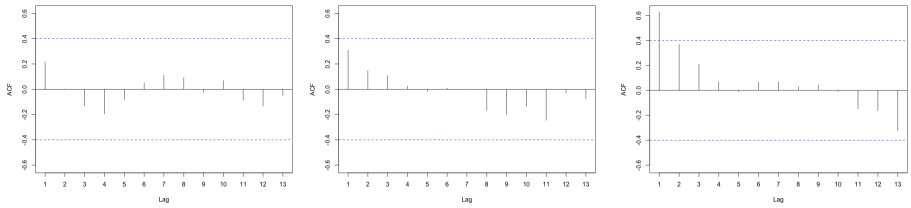

Figure 6.9: Out-of-sample residual ACF plots for the ARDL-PLS model. From left: $h = 3$, $h = 6$ and $h = 9$

Naive forecast		
	h=1	*h=2*
Log model	0.1311	0.1588
Log return model	0.2096	0.2370

Table 6.6: RMSE values of the Naive forecast

6.6 Improvements and expansion

This section provides possible improvements and expansions of the methodology presented in this paper, which would be interesting to investigate in future research.

As aforementioned, I have implemented cointegration constraint in the GA-search, in order to prevent spurious regression. This enables the use of large datasets without the need to manually asses that all possible subsets of candidate predictors satisfy the cointegration requirements. However, the restriction concerning exogeneity is difficult to fulfill. In the exogeneity test implemented in the GA-search, we consider the variables to be exogenous as long as it is not significantly unlikely that this hypothesis is false. Consequently, it is likely that some of the covariates used in the model are not actually exogenous. As the consequences of including endogenous variables in the model is worse than the consequence of failing to include variables that are in fact exogenous, it would perhaps be better to use a test that have endogeneity as the null hypothesis. However, it would probably be more useful to modify the model in such a way that allows for the removal of the exogeneity requirement, as it is generally hard

to find variables that are strictly exogenous. Bentzen & Engsted (2001) states that if the explanatory variables are in fact endogenous, we can include an adequate number of lagged changes in the regression in order to make the model theoretically valid. Thus, a possible improvement of the methodology in this paper is to remove the exogeneity restriction from the GA-search, and include additional covariates, in the form of lagged changes of the existing covariates, as predictors. The removal of the exogeneity requirement in the GA is likely to reduce the computation time of the algorithm. However, the inclusion of the lagged changes will increase the number of covariates in the model, making it less parsimonious and more vulnerable to overfit.

The computational requirements of the GA-search, and the fact that it had to be run for both the ARDL-PLS model and the OLR model, forced me to limit the number of runs computed for each horizon. Thus, in order to increase the empirical basis of the performance evaluation, one could conduct a more comprehensive model testing by running the model many times for each forecast horizon. This would result in a better estimate of the average forecast precision of the ARDL-PLS model, which would create a better basis for the comparison of the ARDL-PLS model and the OLR model. In addition, one could conduct the model testing for all horizons $h \in \{1, 2.., 12\}$.

Lastly, it is possible that one could discover non-linear dependencies between the salmon price and the explanatory variables, as it is arguably unlikely that the optimal dependency of all the explanatory variables are linear. Non-linear dependencies can be identified by examining the scatter plots of the exogenous variables and the salmon price. An alternative method would be to compare a linear model with several polynomial models, for each explanatory variable in the dataset. If one discovers non-linear dependencies, additional predictors that are functions of the original predictors can be added directly in the regression, in order to possibly improve the prediction accuracy of the model. The inclusion of non-linearities in the model would be an interesting expansion of the methodology presented in this paper. However, one should be sure to remove the original linear variables from the regression when adding the corresponding non-linear functions, in order to maintain the parsimonious characteristics of the model.

7 | Conclusion

As the market of salmon farming is becoming more globalized and competitive, the ability to limit unnecessary costs is essential for maintaining market positions. A satisfying price forecast model is an important tool in order to establish functional risk management and operational efficiency. Reliable estimates of future salmon prices can provide vital decision support when determining the optimal timing of salmon harvest, which terms producers should engage in forward contracts, or the required machine capacity. In addition, reliable estimates of future salmon prices are useful from a financial standpoint, as the salmon industry is increasing its presence in capital markets. Price models can be used to improve investments decisions, trading and contribute to better valuation of bonds and stocks

In this study, I develop a prediction model for the salmon spot price, represented by the NQSALMON index. In particular, I apply a general-to-specific approach to compute the 3-, 6- and 9- months ahead forecast of the salmon price. I use a dataset consisting of 17 time series, such as salmon biomass, average sea temperature and prices of alternative protein. The time series are log transformed, and lagged according to industry assumptions regarding when they are assumed to impact the salmon price. The time series combined with the applied lag structure, generate a large set of candidate predictors for the multi-step spot price prediction. The prediction model, which is named ARDL-PLS, combines an ARDL model and a PLS regression. The model enables the use of non-stationary data, and avoids regression problems resulting from the use of intercorrelated predictors. A genetic algorithm (GA) for variable selection is applied in order to determine the best subset of candidate predictors for each forecast horizon. As some of the time series in the dataset are non-stationary, the set of covariates used in the model has to satisfy certain cointegration requirements, which are formulated as optimization constraints in the GA-search. Consequently, all the subsets of predictors selected by the GA-search are cointegrated with the salmon price. The out-of-sample performance of the ARDL-PLS model is compared to the performance of an ordinary least square regression (OLR) model. The OLR model utilize a GA for variable selection, but is applied to log return transformed time series. The resulting predictions from the ARDL-PLS model and the OLR model are computed on 5 different runs of the GA, in order to get a more reliable estimates of the model

performance. Lastly, I apply an ARIMA model and a naive model as reference.

The GA-search is quickly able to find cointegrated subsets of variables that result in a favourable goodness-of-fit. Generally the ARDL-PLS model is able to explain a large proportion of the variance in the salmon price. The respective R^2-values for the 3-, 6- and 9- months ahead out-of-sample forecasts are 58%, 41% and 44%, which is satisfying given the fact that economic time series are considered low signal-to-noise environments. The 9-month ahead predictions are more accurate and less volatile than the 6-months ahead predictions. This is somewhat surprising, but likely to result from the fact that there are more available candidate predictors for sub-models predicting the shorter horizons, due to the aforementioned lag structure. Thus, the variable selection has a larger search space and more combinations to assess, which makes it harder to select appropriate predictors. The predictive accuracy of the ARDL-PLS model is better than the OLR model on all horizons. The 3 month-ahead prediction accuracy of the models are relatively similar, but the ARDL-PLS model excel in the longer horizons. I attribute the performance difference between the OLR model and the ARDL-PLS model with two factor: (1) The return transformation of the data used in the OLR model is known to remove long term predictive information which decreases the model's ability to forecast longer horizons. (2) Intercorrelations among the explanatory variables can result in unstable regression coefficients in the OLR model, as it, in contrast to the ARDL-PLS model, does not transform the predictors into uncorrelated orthogonal components. The prediction accuracy of the ARDL-PLS model is more volatile than to the prediction accuracy of the OLR model. This is likely to be caused by the fact that the ARDL-PLS predicts in logs, instead of log returns, which makes the choice of covariates somewhat more crucial for the ARDL-PLS model, relative to the OLR model. The out-of-sample performance of the ARDL-PLS model is superior to that of the Naive model and the ARIMA model, yielding a forecast error of 56% of the one produced by the Naive model, for the 9-months ahead forecast. It was also observed that the OLR model performed far better than the ARIMA model. Since the two models only differ due to the fact that OLR model use exogenous variables, this result is a proof of concept for the use of exogenous variables in salmon prediction models.

It was observed that some of the regression coefficient signs in the ARDL-PLS model deviate from expectation. A likely explanation is based on the fact that the model transforms the predictors into orthogonal components. Consequently, there exists only an implicit connection between the regression coefficients and

the response variable. However, another explanation is that the model suffers from omited variable bias, which can affect the predictive ability of the model. It is uncertain, if any, which of the two is the case. However, the fact that the prediction accuracy is satisfying for most model runs, indicate that there is limited bias in the model. In addition, "Some econometricians would go as far as to suggest that the statistical adequacy of a model (...), is largely irrelevant if the model produces accurate forecasts" (Brooks 2008).

I propose some possible improvements and expansions of the methodology presented in this paper. The exogeneity constraint in the GA-search can be removed if one includes an adequate number of lagged changes as covariates in the model. This would likely increase the computational speed of the GA and reduce the probability of the regression becoming spurious. Also, it would be useful to perform a more comprehensive model testing, in order to get a better estimate of the ARDL-PLS models predictive ability, and create a better basis for the comparison of the ARDL-PLS model and the OLR model. This can be done by greatly increasing the number of model runs for each forecast horizon, and increasing the number of horizons tested. Lastly, one could investigate the impact of including non-linearities in the model. This can be done by including covariates in the model that are functions of the original predictors.

A | Appendix

A.1 The ARDL-PLS model

Prediction plots for the ARDL-PLS model

Figure A.1-A.3 shows the cross validation predictions and the out-of-sample predictions for the ARDL-PLS submodels $r = 1, h \in \{3, 6, 9\}$. The predictions are separated by the black line.

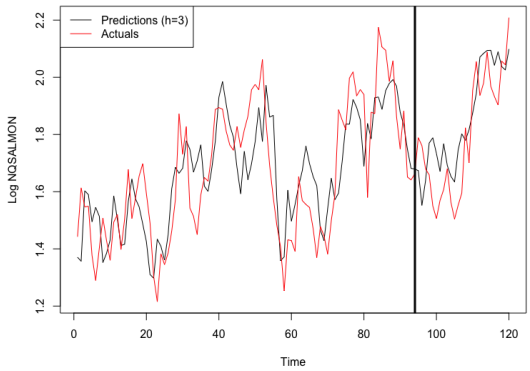

Figure A.1: Plot of the CV and OOS predictions for the ARDL-PLS submodel $h = 3, r = 1$

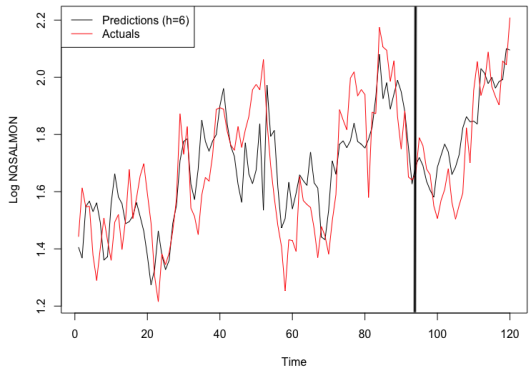

Figure A.2: Plot of the CV and OOS predictions for the ARDL-PLS submodel $h = 6, r = 1$

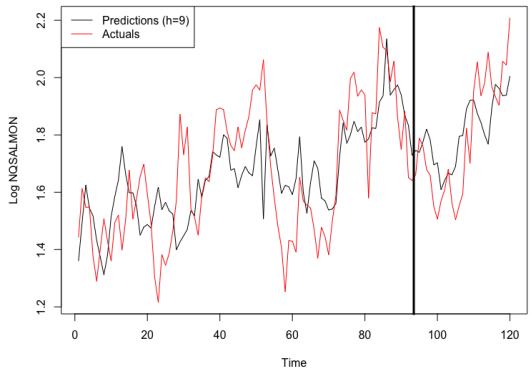

Figure A.3: Plot of the CV and OOS predictions for the ARDL-PLS submodel $h = 9, r = 1$

Regression coefficients for the ARDL-PLS model

Table A.1-A.3 shows the regression coefficients for all ARDL-PLS submodels.

Time series, Unit	**Horizon 3**[1]									
	r=1		r=2		r=3		r=4		r=5	
	Lag	*β*	*Lag*	*β*	*Lag*	*β*	*Lag*	*β*	*Lag*	*β*
Standing biomass, #Individuals (Norway)	6	0.8211	6	0.8219						
Standing biomass of trout, Tonnes (Norway)	6	0.2780	6	0.3806	6	0.7677	6	0.6342	6	0.4319
					9	0.4571	9	-0.3419		
Sea lice treatments, % of fish being treated (Norway)	12	-0.0389	3	-0.012	3	-0.0498	12	-0.3224	12	-0.0261
Meat price index, Index	6	-1.8129	6	-1.7499	6	-1.6647	6	-1.5255	6	-1.2188
Poultry index, Index	6	1.0723	6	1.0147	6	1.5513	6	1.4766	6	1.2214
Trout price, NOK/kg (Norway)	3	0.5836	3	0.6350	3	0.8534	3	0.8105	3	0.8085

[1] All lags are denoted relative to the horizon being forecasted.
To get the number of lags relative to the forecasters point of view, subtract by h months for any horizon h

Table A.1: Regression coefficients for the ARDL-PLS submodels conducting 3-month ahead predictions

Time series, Unit	**Horizon 6**[1]									
	r=1		r=2		r=3		r=4		r=5	
	Lag	*β*	Lag	*β*	Lag	*β*	Lag	*β*	Lag	*β*
Trend-variable[2]					1	0.0081	1	0.0077		
Standing biomass, #Individuals (Norway)	6	1.1927	6	1.1234					9	0.4664
Standing biomass of trout, Tonnes (Norway)	12	0.1297	6	0.2008	9	-0.3460	12	-0.2340	12	-0.1013
					12	-0.0499				
Sea lice treatments, % of fish being treated (Norway)	12	-0.1189	12	-0.0951	12	0.1475	12	-0.1156	12	-0.1311
Meat price index, Index	6	-2.2759	6	-2.4288	6	-2.8222	6	-2.7205	6	-1.8073
Poultry index, Index	6	1.1992	6	1.4191	6	1.6800	6	1.6749	6	1.4471
Trout price, NOK/kg (Norway)	6	0.3472	6	0.4417			6	0.0919	6	0.2937

[1] All lags are denoted relative to the horizon being forecasted.
To get the number of lags relative to the forecasters point of view, subtract by h months for any horizon h
[2] 1 indicates that the trend variable is included

Table A.2: Regression coefficients for the ARDL-PLS submodels conducting 6-month ahead predictions

Time series, Unit	**Horizon 9**[1]									
	r=1		r=2		r=3		r=4		r=5	
	Lag	*β*	Lag	*β*	Lag	*β*	Lag	*β*	Lag	*β*
Trend-variable[2]	1	0.0057	1	0.0055	1	0.0062	1	0.0038	1	0.0055
Standing biomass of trout, Tonnes (Norway)	9	-0.2639	9	-0.3806	9	-0.2203	9	-0.3862	9	-0.3806
Harvest volume of trout, Tonnes (Norway)			9	0.0996	12	-0.0634	9	0.1368	9	0.0996
Sea lice treatments, % of fish being treated (Norway)	12	-0.1512	12	-0.1503	12	-0.1439	12	-0.1141	12	-0.1503
Meat price index, Index	9	-2.0460	9	-1.9849	9	-2.1641	12	-0.1949	9	-1.9849
Poultry index, Index	9	1.3163	9	1.2779	9	1.3699	9	-0.0375	9	1.2779

[1] All lags are denoted relative to the horizon being forecasted.
To get the number of lags relative to the forecasters point of view, subtract by h months for any horizon h
[2] 1 indicates that the trend variable is included

Table A.3: Regression coefficients for the ARDL-PLS submodels conducting 9-month ahead predictions

A.2 The OLR model

The GA used by the OLR model

Table A.4 shows the parameter settings of the GA used by the OLR model, and below is an illustration of the corresponding fitness function. Since the OLR model is computed on stationary data, the cointegration constraints are left out of the fitness function. By leaving out the cointegration requirements, the search space becomes more relaxed and there are more valid solutions. Thus, the mutation possibility is lowered relative to GA-search used by the ARDL-PLS model, as it is much easier for the algorithm to obtain valid solutions.

Parameter	Value	Description
Population size	populationSize= 400	The number of chromosomes in each population. A small population creates a small search space, while a large population slows down the convergence time
Generations/Iterations	numGenerations =400	The number of new populations created, i.e the number of iterations of the algorithm.
Cross over	crossover = "single"	Defines the way parent chromosomes are combined in order to create new offsprings.
Elitism	elitism = 9	Copies the best chromosomes to the next generation, in order to prevent loosing the best found solution. It is found that elitism rapidly increases the performance of the GA.
Mutation probability	mutationProbability=0.05	Randomly changes the genes of new offsprings in order to prevent the GA from falling into local optimums.

Table A.4: Parameter settings for the GA-search used by the OLR model

Finess evaluation of subset S

1. Fitness = (1/SIC) computed on subset S
2. **return** Fitness

Regression coefficients for the OLR model

Table A.5-A.7 shows the regression coefficients corresponding to the OLR sub-models.

Time series, Unit	Horizon 3[1]									
	r=1		r=2		r=3		r=4		r=5	
	Lag	β	Lag	β	Lag	β	Lag	β	Lag	β
Standing biomass, Tonnes (Norway)	3	-1.5157	3	-1.4667	3	-1.3569	3	-1.7188	3	-1.5116
Feed consumption, Tonnes (Norway)	3	-0.1589	3	-0.2031	3	-0.2155	3	-0.1431	3	-0.1391
Standing biomass of trout, Tonnes (Norway)			9	-0.3797						
Sea lice occurrence, #Lice/fish (Norway)					12	0.04565				
Sea temperature, Degrees Celsius (Norway)	6	0.66526	6	0.5562	6	0.6311	6	0.6726	6	0.6127
Meat price index, Index	6	-3.0265	6	-2.9799	6	-29512	6	-2.9443	6	-2.8185
Poultry index, Index	6	1.8092	6	1.8203	6	1.8644	6	1.7213	6	1.6683
Beef price, US cents/pound							12	-0.3149		
Trout price, NOK/kg (Norway)	6	0.2698								

[1] All lags are denoted relative to the horizon being forecasted.
To get the number of lags relative to the forecasters point of view, subtract by h months for any horizon h

Table A.5: Regression coefficients for the OLR submodels conducting 3-month ahead predictions

Time series, Unit	Horizon 6[1]									
	r=1		r=2		r=3		r=4		r=5	
	Lag	β	Lag	β	Lag	β	Lag	β	Lag	β
Standing biomass, Tonnes (Norway)	6	0.6392	9	-0.7805	6	0.6563	6	0.6139	6	0.5755
	9	-0.9587			9	-0.7674	9	-0.8021	9	-0.8207
Standing biomass of trout, #Individuals (Norway)	12	-0.6453	12	-0.8982	12	-0.4007	12	-0.4552	12	-0.5343
Sea lice occurrence, #Lice/fish (Norway)			12	-0.0761						
Meat price index, Index	6	-3.2428	6	-3.3925	6	-2.8768	6	-3.0180	6	3.1802
Poultry index, Index	6	1.9185	6	2.0315	6	1.9398	6	1.8329	6	1.9364
Currency pair, USD/EUR					6	0.6358			12	-0.4973
Trout price, NOK/kg (Norway)	12	-0.3341	12	-0.3324						

[1] All lags are denoted relative to the horizon being forecasted.
To get the number of lags relative to the forecasters point of view, subtract by h months for any horizon h

Table A.6: Regression coefficients for the OLR submodels conducting 6-month ahead predictions

Time series, Unit	Horizon 9[1]									
	r=1		r=2		r=3		r=4		r=5	
	Lag	β	Lag	β	Lag	β	Lag	β	Lag	β
Standing biomass of trout, #Individuals (Norway)	12	-0.9548	12	-0.5927	12	-0.7337	12	-0.8419	12	-0.7760
Standing biomass of trout, Tonnes (Norway)							12	-0.4084		
Sea lice occurrence, #Lice/fish (Norway)	12	-0.8151	12	-0.0865	12	-0.0856			12	-0.0728
Meat price index, Index	9	-2.1029	9	-1.8146	9	-2.3517	9	-2.0395	9	-2.0047
Poultry index, Index	9	1.8492	9	2.1160	9	1.8857	9	1.7754	9	1.8046
Beef price, US cents/pound			12	-0.8156						
Currency pair, USD/EUR	9	1.0529	9	1.7295	9	1.3379	9	0.9113	9	1.1590
					12	-0.7743				
Trout price, NOK/kg (Norway)	12	-0.3469					12	-0.4877	9	-0.1944

[1] All lags are denoted relative to the horizon being forecasted.
To get the number of lags relative to the forecasters point of view, subtract by h months for any horizon h

Table A.7: Regression coefficients for the OLR submodels conducting 9-month ahead predictions

A.3 The ARIMA model

The following section describes the composition of the ARIMA model. In addition, the model tuning and resulting coefficients from the model fitting is displayed.

The ARIMA model

A non-seasonal ARIMA model is presented in Equation A.2. The model combines differencing with autoregression and a moving average model (Hyndman & Athanasopoulos 2014b). The model in Equation A.2 is referred to as an ARIMA(p,d,q) model, where p is the order of the autoregressive part, d number of first differences and q is the order of the moving average part.

$$y'_t = y_t - y_{t-1} \tag{A.1}$$

$$y'_t = c + \phi_1 y'_{t-1} + ... + \phi_p y'_{t-p} + \theta_2 \epsilon_{2-1} + ... + \theta_1 \epsilon_{t-q} + \epsilon_t \tag{A.2}$$

ARIMA tuning

In order to tune the ARIMA model, i.e set the the values for p, d and q, I use the *auto.arima()* function presented in the R package "Forecast", which utilize the Hyndman-Khandakar algorithm for automatic tuning of ARIMA models. The forecast package and the algorithm is presented by Hyndman & Athanasopoulos (2014a). The model resulting from the tuning algorithm is an ARIMA(1,1,0) model, which can be written on the form:

$$y'_t = c + \phi_1 y'_{t-1} + \epsilon_t \tag{A.3}$$

The autoregressive parameter ϕ_1 for the forecast horizons $h = \{3, 6, 9\}$ can be found in table A.8

ARIMA coefficients

	h=3	*h=6*	*h=9*
ϕ_1	-0.1272	-0.1246	-0.1192

Table A.8: ARIMA coefficients

A.4 Time series cross validation

This section discusses the validity of applying cross validation on time series models.

Cross validation (CV) often produce performance estimates superior to a single test set because they evaluate many alternate versions of the data (Kuhn & Johnson 2013). Because of serial correlation and potential non stationarity of time series data, validation of time series models are often done by an out-of-sample (OOS) evaluation. However, Bergmeir et al. (2018)) shows that in the case of purely autoregressive model, the use of standard K-fold CV is theoretically valid as long as the models considered have uncorrelated errors. In addition Rob Hyndman argues that although Bergmeir et al. (2018) considers univariate AR models, it extends naturally to multivariate AR models"(Hyndman 2018). In addition, Bergmeir & Benítez (2012) conducts an extensive study on the use of CV and OOS on time series, and found that the use of CV led to more robust model selection although the procedure might not be theoretically valid. Based on the abovementioned arguments, I find it reasonable to use cross validation as a tool to tune the ARDL-PLS model.

Bibliography

Asche, F., Misund, B. & Oglend, A. (2016), 'The spot-forward relationship in the atlantic salmon market', Aquaculture Economics & Management **20**(2), 222–234.

Bentzen, J. & Engsted, T. (2001), 'A revival of the autoregressive distributed lag model in estimating energy demand relationships', Energy **26**(1), 45–55.

Bergmeir, C. & Benítez, J. M. (2012), 'On the use of cross-validation for time series predictor evaluation', Information Sciences **191**, 192–213.

Bergmeir, C., Hyndman, R. J. & Koo, B. (2018), 'A note on the validity of cross-validation for evaluating autoregressive time series prediction', Computational Statistics & Data Analysis **120**, 70–83.

Bloznelis, D. (2016), 'Salmon price volatility: A weight-class-specific multivariate approach', Aquaculture economics & management **20**(1), 24–53.

Brooks, C. (2008), Introductory Econometrics For Finance, 2 edn, Cambridge University Press.

Coleman, L. (2012), 'Explaining crude oil prices using fundamental measures', Energy Policy **40**, 318–324.

Diebold, F. X. (2014), Forecasting in Economics, Business, Finance and Beyond.

Evans, M. K. (2002), Practical business forecasting, John Wiley & Sons.

Fedorová, D. (2016), 'Selection of unit root test on the basis of length of the time series and value of ar (1) parameter', STATISTIKA **96**(3), 3.

Fuentes, J., Poncela, P. & Rodríguez, J. (2015), 'Sparse partial least squares in time series for macroeconomic forecasting', Journal of Applied Econometrics **30**(4), 576–595.

Groen, J. J. & Kapetanios, G. (2009), 'Revisiting useful approaches to data-rich macroeconomic forecasting'.

Gu, G. & Anderson, J. L. (1995), 'Deseasonalized state-space time series forecasting with application to the us salmon market', Marine Resource Economics **10**(2), 171–185.

Guttormsen, A. G. (1999), 'Forecasting weekly salmon prices: risk management in fish farming', Aquaculture Economics & Management **3**(2), 159–166.

Harrell, F. E. (2015), Regression Modeling Strategies, 2 edn, Springer.

Hyndman, R. J. (2018), 'Time series cross-validation: an r example'. Accessed: 2018-04-04.
URL: *https://robjhyndman.com/hyndsight/tscvexample/*

Hyndman, R. J. & Athanasopoulos, G. (2014a), 'Forecasting: Principles practice'. Accessed: 2018-04-04.
URL: *https://www.otexts.org/fpp/2/6*

Hyndman, R. J. & Athanasopoulos, G. (2014b), Forecasting: principles and practice, OTexts.

Johansen, S. (1991), 'Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models', Econometrica: Journal of the Econometric Society pp. 1551–1580.

Kennedy, P. E. (2005), 'Oh no! i got the wrong sign! what should i do?', The Journal of Economic Education **36**(1), 77–92.

Kuhn, M. & Johnson, K. (2013), Applied predictive modeling, Vol. 26, Springer.

Lusk, J. L. & Aaron, A. (2016), 'No title'.

Marine Harvest (2017), 'Salmon farming industry handbook'. Accessed: 2018-03-20.
URL: *http://marineharvest.com/globalassets/investors/handbook/salmon-industry-handbook-2017.pdf*

Misund, B. (2017), 'Financial ratios and prediction on corporate bankruptcy in the atlantic salmon industry', Aquaculture Economics & Management **21**(2), 241–260.

Misund, B. & Asche, F. (2016), 'Hedging efficiency of atlantic salmon futures', Aquaculture Economics & Management **20**(4), 368–381.

Nkoro, E., Uko, A. K. et al. (2016), 'Autoregressive distributed lag (ardl) cointegration technique: application and interpretation', Journal of Statistical and Econometric Methods **5**(4), 63–91.

Obitko, M. (2014), 'Introduction to genetic algorithms'. Accessed: 2018-04-04.
URL: *http://www.obitko.com/tutorials/genetic-algorithms/about.php*

Oglend, A. (2013), 'Recent trends in salmon price volatility', Aquaculture Economics & Management **17**(3), 281–299.

Oglend, A. & Sikveland, M. (2008), 'The behaviour of salmon price volatility', Marine Resource Economics **23**(4), 507–526.

Sandaker, K., Mjaugeto, P. O. W. & Steinshamn, K. B. (2016), 'Forecasting the atlantic salmon spot price using a general-to-specific regression approach'.

Sandaker, K., Mjaugeto, P. O. W. & Steinshamn, K. B. (2017), Predicting the distribution of the atlantic salmon spot price using quantile regression, Master's thesis, NTNU.

Seafish (2017), 'Aquaculture'. Accessed: 2017-12-15.
URL: *http://www.seafish.org/industry-support/aquaculture*

Sjømat Norge (2011), 'Aquaculture in norway'. Accessed: 2017-12-15.
URL: *https://sjomatnorge.no/wp-content/uploads/importedfiles/Aquaculture%2520in%2520Norw*

The Nasdaq Group Inc. (2016a), 'Nqsalmon annual summary 2016'. Accessed: 2018-03-20.
URL: *http://fishpool.eu/wp-content/uploads/2017/05/NQSALMON-Annual-Summary-2016.pdf*

The Nasdaq Group Inc. (2016b), 'Rules for the construction, maintenance and use of the nasdaq salmon index – nqsalmon'. Accessed: 2018-03-20.
URL: *https://salmonprice.nasdaqomxtrader.com/NQSALMONRules160425.pdf*

Vukina, T. & Anderson, J. L. (1994), 'Price forecasting with state-space models of nonstationary time series: case of the japanese salmon market', Computers & Mathematics with Applications **27**(5), 45–62.

Wold, H. (1966), 'Estimation of principal components and related models by iterative least squares', Multivariate analysis pp. 391–420.

Wold, H., Johansson, M. & Cocchi, M. (1993), 'Pls–partial least-squares projections to latent structures', In H Kubinyi (ed.),"3D QSAR in Drug Design," volume 1, pp. 523–550. .

Wold, S., Ruhe, A., Wold, H. & Dunn, III, W. (1984), 'The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses', SIAM Journal on Scientific and Statistical Computing **5**(3), 735–743.

Ye, M., Zyren, J. & Shore, J. (2005), 'A monthly crude oil spot price forecasting model using relative inventories', International Journal of Forecasting **21**(3), 491–501.