Bacheloroppgave

NTNU
Norges teknisk-naturvitenskapelige universitet
Fakultet for økonomi
NTNU Handelshøyskolen

Johannes Risbakk

Predicting the pricing of salmon using Related commodities.

Bacheloroppgave i Business Administration, Business Analytics
Veileder: Denis Becker
August 2023

NTNU
Kunnskap for en bedre verden

Johannes Risbakk

Predicting the pricing of salmon using Related commodities.

Bacheloroppgave i Business Administration, Business Analytics
Veileder: Denis Becker
August 2023

Norges teknisk-naturvitenskapelige universitet
Fakultet for økonomi
NTNU Handelshøyskolen

NTNU
Kunnskap for en bedre verden

PREDICTING THE PRICING OF SALMON USING RELATED COMMODITIES.

Bachelor’s thesis in Business Administration, Business Analytics

Supervisor: Denis Becker
Norwegian University of Science and Technology
Faculty of Economics and Management
NTNU Business School

Johannes Risbakk
August 2023

Abstract

Norway is the world's largest salmon producer and salmon is one of Norway's most important exports. For salmon farmers and market participants, salmon price prediction is of great value as this can help farmers adjust their production in both the short and long term, and there are large economic rewards for providing good predictions of salmon price. In this article I am trying to predict the price of Norwegian salmon based on historical prices, other related commodities like fishmeal and shrimp, and the NOK-EUR exchange rate.

Four different models were employed to predict the salmon prices. As a baseline a naïve no-change forecast was created to test the other, more complex, models against. An ARIMA was used to serve as a univariate prediction baseline to see if the multivariate models could improve the predictions with the use of the commodity prices and the NOK-EUR exchange rate. The two multivariate models were a k-nearest neighbors (KNN) approach and a Long Short-Term Memory (LSTM) deep learning approach. The naïve no-change model was superior to the other models on all but one accuracy metric, hit rate, where KNN had the highest score.

Sammendrag

Norge er verdens største lakseprodusent, og laks er en av Norges viktigste eksportvarer. For oppdrettere er det av stor verdi å kunne predikere laksepriser, da dette kan hjelpe oppdretterne med å justere produksjonen på både kort og lang sikt. Gode prediksjoner av lakseprisen vil og stor økonomisk gevinst for alle markedsaktører. I denne artikkelen prøver jeg å forutsi prisen på norsk laks basert på historiske priser, andre relaterte råvarer som fiskemel og reker, og NOK-EUR valutakurs.

Fire forskjellige modeller ble brukt for å predikere lakseprisene. Som en baseline ble en naïve no-change modell opprettet for å teste de andre mer komplekse modellene mot. En ARIMA ble brukt som en univariat baseline for å se om de multivariate modellene kunne forbedre prediksjonene ved å bruke de andre råvareprisene og NOK-EUR valutakursen. De to multivariate modellene var en k-nearest neighbors (KNN) tilnærming og en Long Short-Term Memory (LSTM) dyp læring tilnærming. Den naive modellen presterte bedre enn alle de andre modellene på alle nøyaktighetsmetrikkene bortsett fra treffrate, hvor KNN hadde den høyeste scoren.

Table of Contents

Table of Figures

Table of Tables

1. Introduction

The global food industry is rapidly evolving, and the seafood sector is no exception. A growing global population has naturally been followed by increasing demand for food and seafood. Salmon farming is one of the most important export industries in Norway who historically has been the largest producer of salmon in the world, and the Norwegian share of global production has for the most part been above 50% (Iversen et al., 2020). This amounts to over one million metric tons over a year. Nearly all of the salmon farmed in Norway is exported and Europe is the largest importer of Norwegian salmon. Marine Harvest (now MOWI) reported in 2018 that only around 3 to 4 % of the produced salmon was consumed locally and the rest was exported (Marine Harvest ASA, 2018). Some of the other salmon farming countries are the UK, Canada and Chile and the global production of salmon has been increasing substantially from 230 000 metric tons in 1990 to 2.2 million tonnes in 2018 (Iversen et al., 2020).

Norwegian salmon is traded both on the spot market but also by the use of forward contracts. The price of salmon is a result of supply and demand, and the supply of salmon is influenced by among other things regulatory factors, biological and environmental factors, production productivity, price of input-factors, price of transportation, and exchange rate (Bloznelis, 2018). One of the most important regulatory factors is limitations on maximum allowable biomass, meaning the total amount of fish allowed to keep in one facility. The salmon producer is required to slaughter to stay within the total allowable biomass or be rewarded with fines. The growth of the biomass is decided largely by biological and environmental factors. Some of these are water temperature which influence salmon growth, and yearly cycles which cause the biomass of salmon to naturally peak in September which is contributing to create seasonal variations in price. This is also influenced by sexual maturation which, depending on among other things feed, happen in the end of summer or start of autumn and lead to lower quality of flesh. This cause producers to often want to slaughter the fish before this happens to deliver a higher quality product (Pall, Andersson, & Taranger, 2006).

Disease and parasites are also important factors and have the potential to create supply shocks as seen in Chile in the late 2000s when the salmon disease infectious salmon anemia (ISA) caused the total biomass in Chile to drop substantially leading to global impact on the trade of salmon (Asche et al., 2009). Sometimes it is possible to slaughter and sell sick fish as the diseases often aren’t dangerous for human health and might not lead to worse quality fish, other times it is not possible and the fish has to be discarded. Demand for salmon is not constant throughout the year but show signs of seasonal trends. Salmon sees increased demand around the holidays like Christmas (Bloznelis, 2018) and an explanation is that salmon is a more expensive source of protein than many others and might therefore see increased demand related to special occasions where consumers are willing to spend more money on food.

Forecasting of prices are an important activity for many actors within different commodity markets and also within the salmon market. Salmon farmers would like to optimize biomass and slaughtering based on the prices in the market in order to maximize profits. Traders, processors and retailers will be sensitive to short term fluctuations in the salmon price when trying to minimize the purchasing cost of salmon and trying to optimize revenues from reselling it. And of course, forecasting of price is fundamental for speculators looking to make a profit within the salmon market. Better forecasts can reduce inefficiencies in the production of salmon and lead to a more efficient market and in the end better prices for the consumer.

Bloznelis (2018) presents different factors influencing how producers and market participants in the shorter and longer term can receive potential benefits from different forecast horizons. A shorter-term forecast (e.g., days, weeks or a few months) of salmon prices can allow farmers and slaughtering facilities to adjust slaughtering volume to match expected market movements. Feeding intensity can also be adjusted, and this can to a limited extent adjust short term production of salmon. In the longer term (over the course of a year) forecasts can be useful in planning of production as salmon farmers can influence production to a large extent by planning the stocking of their pens and thereby total biomass. Longer term forecasts have historically been provided by several actors such as private analytics companies, salmon farming companies, and bank analysts. There has from a research perspective been less interest in short term forecasts (Bloznelis. 2018).

The ambition of this paper will be to predict the export prices of Norwegian salmon using related commodities and macroeconomic factors. I want to investigate if a multivariate approach based on LSTM or KNN is able to provide better forecasts than a univariate time series approach, and if either are able to beat a naïve no-change model. In the next chapter I will introduce some theory relevant that potentially can influence the price of farmed salmon. I will then introduce the variables I believe can be useful in creating multivariate analysis, before I further precent the methods that will be used for the prediction. Following this I will present the results and have a short discussion that discusses the value of this paper, and some limitations of this paper, before ending with a short conclusion.

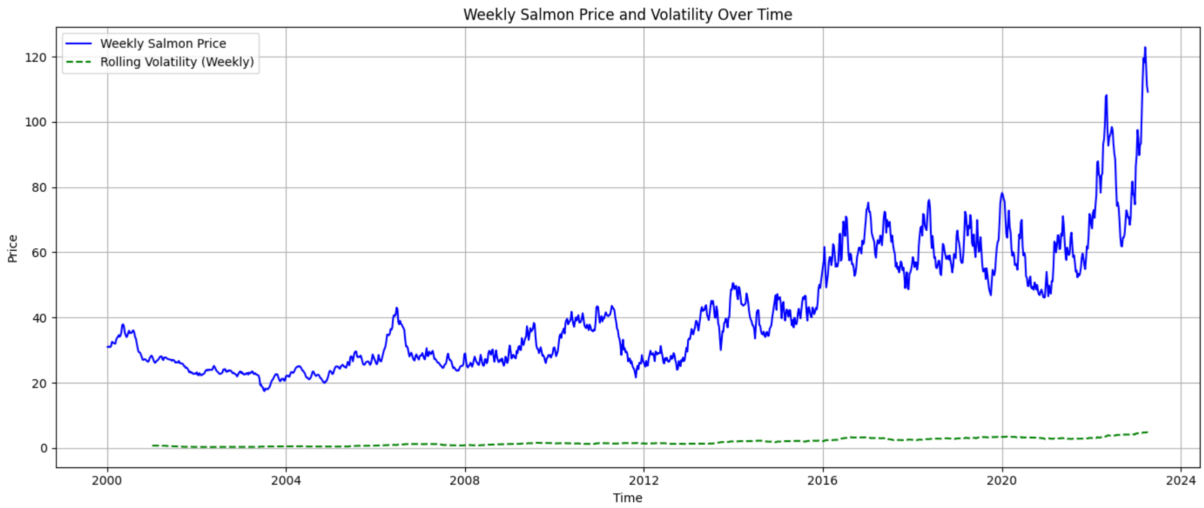

Figure 1: Salmon Price and Volatility over time

2. Theory

The cost of producing farmed salmon has increased the last years from an all-time low in 2001 (Asche and Roll, 2013). In the decades before there was a substantial decrease in unit production costs due to research and development and increased productivity. After this period, the price of salmon feed has increased faster than improvements in productivity leading to an overall increase in unit production costs. Analysis of salmon prices find indications that salmon prices have become largely cost driven (Asche and Oglend, 2016). The prices for the raw feed material very a large determinator of the production and sales price of farmed salmon. This support the hypothesis that the salmon market has matured into an industry where costs are more related to input-factors and less to productivity of production. This relationship was largest when salmon prices were denoted in US dollars which is the currency most used in the commodities markets for the raw feed material, but it was also present when looking at the salmon price in Norwegian kroner and in Euro. Asche and Oglend (2016) found the correlation between prices of the input factors soybean meal, fishmeal, and wheat were increasing throughout the years prior to the study. Their analysis discovered that salmon prices had a stronger relationship to protein-based input-factors and they suggest that this is due to a common link to certain market factors e.g., aggregate protein demand. The increasingly globally integrated commodity market has to a certain extent led to prices being more stable as they are less influenced by local conditions and the stability of the market has bade it easier for producers to switch between different input-factors. This has resulted in an increased observable pass-through effect from input-factors to unit production cost (Asche and Oglend, 2016).

The feed used in salmon farming has traditionally consisted of a combination of protein, which has usually been in the form of maritime protein from fish meal, carbohydrates, fat in form of fish oils and vegetable oils and micronutrients and pigment (Skretting, 2011). Even though fishmeal contains certain micronutrients which vegetable protein does not contain there have been efforts to reduce dependence on marine raw materials in salmon feed due to, among other things, increasing prices and worries regarding the sustainability of the marine raw materials (Asche and Oglend, 2016). The use of fishmeal in salmon feed has seen a substantial drop in recent decades. In the start of the 1990s it was normal that it accounted for around 50 % of the salmon feed while in 2008 Tacon and Metian found it to be as low as 15 % of the feed (2008). The total supply of fishmeal has remained relatively consistent globally, however the demand for marine protein has increased leading to increased competition and prices (Tveterås and Tveterås, 2010) and has driven aquaculture to use vegetable proteins to a larger extent.

As the world changes, the relationships between different commodities change as well. An example of this is how the relationship between corn, soybean and energy prices has gotten stronger as biofuel has been used to a larger extent (Nazlioglu and Soytas, 2011). At the same time many different food industries rely on the same input-factor such as soybean-meal, wheat, fishmeal etc. The aggregate demand for food can also increase when one food production industry uses another food in order to create its product leading to the price of both increasing and all of this can lead to increased relationship between various commodities (Asche and Oglend, 2016). This results in prices being correlated on the basis of causes outside the specific supply chain. Price correlation on the basis of a common factor depend on the pass-through rate of the input-factor prices to unit production costs. In the case of increased aggregate demand, a strong pass-through rate will increase the price both indirectly through increased unit production costs due to input-factors increasing in price, and directly through increased demand. This will result in less incentive to increase production as margins in the output industry remain relatively stable. Salmon's relationship with other commodities is established through both food substitution and input factors, as feed for salmon aquaculture constitutes the most significant portion of production expenses.

It is not only the price of salmon that has increased, but also the volatility. Salmon price volatility more than doubled from 2009 to 2019 (Asche, Misund and Oglend, 2019). Before this it was fairly stable for being a commodity. Oglend (2013) argued that the increase in volatility was related to food prices relevant to salmon, like oils, meats, cereals and fishmeal. Asche and colleagues (2019) argued that the main reason for this increase in volatility were the short-run elasticity of supply and support the hypothesis by making three observations. The first is that there has been integration of the upstream supply chain into fewer entities of larger size. The second is the importance placed on fixed harvest schedules in order to cater to the consumers' desire for stability. And the last reason is the combination of restrictions limiting new production capacity and the promotion of "race to raise" (which they define as" a fear of not being able to capitalize on exceptionally, in an historical context, high prices that are not expected to persist") harvest policies (Asche, Misund and Oglend, 2019).

At the same time, Oglend and Sikveland (2008) found that an increase in volatility was correlated with higher prices, meaning that periods of greater expected profits pay the price of greater risk. This is in line with previous theories on the behavior of commodity prices that argues that there should be a positive relationship between price and volatility (Chambers and Bailey, 1996; Pindyck, 2001). It is believed that when supply is limited both price and volatility will increase in competitive markets for storable commodities. Scarcity implies that there are insufficient reserves to mitigate fluctuations in demand. This situation can lead to speculative stock shortages or a substantial convenience yield on the remaining inventory. Volatility is also know to vary throughout the year and it reaches its peak prior to the harvest or production period (Peterson and Tomek, 2005). This is a trend seen in seasonally produced commodities like for example wheat and corn. The increasing price and volatility of the salmon market also brings with it increasing uncertainty and risk for the industry, consumers, and market participants. This further contributes to the importance of using forecasting models attempting to reduce this uncertainty and risk for all involved.

3. Data

3.1. Variables

3.1.1. Salmon prices

Salmon prices are gathered from the stats bank of the Norwegian center for statistics (*statistisk sentralbyrå*) (SSB). It is weekly data of the price per kilo of exported fresh salmon.

The data goes all the way back to the start of the year 2000. In the univariate analysis such as the ARIMA I will make use of the weekly salmon prices alone while for the multivariate analysis I will use the monthly salmon prices in addition to the other variables I will introduce below. Except for the exchange rate, all of these other factors are provided on a monthly data. The reason for using the weekly data in the univariate analysis and monthly in the multivariate analysis is that using the most granular data will give me more datapoints and possibly a more accurate result. However as only the salmon price and the exchange rate are available on a weekly level, I transform these into monthly data for the multivariate analysis by using whichever data point that is closest to the middle of the month.

3.1.2. Fishmeal

Fishmeal is a commodity made from reduction fisheries where fish is dried and grounded. Peru is the largest producer and exporter of fishmeal followed by Chile, while in Europe, Norway and Denmark are amongst the biggest exporters. The fish used to make fishmeal stems from fisheries of small pelagic fish, which is migrating fish that inhabit the surface layers of the sea. Pelagic fish is also used for human consumption. The amount that goes to human consumption is relatively stable and it is the surplus fish that is used for fishmeal. The amount left after human consumption can vary greatly, partly because of the El Niño event, and the supply of fishmeal will therefore vary accordingly (Asche and Oglend, 2016). In addition to this, there is little room for expansion as the Food and Agriculture Organization (FAO) categorize pelagic fisheries as fully or over exploited (Food and Agriculture Organization, 2022). Most fishmeal is used by aquaculture, but other industries like pig and poultry farming also consume a substantial part. The prices of the fish meal is gathered from the “pink sheet” of the World Bank. It is monthly data on the prices of exported Peruvian (German and Danish) fish meal (65% protein) (FOB).

3.1.3. Soybean meal

Soybean meal is a vital commodity in many food industries, including the aquaculture industry Soybean meal is serving as one of the primary ingredients in the formulation of salmon feed. Soybean meal is a by-product of soybean oil extraction and is rich in both protein and essential amino acids, which makes it a desirable ingredient for aquafeeds. Brazil and the United States has historically been the largest soy producers and it typically contains about 44-48% crude protein. As the industry seeks alternatives to fish-based feed ingredients, soybean meal has emerged as an important protein source in the diet of farmed salmon. Aas and colleagues (2019) found salmon feed for Norwegian salmon to consist of 19% soy.

Partly because of the extensive use in animal feed, the global demand of soybean meal has increased substantially. As a result of this the price of soybean meal is influenced by factors such as demand from livestock industry, global soybean production and trade policies and geopolitical events. At the same time due to the demand for soybean meal for animal feed, the price of soybean meal can influence the price of animal products. The data gathered on soybean meal prices comes from the World Bank pink sheet and is Brazilian soybean meal consisting of 48 % protein (CIF).

3.1.4. Wheat

Wheat is an essential commodity and is used in various different foods and animal feeds. It serves as a crucial foundation for a multitude of food products such as bread, pasta, cereals and beverages like beer. Recently the wheat market has been disrupted by its biggest producer Ukraine being invaded by Russia. This caused a two-thirds increase in price from the third quartal of 2021 to the second quartal of 2022 (Arndt *et al.*, 2023).

The use of wheat in aquaculture has increased and it has become one of the principle ingredients of salmon feed. The wheat content in Norwegian salmon feed was found in a study to be 8.9 % of wheat carbohydrates and 9 % wheat protein in the form of wheat gluten adding up to a total of 17.9 % (Aas, Ytrestøyl and Åsgård, 2019). The wheat price data used in this paper is gathered from the Pink Sheet of the World Bank and it is the monthly export price for Canadian Wheat.

3.1.5. Shrimp

55% of all shrimp consumed globally is farmed according to World Wild Life (WWF) (WWF.org, 2023) and farmed shrimp consume some of the same food as farmed salmon does, like fishmeal and soybean meal. Although the data used in the prediction models in this paper stems from the prices of wild caught shrimp, it is natural to assume that the price of farmed shrimp will influence that of wild caught shrimp. Shrimp is like salmon also a source of marine protein, and one can imagine that there is a degree of substitution effect between the two. The shrimp prices used in this study comes also from the World Bank pink sheet. It is data for wholesale of frozen headless American shrimp (26 to 30 count per pound).

3.1.6. Nok-Euro exchange rate

Aside from the commodities discussed above, currency exchange rates are likely to influence the salmon price. As mentioned, the vast majority of Norwegian salmon is being exported. The farmers home currency is NOK, and a lot of costs related to labor, equipment, facilities etc., will be paid in this currency. Europe is the biggest market for Norwegian salmon (with an export share of about 60%) meaning that most exports are invoiced in Euros, and a lot of consumers face retail prices denoted in Euros. Exposure to the U.S. dollars can also be relevant as the commodities used for salmon feed are often traded in U.S dollar and outside of Europe most exports are invoiced in U.S. dollars (Straume, 2014). A study from 2016 found the EUR/NOK exchange rate to be an economically significant determinant of the price of Norwegian salmon between 2007 and 2013 (Bloznelis, 2016).

The data for the exchange rate was gathered from the International Monetary Fund (IMF) and is the SDR per Norwegian Kroners and Euros for each weekday in the period 01.01.2000 to 31.07.2023. The exchange rate data is transformed into weekly data by taking the last price of each week (friday) and is transferred into monthly frequency by taking the day that is closest to the 15th in order to match the salmon prices.

3.2. Dataset

The total dataset consists of 10 different variables. Two of these are datetime variables. One for dates on a weekly interval and one for dates on a monthly interval. There are three weekly variables in total, the time data, salmon price data and NOK to EURO currency exchange data and each of these variables have a total of 1214 entries with weekly data from 7th January 2000 until 7th July 2023. All the other variable (fishmeal prices, soybean meal prices, wheat prices and shrimp prices) are on a monthly interval and salmon prices and NOK to EURO exchange rate have been transformed into monthly data for the multivariate analyses. This means that there are in total 7 monthly variables with 279 entries each from 15th January 2000 until 15th July 2023.

Table 1: Dataset

Variable	Data type	Entries
‘week’	datetime64[ns]	1214
‘salmon_weekly_price’	float64	1214
‘nok_eur_rate’	float64	1214
‘month’	datetime64	279
‘salmon_monthly_price’	float64	279
‘monthly_nok_eur_rate’	float64	279
‘fishmeal’	float64	279
‘soybean_meal’	float64	279
‘shrimps’	float64	279
‘wheat’	float64	279

4. Method

There are two common ways in which model selection within time series forecasting usually is done. One is that one may pic the model with the lowest AIC and the other is that use timeseries cross-validation (Hyndman and Athanasopoulos, 2014) and select the model with the lowest root squared mean forecast error (RMSE) (Bloznelis, 2018). A paper from 2016 discussed best practices for prediction of salmon prices (Bloznelis, 2016). In this paper he concludes that cross-validation is best for forecasting models using regularization-based multivariate time series models, k-nearest neighbors method and artificial neural networks, and that using AIC is preferable for most other models like the ARIMA. I will in this paper use the model selection practices he suggested.

4.1. Seasonal adjustment

As discussed in the previous sections, the salmon farming as an industry exhibit seasonal characteristics on both supply and demand level. The seasonality of supply and demand does not overlap and show different yearly patterns which in turn create seasonal patterns in price. Both the weekly and monthly salmon price data is turned into seasonally adjusted data by performing a seasonal decomposition with a multiplicative. The decomposition was performed using the **seasonal_decompose** function from the **statsmodels.tsa.seasonal** module in python. The period was set to 52 weeks as there are a fractional number 52.1429 weeks in a year. I chose to use the multiplicative model for seasonal decomposition as when prices of this series increased volatility also increased, and this would fit the multiplicative model better than the additive. This model can be represented as:$Y_t = T_t * S_t * R_t$ where:

- Y_t is the original series at time *t*.
- T_t is the trend-cycle component at time *t*.
- S_t is the seasonal component at time *t*.
- R_t is the residual at time *t*.

After using this model, the seasonally adjusted time series can be calculated by dividing the original series on the seasonal component.

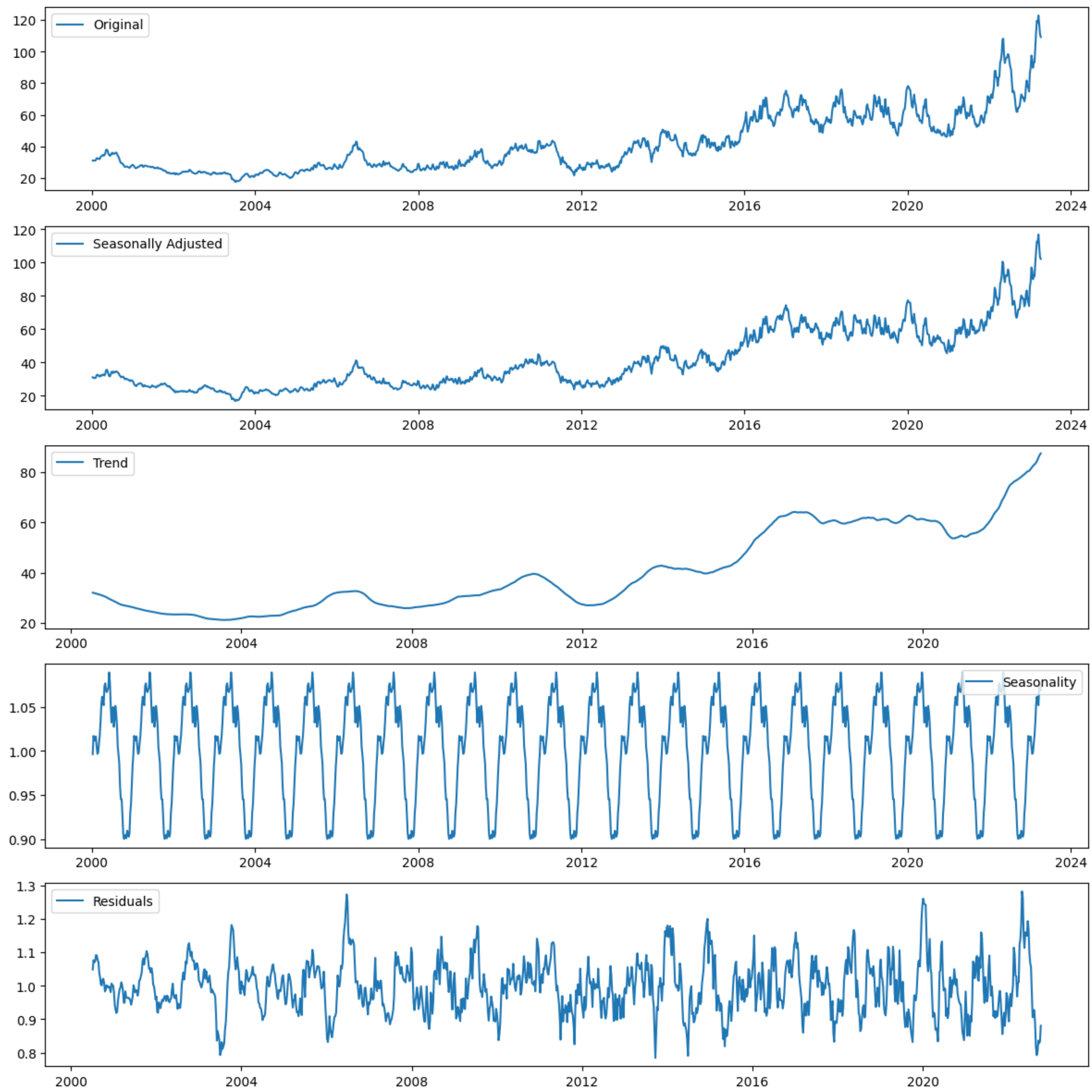

Figure 2: Decomposition data for salmon price

4.2. Transformation

In order to deal with the fact that the data appears to have non-normal and heteroskedastic residuals I transform the non-stationary data to make it stationary. As the timeseries has an increasing variance logarithmic or square root transformation might be best suited. These are often used to stabilize variance, especially when there's a multiplicative or exponential growth pattern (Salles *et al.*, 2019). After a statistical analysis of the different transformations. The model with the logarithmic transformation had the lowest (most negative) AIC, which would suggest the best fit among the two, although it had a couple non-significant coefficients which could be of concern. However, as this is a prediction model I chose to go with the log transformed timeseries which seems to provide the best fit to the data even though is shows some insignificant coefficients. The square-root transformed model with its significant coefficients might be better suited for an explanatory approach.

4.3. Models

4.3.1. Naïve no change

A naïve no-change forecast is a forecast that predicts that the value of t periods ahead will be the same as the previous observed value (t-1). This can be showed mathematically as:

$$\hat{Y}_{t+1} = Y_t$$

With $\hat{Y}_{t+1}$ representing the forecasted value for the coming period, and Y_t is the most recent observed value.

This will serve as a natural benchmark to test the models against especially as naïve no-change forecasts are known to often outperform prediction models, particularly at the short term, in financial studies (Kilian and Taylor, 2021).

4.3.2. ARIMA

ARIMA is an abbreviation for autoregressive integrated moving average and is a statistical model that estimates using maximum likelihood estimation (MLE). MLE identifies which parameters will make the observed data most likely to occur, given a particular model which is done by maximizing the likelihood function which is a measure of how well the model explains the observed data. The ARIMA model consists of 3 main parts: autoregressive (AR), differencing (I), and moving average (MA). The AR part is capturing the linear dependence between a lagged value of a time series and an observation:

$$AR(p): X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \epsilon_t$$

Where:

- X_t is the time series at time t.
- c is a constant.
- ϕ_1 are the autoregressive coefficients.
- p is the number of lags.

The next part is the differencing part (I) which reprecent the order of differencing used to make the time series stationary:

$$I(d): Y_t = X_t - X_{t-d}$$

Where Y_t is the differenced time series after d differences.

The last part is the MA component which measures the relationship between an observation and a residual error from previous a time point:

$$MA(q): X_t = \mu + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \cdots + \theta_q\epsilon_{t-q}$$

Where:

- μ is the mean of the time series.
- $\epsilon\epsilon_t$ is the residual error at time t.
- θ_i are the moving average coefficients.
- q is the number of lagged residuals.

When performing an ARIMA it is first necessary to check for stationarity, meaning that statistical properties like the variance, mean and autocorrelation remains consistent over time. Stationary data are beneficial as they facilitate a more accurate modeling and forecasting when using ARIMA. Forecasts will be modeled using the logarithmically transformed seasonally adjusted data. This is used as the salmon price data show large increases in absolute value and standard deviation over time. The use of logarithmic transformation will address some of these challenges by moderating the effects of increasing variance (heteroskedasticity) by stabilizing it, transforming any exponential growth pattern to a linear one and making multiplicative seasonal effects and shocks become additive. The logarithmically transformed seasonally adjusted data exhibit skewness and excess kurtosis closer to a normally distributed random variable, however the data still do not perfectly follow a normal distribution and retain signs of heteroskedasticity.

The autocorrelation of the seasonally adjusted and transformed time series showed a strong positive autocorrelation with significant results in the first 3 lags whereas and the partial autocorrelation showed an initial strong spike in the first lag followed by a less strong but significant spike in the next two lags also. The results of the autocorrelation and partial autocorrelation plots serve as a good starting point for the p and q terms (term) of the ARIMA model. The time series was further checked for stationarity with an Augmented Dickey-Fuller

(ADF) test which comes out non-significant meaning that the timeseries appears to be non-stationary. After a first order differencing the ADF test is significant and indicate stationarity. This indicates that (3,1,3) is a good starting point for this ARIMA model.

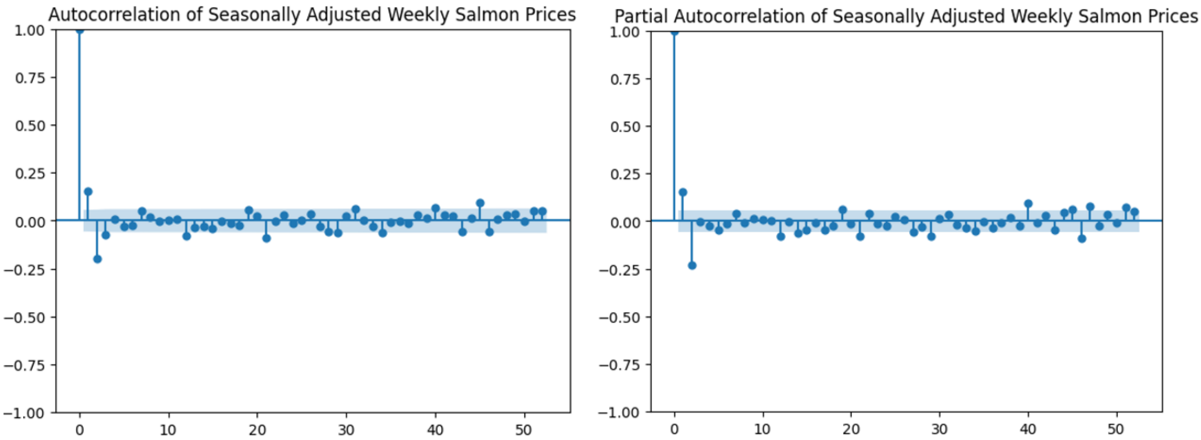

Figure 3: Autocorrelation (AFC) and partial autocorrelation (PAFC) of seasonally adjusted salmon price

While finetuning the ARIMA model several metrics was used in order to evaluate the performance. These were Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) and Hit Rate. As the main focus was to achieve accurate predictions, most weight was given to MAE, MSE, RMSE, MAPE, and Hit Rate. No further optimizations were found and the model used for predictions was a (3,1,3) logarithmically transformed and seasonally adjusted ARIMA.

4.3.3. LSTM

Long Short-Term Memory (LSTM) networks are a special kind of recurrent neural network (RNN) architecture designed for sequence prediction problems (Yu *et al.*, 2019). One thing that separates LSTM from RNNs is that LSTM networks has mechanisms built is that that control how information is memorized or abandoned throughout time (Li and Becker, 2021) The LSTM has memory cells that are able to maintain information over longer time periods that can mitigate vanishing gradient problem associated with traditional RNNs .This can make them particularly useful for time series forecasting, natural language processing, and modelling of other sequential data.

At the center of this are three primary gates: an input gate (i_t), a forget gate (f_t), and an output gate (o_t). These gates control the flow of information in the model and ensure that relevant information is retained over time. The forget gate determines the amount of earlier information that gets discarded from the memory cell (c_t), the input gate determines the amount of new information to get stored in the memory cell, and the output gate specifies the amount of information from the memory cell to output to the hidden state (h_t). The memory cell is a component that holds the networks internal state. This can be shown as:

$$i_t = \zeta(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i)$$

$$f_t = \zeta(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f)$$

$$o_t = \zeta(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_o)$$

$$c_t = f_t \otimes c_{t-1} \otimes Tanh\ (W_{xc}x_t + W_{hc}h_{t-1} + b_c)$$

$$h_t = o_t \otimes (c_t)$$

ζ represent the sigmoid function ensuring values between 0 and 1, allowing the gates to perform binary-like operations and Tanh represent the hyperbolic tangent function, scaling values between -1 and 1, primarily used for the memory cell to keep the network activations normalized. $\otimes$ is denoting the element-wise product, which allows, based on the gate activations, for the selective update or retention of information. This can be visualized like below in figure 4:

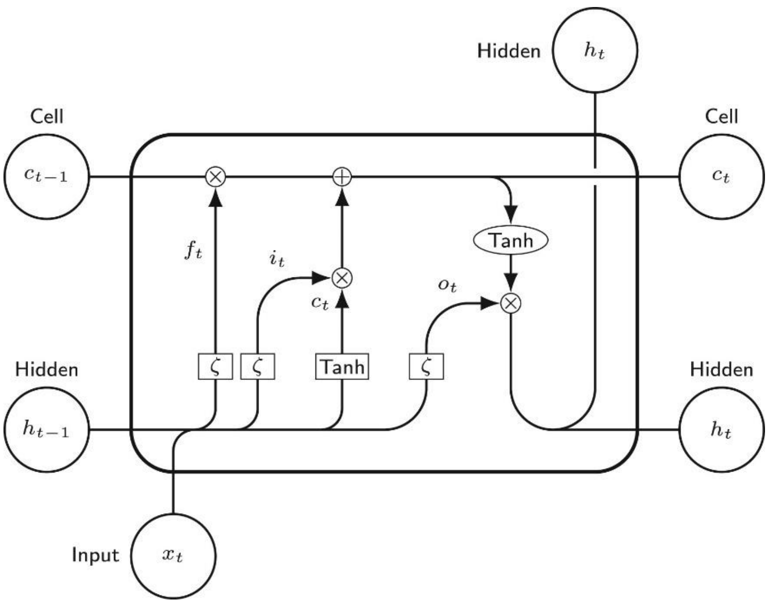

Figure 4: Architecture of a LSTM cell (from Li and Becker, 2021)

Data Preprocessing, architecture and model evaluation

The dataset was first cleaned by dropping missing values. The variables used were salmon price, NOK-EUR exchange rate, fishmeal price, soybean meal price, shrimp price, and wheat price. The features were then scaled by the use of **`MinMaxScaler`** from the scikit-learn library, adjusting all the selected features to have values between 0 and 1. The data was organized into chunks, each covering a specific time duration. And for each chunk, I used its data to predict the salmon price at the end of that period.

The LSTM model used is a sequential model consisting of:

- An initial LSTM layer with 250 units, set to return sequences.
- A second LSTM layer with 250 units.
- A dense output layer with a single unit (representing the predicted price).

Multiple model architectures were investigated such as multiple layers, extra dense layers, more or less units, dropout layers, or combinations of these, however this architecture were the most efficient at predicting the salmon prices. The model uses the Adam optimizer and is trained using mean squared error (MSE) as a loss function. Adam was used as it is considered to perform well with less memory requirement (Kingma and Ba, 2015) which was essential due to hardware limitations.

In order to examine the model's performance and generalization ability, I employed time series cross-validation. First the dataset was split into several non-overlapping and consecutive subsets, using earlier data points for training and later data points for testing. This was done using the **TimeSeriesSplit** method from the scikit-learn library, making a set of train and test indices for each split the model was trained on the training set, made predictions on the test set and various metrics like RMSE, MAE, MASE, MAPE, and hit rate were calculated. After iterating through the all the **TimeSeriesSplit**, an aggregated result was calculated, presenting the model's overall performance on all the test data. In addition to this, the LSTM model was finally trained on all available data except the very last split consisting of the last 20% percent of the salmon price time series. It was then evaluated on this last test set in order to evaluate the model's predictive capabilities and, providing insights into its likely future performance.

4.3.4. *K-Nearest Neighbors*

K-nearest neighbors (KNN) is a nonparametric pattern recognition technique used for regression and classification (Bloznelis, 2018). KNN is based on the logic that patterns of sequences and events repeat over time. If we have encountered a sequence e.g., 1,2,3,4,5 many times before and we see 1,2,3,4 we can often expect that 5 will follow. When performing a KNN the time series are transformed into into a sequence of vectors using a sliding window approach. In this way KNN can notice patterns in the multivariate dataset used in this paper. The similarity between two sequences is measured by a distance function, in my case using Euclidian distance, and more similar sequences are considered closer and more dissimilar sequences are further apart. For each data point x with p with features, $x = (x_1, x_2, \dots, x_p)$ and an additional data point, $y = (y_1, y_2, \dots, y_p)$ the Euclidean distance d between these two points is given by the formula (Peterson, 2009):

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2}$$

In the context of regression analysis, KNN predicts the output of a new data point based on the average of its k nearest neighbors from the training set. The output $\hat{y}$ for an input point x can expressed as:

$$\hat{y}(x) = \frac{1}{k} \sum_{i \in N_k(x)} y_i$$

Where $N_k(x)$ is the is the set of k data points that are nearest to x

For the use in this paper the variables were the same as in the LSTM mode: price, NOK-EUR exchange rate, fishmeal price, soybean meal price, shrimp price, and wheat price. To determine the optimal value of k, a grid search approach with 10-fold cross-validation was used. The grid search was set up to span a pre-defined range of k values, and the scoring metric used to evaluate performance was the mean squared error (MSE). The optimal value of k was found to be $k = 10$.

To allow for more memory in the model I incorporated lagged versions of all the variables. I do this by allowing the lag to be between 2 and 9 and then determining the optimal lag length through cross-validation. The model was iteratively trained on a growing amount of the training data and validated on the subsequent few data points, which is sometimes referred to as "rolling" cross-validation. At each lag length the RMSE was calculated across all the validation sets. For each variable, the lag that gave the lowest RMSE was used for the KNN model to make predictions.

Table 2: Optimal lag for variables in KNN model

Variable	Optimal lag	RMSE
Salmon price	9	16.57
Fishmeal	9	18.45
Soybean meal	1	19.47
Shrimp price	1	19.33
Wheat	1	19.51
NOK-EUR exchange rate	1	19.87

4.3.5. Performance Assessment

It is important to address how to measure and compare these models. One can evaluate the models used based on the economic results of using these in the market and based on a purely statistical perspective. In this paper I focus on the statistical results and just comment on the potential economic performance of these models. From a statistical point of view there are various ways of considering the performance of the models. One can for example see the forecast accuracy related to the average realized value of the actual time series, accuracy related to a benchmark by a naïve model, forecast accuracy related to other forecasts, or by examining the predictability of the forecast errors (Bloznelis, 2018). In this paper I will stick to the first two mentioned above. Traditional measures of forecast accuracy includes measures like mean absolute error (MAE), root mean squared error (RMSE), mean absolute scaled error (MASE), mean absolute percentage error (MAPE) and hit rate and these will be introduced in the coming section.

4.3.6. Measures of Forecast Accuracy

Different forecasting accuracy measures can lead to different conclusions and it is therefore important to be conscious of why one is using the forecasting measure one is using. There are no general consensus within the field of which accuracy measures are best and each of them has their own strengths and weaknesses (Bloznelis, 2018). It can be beneficial to report various accuracy measures as they report on different loss functions and employ different methodology which can give a more rounded picture of the forecasting models used, especially when combining different forecasting methods (Chen, Twycross and Garibaldi, 2017).

Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) is a measure of the average magnitude of the errors between the predicted values and the actual values. This model does not consider the direction of these errors. MAE represents the average errors in the same units as the data used and smaller errors indicate better predictions.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

y_i represent the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations.

Root Mean Squared Error (RMSE)

The Root Mean Squared Error (RMSE) is a measure that evaluates the square root of the average of squared differences between forecasted values and actual values. It is often used to examine the difference between the values that a model predicts and the actual values. As with the MAE, lower values indicate a better fit. This model is more sensitive to outliers than MAE as is squares the errors.

$$RSME = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Mean Absolute Scaled Error (MASE)

The Mean Absolute Scaled Error (MASE) is a measure of using the MAE from a naïve forecast method to scale the MAE. This can be especially useful when evaluating models on datasets with varying degrees of volatility as the naïve forecast might produce large errors, especially if the time series undergoes abrupt changes. MASE uses these errors to scale the errors of the model of interest, and accounts therefore for the inherent difficulty in predicting such datasets. A MASE of 1 indicates that the model is performing as well as a naïve forecast, values below indicate better performance, and values above indicate worse.

$$MASE = \frac{MAE}{\frac{1}{n-1}\sum_{i=2}^{n}|y_i - \hat{y}_{i-1}|}$$

Mean Absolute Percentage Error (MAPE)

The Mean Absolute Percentage Error (MAPE) presents forecasts errors as percentage which makes this a scale independent accuracy measure. An advantage of MAPE is that it is easy to understand as it gives the error in terms of percentage. One disadvantage is that it is undefined for $y_i = 0$ and can be misleading as errors when y_i is small can make big impacts on the measurement.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| * 100$$

Hit Rate

Hit rate is often used in classification problems. It measures the proportion of correct predictions in the total predictions made. A higher hit rate indicates better model performance. It's particularly useful when the actual "hit" or "correct prediction" has significant implications. In this case, a high hit rate, meaning that the model predicts direction significantly better than chance, this could be an indicator of a model with good economic value.

$$Hit\ Rate = \frac{Number\ of\ Correct\ Predictions}{Total\ number\ of\ Predictions}$$

5. Results and discussion

5.1. Naïve no-change forecast

The time series representing weekly salmon prices were analyzed with the naïve no-change forecast as described above. This approach which assumes that future prices will remain consistent with the most recently observed prices were compared against the actual data and these were the results:

MAE was found to be 2.77, the MSE was 12.88, the RMSE was 3.59, The MAPE was 4.09% and the MASE was 1.13. Hit rate was calculated to be 56.84 % which in a naïve forecast effectively is a measure of how often direction of change remains consistent from one observation to the next. This will serve as a benchmark and will be compared with the more advanced modeling techniques further down.

Table 3: Accuracy measures for naïve no-change model

Accuracy measure	Value
MAE	2.77
RMSE	3.59
MAPE	4.09%
MASE	1.13
Hit Rate	56.84%

5.2. ARIMA

The forecasted values from this ARIMA were first transformed back to the original scale using inverse transformation in order compare the results to the results of the other forecasts. The ARIMA (3,1,3) was selected based on a iterative process of fitting and evaluating orders as mentioned above, and was proved to be the best fit that was found for this model and this model was used to predict the values of the salmon price test set. These were the results: MAE was found to be 10.07, RMSE was found to be 13.41, MAPE was 14.52 and MASE was 1.74 which is a large increase from the naïve no change forecast indicating this model is doing worse at predicting the future values than the naïve no-change forecast. Hit rate was also lower and worse than 50% indicating that this model was worse than chance as predicting direction of change.

Table 4: Accuracy measures for ARIMA

Accuracy measure	Value
MAE	10.07
RMSE	13.41
MAPE	14.52%
MASE	1.74
Hit Rate	46.26%

.

Differencing and logarithmic transformation were important in processing the time series used for this model, however there were some indications of non-normality and hetroskedacity even after these adjustments. Some of the reasons why this model seems to not be performing as well as the naïve forecast can be that the seasonality adjustments and transformations were not sophisticated enough. One possible solution to this could be to employ more sophisticated seasonality adjustments like Bloznelis (2018) with using a regression with ARMA errors, with Fourier terms as regressors. This method might be more suited at handling the fractional number of weeks in a year, and combining this with dummy variables for the holidays like Christmas and easter where we see movements in salmon price, this could better measure the seasonality of the series. It is also possible to use Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models to account for the heteroskedasticity observed in the residuals of the model. GARCH models are particularly useful for modeling financial time series where volatility changes over time. It captures the time-varying volatility which can be an essential aspect in forecasting salmon prices.

5.3. KNN

The KNN was as mentioned above trained with $k = 10$ and up to 9 lags for the variables. This resulted in a model with the following results: MAE were 11.67, RMSE were 16.69, MASE were 0.98 and MAPE were 15.65%. Except for the MASE, this is higher than the other models indicating a worse overall prediction. The KNN however has a hit rate of 61.11% which is the highest of the models. This is an indication that even though the model is not making accurate predictions it is better at predicting the direction of change. This might be due to KNN models tending to capture trends and patterns without necessarily accurately predicting exact values.

Table 5: Accuracy measures for k-nearest neighbors

Accuracy measure	Value
MAE	11.67
RMSE	16.69
MAPE	15.65%
MASE	0.98
Hit Rate	61.11%

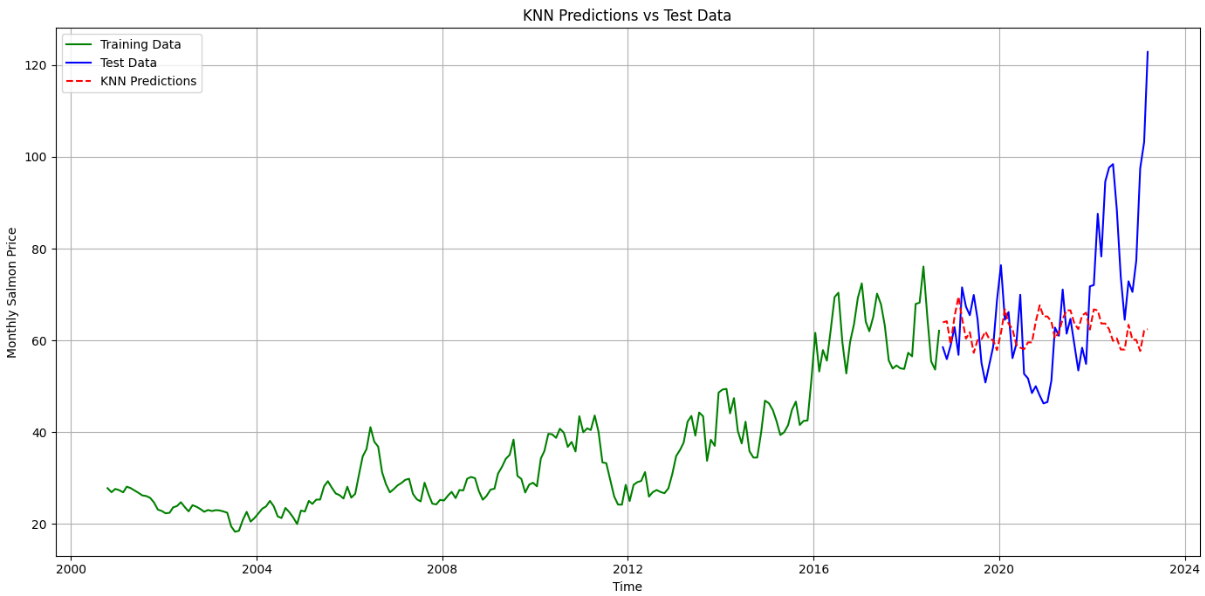

Figure 5: k-nearest neighbors prediction

5.4. LSTM

The results of the LSTM are presented both for the aggregated result of the LSTMs cross-validation process and for performance on the test set. The first table below shows the aggregated results and one can see that even here which includes the results for the sets used for training, the model performs worse than the naïve forecast on all metrics.

Table 6: Accuracy measures for aggregated LSTM model

Accuracy measure	Value
MAE	7.06
RMSE	9.33
MAPE	15.56%
MASE	1.81
Hit Rate	49.3%

The table below shows the result of the LSTM on the test set. One can see here that it scores quite a bit worse than the aggregated results on the measures MAE and RMSE.

Table 7: Accuracy measures for LSTM model on test set

Accuracy measure	Value
MAE	10.60
RMSE	13.73
MAPE	14.46%
MASE	1.45
Hit Rate	41.86%

Although this might be expected, this difference between the aggregated results of the LSTMs cross-validation process and the results on the test set might indicate that the model was overfitting. However, regularization techniques such as dropout layers in the LSTM were applied with worse all over results. Gathering more data to have a larger dataset to train the model on could be a solution that could potentially alleviate this. The increase of the RMSE and MAE metrics in the test set indicates that the test set might include new patterns or volatilities that the model wasn't trained on. When looking at the plot of the salmon price above, one can see the increase of volatility price and volatility and it is evident from the plot below that the model underpredicts the amount of volatility in the end.

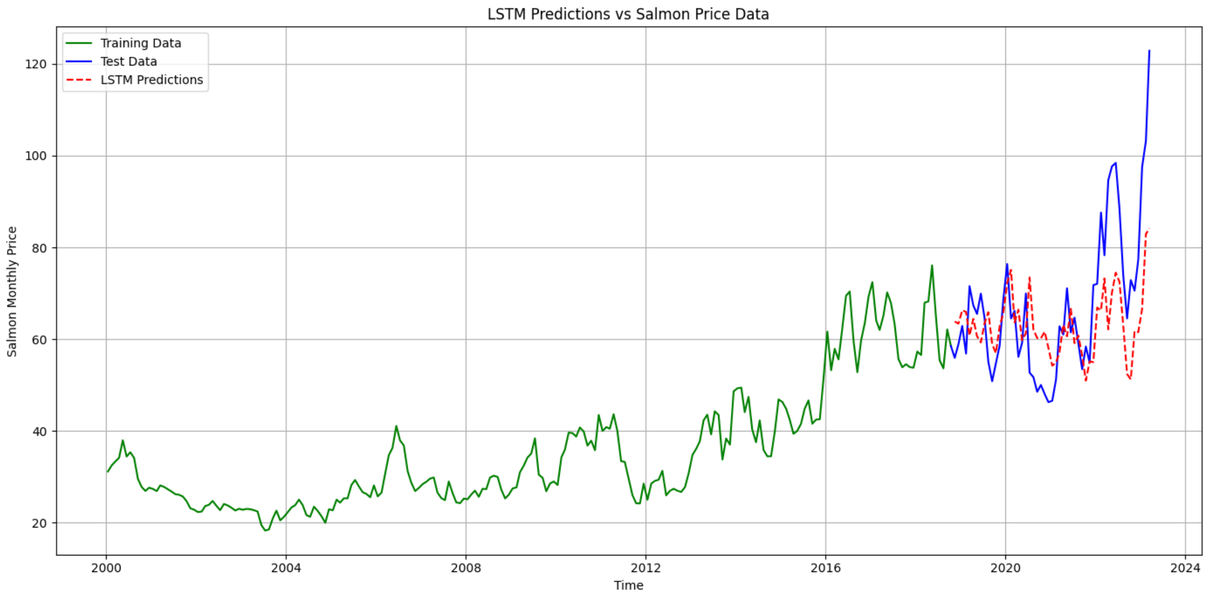

Figure 6: LSTM prediction

The prediction of the LSTM seems to be lagging a bit after the test set and my hypothesis for why is that the, as a lot of financial data, this time series has autoregressive properties. The model might follow a logic thinking that a strong way to predict the next value is that it will be similar to the current one.

5.5. Economic value of forecasts

With the accuracy measures presented above for the different models, there is little to suggest that these models would provide value in the market. The naïve no-change forecast consistently performed better than all the other models. The one exception to this was the KNN model which was able to show a hit rate of 61.11 %. Hit rate is important as a predicting method that consistently predicts the direction of the salmon price will be able to provide significant economic gain for salmon farmers and traders. The hit rate of the KNN model means that there are indications that it might be better than chance at predicting the future directional change of the salmon price. However, this is only tested within this dataset and cannot say anything about its future performance as this might be due to chance, or the future behavior of the salmon price might be different to the behavior captured in the time series used for this analysis.

5.6. Limitations and Implications for Future Work

This analysis attempted to forecast salmon price predictions based on solely macroeconomic factors and commodities. This leaves out important factors related to salmon production like environmental factors (e.g., temperature), regulatory factors (e.g. maximally allowed biomass at a farming site), biological factors (e.g. disease and sexual maturation), time of year (does it? seasonality) and share prices of salmon companies which there has been found evidence for to have a strong relationship with salmon prices.

The salmon future prices can also be seen as a price prediction from 1 month ahead all the way to 60 months ahead on Fish Pool and could have been valuable to include in this kind of analysis, however I was not able to use this data due to the costs of acquiring the historical salmon future prices from Fish Pool. However, this data could be useful in the models, and it is something worth exploring.

Short term financial timeseries forecasts rarely beats a naïve no-change forecasts due to the efficiencies of the markets (Bloznelis, 2018). While seasonality can be a dominant factor, in the short term it can be dominated but unpredictable shock and make it harder to predict. Bloznelis (2018) argues that targeting the seasonal component specifically might be one way towards increased accuracy.

6. Conclusion

I wanted to investigate whether or not I was able to predict the export prices of Norwegian salmon using related commodities and macroeconomic factors. I wanted to investigate if a multivariate approach based of LSTM and KNN is able to provide better forecasts than a univariate time series approach, and if either are able to beat a naïve no-change model. From the results we see that the model that consistently performs the best on the accuracy measures chosen is the naïve no-change model. With the one exception being that the K-nearest neighbors model has a higher hit rate than the naïve model, indicating that it might be better at predicting future directional change of the salmon price than the naïve model. A potential reason for the naïve model being superior is the statistical properties of the dataset. There were still indications of non-stationarity and heteroskedacity after seasonally adjusting and transforming the dataset.

It could be advisable in future studies to use more advanced seasonal adjustments with dummies for periods with changed statistical properties or models like GARCH which can deal with the heteroskedacity of the time series. Future studies might also use larger dataset as one limitation here is that the use of monthly data for commodity prices led to a dataset with relatively few (279) entries for each time series. This can have limited the prediction accuracy of the LSTM which has less data to train on and might be more prone to overfitting with a dataset this size. It is however not unusual that naïve models perform better than more advanced statistical model on predictions on financial data particularly at the short term (Kilian and Taylor, 2021). This is, according to Bloznelis (2018) due to market efficiency.

7. References

Aas, T.S., Ytrestøyl, T. and Åsgård, T. (2019) 'Utilization of feed resources in the production of Atlantic salmon (Salmo salar) in Norway: An update for 2016', *Aquaculture Reports*, 15. Available at: https://doi.org/10.1016/j.aqrep.2019.100216.

Arndt, C. *et al.* (2023) 'The Ukraine war and rising commodity prices: Implications for developing countries', *Global Food Security*, 36. Available at: https://doi.org/10.1016/j.gfs.2023.100680.

Asche, F. *et al.* (2009) 'The salmon disease crisis in Chile', *Marine Resource Economics*, 24(4). Available at: https://doi.org/10.1086/mre.24.4.42629664.

Asche, F., Misund, B. and Oglend, A. (2019) 'The case and cause of salmon price volatility', *Marine Resource Economics*, 34(1). Available at: https://doi.org/10.1086/701195.

Asche, F. and Oglend, A. (2016) 'The relationship between input-factor and output prices in commodity industries: The case of Norwegian salmon aquaculture', *Journal of Commodity Markets*, 1(1). Available at: https://doi.org/10.1016/j.jcomm.2015.11.001.

Asche, F. and Roll, K.H. (2013) 'DETERMINANTS OF INEFFICIENCY IN NORWEGIAN SALMON AQUACULTURE', *Aquaculture Economics and Management*, 17(3). Available at: https://doi.org/10.1080/13657305.2013.812154.

Bloznelis, D. (2016) 'Salmon price volatility: A weight-class-specific multivariate approach', *Aquaculture Economics and Management*, 20(1). Available at: https://doi.org/10.1080/13657305.2016.1124936.

Bloznelis, D. (2018) 'Short-term salmon price forecasting', *Journal of Forecasting*, 37(2). Available at: https://doi.org/10.1002/for.2482.

Chambers, M.J. and Bailey, R.E. (1996) 'A theory of commodity price fluctuations', *Journal of Political Economy*, 104(5). Available at: https://doi.org/10.1086/262047.

Chen, C., Twycross, J. and Garibaldi, J.M. (2017) 'A new accuracy measure based on bounded relative error for time series forecasting', *PLoS ONE*, 12(3). Available at: https://doi.org/10.1371/journal.pone.0174202.

Hyndman, R.J. and Athanasopoulos, G. (2014) 'Optimally Reconciling Forecasts in a Hierarchy', *Foresight: International Journal of Applied Forecasting* [Preprint], (35).

Iversen, A. *et al.* (2020) 'Production cost and competitiveness in major salmon farming countries 2003–2018', *Aquaculture*, 522. Available at: https://doi.org/10.1016/j.aquaculture.2020.735089.

Kilian, L. and Taylor, M.P. (2021) 'Why is it so Difficult to Beat the Random Walk Rorecast of Exchange Rates?', *SSRN Electronic Journal* [Preprint]. Available at: https://doi.org/10.2139/ssrn.356266.

Kingma, D.P. and Ba, J.L. (2015) ‘Adam: A method for stochastic optimization’, in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

Li, W. and Becker, D.M. (2021) ‘Day-ahead electricity price prediction applying hybrid models of LSTM-based deep learning methods and feature selection algorithms under consideration of market coupling’, *Energy*, 237. Available at: https://doi.org/10.1016/j.energy.2021.121543.

Marine Harvest ASA (2018) ‘Salmon Farming Industry Handbook’, *Processing* [Preprint].

Nazlioglu, S. and Soytas, U. (2011) ‘World oil prices and agricultural commodity prices: Evidence from an emerging market’, *Energy Economics*, 33(3). Available at: https://doi.org/10.1016/j.eneco.2010.11.012.

Oglend, A. (2013) ‘RECENT TRENDS IN SALMON PRICE VOLATILITY’, *Aquaculture Economics and Management*, 17(3). Available at: https://doi.org/10.1080/13657305.2013.812155.

Oglend, A. and Sikveland, M. (2008) ‘The behaviour of salmon price volatility’, *Marine Resource Economics*, 23(4). Available at: https://doi.org/10.1086/mre.23.4.42629677.

Peterson, H.H. and Tomek, W.G. (2005) ‘How much of commodity price behavior can a rational expectations storage model explain?’, *Agricultural Economics*, 33(3). Available at: https://doi.org/10.1111/j.1574-0864.2005.00068.x.

Peterson, L. (2009) ‘K-nearest neighbor’, *Scholarpedia*, 4(2). Available at: https://doi.org/10.4249/scholarpedia.1883.

Pindyck, R.S. (2001) ‘The dynamics of commodity spot and futures markets: A primer’, *Energy Journal*, 22(3). Available at: https://doi.org/10.5547/ISSN0195-6574-EJ-Vol22-No3-1.

Salles, R. *et al.* (2019) ‘Nonstationary time series transformation methods: An experimental review’, *Knowledge-Based Systems*, 164. Available at: https://doi.org/10.1016/j.knosys.2018.10.041.

Straume, H.M. (2014) ‘Currency invoicing in norwegian salmon export’, *Marine Resource Economics*, 29(4). Available at: https://doi.org/10.1086/678930.

Tacon, A.G.J. and Metian, M. (2008) ‘Global overview on the use of fish meal and fish oil in industrially compounded aquafeeds: Trends and future prospects’, *Aquaculture*, 285(1–4). Available at: https://doi.org/10.1016/j.aquaculture.2008.08.015.

Tveterås, S. and Tveterås, R. (2010) ‘The global competition for wild fish resources between livestock and aquaculture’, *Journal of Agricultural Economics*, 61(2). Available at: https://doi.org/10.1111/j.1477-9552.2010.00245.x.

Yu, Y. *et al*. (2019) ‘A review of recurrent neural networks: Lstm cells and network architectures’, *Neural Computation*. Available at: https://doi.org/10.1162/neco_a_01199.

NTNU

Kunnskap for en bedre verden