# Forecasting freight rates for tipper trucks in Russia

Presented by
Iryna Fedenko 2004060
Master of Business Analytics and Operations Research

Supervised by
prof. dr. Goos Kant


Tilburg University
School of Economics and Management
August 22, 2022

# 1 Abstract

Cargill's Profits and Losses from trading highly depend on its possibility to accurately estimate future transportation costs. Therefore, Cargill would like to have a data driven model to forecast the freight rates. In this study, we develop the methodology to forecast tipper truck rates in Russia. The thesis examines different input variables to find the best combination of predictors. Furthermore, we develop the methodology to convert delivery-based data on freight rate into a weekly time series target variable to represent the average weekly freight rate. Then, the study tests the performance of Multiple Linear Regression against the Autoregressive Integrated Moving Average model using Cross-Validation. The results suggest that ARIMA fails to predict the future freight rates, while MLR captures around 85% of variation in target variable. Lastly, we develop a methodology to create protection against uncertainty and external risks. Risk assessment is a process of evaluation of risks imposed by different scenarios, identification of the probability of occurrence of those scenarios, and the magnitude of their effect on the metrics of interest. The methodology should help to position with higher precision the value of the target variable within the prediction interval suggested by the model, given external risk factors.

## 2 Acknowledgements

First of all, I want to express my sincere gratitude to my thesis supervisor, prof. dr. Goos Kant. He supported me throughout the whole project with his immense knowledge and experience, motivated me to do the best work possible, and contributed with his ideas towards the improvement of the research. His clear guidance at every step of the project helped me to understand and approach the problem in the best way.

Furthermore, I am very grateful to my supervisors from Cargill - Rob Moen and Michel Roulet. Their extensive expert knowledge allowed me to deeply understand the problem and the specifics of the market in a short period of time. At the same time, their innovative ideas, passion, enthusiasm and very supportive attitude allowed me to enjoy the process of writing the thesis.

# Contents

# 3 Introduction

## 3.1 Problem definition

Cargill is an American multinational company, which provides a range of services in the food and energy sector. Among Cargill's major business activities are trading and transportation of raw food, energy, and metal commodities, manufacturing of food ingredients, harvesting crops and livestock, and even provision of financial services, such as risk management. Cargill Transportation and Logistics (CTL) team is an entity within Cargill, and its goal is to provide efficient and timely transportation of raw commodities and finished products across and within countries. Depending on the product transported and the locations of origin and destination, CTL charters a various range of transport, including cargo ships, rail transport, tippers, and refrigerator trucks. At the same time, transportation expenses account for around 5.5% of Cargill's revenues with approximately $6.1 billion of global annual spending across five modes, including $3.6 billion of expenses going to transportation by truck.

The problem this thesis is aiming to solve can be formulated as follows. Cargill often competes with other companies to purchase the grain commodities from farmers, and the bidder who gives the highest price for the grain will get the contract to purchase the grain. Normally, the contract also includes a promise to deliver grain to a certain location in the future; therefore, transportation cost has to be accounted for already when Cargill makes a price offer for grain to the farmer. If the future freight cost is underestimated, Cargill will bear losses for paying higher shipping rates than expected. If the transportation costs are overestimated, Cargill will bid a lower price to the farmer and will risk losing a contract, hence potential profit. In such a way, Cargill's Profits and Losses from these business activities highly depend on the possibility of accurately estimating future transportation costs. Right now, the prediction of the freight rate is based on expert knowledge of the freight market fluctuations. Experts in the field estimate the range of freight cost for a certain period and direction based on their experience and personal expectations about future trends. These estimations are not based on statistical data and are subject to human mistakes. Therefore, Cargill would like to have an accurate data-driven predictive model for the freight rates, and then have the rule to convert the results of this model into understandable for decision-makers metrics.

We narrow down the scope of the project to CASC (Cargill Agricultural Supply Chain) in Russia, and the transportation of grain by tipper trucks. Narrowing the scope is necessary, since freight fluctuations will be different depending on the type of transport, region, and even product transported. For example,

grain is transported by tipper carriers, and the tipper trucks for grain are not likely to be used for transportation of non-food bulk commodities such as coal and fertilizers due to sanitary requirements. Therefore, the goal of this project is to design an accurate model to forecast tipper freight rates in Russia while taking into account the uncertainty in the market. This thesis utilizes historical data on Cargill's transportation cost from January 2020 until December 2021 along with other exogenous variables to explain weekly fluctuations in the freight rate. The performance of the time series Multiple Linear Regression (MLR) model is compared to the one of Autoregressive Integrated Moving Averages (ARIMA).

The rest of the paper is structured in the following way: Sections 3.2-3.4 describe the specifics of the truck transportation market, and briefly analyze its current and future trends in Russia; Section 4 gives an overview of the existing literature on transportation rate forecasting; Section 5 evaluates different methodologies available for forecasting, and in short, introduces the methodology applied in this paper; Section 6 discusses data collection, data clean-up, and data analysis; Sections 7 describes MLR methodology and its assumptions; the results of MLR and ARIMA performance are presented in Sections 8 and 9 respectively; in section 10 of the thesis we examine the importance of data quality; in Section 11 we present the methodology for the risk protection, and finally, a short conclusion is derived in Section 12.

## 3.2 Trucking transportation market

### 3.2.1 For-hire and private carriers

We distinguish between for-hire and private carriers. For-hire carriers transport the products for third parties and are not involved in the production or processing of products they are transporting. Such carriers can be large entities operating a fleet of trucks, or small private enterprises with one or a few trucks. On the other hand, private carriers are the ones who operate the trucks to transport their products, or the products they need for their business. Cargill is mainly working with for-hire trucking companies. However, in some cases, Cargill agrees with the grain elevators to transport the grain by an affiliated person of the elevator. In such situations, the contract is issued on an FCA (Free Carrier Agreement) basis, with the carrier being approved on a date when the contract is signed at a fixed rate. The affiliated person can be the supplier himself, or the supplier can agree with for-hire transportation companies.

### 3.2.2 Spot VS Contract rates

The payment agreement between the buyer and the supplier can be of two types: contract or spot. Under contract agreement, the carrier agrees to deliver products during a specified time under certain conditions. In general, contracts are signed on annual basis and specify the delivery volume in a given direction over a certain period. The contract provides insurance against market or seasonal freight rate fluctuations, as well as shortages of transport during pick demand. In contrast, the buyer is paying the spot rate price when the transport and its price are arranged at the moment when the buyer needs to deliver the products. Spot freight rates are characterized by relatively high volatility due to their dependency on exogenous factors such as demand, supply, and fuel prices. In Russia transportation is predominantly arranged shortly before the delivery takes place, so Cargill mainly pays the spot rate. However, as mentioned in 3.2.1, in some rare cases Cargill has FCAs with some elevators. This is mainly the case when a special transport which meets particular technical requirements on a certain direction is needed.

### 3.2.3 LTL and FTL

Shippers typically make a distinction between Less than Truckload (LTL) and Full Truck Load (FTL) shipments. LTL shipment refers to the case when multiple deliveries or pickups take place in one route. Often the carrier makes stops during the shipment to pick up and drop off the products. In contrast, in FTL shipment a truck carries only one shipment for one customer. FTL is generally considered to be faster and less risky since the shipment directly follows from the point of origin to the destination location. In this project, only FTL shipments are considered.

## 3.3 About Russian transportation market

The size and terrain relief of Russia requires the engagement of various means of transportation, although road remains the main mean of freight delivery on both national and international levels. In 2021 transportation by trucks accounted for almost 50% of total freight, with rail and pipeline being the second most used means constituting around 13% each of overall transportation ("Russia in the Global Transport and Logistics System: the Main Vectors of Development", 2022). Truck and rail transportation services are mainly used for the transit of products on a national level, while cargo ships transport the majority of products that go to and from international markets. The decision to transport

products by rail or road is taken depending on the distance. In general, the analysts determine the break-even point - a distance for which the price of transportation is the same for both train and truck. In Russia, the break-even point is set at around 500-600 km. The product will be transported by truck whenever the distance is below the break-even point, and by rail otherwise.

The Russian trucking market is operated by big carrier companies, which hold a large fleet of trucks, as well as small private entities owning only one or few trucks. In 2017 around 20000 transportation companies with an average fleet of 30 trucks and around two million private truck owners were operating in Russia (ATI.SU, 2021). The recent emergence of digital apps significantly simplified the process of matching shippers with clients, especially for small private entrepreneurs. Overall, the Russian trucking market is highly fragmented with a lot of small participants in the freight transportation industry.

The level of maturity of transportation services depends on the quality of infrastructure (road and rail quality and network, port infrastructure, terminals) and the level of modernization of the supply chain. Despite Russia being the largest country in the world, the road network of Russia is smaller than the one of the US, Japan, and France. However, over time there has been a positive trend in infrastructure development, road quality, and maturity of supply chain services due to government investment in modernization and digitization. According to OECD, there has been a steady improvement in road quality and maturity of the logistic chain over the past 10 years. Recently, the Russian government announced a plan to invest $88 billion into the modernization of airports, highways, railways, and ports by 2030. An additional $66 billion is allocated to the "Safe and High-Quality Highways program" ("Russia - Construction and Infrastructure", 2020).

There has been a steady increase in the volume of freight transportation since 2010, except for the pandemic in 2020, but the upward trend continued in 2021 and was expected to continue further. In addition, a large-scale infrastructure policy will most likely lead to changes in the quantity and geography of transportation in the future. However, military conflict with Ukraine is likely to cause some disruptions and structural changes in international and domestic transportation markets, posing huge uncertainty on how the market will evolve. At the time when we write this thesis, there are constant changes in national and international regulations, some of which are impossible to predict. Western sanctions on imports and exports to and from Russia have already impacted the operations of transportation companies. Temporary withdrawal of some major international companies from the Russian market and ban on truck operation abroad can cause reorientation of trucks from international to the domestic market. In addition, there has already been an increase in the prime cost of transportation

by 10-12% since the beginning of 2022, and the cost of service of freight transport by 60% (ATI.SU, 2021). These events can cause structural changes in the demand for and supply of tipper trucks. Such uncertainty poses challenges for the decision-makers within Cargill and requires close monitoring of new policies and regulations to estimate their effect on the operational strategy of the company.

### 3.4 Trends in the agriculture of Russia

Demand for freight is an important factor in the formation of freight rates. Since we narrow down the scope of this project to forecasting the transportation price of agricultural commodities, trends in the agricultural market will significantly affect the demand for tipper truck shipments. The agricultural market of Russia is characterized by high seasonality and cyclicality. The seasonality of sales reflects the harvesting pattern of the crop, which differs depending on the grain. Wheat is harvested in June and July. Since it is the main crop exported by Russia, during these two months we generally observe the biggest increase in grain exports from ports.[1] At the same time, harvesting creates additional demand on tipper trucks, since grain must be transported from fields to grain storage centers. During the harvesting season, which starts in July and ends in October, exporters of grain have to compete with farmers for transport. Since harvesting involves multiple short deliveries over small distances, transporters usually find such orders more attractive, which drives the freight rate for grain transportation to ports even higher.

Apart from planting and harvesting patterns, the seasonality of exports of grain highly depends on tariff rates and quotas on exports established by the Russian government. By tightening and relaxing export conditions the government can regulate the national prices of grain, as well as stocks of grain commodities inside the country. While quotas directly limit the volumes of exports, tariffs make exports less profitable. Before 2021 there were no restrictions on grain exports from Russia. However, as of February 2021 and until the 1st of July 2021 a tariff quota was imposed: it was allowed to export a maximum of 17.5 mln tons of grain (wheat, rye, barley, and corn), and everything above that value was taxed at 50% rate. The government replaced the tariff quota with the mechanism of grain damper as of June 2021, and the system is still in place on the day we write the thesis. The quota is calculated based on world prices of grain. The tariff rate is $0 when the world price goes below $200 per ton for wheat ($185 for corn and barley), and 70% of the difference between the world price and the threshold price when the price is above $200 per ton. The tariff is calculated weekly based on the price from the

[1] In 2020 and 2021 wheat export constituted about 85% of total grain export volume.

week before. Overall, total exports are likely to fluctuate in response to changes in governmental policy and harvesting patterns.

# 4 Literature Review

In light of recent advances in statistical and machine learning methods, businesses have become more interested in incorporating mathematics into their day-to-day decision-making process. While expert knowledge still plays a pivotal role in conducting business operations, statistical and machine learning techniques become essential tools to guide and support experts' decisions. In particular, there has been a growing interest in projecting historical data into mathematical models to deliver predictions of various price indicators. From the business perspective, accurate knowledge of the future costs and revenues in face of uncertainty allows to maximize the profits from activities through more efficient strategy formation, provision of better services to clients, possibility to outcompete other players in the industry. Numerous literature is dedicated to the development, assessment, and application of statistical and machine learning models to forecast different price indicators for various time horizons. Time-series regressions, such as ARIMA and Prophet, are typically employed when decision makers are interested in short and medium-term price predictions (Dhaval & Deshpande, 2020; Contreras, Espinola, Nogales & Conejo, 2003; Garlapati et.al., 2021; Güleryüz & Özden, 2020). In some cases, machine learning techniques have demonstrated a good predictive power for both spot and contract market prices (Wanjawa & Muchemi, 2014; Ho, Darman & Musa, 2021; Güleryüz & Özden, 2020). On the other side, some authors try to do medium-range, long, and very long-range price forecasts. Typical techniques employed are decision trees, neural networks, and multiple linear regressions (Liu, Hu, Li & Liu, 2017; Kotur & Žarković, 2016; Ismail, Yahya & Shabri, 2009).

Transportation cost is one of the major components in cost formation, and there exists an extensive list of literature on the prediction of trucking freight rates. The forecasts are developed for spot and contract rates, different types of trucks (LTL and FTL), route-based or general model-based, long term and short-range predictions. Some researchers are employing more conventional approaches. The paper by Budak et al (Budak, Ustundag  Guloglu, 2017) forecasts spot freight rates using two methods: Artificial Neural Networks (ANN) and quantile regressions. The two methodologies are applied firstly, to the route-based model by assessing the cost for each route separately and then to a general model which combines all the routes in one model. The results of the paper demonstrate that both approaches can provide an accurate estimation of the future freight rate. However, which method is better depends

on whether we consider the general or route-based model. ANN shows better forecasting power for route-based prediction with 33% of the predicted observations having the Mean Absolute Percentage Error (MAPE) of 0. On the other hand, quantile regression is superior when the general model is used. The paper "Forecasting short-term trucking rates" utilizes US nationwide daily data on long-haul truck transportation to estimate spot and contract dry van freight rates (Bai, 2018). The author incorporates the hybrid neural network for times series, known as NAR(X) LTSM neural network to forecast transportation costs, and compares this to the traditional time series ARIMA(X) model. The paper by Miller (Miller, 2018) utilizes the ARIMA time series framework to forecast full truckload prices at the macro-level, including national spot freight for general and refrigerator full truckload transport. The authors conclude that ARIMA can capture the dynamics of the evolution of spot truckload rates.

Instead of using the time series method, which typically relies on autoregressive properties of time series, some authors explain freight rate with other exogenous variables. In the study "Developing a forecasting model for freight prices in Poland" the author uses Multiple Linear Regression to forecast trucking freight rates in Poland (Spreeuwenberg, 2022). The model is examined under different specifications and at various time horizons. The author uses Cross-Validation to select the best predictive model. Another regression-based approach is employed in the paper "Modeling net rates for expedited freight services" (Smith, Campbell  Mundy, 2007), in which a few specifications of multivariate regression are tested to examine which factors contribute the most to the formation of customer-lane revenues. Both papers demonstrated that the evolution of freight rate can be explained with a case-specific set of predictors.

Research suggests that under certain assumptions conventional techniques can be applied to design an accurate predictive model for the trucking freight rate. However, some researchers focused on incorporating new ideas to improve the performance of classical models. The paper "Short-Term Truckload Spot Rates' Prediction in Consideration of Temporal and Between-Route Correlations" (Xiao, Xu, Liu, Yang  Liu, 2020) develops an advanced statistical regression model to predict the short-term route-specific truckload rates in China. The study introduces the weighted lagged coefficient matrix (LagWMR) which incorporates the correlations on the adjacent routes and time-lagged correlations as an additional exogenous variable in multivariate linear regression. The paper then compares the method with machine learning (LGB) and time series approaches, and concludes, that LagWMR outperforms the two. Estimating and Less-than-Truckload market rates (Özkaya, Keskinocak, Roshan Joseph & Weight, 2010) provide an alternative regression-based approach to explain the transportation cost. Along with 'tangible' factors (i.e.the observable data, such as volumes and distances), 'intangible'

variables, whose values are derived from expert knowledge, are added to the model. These additional factors, which include desirability, negotiation power, perceived freight class, and economic value, significantly improve the power of the model to explain the variation in data. Swenseth and Godfrey employ a completely different approach in their research paper (Swenseth  Godfrey, 1996). It is generally believed that we can estimate freight rate by having it as some function of weight and distance. Swenseth and Godfrey test this hypothesis by comparing five different functional forms, and then assessing their accuracy. The obvious drawback of this approach is its inability to capture relevant market conditions, making it nonfunctional in an uncertain environment.

The majority of research is focused on designing a time series model with the historical values of the series entering as the main input to the model. In addition, the size of the data set used is usually relatively large. This thesis adds to the existing literature by emphasizing the effect of exogenous variables such as demand and price of gas for the spot market freight rate formation, rather than assuming that future freight rate can be explained with the historical values of freight. The thesis also shows that with small data set traditional time series technique ARIMA fails to predict the freight rate. Furthermore, we use quantile regression to test the hypothesis that the elasticity of freight to demand is non-constant across its percentiles. Lastly, the study suggests the methodology for a simple and user-friendly risk assessment process to potentially improve the prediction accuracy.

In the next chapter, we will review the literature to select the best methodology for our problem.

# 5 Methodology Review

## 5.1 Methodologies overview

An accurate methodology design is crucial to deriving robust predictions. There exists a wide range of statistical and machine learning forecasting methods. However, their applicability highly depends on such factors as the size of the training set, the presence and number of predictors, the complexity of the relationship between the variables, and historical and future trends of the indicators. This section presents a closer look at different methodologies and discusses when one can and cannot apply the given method for forecasting purposes.

Autoregressive Moving Averages is a class of models designed specifically for time series analysis. Its design is based on Wold's decomposition theorem, which states that any stationary time series can be

described as a sum of one deterministic and one stochastic time series. The deterministic time series is represented by the Autoregressive (AR) part of ARMA and it models a linear combination of a certain number of lags of time series. The stochastic time series is the Moving Averages (MA) part, and it aims to explain the error term of the current value of the series through error terms in the previous periods. ARMA generally shows good predictive power when modeling not seasonal, stationary time series, and when the outcome is not formed by only white noise. However, there exist variations of the ARMA model, which allow accounting for non-stationarity (ARIMA), seasonality (SARIMA), and additional predictors (ARIMAX). These methodologies have been applied successfully in the research literature for time series forecasting (Contreras, et.al., 2003; Ariyo, Adewumi & Ayo, 2014; Azari, 2019). Facebook Prophet is another approach, which similarly to ARMA, is designed to make predictions using time series data. The algorithm first decomposes the series into seasonal, holiday, and trend effects, and then calculates the sum of these effects (Jha & Pande, 2021; Kaninde et.al, 2022). Both ARMA and Prophet models are user-friendly and have some degree of interpretability due to their parametric form. However, historical values of the series are the main or even the only input of the model. These methods rely on the assumptions that past values capture the majority of the information about the variation in the series, and, therefore, will be unable to capture the turning points and exogenous shocks.

In some cases, traditional time series techniques have low predictive power as they fail to capture complex non-linear relationships. In recent years there has been a growing interest in making forecasts with Artificial Neural Networks. Artificial Neural Network is a deep machine learning computing technique, designed to deal with complex, highly non-linear problems. Its design was inspired by the brain operation of humans and animals: a typical Neural Network is composed of a series of interconnected layers (nodes), which similar to signal-transmitting neurons in human brains, transmit the data from one layer (node) to another within the network. At each node, the algorithm applies a weighted function to the input data and passes the output of this function onto the next node. The final node returns the predicted value. The process is repeated multiple times, and at each iteration, the weights are updated using optimization algorithms to reduce the training error. This iterative process is known as the backpropagation algorithm, and it assures that the local optimum is achieved in an efficient and timely manner.

When forecasting time series, more advanced ANN can be applied. Nonlinear Autoregressive (NAR) Neural Network includes lagged values of the series into the model while accounting for the error term associated with the predicted value. Long Short Term Memory (LSTM) Neural Network is a recurrent neural network, which can incorporate new information and adjust the model accordingly when new

data arrives. Unlike the other recurrent networks, which remember only recent information, LSTM NN can keep the information for a relatively long time.

NN techniques have shown good performance when applied to forecasting various price indicators using time series data (Ho, Darman Musa, 2021; Güleryüz, Özden, 2020; Bai, 2014). However, calculations behind the neural network are a black box, as the user can only see the final output of the network. This quality makes neural networks non-interpretable, which poses a series of issues for decision-makers when applying the results of the model in the real world. General users will have to fully rely on the results provided by the NN without understanding the mathematics behind those results. In addition, training the model requires a large data set, typically thousands of data points, which poses major limitations on its applicability. Among other limitations of the network is its reliance on the experience of its developer. The user has the flexibility to choose certain functions at certain nodes, and the right choice significantly depends on the user's experience.

Sometimes, we are interested in very long-range predictions or believe that the current value of the dependent variable is formed by the influence of other exogenous factors - predictors. In such cases, linear and polynomial regressions are widely applied techniques for the prediction and estimation of causal relationships, and under certain assumptions, they provide a very robust inference and show high predictive power (Patil  Sahu, 2015; Velonias, 1987; Ismail, Yahya & Shabr, 2009; Uras et.al., 2020). The benefit of the regression approach is in its low computational power and ability to provide a reliable and unbiased result with a relatively small training sample. Furthermore, because of the model's simplicity, the results are easy to interpret and apply. However, regressions typically assume some constant functional form of the relationship between variables. If the training data is not accurate or incomplete, the predictions will be inaccurate.

A further brief overview of possible methodologies is summarized in Table 1.

### 5.2 Methodology used in our research

The choice of the best methodology to forecast tipper truck freight rates in Russia is based on a careful analysis of the problem and available data. Two main constraints for the model choice are the requirements of Cargill to have an interpretable, user-friendly model, and limited data quantity. These limitations make neural networks and decision trees unsuitable for our problem. Furthermore, after talking to experts the conclusion was made that the freight rate is significantly driven by supply and demand. Given this information and the small data set to train the model, multiple linear regression

was chosen as the best methodology to meet our purposes. In addition, the predictive power of ARIMA was tested when the train set is small, and the performance was compared to the one of the MLR. **??** presents the methodology used in this paper.

| Methodology Summary | | | |
|---|---|---|---|
| Model Type | Time scope | Pros | Cons |
| Deep Learning (based on NN) | Very short, short, medium, long. | Good performance for short term predictions. Does not imply linearity, therefore, can solve highly non-linear problems in a robust and efficient way. | Require a lot of data to train the model. Lack of interpretability: we only specify the model parameters, but the calculations go behind the scene. Convergence to global optimum is not guaranteed (depends on the starting point). |
| Times series (ARIMA, Prophet, Exponential smoothing) | Very short, short, medium. | Simplicity, we can explain current values just with lags and/or lags of error terms. Low computational power. | Assumes that lagged values and lagged error terms capture enough of information to explain future values. If significant changes occur in the future and historical data does not capture it, the forecast is wrong. |
| Linear and polynomial regressions | Short, medium, long, very long. | Very easy and intuitive interpretation of the results. Can be used to establish causal relationships. Low computational power. Asymptotic properties achieved even with small dataset under certain assumptions. | In most cases requires parametric form. Assumptions include independence of the observations. |
| Support Vector Regressions | Short, medium, long. | Robust to outliers. Can be used for highly non-linear data. | Kernel function is determined based on the belief about the structure of the relationship. Does not work with too many observations. |
| Decision Trees, Random Forest | Short, medium, long. | Can handle linear and non-linear problems. Good interpretability with simple regression tree. | A lot of data is required to train a good decision tree. It is easy to overfit the model, especially if the dataset is small. |

**Table 1:** Methodology review.

# 6 Data

To train the model we collect the dependent variable - transportation cost in Russia, and a set of predictors which should explain the transportation cost.

## 6.1 Dependent variable

The data on the historical freight rate used in the project is SAP data on Cargill's transportation transactions in Russia from January 2020 until December 2021, and it was collected and provided by Cargill. The data set is in the form of separate data points for each shipment, and it includes information on departure and arrival dates, origin and destination locations (location code, location enterprise name, country, state, city, and zip code), net weight and cost of each shipment, carrier name on a given transaction, and a set of other variables which are omitted in this study due to their irrelevance. Figure 1 presents the plot of the origin-destination locations. Bubbles represent the delivery locations, with the size of the bubble being proportional to the total volume sent to the given location. From the map, it is clear that there are 3 major destination points (they are numbered on the map), and a few minor delivery locations. The origin locations are distributed relatively evenly across the Southwestern region of Russia. Location 1 is an oil seed crush plant owned by Cargill. It is located in the city Novoanninskyy and accounts for the majority of the volume transported. The second location is the port on the Azov sea next to the city Rostov-on-Don, and Location 3 is the port in Novorossiysk situated on the Black sea. Summary statistics for grain transportation to the three locations are presented in Table 2. Less than half of the grain transported by Cargill goes to the two ports together for export, and the rest of the volume is driven by deliveries to the factory and some smaller locations nearby. Average distance and cost per tone are the highest for Location 3, and Location 1 has the highest origin-destination variation.

| | Port Novorossiysk | Port Rostov | Plant |
|---|---|---|---|
| Total Volume (Jan, 2020-Dec, 2021) | 1,223,484,760 | 418,104,080 | 2,193,293,240 |
| Average distance, km | 330 | 217 | 220 |
| Number of routes | 151 | 45 | 473 |
| Average cost/ton | 1426 | 1014 | 1095 |

**Table 2:** Summary statistics for different locations

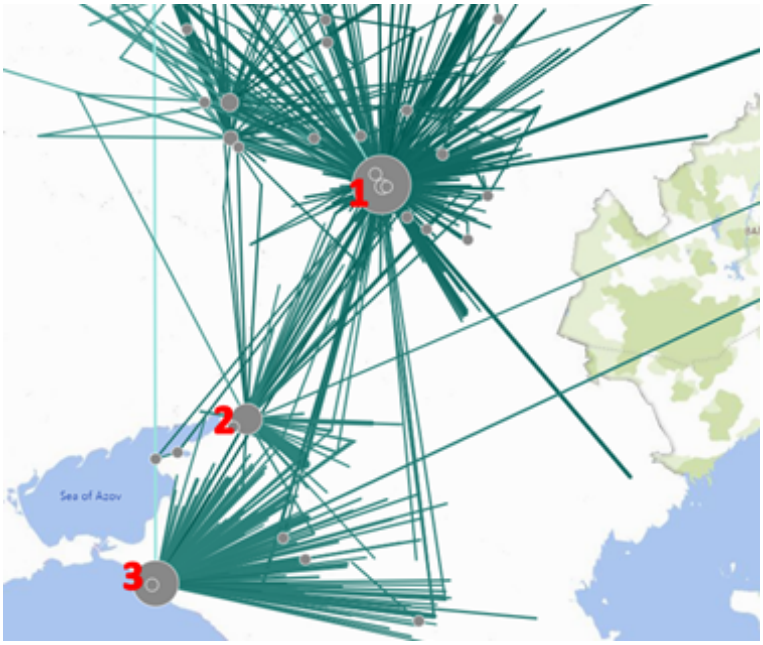



**Figure 1:** Origin-Destinations map

### 6.1.1 Data clean-up

Since the data is collected manually, some mistakes could be made while inputting the data. Therefore, the next step involves outliers removal. In general, we should expect a linear relationship between cost per ton and distance, and observations that significantly deviate from the trendline should be removed from the analysis. Figure 11a in Appendix B is a plot of cost per tone against distance before the outliers removal process. We can observe that some points do not follow the general trend in freight rate and, therefore, can bias the model. To filter these outliers we implement the following steps:

1. Plot a regression line and extract the coefficients of the regression.

2. Using the coefficients (slope and intercept) from the regression, fit the data points according to the regression equation.

3. Calculate the difference between the fitted and the actual values.

4. Derive the interquartile range of the difference.

5. Only keep the data that falls in the interval [Q1-1.5*IQR, Q3+1.5*IQR] (Yang et.al., 2019).

Figure 11b in Appendix B shows the relationship between the distance and cost per ton after removing the outliers. $R^2$ which measures the goodness of fit of the data is improved from 0.45 to 0.52. There is still a significant variation in data: for the same distance, the transportation cost can be lower or

higher. However, the variation is monotone across different distances and is mainly caused by seasonal fluctuations in freight rates. Overall, in this stage, we remove 2.9% of the data for port Novorossiysk, 11% for port Rostov and 4.8% for the plant in Novoanninskyy. In total, we remove around 4.5% of all the observations from the original dataset.

### 6.2 Independent variables

We collect a set of predictor variables externally. The variables include:

- **Port line ups.** We extract the data on port lineups from Zenith. This is the data on the volume of grain that departs from Russian ports with cargo ships on a given day, and it is used as a proxy for demand for tipper trucks. The data includes lineups at all Russian ports for commodities, including various grain commodities, oils, coal, and fertilizers. First, we filter relevant ports and commodities for analysis. We only include bulk grain commodities, since we want to measure the demand for tipper trucks used to transport grain. Secondly, we convert the data into moving averages to correct the following two issues: (i) we used the date of departure of the cargo to estimate the total volume delivered which does not exactly coincide with the date of the departure of the truck; (ii) we include total weekly lineups, however, there can be a significant difference between volume delivered to ports on the first and the last day of the week.

  Figure 2 maps the location of the ports from which Russia exported grain during 2020 and 2021. The size of the bubble indicates how much grain was exported from the port in relative terms. We indicate Novorossiysk and Rostov locations with red color. From the map, it is clear that the majority of grain volume was transported to or close to the port of Novorossiysk, and less to the port of Rostov. Later, we try different combinations of commodities and ports to see whether different types of grains exports and locations of exports have a different relationship with the freight rate.

- **Diesel prices.** Diesel prices are derived from Yandex.ru. The website publishes historical data on daily retail prices of diesel. We use the price on the first day of each week to have a weekly diesel price. Since diesel is one of the costs incurred by the transportation companies, it should significantly positively affect the freight rates.

- **Commodity prices.** The data is received from Cargill and includes weekly commodity prices for corn, wheat, rice, barley, and coal. Commodity prices affect the decision of the owner of


**Figure 2:** Russian ports

commodities to sell the grain in such a way increasing demand for transport.

- **Inflation (CPI) and Producer price index (PPI).** The data is received from Cargill. The data for CPI and PPI is monthly and is plotted on Figure 15 in Appendix B. We convert PPI and CPI from monthly to weekly data by assuming that both indexes change linearly from week to week (linear interpolation).

- **Dummies.** We create 3 additional dummy variables: for harvesting season (from July till October), for January of 2021, and the weeks before and after New Year's Eve. A dummy for harvesting is created to account for possible additional demand for trucks during the harvesting season. From July till October farmers order tipper trucks to deliver grain from the fields to the storage points. This causes additional demand for tipper trucks. We also control for January 2021. During this time there were storms in the Black sea, which caused delays in cargo ship departures. Finally, the New Year period generally has a lower supply of trucks, since drivers often go on holidays during this period. Therefore, we control for the two weeks before the New Year and two weeks after New Year's Eve.

### 6.3 Target variable

The target variable is defined as the variable whose values are predicted by the model. In this thesis, we want to estimate the trucking freight rates with historical data on Cargill's transportation costs in Russia. Therefore, the target variable should accurately represent the variation in costs of transportation. The data on freight rate is recorded daily with each shipment entering the data set as a separate data point. We want to create a time series model, therefore, the transactions-based data should be transformed into the periodic variable where the resulting observation is representative of the average freight rate in a given period. We choose to create a weekly time series model. The decision is based on the quantity and quality of data, the number of missing data (in our case periods with no deliveries), and the variability of freight rate over time.

The cost of transportation on some routes can significantly deviate from the average trend because of some route-specific factors. For example, a truck might have to take a boat which significantly increases the freight rate. Our goal is to capture the average trend, therefore, these points should be removed from the analysis. Since there is little variation in the weekly data, we remove the outliers by month. The procedure is as follows:

1. Group all the observations by month.
2. Plot a regression line and extract the coefficients of the regression.
3. Using the coefficients from regression, fit the data points according to the regression equation.
4. Find the difference between the fitted and true value of the observation.
5. Sort the data according to the difference with fitted values.
6. Remove top and bottom 10% of the observations sorted by the difference (overall, 20%).

10% is chosen based on the examination of the data after trying different shares of outliers removal. We want to have a good fit of the data ($R^2$), but also preserve some variation in data. Based on this trade-off, the removal of 20% of the data was chosen as optimal. Overall, in steps 1 and 2 around 24% of all the data was removed. Next, we remove duplicating data: when the delivery has the same cost per ton over the same route. Eventually, we end up with 1026 relevant observations, on average 10 per week.

It was not possible to use the weekly average cost per ton per distance for the analysis because of data quality issues. We observe large weekly volumes of grain moved by Cargill both in terms of tonnage

and the number of trucks moved. However, in some weeks there is little variation in the observations. In particular, multiple trucks can operate on the same route and at the same cost. In addition, in some periods we observe a high variation in the cost of transportation. The difference can be explained by such factors as road quality, route complexity, weather conditions, arrangements with transportation companies, and other factors that cannot be controlled. Therefore, we used the following methodology to create a target variable:

**Step 1: Create a theoretical target.**

Generate a linear regression between distance and cost per ton, and extract the coefficients (intercept and slope). Use these coefficients to derive a theoretical cost per ton for a range of distances from 0 to 1000 km using the equation:

$$targ\_theor_i = intercept + slope * i, \quad i = 1, 2, \ldots, 1000. \tag{1}$$

Where i is the distance in kilometers.

**Step 2: Difference to the base value.**

The base value is chosen as $targ\_theor_M$, where M is approximately the average distance of the route in the sample[2]. Next, we calculate the theoretical difference to the base value as:

$$dif_i = \frac{targ\_theor_i}{targ\_theor_M - 1}, \quad i = 1, 2, \ldots, 1000. \tag{2}$$

**Step 3: convert cost per ton to the base value using theoretical difference.**

For each true observation j from the dataset with distance d, we obtain the $target_j$ as:

$$target_j = \frac{obs_j}{1 + dif_d}, \quad \forall j, \tag{3}$$

where $dif_d$ is derived from Step 2.

**Step 4. Find a maximum of each week.**

To derive the final unique target for each week, we use the maximum of all j for every week. Maximum target value results in the best correlation with the lineups at the ports, when compared to using average, upper quartile, and minimum. This observation can be explained by the fact that maximum

[2] M = 300 for this dataset

better reflects the market value of the freight rates since it captures the mark-up the customer is willing to pay in the period of high demand. Furthermore, sometimes Cargill makes contracts with the farmers for the purchase of grain under the conditions that the seller also takes the responsibility for the delivery of the commodity to the destination point. In such cases, the rate that Cargill pays to the farmer can be fixed over a certain time, and will not reflect the fluctuations in the external factors affecting the rate. If included in the average, these observations can contaminate the data. However, the maximum target will be less sensitive to the short-term fixed freight rates, given that Cargill is a profit maximizer and will not make a contract where they overpay the market price. Given the low variation in data, the maximum will better reflect the change in the market value of the truck freight rates.

### 6.4 Assumption

Before we introduce the model, it is important to mention the assumptions used in this thesis.

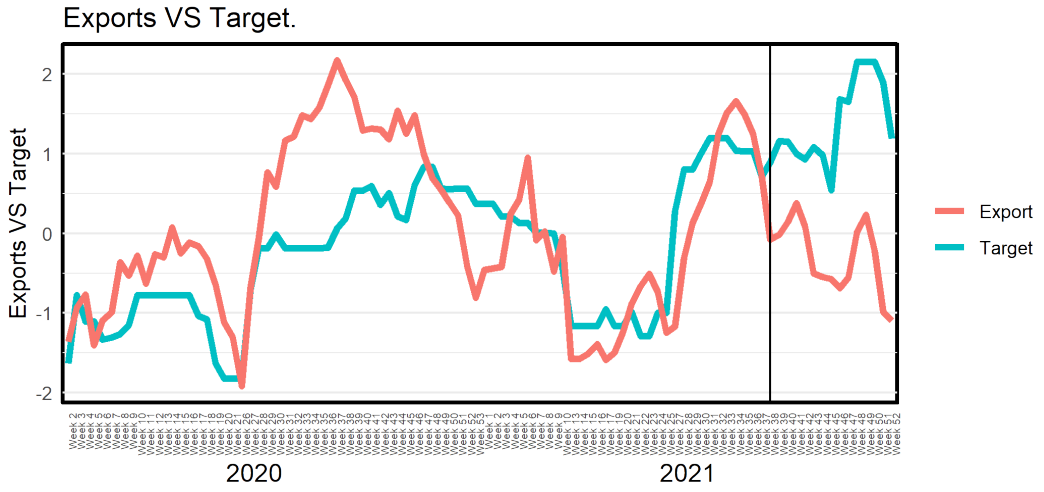
**A1. Freight rates are mainly demand and supply driven.**

This assumption can be informally checked by looking at the relationship between demand for trucks (estimated by the moving average of port lineups) and freight rate over time. Figure 3 a and b are both line plots of the two variables over time[3]. The target variable on Figure 3a is a 'raw' target, whose derivation is described in section 6.3. From the plot, it is clear that the lineups and target variable display similar fluctuation patterns. However, starting from week 38 of 2021 (indicated by the vertical line), the target is significantly higher than exports. Figure 16 in Appendix B indicates that there was an increase in diesel prices around that time, so we eliminate the effect of diesel from the target variable to see whether the rise in the price of diesel could cause an abnormally high freight rate in the second half of 2021. First, we regress diesel prices on the target variable and save the residuals. These residuals represent the freight rate, controlled for diesel prices. The residuals are plotted along the weekly export volumes on Figure 3b. The two variables seem to correlate much better after the effect of diesel on freight is eliminated with the correlation coefficient increasing from 0.47 to 0.77.

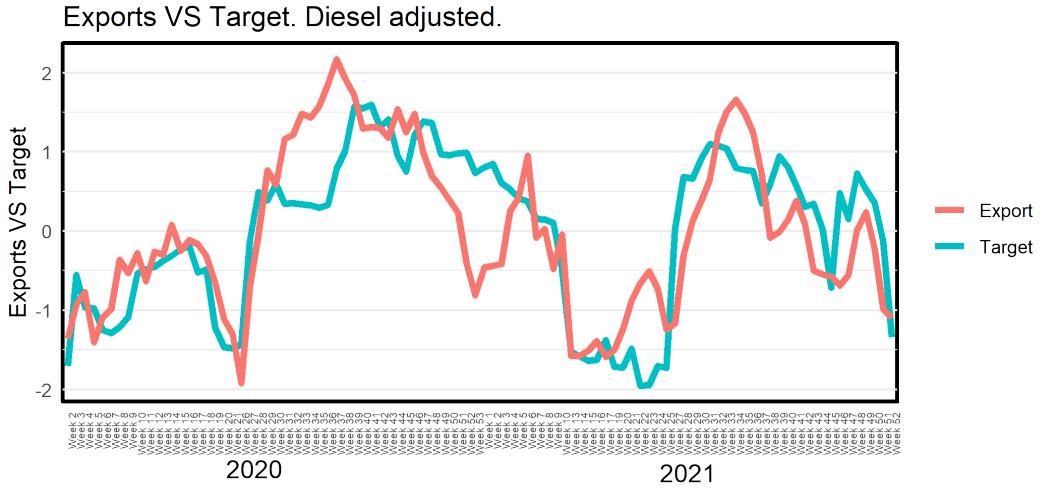
**A2. The supply of trucks is fixed.**

In reality, supply is unobserved, since there is no available database that keeps track of the number of trucks in the industry on a weekly basis. However, this assumption should hold in the short run. The demand for transportation varies seasonally with pick demand occurring in the summer during

[3]Both variables have been standardized to resolve the problem of different scales. Standardization preserves the exact variation in data.




**(a)** Export vs target. Not diesel adjusted




**(b)** Exports vs target. Diesel adjusted

**Figure 3:** Export vs Target

harvesting and high grain export seasons. However, the cost associated with acquiring the trucks for just a few months and disposing of them later is relatively high, especially for smaller players in the trucking industry which dominate the market. It is more likely that some drivers switch to other activities during low season with trucks being idle, or rented out. However, this should also be reflected in lower freight costs, as drivers would stay in the industry if there was enough demand. There is a bigger concern when it comes to the long run. For example, some extreme events like Covid and now conflict with Ukraine can permanently affect the structure of the market. Such events are difficult to account for in the model without having accurate data. One possible option for partial control for such effects is to include dummy variables.

**A3. Trucks operating in one region do not operate in other regions.**

Assumption 3 was discussed and confirmed by the Russian CTL team. The distance between locations in two neighboring regions can reach more than 1000 km. Therefore, the majority of trucks are operating only in a specific region. For example, the truck which usually operates in Location 3 will be very unlikely to drive 1000 km North to make a delivery inside the region of Location 1.

### 6.5 Further analysis of the market

Assumption 1 is that the freight prices are demand and supply driven. Given that the supply of trucks is fixed in the short run (Assumption 2), freight rates should significantly respond to fluctuations in demand on tipper trucks. Since Location 1 (Novoaninsskyy) is a sunflower seed crush plant that belongs to Cargill, and the production of oil does not show huge seasonality, the demand for raw commodities is expected to be stable over time. Contrary, both Rostov (Location 2) and Novorossiysk (Location 3) destinations are ports, and the demand for transportation in their neighborhood highly depends on the amount of grain exported from these ports. Therefore, we should expect different volatility of freight rates on transportation to the ports compared to the plant. To check this assumption we investigate the data graphically. Figure 12 in Appendix B shows total weekly volumes sent to three locations over 2020 and 2021. The first observation is that a much higher volume is delivered to Location 1, and the fluctuations in the volume are smaller than for Locations 2 and 3. Secondly, in some weeks very little or no grain is moved across all three routes. While Locations 2 and 3 "zero delivery" periods coincide, it is not necessarily the case for Location 1. We can conclude that the total volume transported is more stable for Location 1 relative to Locations 2 and 3. However, contrary to our expectations, it is also not constant over time. Next, on Figure 4 we plot the average weekly cost per ton per one kilometer for grain transportation to the plant (orange line) and both ports together (blue line). Unfortunately, the plot does not provide evidence that the volatility of freight rate is higher for deliveries to the ports, compared to the deliveries to the plant. However, despite the presence of a similar trend in some periods, the fluctuation of average freight cost in Location 1 does not resemble the variation of freight rate to ports well enough to be included in the model. Furthermore, the distance between Rostov and Novoanninsky is 550 km, and between Novorossiysk and Novoanninsky - is 950 km. A lot of origins for the delivery of grain to Novoanninsky are north of it, with no intersection with origin locations for transportation to the port. Large distances between the destination locations and distribution of origin locations imply that regional differences in transportation companies are present. In particular, trucks

operating in one region are not likely to operate in another one (Assumption 3).

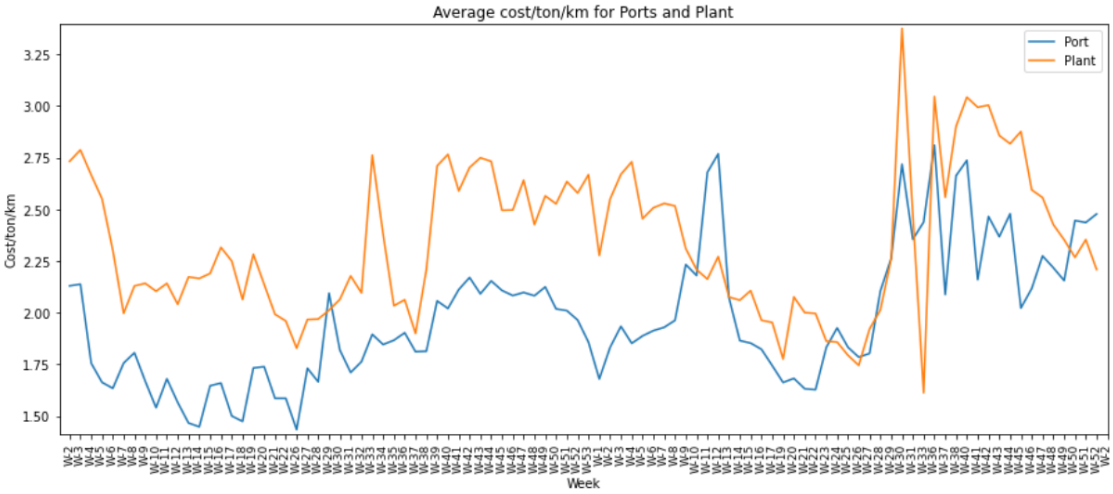



**Figure 4:** Cost/ton, plant VS ports

Next, we compare the weekly variation in freight rate for two ports. Figure 13 in Appendix B presents the average weekly cost per ton for two locations: Novorossiysk (blue line) and Rostov (orange line). In 2021, the transportation costs for Rostov and Novorossiysk destinations almost do not resemble each other. However, during this year we mainly have little variation in data: in some weeks the delivery happens over one or just a few routes (Figure 14 in Appendix B). This is particularly relevant for Rostov, as it is not the main destination point. A better resemblance between the two lines is observed in the first year of the analysis. Despite no robust quantitative proof of a similar trend in freight rate for the ports, we still decide to make one general model for both ports. There is a significant interaction between the origin locations for the deliveries to Rostov and Novorossiysk. Therefore, if we exclude the data for port Rostov from the model, Assumption 3 will not hold anymore: trucks operating in the direction of port Novorossiysk most likely also deliver grain to the port of Rostov.

Overall, we conclude that two forecasting models should be created: one model to predict trucking freight rates to the plant, and a separate model to forecast freight rates for the deliveries to the ports. Following the preference of Cargill, in this study, we focus on creating a model for the second case.

# 7 Linear Regression

Linear regression is a statistical approach used to model a linear association between an exogenous variable (the variable of interest) and one or more endogenous variables (predictors). Linear regression tries to estimate the conditional mean of the variable of interest as an affine transformation of predictor variables, such that the estimation error, represented as some function of true and fitted values, is minimized.

Suppose $y = [y_1, \ldots, y_n]$ is a row vector of a true dependent variable, and X is the n*p matrix of sample data, where n is the sample size (i.e. number of data points) and p is the number of predictor variables. Linear regression tries to estimate y according to the following equation:

$$\tilde{y} = X * \tilde{\beta} + \tilde{\epsilon}, \tag{4}$$

where $\tilde{y} = [\tilde{y}_1, ..., \tilde{y}_n]$ are the fitted, or predicted values of the true y, $\tilde{\beta} = [\tilde{\beta_1}, \ldots, \tilde{\beta_p}]$ is a row vector of coefficients estimated by the regression model, and $\tilde{\epsilon} = [\tilde{\epsilon_1}, ..., \tilde{\epsilon_n}]$ is the error term from the regression.[4] The error term is the difference between the true $y_i$ and its predicted value $\tilde{y_i}$. Mathematically speaking, the error term has the following equation:

$$\tilde{\epsilon_i} = y_i - \tilde{y_i} = y_i - \tilde{\beta_0} - x_i^T * \tilde{\beta} = y_i - \tilde{\beta_0} - \tilde{\beta_1} * x_{i1} - \tilde{\beta_2} * x_{i2} - ... - \tilde{\beta_p} * x_{ip} \tag{5}$$

To explain it graphically, we should consider Figure 5. Suppose we want to estimate the best linear fit between x and y. The unique pairs of x and y values are plotted as dots and suppose the line is fitted according to the equation (4). The error is the vertical difference between the dot, or true value of y coordinate, and the value of y coordinate on the line which is obtained through the vertical projection of the dot on that line. The projected value is the fitted value of the regression $\tilde{y_i}$.

There are several methods to estimate true $\beta$. The most common are least squares estimation techniques, which include OLS (Ordinary Least Squares), GLS (Generalized Least Squares), and WLS (Weighted Least Squares). In this study, we apply the OLS technique, which estimates true $\beta$ by minimizing the sum of squares of residuals. The function is as follows:

[4] True $\beta$ and $\epsilon$ are unknown since we use sample data. If we had population data, we could estimate true $\beta$ and $\epsilon$.

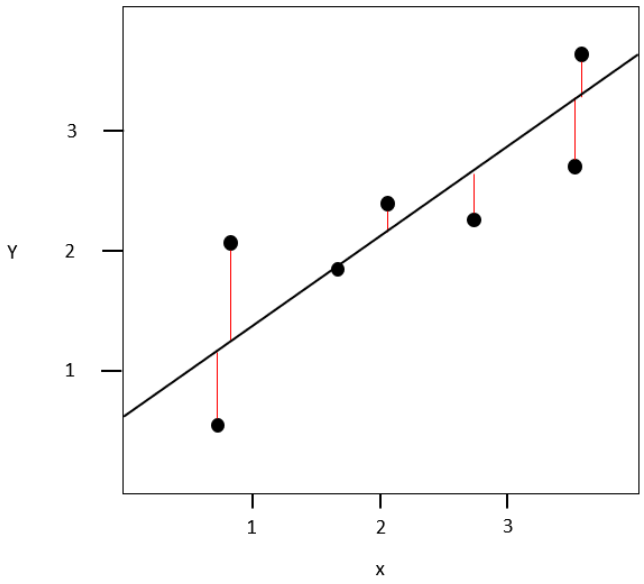



**Figure 5:** Linear Regression example.

$$\beta = \arg\min_\beta \sum_i^n \epsilon_i(\beta)^2 \tag{6}$$

Or, in matrix notation:

$$\begin{aligned} \beta = \arg\min_\beta \epsilon^T\epsilon = \arg\min_\beta (y - X\beta)^2 = \arg\min_\beta (y - X\beta)^T(y - X\beta) \\ = \arg\min_\beta (y^T y - 2(X\beta)^T y + \beta^T X^T X\beta) \end{aligned} \tag{7}$$

According to Gauss-Markov, OLS is the Best Linear Unbiased Estimator (BLUE) under certain assumptions. The BLUE property implies that the estimator provides consistent results while producing the lowest possible variance. In particular, if the Gauss-Markov assumptions hold, OLS outperforms other methods for estimating the unknown parameters of the regression. We present the assumptions of OLS below, and later in the results section, we validate that they indeed hold for our data.

## 7.1 Assumptions of OLS

### 1. Linearity

The assumption implies that the variable of interest is a linear combination of the independent variables.

The assumption is not strict, and in fact, relationships with various functional forms can be estimated. However, when applied in its conventional form ($y = X\beta + \epsilon$) to the variables with a nonlinear relationship, the model will find the best linear fit to a nonlinear function. This will result in significant underfitting. The model will have low variance, i.e. the coefficients will not change significantly when applied to a different sample, but might have a high bias. Bias can be reduced by including one or more polynomial terms of the explanatory variable in the equation.

The expert knowledge and genuine belief about the relationship between the two variables help to identify the correct model. However, the linearity assumption can be more formally verified by making partial plots of the predictor versus the predicted variable.[5] We investigate the presence of the nonlinear relationship between exports and freight rate in the results section. The other variables are expected to be related linearly to the target variable.

**2. No multicollinearity**

Multicollinearity arises when one or more independent variables are highly correlated with each other. Multicollinearity makes the model more sensitive to small changes in data parameters, sometimes to the extent, that the coefficients switch signs. This can result in an inaccurate model which is difficult to interpret and apply. Variance inflator factor (VIF) is generally used to test for the presence of multicollinearity. If the issue is detected by VIF, one or more of the variables with high correlation should be eliminated from the model. If elimination of variables from the equation is not desirable, another estimation technique called Partial Least Squares (PLS) can be applied. This technique allows for a high degree of collinearity between variables in the dataset. The main drawback of the PLS is that it is generally regarded as a predictive technique and not an interpretive one, which in some cases might be not acceptable (Pirouz, 2016). In our project for each specification of the model, we conduct a VIF test to check whether the specification suffers from multicollinearity[6].

**3. Independence of error terms**

Also known as autocorrelation or serial correlation, the assumption implies that errors are randomly distributed, and, therefore, are independent of each other across time (in case of time series) or clusters (in case of cross-section data). If the assumption is violated, coefficients are still unbiased, consistent, and asymptotically normally distributed. However, OLS is no longer BLUE, which implies that we can find a more efficient estimator with lower variance. The confidence intervals provided by OLS may be

[5]Partial plot is a scatter plot between two variables, where the effect of other variables is controlled for.
[6]R has a built-in function for the VIF test.

wider than they are, and when testing the NULL hypothesis of the coefficients to be different from 0, we can falsely reject it. Time series often correlate across time. This may lead to the violation of the independence of residuals assumption. A simple method to test the presence of serial correlation between the residuals is by plotting residuals over time. The presence of seasonal or/and time trends indicates an autocorrelation issue. A more formal way to test autocorrelation is with the Breusch-Godfrey test.

The easiest way to deal with the problem of autocorrelation is to use Newey-West standard errors. The method does not directly resolve the serial correlation. Instead, it allows to correct for the inconsistency of the standard errors in the presence of serial correlation and heteroskedasticity. Instead of correcting for the standard errors, we could solve the issue of serial correlation directly in the model. If the time series is AR(1), i.e. the present values correlate with the preceding values, Prais Winsten's estimation can be used to eliminate the possible problem. In addition, GLS model does not assume independence of the error terms, therefore, can deal with autocorrelation. Both methods attach a transformation matrix G to a linear regression equation $Gy = GX + G\epsilon$, which allows accounting for non-constant variance and autocorrelation of the error terms. We can also apply OLS regression to our model if we add the lags as additional exogenous variables. The drawback of this approach is the time horizon of the forecast: we can not predict the freight rate two months in advance without knowing the freight rate value 1 month and 3 weeks ahead.

We check serial correlation by plotting error terms over time. For all specifications, there is no clear trend or seasonal fluctuations of error terms. However, we also present the results of the model with Newey-West standard errors.

**4. Homoscedasticity**

Homoskedasticity implies that the variance of the error terms is constant across observation or time, and does not depend on the size of the data point or/and time. Similar to the case with serial correlation, when the assumption is violated (i.e. error terms are heteroskedastic), coefficients are still unbiased. However, the inference can no longer be accurate. The most common remedy against heteroskedasticity is to use robust standard errors or Newey-West standard errors. GLS and WLS also do not assume heteroskedasticity, and therefore, can be applied when error variance is not constant.

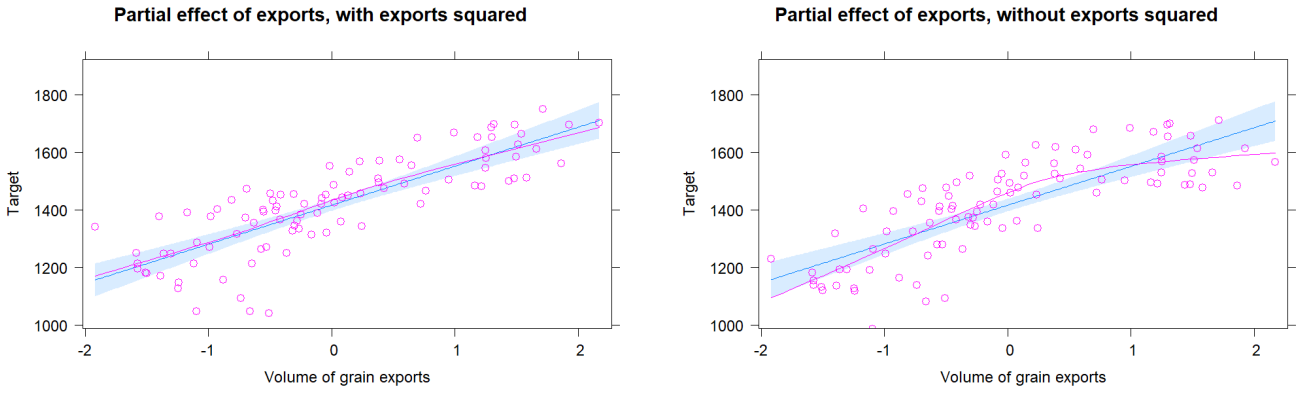
The most common way to test for the presence of heteroskedasticity is to plot residual versus fitted values. When the errors are homoskedastic, the variance of residuals will be constant across different values of predicted values. After training the model, we make a plot of residuals versus fitted values

and conclude that errors are homoskedastic.

# 8 Results

## 8.1 OLS

This section presents the results of the OLS regression. The regression is estimated under twelve different specifications. Since there exists a strong relationship between exports and target, the volume of grain exports always enters the equation as one of the exogenous variables. It is possible, that the strength of the effect of export volume on target is non-linear: freight rate can be more or less responsive to demand depending on how large the value of freight rate already is. Therefore, we investigate the linearity of the relationship between freight rate and port line-ups with a partial effect plot on Figure 6. A partial effect plot allows capturing the relationship between the two variables while controlling for other regressors in the model. Figure 6b indicates that there is some non-linearity in the relationship between exports and freight rate. However, if we control for the exports squared, the relationship becomes more linear Figure 6a. To see whether polynomial regression indeed performs better than linear all specifications are investigated under two different assumptions: (i) the relationship between exports and freight is linear; (ii) the relationship between exports and freight is second-degree polynomial.




**(a)** Partial effect. Linear.

**(b)** Partial effect. Non-linear.

**Figure 6:** Partial effect of exports on target.

It is important to note that when the second order polynomial term for export of grain is added to the equation we cannot interpret the coefficients for grain exports while keeping the coefficient of export

squared constant. To understand the effect, consider the equation:

$$y = \beta_1 x + \beta_2 x^2 \tag{8}$$

After taking the derivative on both sides we get:

$$dy/dx = d(\beta_1 x + \beta_2 x^2)/dx = \beta_1 + 2\beta_2 x \tag{9}$$

So the effect of x on y, in this case, depends on the size of the coefficients $\beta_1, \beta_2$ obtained by OLS and on the current value of x. When $\beta_2$ is positive, the effect of x on y increases as x gets larger. When the sign of $\beta_2$ is negative, as x increases, the strength of its relationship with y diminishes.

Furthermore, the variables for "Export grain", "Export grain$^2$", and "Price of diesel" were standardized in all specifications. Standardization rescales each value of the variable such that it is standard normally distributed with a mean of 0 and a standard deviation of 1. This is done by subtracting the mean from each value and dividing the outcome by the standard deviation of that variable. In mathematical terms the equation to standardize exports of grain is as follows:

$$x\_stand_i = \frac{x_i - \text{mean}(x)}{\text{std}(x)} \tag{10}$$

$$x\_stand_i = \frac{x_i - \mu}{\sigma}, \tag{11}$$

where $\mu$ and $\sigma$ are the averages and standard deviation of variable x respectively.

Standardization of exports allows for avoiding multicollinearity problems when the variable for exports grain squared is added to the equation. The variable for diesel prices is standardized to avoid negative intercepts caused by the high absolute value of diesel in rubles.

Interpretation of the coefficients changes after standardization. Coefficient $\beta_i$ in front of the standardized variable $X_i$ has the following meaning: a one standard deviation increase in $X_i$ on average increases the target variable by $\beta_i$, keeping all other variables constant. Furthermore, without standardized variables, the value of constant means the base value of the target variable when all the predictors have the value of 0. Standardized values of diesel and exports are 0 when both of them are equal to their average values. Therefore, constant has to be interpreted as the value of target when exports and diesel

are equal to their average values, and all other variables are equal to 0.

Overall, we apply OLS to six specifications with exports squared, and the same six specifications without a squared term. The first model is the largest one, and it includes 11 exogenous variables. The equation of the big model is as follows:

$$Target = \beta_0 + \beta_1 ExportGrain + \beta_2 ExportGrain^2 + \beta_3 ExportGrainOtherPorts + \beta_4 DieselPrice + \\ + \beta_5 Wheatprice + \beta_6 PPI + \beta_7 ExportCoalandFertilizers + \beta_8 ExportCorn + \beta_9 January2021 + \\ \beta_{10} Harvesting + \beta_{11} NewYear + \epsilon \tag{12}$$

After performing the VIF multicollinearity test on this model, we identify a few variables which are likely to cause a multicollinearity problem. In particular, the VIF test suggests that diesel, wheat prices, and PPI are highly correlated with each other, which can lead to the bad performance of the model. Therefore, the second specification eliminates wheat price from the model, and the fourth specification additionally excludes PPI. The third specification is the same as the second one apart from one change: the variable "Price of diesel" is replaced with PPI. [7] According to the VIF test the combination of variables from Model 4 does not lead to multicollinearity.

The last two specifications - Model 5 and Model 6, exclude the variable "Export corn" since corn is already included in the variable "Export grain". Model 6 differs from Model 5 by the presence of the interaction terms.[8] Interaction term should be included in the model when there is a belief that the effect of some independent variable $X_i$ on the dependent variable $y$ changes depending on the size of another independent variable $X_j$. If this is the case, it makes little sense to interpret the effect of $X_i$ while keeping $X_j$ constant. Interaction term $X_i * X_j$ allows determining the change in the amplitude of the effect of $X_i$ when $X_j$ increases by one. We want to investigate whether the effect of export volume on target changes during New Year time and harvesting when demand is the highest. Therefore, interaction terms of exports with dummy variables "Harvesting" and "New Year" are added to the regression.

Outcomes of all the specifications described above can be viewed in Appendix A in Table 13 (exports squared included) and Table 14 (excluding exports squared), and Table 15 and Table 16 show the same

[7] Variable 'Export other ports' is removed from the fourth specification onward as its coefficient is statistically insignificant. The absence of a relationship is explained by the fact that tipper trucks that are used for grain transportation cannot be used for non-food commodities.

[8] Since no additional variables are added, Model 5 and Model 6 do not suffer from multicollinearity.

results but with Newley-West standard errors, to account for possible serial correlation. In this section, we present the results of three models: Model 3, Model 5, and Model 6.

| | Model 5.1 | Model 5.2 |
|---|---|---|
| Dependent variable | Coefficient | Coefficient |
| Constant | 1151*** | 1061.9*** |
| Export grain | 135.7*** | 134.9*** |
| Export grain$^2$ | -46.9*** | - |
| Price of diesel | 165.0*** | 180.6*** |
| Export grain other ports | -0.66 | 0.41 |
| Exchange rate | 4.34 | 4.34 |
| January 2021 | 142.3*** | 188.7** |
| Harvesting | 46.4 | 29.7 |
| New Year | 110* | 120.1 . |
| R-squared | 0.84 | 0.81 |
| *Note:* | | *p<0.1; **p<0.05; ***p<0.01 |

**Table 3:** Results, Model 5.

| | Model 6.1 | Model 6.2 |
|---|---|---|
| Dependent variable | Coefficient | Coefficient |
| Constant | 1198*** | 1195.6*** |
| Export grain | 175.8*** | 185.0*** |
| Export grain$^2$ | -16.3** | - |
| Price of diesel | 172.5*** | 177.1*** |
| Export grain other ports | 0.83 | 0.35 |
| Exchange rate | 3.1 | 2.82 |
| January 2021 | 100.7 | 112.1 . |
| Harvesting | 83.3** | 95.2** |
| New Year | 136.0* | 143.8* |
| Export grain:New Year | -250.1 | -262.8 . |
| Export grain:Harvesting | -112.6** | -139.5*** |
| R-squared | 0.86 | 0.81 |
| *Note:* | | *p<0.1; **p<0.05; ***p<0.01 |

**Table 4:** Results, Model 6.

As expected, the volume of grain exports has a strong positive effect on the target variable in all specifications, and it is statistically significant at less than 1% significance level.[9] There are some fluctuations in the size of the coefficient under different specifications, but the changes are not economically significant. On average, one standard deviation (=217 thousand tones) increase in export volume increases the target variable by around 135-180 units. In other words, a one thousand tone increase in exports

[9]Statistical significance helps to understand whether the effect is indeed different from zero. The smaller the p-value, the more precise the value of the coefficient is.

increases the average target by 0.62-0.0.83 rubles. The coefficient for squared exports of grain is negative and statistically significant at a 1% level for all the models, apart from Model 6. Diesel also has a strong positive coefficient, as expected. Since the diesel variable was standardized, the interpretation of its coefficient is as follows: a one standard deviation increase in diesel price (=1.6 rubles) increases the target variable by approximately 175 units, keeping all else constant. In other words, if the price of diesel increases by 1 ruble, the target variable increases by around 110 units. Variable "Exchange rate" has a positive effect, as expected, however, the coefficient is not statistically significant. All dummy variables enter the equation with a positive coefficient, however, the statistical significance depends on the specification of the model, as well as whether we use normal or Newey-West standard errors.

In addition, below we present the results of Model 3, where diesel price is replaced by the PPI index. This specification of the model does not show good performance: the fit of the model is much lower ($R^2$ = 0.78 compared to $R^2$ of 0.84 in Model 5.1). Furthermore, the coefficient for the variable "Harvesting" is negative, which is counterintuitive. During the harvesting period, there is an additional demand for trucks, which should increase the freight rate. Therefore, we conclude that the model with PPI is not good for predicting the target.

| | Model 3.1 | Model 3.2 |
|---|---|---|
| Dependent variable | Coefficient | Coefficient |
| Constant | -173.5 | -411.3 |
| Export grain | 152.0*** | 153.3*** |
| Export grain$^2$ | -49.4*** | - |
| PPI | 9.08*** | 10.0*** |
| Export grain other ports | -1.08 | 2.49 |
| Exchange rate | 4.62 | 4.51 |
| Export corn | -0.86** | -0.94* |
| January 2021 | 81.9 | 129.8 . |
| Harvesting | -25.6 | -49.2 |
| New Year | 211.9* | 238.0** |
| R-squared | 0.78 | 0.74 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 | |

**Table 5:** Results. Model 3.

After models with different sets of exogenous variables are trained, their performance is measured with Cross-Validation. K-fold Cross-Validation is a process of out-of-sample testing. First, it partitions the initial sample into K equally sized non-intersecting subsets. At each iteration, one of these subsets is held out, and the model is trained on the remaining data. The held-out data is used to test the performance of the model. The process is repeated K times, and a new subset is held out each time.

The average of the test error is then calculated as a measure of the performance. Figure 7 below illustrates the process of 3-fold Cross-Validation on a sample space of 9 observations with each circle being a separate observation.

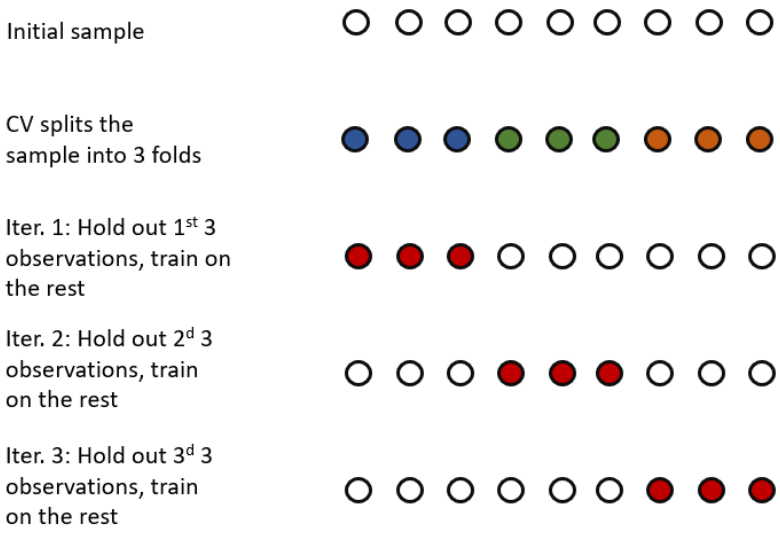



**Figure 7:** Cross Validation process

The models are evaluated based on three different metrics: Mean Squared Error (MSE), R-squared, and Mean Absolute Error:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{i=n}\epsilon_i^2, \quad \text{R}^2 = 1 - \frac{\sum_{i=1}^{i=n}\epsilon_i^2}{\text{Var}(y)}, \quad \text{MAE} = \frac{1}{n}\sum_{i=1}^{i=n}|\epsilon_i|, \tag{13}$$

where $\epsilon_i$ is the error associated with each test observation from the test set of size n, and Var(y) is the variance of the out-of-sample true output.

Results of 10-fold Cross-Validation in Table 6 suggest that the first model has the best performance out of all independently of whether the polynomial term is included or not. However, this specification suffers from a multicollinearity problem (as mentioned earlier); therefore, it might result in misleading inference. The second best models are Model 4 and Model 5, they have similar out-of-sample performance, and can be considered to be the most suitable for the application. The inclusion of the interaction term in Model 6 slightly impairs the out-of-sample performance. This suggests that interaction terms results in overfitting, and, therefore, should be excluded from the model. Lastly, the performance of the model improves in all specifications when the variable for export squared is added

to the equation. Overall, we can conclude that Model 5.1 is the best for our problem. It can explain 84% of the variation in the target variable, and the average error rate is around 5.3%.

| With Polynomial 2 term | | | | | | |
|---|---|---|---|---|---|---|
| | Model 1.1 | Model 2.1 | Model 3.1 | Model 4.1 | Model 5.1 | Model 6.1 |
| MSE | 88.5 | 98.4 | 120.5 | 95.6 | 96.0 | 105.7 |
| R-squared | 0.86 | 0.83 | 0.76 | 0.84 | 0.84 | 0.82 |
| MAE | 72.7 | 79.3 | 94.5 | 76.9 | 78.2 | 78.8 |
| Without Polynomial 2 term | | | | | | |
| | Model 1.2 | Model 2.2 | Model 3.2 | Model 4.2 | Model 5.2 | Model 6.2 |
| MSE | 91.2 | 106 | 129.5 | 102.9 | 103.1 | 113.2 |
| R-squared | 0.85 | 0.81 | 0.74 | 0.82 | 0.82 | 0.79 |
| MAE | 74.2 | 86.8 | 103.7 | 85.2 | 85.0 | 86.3 |

**Table 6:** 10-fold Cross Validation results

### 8.2 Quantile regression

OLS gives us a good estimate of the effect of exogenous variables on the conditional mean of the variable we are interested in. However, it assumes that the elasticity of the response variable is identical across different percentiles, and this assumption often fails. Unlike OLS, quantile regression investigates conditional percentiles of the response variable, and in such a way allows to capture the non-constant effect of independent variables on the conditional values of dependent variables (Koenker, Hallock, 2001). This property makes it a much more accurate method in the presence of outliers and allows it to deal with nonlinear relationships between variables.

In terms of methodology, the main difference with OLS is that instead of minimizing the sum of squared residuals, quantile regression minimizes median absolute deviation. Mathematically, the equation of MAD to estimate the effect on $\tau$th percentile is:

$$\text{MAD} = \sum_{1}^{n} \rho_\tau(y_i - (\beta_0(\tau) + \beta_1 x(\tau)_{i1} + ... + \beta_p(\tau)x_{ip})), \quad i = 1, ..., n \tag{14}$$

where the function of $\rho$ is to apply a certain weight to the error, which depends on the percentile and error sign. Giver error $\epsilon$, $\rho$ is determined as follows:

$$\rho(\epsilon) = \tau \max(\epsilon, 0) + (1 - \tau) \max(-\epsilon, 0) \tag{15}$$

For each specified value of $\tau$ quantile regression will provide a different set of coefficients unless the

elasticity is perfectly uniform across different percentiles.

We estimate the effect of a small set of exogenous variables (Model 5.1) on freight rate using quantile regression. Examination of different percentiles of the regression allows us to understand how the elasticity of the target variable changes for different levels of the target. For example, freight rates could be less sensitive to changes in exogenous variables at the extremes. For the lower rate, there are certain fixed costs (labor costs, amortization cost, fixed cost of handling the transaction for the transportation company like paperwork and taxes), which create a lower bound below which the company will never supply services. On the other hand, there could also be an upper bound for the rate which the supplier of transport will be able to charge. No matter how high the demand is, a customer could be more likely to wait and pay the backlog cost rather than pay an abnormally high price for the transportation.

The results of quantile regression are presented in Table 7. The coefficient of 'Export grain$^2$' is negative, so there is a diminishing effect of exports on the target variable at every quantile of the data. For example, the increase of one ton of exports from 10 to 11 tones will increase freight stronger than an increase from 20 to 21 tons. Results suggest that the diminishing effect is the strongest for the 60% quantile. At the same time, we observe a reduction in the size of the effect of the linear term of exports on freight rate for higher percentiles.

Similarly, diesel prices constitute less to the formation of transportation costs as we move up the quantile. This result is intuitive: when demand for trucks is low, transportation companies will be willing to work at their marginal cost, which is composed of diesel price, labor, and some fixed operational expenses. However, during periods of high demand transportation companies will be less sensitive to changes in the prices of diesel since the mark-up over the marginal cost is sufficient to cover some small increase in cost.[10]

Interestingly, the coefficient of the variable 'Export grain other ports' flips sign and gains statistical significance as we move to higher quantiles. This could be explained as follows: when demand is high, trucks which previously moved to the ports in the east will be more likely to travel to Novorossiysk and Rostov as well. This can reduce the freight rate.

[10]Partial effect plot of diesel (Figure 17 in Appendix B) on target does not reveal a clear non-linear relationship.

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | max_tar_new | | | | |
| | 20% | 40% | 60% | 80% | 90% |
| Export grain | 91.735*** | 135.330*** | 130.595*** | 112.736*** | 114.761*** |
| | (34.129) | (19.384) | (19.714) | (18.723) | (18.983) |
| Export grain$^2$ | −31.294* | −57.858*** | −60.272*** | −37.072*** | −49.293*** |
| | (18.557) | (13.600) | (13.582) | (13.794) | (14.870) |
| Diesel price | 198.401*** | 166.571*** | 155.800*** | 153.516*** | 147.810*** |
| | (18.704) | (15.689) | (13.453) | (10.015) | (11.865) |
| Exchange rate | 6.667* | 2.935 | 3.970 | 5.509 | 2.987 |
| | (3.888) | (4.067) | (3.203) | (4.591) | (3.803) |
| Export grain other ports | 1.546 | −1.681 | −2.772*** | −3.452*** | −3.918*** |
| | (1.363) | (1.181) | (1.030) | (0.657) | (0.702) |
| January 2021 | 172.466* | 138.752** | 100.812** | 66.388** | 25.492 |
| | (88.214) | (64.799) | (49.442) | (31.622) | (33.465) |
| Harvesting | 149.224* | 36.550 | 21.523 | −5.427 | −18.231 |
| | (87.261) | (45.194) | (32.424) | (23.115) | (26.104) |
| New Year | −133.586 | 187.134 | 115.490 | 82.730 | 49.491 |
| | (218.766) | (176.991) | (133.256) | (98.396) | (76.129) |
| Constant | 754.423** | 1,306.283*** | 1,327.770*** | 1,274.151*** | 1,518.728*** |
| | (300.056) | (320.797) | (246.502) | (331.914) | (274.843) |
| Observations | 96 | 96 | 96 | 96 | 96 |
| *Note:* | | | | | *p<0.1; **p<0.05; ***p<0.01 |

**Table 7:** Quantile regression results

# 9 ARIMA

## 9.1 Background information

Section 7 demonstrates that OLS provides good results when the aim is to predict the freight rate. However, OLS treats the data as a random sample rather than time series data. We have to impose the assumption that there is no serial relationship within data across time. ARIMA on the other hand is designed to model time series data, and, therefore, can capture time dependencies between series of data points. In this section, we apply the ARIMA framework to the same data set and compare its performance to OLS.

ARIMA requires the specification of three parameters - p, d, and q: p stands for the number of AR terms, or how many lags are used to predict the current value of the time series; q specifies the order of

MA terms, i.e. the number of lagged error terms; d specifies how many times series has to be differenced to convert a non-stationary time series into a stationary one. Time series is stationary if its statistical properties, such as mean, and standard deviation, do not change over time. ARIMA imposes a strict assumption of stationary, so whenever we have non-stationary time series, the order of d must be greater than 0. Other parameters can be specified within the model by testing the relative performance of ARIMA under various parameter specifications. This can be done manually, or by applying a build-in function "autoarima" which fits different combinations of several lags, lagged errors, and differencing terms, and returns the best model based on AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion) values. Both AIC and BIC aim to select the best model by making a trade-off between complexity (the size of the model in terms of the number of parameters), and its goodness-of-fit (how well the model fits the training sample). Because higher complexity implies a higher chance of over-fitting the model, and both criteria penalize the size of the model, asymptotically they are equivalent to Cross-Validation. However, in small samples the convergence is not guaranteed, therefore, AIC and BIC might fail to choose the best model. In this section, we test both the model returned by the "autoarima" function based on AIC, as well as try a set of different ARIMA models manually. To measure the performance, first, we train the model on the training set (first 80 observations) and then use the test set (last 16 observations) to obtain the Mean Absolute Error of the ARIMA(p, d, q) forecast.

### 9.2 Results

After applying Autoarima to our train set the function returned (p, d, q) = (0,1,0) as the best model. The result implies that the series is non-stationary, and its fluctuations can only be explained by a process called random walk: the movement of one time differenced time series does not exhibit any particular pattern, but is rather. The results of different ARIMA specifications in Table 17 in Appendix A support the inability of ARIMA to model freight rates in Russia. The MAE is much higher when compared to OLS. Furthermore, when we plot foretasted values, they look like a horizontal line, failing to capture any fluctuation in true freight rate (Figure 8). It is important to note that unlike OLS, which we tested with Cross-Validation, ARIMA is tested only on the last 16 observations. Cross-Validation is generally a more accurate test metric, as it minimizes the possibility to obtain a certain level of error only because of a fortunate split of data.

We can explain the results as follows. ARIMA is good at capturing seasonal and general trends, or

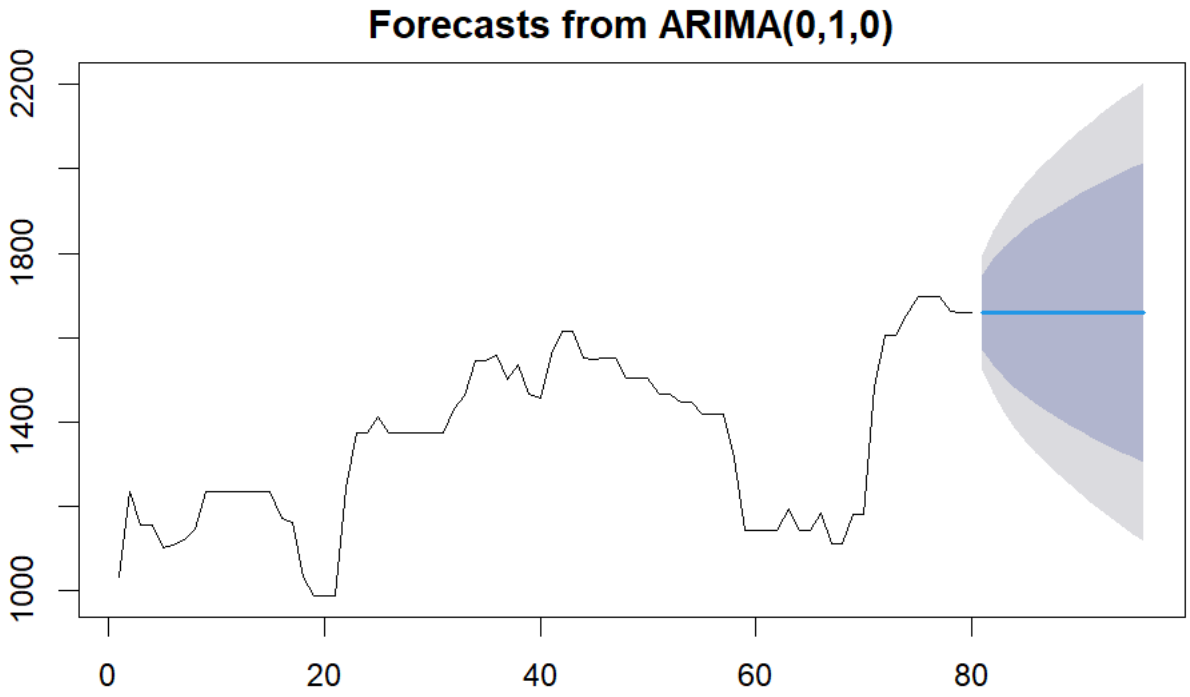



**Figure 8:** ARIMA (0,1,0)

when series show a pattern over time. However, ARIMA will most likely fail to identify the outliers, especially if they are caused by exogenous shocks. It is expected that freight rates show some degree of seasonality with rates increasing during the harvesting season. However, these fluctuations are highly dependent on the volume of grain exports, and the factors that determine them do not necessarily show a clear pattern. For example, tariffs and quotas change on weekly basis, and they affect suppliers' decision to sell grain now, or wait for a lower tariff and sell later, irrespective of the harvesting season. Diesel price is another crucial determinant in cost formation in the transportation industry. Russia has a national policy to regulate diesel prices artificially; therefore, the price is not likely to show any time or seasonal patterns. Overall, we can conclude that the Russian freight market is mainly driven by exogenous factors and shocks, which time series models like ARIMA are not able to capture.

# 10 Data quality

Data quantity and quality are two important factors when designing a forecasting model. We can define data quantity as the number of observations available to train and test the model. Data quality is a broader concept, and multiple criteria should be assessed to determine how good the data is:

(i) Accuracy: Data should reflect actual real-world scenarios. If a bias or defect is present in the data, the model that relies on supervised learning will learn the bias, and will most likely provide a false conclusion.

(ii) Completeness: Are all the required measures observable, or the dataset contains a lot of missing values that limit the ability to analyze and use the data? Too many missing values for the essential variables may impose a problem if we cannot replace them, or when the model is poor at dealing with them.

(iii) Consistency and uniqueness: Data should be uniform across networks and applications. The indexes must be unique, when present. For example, we cannot use data on volume which is measured in kilograms and pounds within the same model.

(iv) Validity: the data should be collected in line with specific rules, and must be reconciled in the right format.

**Why quantity is important**

The quantity of observations is an important factor when designing the methodology of the forecasting model. Although some regression techniques like OLS do not require large training samples, they pose strict assumptions on the properties of sampled data. Furthermore, in recent years deep learning techniques like Neural Networks have been gaining more popularity for real business applications due to their ability to capture and learn more sophisticated dependencies in the data. However, because of their complex structure, a significant amount of data is required to train the model. For example, methods like ARIMA, regression trees, and NN may provide wrong results as they can fail to capture certain relationships, like trend and seasonality, produce overfitting or wide confidence interval. Therefore, with short time series data, we have a limited range of techniques to be used for forecasting. At the same time, there is an extensive list of literature that shows that such techniques give accurate forecasts, although empirical evidence of their superiority over linear regression is contradictory. Paper by Altay and Satman finds evidence of the better performance of ANN compared to regression methods for forecasting the stock market (Altay, Satman, 2005). Better neural network performance is observed for

the prediction of the direction of stock movements, while out-of-sample test statistics are the same across methods. Similarly, artificial neural networks outperformed both decision trees and linear regression when forecasting the academic performance of students (Ibrahim, Rusli, 2007). In contrast, Kim compares the performance, measured by RMSE, of ANN, decision tree, and linear regression with 60 simulated examples (Kim, 2008). After testing different numbers and combinations of the exogenous variables, the author concludes that for continuous variables linear regression outperforms the other two methods. ANN is better when more than one categorical variable is used. Overall, while some papers suggest the superiority of ANN over linear regression techniques, this conclusion is not uniform. Therefore, the benefit of using a different methodology can be limited when the cost of extracting more data is high.

OLS is unbiased and consistent when the assumptions from section 7.1 hold for the sampled dataset. Consistency implies that the estimation technique will estimate true beta as the sample size goes to infinity. This property also means that the probability of obtaining true coefficients gets closer to one and the variance of the estimator converges to 0 as the sample size increases. This is because unbiasedness only guarantees that in-sample $\tilde{\beta}_i$ is equal to true population $\beta_i$ in expectation: if we drew multiple samples and calculated their mean, we would get a true model. However, in reality, we have to deal with limited or just one sample, and a larger sample size suggests a higher chance of obtaining a true model, which is crucial for predictions. Furthermore, supervised learning techniques use historical data to capture trends and relationships within the model. Sometimes, if the timeline is too short, the model might not capture essential for the prediction events.

To test whether increasing the sample size increases the accuracy of the model, we check the performance of the small specification of the OLS model for a different number of observations: 30, 50, 70, and 90. The in-sample and out-of-sample performances are measured and compared, and the results are presented in the table. We use an in-sample R-squared to test the goodness-of-fit and then look at the Cross-Validation results to measure the out-of-sample performance. We present the results in Table 8. Although the goodness-of-fit decreases as sample size increases, both mean squared error and mean absolute error decrease as n gets larger, supporting the consistency argument. Overall, with the average target in the sample having a value of 1417, the Mean Absolute Error with 30 data points will be around 6.8%, and the error decreases to 5.4% when 90 observations are used to train the model. The results suggest that an increase in sample space does not significantly improve the performance of the OLS model. Therefore, if the costs of obtaining more data are large, the added value of the performance improvement of the OLS will most likely not cover them.

| | Sample size | | | |
|---|---|---|---|---|
| Performance measure | 30 | 50 | 70 | 90 |
| R-squared train | 0.83 | 0.85 | 0.83 | 0.82 |
| R-squared test | 0.86 | 0.82 | 0.81 | 0.83 |
| MSE | 113 | 100 | 103 | 97 |
| MAE | 97 | 87 | 84 | 76 |

**Table 8:** 10-fold Cross Validation results

# 11 Risk protection

Section 8 demonstrates that with the available data we are able to explain around 83% of the variation in the target variable. In other words, our model can provide a good prediction of the evolution of the freight rates in the Russian tipper truck market. However, the model is trained on the historical data, and in the future other risks may occur which were not present in the past. Furthermore, some relevant variables could be omitted from the model because of a lack of data. Figure 9 below plots predicted versus true values along with the 90% prediction interval. 90% prediction interval provides upper and lower bounds for the possible value of freight rate such that the probability of obtaining the true value within the interval is equal to 0.9. Majority of predicted values on Figure 9 fall within the prediction interval. However, they do not exactly coincide with the true target values: some external factors unaccounted by the model can make the freight rate deviate from the average trend forecasted by the model. In order to improve the accuracy of the prediction, we conduct a risk assessment. Risk assessment is a process of evaluation of risk imposed by different scenarios, identification of the probability of occurrence of those scenarios, and the magnitude of their effect on the metrics of interest. The final goal is to position with higher precision the value of the target within the prediction interval suggested by the model, given external risk factors.

## 11.1 Methodology for risk assessment used in our thesis

1. Identify the factors that can create upward or downward pressure on the truck freight rate.

2. Determine the relative importance of the effect of each factor by using expert knowledge and previous experience of people working in the field. Relative importance will not change in the short run.

3. Survey the experts on how (in which direction and how strong) a certain factor is expected to affect transportation cost. This survey has to be conducted for every forecasting period.

4. Combine the relative importance of the effect and the expected direction and strength of the effect to

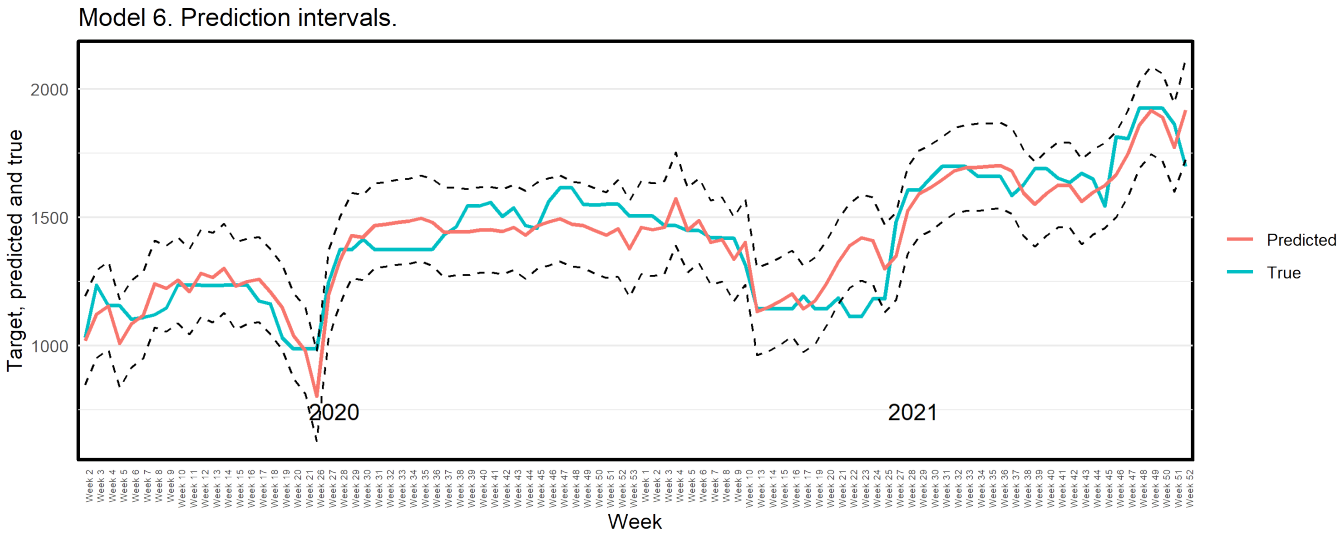



**Figure 9:** Model 6. 90% prediction interval

derive the conclusion on the risk protection in a given period. Derive the predicted value and prediction interval from the model. Given the risk protection identified in step 4, put a value above, on, or below the predicted value within the prediction interval.

Next, we will explain the methodology in more detail.

**Step 1. Relevant factors**

To simplify the identification of the strength of the risk for the experts we identify a few major relevant factors, and then further split the main factors into a few sub-factors.

***Factor 1: Governmental policy and Geopolitical situation***:

- **Inflation** or expectations of inflation excluding fuel prices. May include labor costs, maintenance, and service cost, which constitute a large share of the transportation costs.
- **Tariffs and quotas** on grain export. Elimination or reduction of tariffs on grain exports for a certain period of time will be followed by high export volumes, and, therefore, demand for trucks.
- **Taxes/subsidies/regulations**: affect the costs incurred by the shippers.
- **State of economy**: may affect transportation costs through labor costs, and profit margins.
- **International sanctions or blocks:** the risk of transportation at the Black sea or international bans can affect the possibility of delivering grain by cargo ships.

***Factor 2. Weather:***

- **Weather on the roads**: snow and rain might lead to longer travel time, the necessity to take a different route, and difficulty to get to farms.
- **Rain**: affects the harvesting decision of farmers. Grain cannot be harvested during rainy periods.
- **Weather at ports**: storms delay the departure of cargo ships.

***Factor 3. Harvest:***

- **Volume of harvest** directly affects the demand for trucks, both through the export of grain and through the need to use trucks for harvesting itself.

***Factor 4. Supply:***

- **Truck availability**: sanctions may create problems with truck replacement and maintenance.
- **Driver supply**: availability of drivers affects labor costs.
- **Productivity of trucks**: includes waiting/traveling/loading times, distance traveled.

**Step 2. Relative importance table**

After identifying four main factors and corresponding sub-factors, we conduct a survey in order to determine the relative importance of a given risk measure. One survey is conducted to determine the relative factor importance, and then for each factor, we identify the relative importance of the sub-factor within the group. Table 9 and Table 10 represent the survey results. The last column of Table 10 presents the final weight of each sub-factor for the total effect, which is calculated by multiplying "within the group" weight $w$ with the factor weight $W$ to which given sub-factor belongs.

| Factor | Importance rank | Weight |
|---|---|---|
| Supply | 1 | 0.36 |
| Harvest | 2 | 0.33 |
| Weather | 4 | 0.19 |
| Governmental policy and Geopolitics | 3 | 0.11 |

**Table 9:** Factor relative importance

| **Governmental policy and Geopolitics** | | | |
|---|---|---|---|
| Sub-factor | Importance rank | Weight within the group | Total weight, T |
| Inflation | 4 | 0.13 | 0.04 |
| Tariffs and quotas | 1 | 0.33 | 0.05 |
| Taxes, subsidies and regulations | 3 | 0.2 | 0.04 |
| State of economy | 2 | 0.27 | 0.04 |
| International bans | 5 | 0.07 | 0.03 |
| **Weather conditions** | | | |
| Ports | 3 | 0.17 | 0.02 |
| Road | 2 | 0.33 | 0.04 |
| Field | 1 | 0.13 | 0.05 |
| **Harvest** | | | |
| Harvest volume | - | 1 | 0.33 |
| **Supply** | | | |
| Trucks availability | 1 | 0.5 | 0.26 |
| Driver availability | 3 | 0.17 | 0.05 |
| Productivity of trucks | 2 | 0.33 | 0.05 |

**Table 10:** Sub-factor within group relative importance

### Step 3. Strength and direction of effect

In the next step, we conduct the survey on how the given sub-factor is expected to affect the freight rate for the period of interest. For each sub-factor, we ask whether the expert expects it to affect the freight rate very downwards, downwards, neutrally, upwards, or very upwards. The qualitative values are converted into a simple scale from -2 to 2 presented in Table 11.

| Very down | Down | Neutral | Up | Very up |
|---|---|---|---|---|
| -2 | -1 | 0 | 1 | 2 |

**Table 11:** Caption

### Step 4. Determine the protection level

The total effect of the risk factors on the freight rate is calculated with weighted average formula:

$$Protection_t = \sum_j T_j * e_{j,t}, \qquad j \in I \tag{16}$$

where $T_j$ represent the overall weight of a sub-factor j, $I$ is the set of all the sub-factors considered in the methodology, and $e_{j,t}$ is the expected strength of the effect of factor j in time period t presented in Table 11.

Lastly, we collect the predicted value, lower and upper bounds of the target variable, and put them

on a scale (left-hand side on Figure 10). The scale for risk protection is on the right-hand side of the Figure 10. Given the risk protection derived above, we plot it on a scale. The corresponding value of the target value within the confidence interval is the risk-protected predicted target.

Mathematically, the value is calculated with the equation:

$$\text{Bias} = \text{Protection}/2 * (\text{Upper bound} - \text{Predicted value}) \tag{17}$$

when bias is positive, and:

$$\text{Bias} = \text{Protection}/2 * (\text{Predicted value} - \text{Lower bound}) \tag{18}$$

when the bias is negative. The final forecast of the target variable is calculated by adding the predicted value from the model and the bias together.

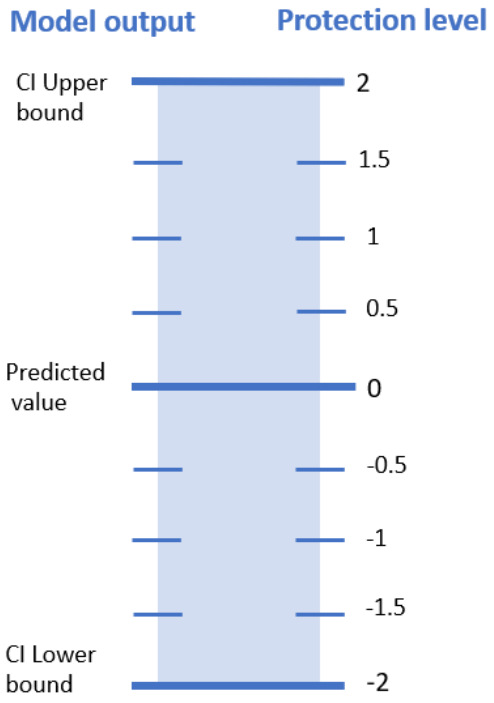



**Figure 10:** Risk protection scale

## 11.2 From target to freight rate

As we explained in the earlier sections, the target value is created to represent the fluctuations in the freight rate on all routes in the region of interest every week. Therefore, the risk protected predicted

value above is the base prediction for the average distance in the sample. To derive the prediction of the route-specific freight rate we have to reverse the steps in section 5.3:

1. Collect the data on the difference to base value from Step 2 of section 5.3. The difference is calculated for every possible distance in the range from 40 to 1000 km.

2. Given the risk-adjusted predicted target variable value p, the freight cost per ton for distance i ($cost_i$) is calculated using the equation:

$$cost_i = p * (1 + perc\_dif_i) \tag{19}$$

for any distance i measured in kilometers.

Lastly, we want to create a user-friendly application that would be able to give an estimation of the freight rate for a given origin and destination location. Therefore, we create an excel spreadsheet where the user is guided to input the following data:

1. For each week of interest input the data from the model, which includes predicted value, upper and lower bound.

2. Input the expected size and direction of the sub-factors effects.

3. Input the desired origin, destination zip codes, and week number. In addition, the user is allowed to input route-specific protection.

4. Collect the output for the risk-adjusted estimated freight rate for a given week and route.

Furthermore, the user is asked to maintain the application by continuously updating the origin-destination data, which is not currently present in the data set. If a certain origin-destination location is not found, or the information of model output is absent, the error will show. In such a way the user will be guided to update the information needed for the prediction of the freight rate.

The goal of the methodology above is to improve the precision of the freight rate prediction by taking into account expert knowledge. However, whether the risk assessment process increases the prediction accuracy has to be tested in the future. For example, the importance of some factors can be under of overestimated, or other relevant risks can be omitted from the model. Therefore, it is crucial to record the output of the methodology at each stage and then backtest the performance of the methodology in order to identify mismatches and inaccuracies.

# 12 Conclusion

This thesis develops a freight rate forecasting model for tipper trucks in Russia using the historic data on transportation costs collected by Cargill. Accurate knowledge of the future freight rate is an important determinant of Cargill's profitability. If trasnoprtaion costs are overestimated, Cargill will bet a lower price for grain than its competitors and will lose the contract for grain. On the other hand, when costs are underestimated, Cargill will pay a higher price than the budget allows and will bear losses.

In this thesis, we investigate through a thorough literature review the possibility to apply different statistical and machine learning techniques to predict the freight rate given the availability of data. Multiple linear regression with the OLS estimation technique is chosen as the best fitting methodology, and it is then compared to ARIMA time series model. We test the performance of the MLR with different sets of inputs based on the out-of-sample $R^2$, MAE, and MSE. The results suggest that the volume of grain exports, diesel prices, exchange rate, and a set of dummy variables explain around 85% of the variation in freight rate. At the same time, we conclude that given a small data set and the fact that freight rate fluctuations are formed through external factors, ARIMA fails to produce a good forecast.

This study adds to the existing research in a few different ways. First of all, the study presents the methodology to convert contract (or delivery-based) data set into target variable which would represent the average weekly freight rate in the region of interest. In addition, the method deals with insufficient variation in origin-destination combinations on the weekly basis. Secondly, the thesis shows that ARIMA might be redundant to forecasting time series when the outcome variable is a combination of fluctuations in the predictor. In this case, the results of ARIMA suggest that the outcome variable fluctuation is just a random walk. Finally, we design a simple and user-friendly methodology which potentially improves the prediction accuracy through the risk assessment process. We select the main external factors not accounted for by the model and then quantify the qualitative expectations of their effect on freight rate. However, the risk assessment methodology has to be backtested to identify whether the process indeed enhances the prediction power of the model.

However, the study has some limitations. The model is designed to forecast the weekly average freight rate for a certain base distance. Although we can estimate the freight for any distance using the average historical relationship between cost per tone and distance, the method does not account for route-specific characteristics. Therefore, the user has to manually adjust the predicted value using his or her experience from the past. Secondly, the period of the prediction is only two years, which might

not completely capture the average effect of the predictors on the outcome variable. These factors together may lead to higher forecasting errors.

## 12.1 Further research

When designing the model we faced some limitations like low variability of data and prediction of the freight for the average distance in the sample. It might be interesting to extend the research with the following approach:

- Extend the model to capture the route-specific characteristics. If it is impossible to incorporate it directly into the model, research why the freight rate on a certain route differs from another for the same driving distance.
- The model designed in this study can only be applied to the deliveries of grain to ports in Rostov and Novorossiysk. It would be interesting to see whether a given model can be extended to the deliveries to the other major location - Novoannynskiy, on different truck types and countries.
- In this study we assume that tipper freight rates are independent across regions. However, one can investigate whether the freight rates in adjacent regions are correlated.
- So far the model uses the data on transportation rates collected by Cargill. Aggregating Cargill's records with external data can provide more accurate insights about the average development of market freight rates.

# 13 Appendix

## 13.1 Appendix A. Tables.

| Variable name | Explanation |
|---|---|
| exp_grain_3_ma_st | Moving average of total export of grain from 3 ports in the area of Novorossiysk, standardized. |
| exp_grain_3_ma_st_sq | Moving average of total export of grain from 3 ports in the area of Novorossiysk squared, standardized. |
| exp_grain_east_v_ma | Moving average of total export of grain from ports on Caspian sea and Volgograd region, standardized. |
| p_diesel$_s t$ | Retail price of diesel, standardized. |
| p_wheat | Global price of wheat. |
| ppi | Producer price index. |
| exp_coal_5_fert | Total exports of coal and fertilizers from ports near Novorossiysk and Rostov. |
| exp_corn_3 | Total export of corn from 3 ports in the area of Novorossiysk |
| jan_21 | Dummy for January 2021. |
| harvesting | Dummy for harvesting season (from July till October). |
| new_year | Dummy for New Year. |

**Table 12:** Independent variables explanation

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Target variable | | | | | |
| | Model 1.1 | Model 2.1 | Model 3.1 | Model 4.1 | Model 5.1 | Model 6.1 |
| exp_grain_3_ma_st | 115.523*** | 132.567*** | 151.974*** | 132.572*** | 135.697*** | 175.831*** |
| | (13.837) | (14.733) | (17.599) | (14.004) | (14.016) | (18.078) |
| exp_grain_3_ma_st_sq | −41.680*** | −46.413*** | −49.419*** | −46.708*** | −46.913*** | −16.328 |
| | (10.296) | (11.343) | (13.824) | (11.096) | (11.207) | (14.255) |
| p_diesel_st | 160.178*** | 162.999*** | | 164.265*** | 165.025*** | 172.494*** |
| | (22.596) | (25.015) | | (11.613) | (11.720) | (11.316) |
| exch_r | 2.751 | 5.374* | 4.626 | 5.438* | 4.339 | 3.138 |
| | (2.903) | (3.148) | (3.838) | (3.102) | (3.061) | (2.932) |
| exp_grain_east_v_ma | −0.727 | −0.358 | 1.080 | −0.414 | −0.656 | 0.008 |
| | (0.803) | (0.885) | (1.045) | (0.759) | (0.752) | (0.738) |
| exp_coal_5_fert | −0.004 | 0.011 | −0.012 | | | |
| | (0.042) | (0.046) | (0.056) | | | |
| ppi | −6.511*** | 0.005 | 9.082*** | | | |
| | (2.063) | (1.618) | (1.004) | | | |
| exp_corn_3 | −0.349 | −0.481 | −0.862** | −0.464 | | |
| | (0.268) | (0.295) | (0.352) | (0.280) | | |
| p_wheat | 1.358*** | | | | | |
| | (0.303) | | | | | |
| jan_21 | 91.918* | 131.499** | 81.883 | 129.445** | 142.293*** | 100.690 |
| | (49.526) | (53.968) | (65.168) | (52.477) | (52.422) | (64.626) |
| harvesting | 88.869*** | 38.468 | −25.580 | 38.735 | 46.403 | 88.302*** |
| | (31.129) | (32.139) | (37.323) | (30.480) | (30.429) | (31.758) |
| new_year | 138.482** | 103.570 | 211.916*** | 99.709* | 110.126* | 135.974** |
| | (57.659) | (63.269) | (74.459) | (59.065) | (59.318) | (56.741) |
| exp_grain_3_ma_st:jan_21 | | | | | | −250.094 |
| | | | | | | (157.145) |
| exp_grain_3_ma_st:harvesting | | | | | | −112.553*** |
| | | | | | | (35.312) |
| Constant | 1,320.111*** | 1,065.644*** | −173.519 | 1,071.849*** | 1,151.021*** | 1,198.234*** |
| | (306.528) | (333.580) | (334.290) | (238.932) | (236.462) | (225.135) |
| Observations | 96 | 96 | 96 | 96 | 96 | 96 |
| $R^2$ | 0.878 | 0.849 | 0.772 | 0.849 | 0.844 | 0.863 |
| Adjusted $R^2$ | 0.861 | 0.829 | 0.746 | 0.833 | 0.830 | 0.847 |
| Residual Std. Error | 88.003 | 97.462 | 118.878 | 96.356 | 97.322 | 92.204 |
| F Statistic | 49.872*** | 42.874*** | 28.845*** | 53.603*** | 58.776*** | 53.577*** |
| *Note:* | | | | | | *p<0.1; **p<0.05; ***p<0.01 |

**Table 13:** Results. Export grain squared.

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Target variable | | | | | |
| | Model 1.2 | Model 2.2 | Model 3.2 | Model 4.2 | Model 5.2 | Model 6.2 |
| exp_grain_3_ma_st | 114.576*** | 133.283*** | 153.266*** | 131.643*** | 134.852*** | 185.033*** |
| | (15.049) | (16.038) | (18.762) | (15.289) | (15.274) | (16.224) |
| exp_grain_east_v_ma | 0.382 | 0.928 | 2.491** | 0.652 | 0.408 | 0.352 |
| | (0.821) | (0.900) | (1.031) | (0.781) | (0.771) | (0.675) |
| p_diesel_st | 163.615*** | 167.161*** | | 179.711*** | 180.563*** | 177.106*** |
| | (24.561) | (27.210) | | (12.030) | (12.116) | (10.594) |
| exch_r | 2.424 | 5.281 | 4.506 | 5.470 | 4.340 | 2.818 |
| | (3.156) | (3.427) | (4.092) | (3.387) | (3.336) | (2.924) |
| p_wheat | 1.484*** | | | | | |
| | (0.328) | | | | | |
| ppi | −6.531*** | 0.664 | 10.032*** | | | |
| | (2.244) | (1.753) | (1.032) | | | |
| exp_coal_5_fert | 0.006 | 0.024 | 0.001 | | | |
| | (0.045) | (0.050) | (0.060) | | | |
| exp_corn_3 | −0.391 | −0.542* | −0.937** | −0.477 | | |
| | (0.291) | (0.320) | (0.375) | (0.305) | | |
| jan_21 | 129.315** | 177.726*** | 129.834* | 175.325*** | 188.745*** | 112.102* |
| | (52.926) | (57.452) | (68.000) | (56.048) | (55.838) | (63.969) |
| harvesting | 75.353** | 17.985 | −49.170 | 21.892 | 29.702 | 95.245*** |
| | (33.665) | (34.562) | (39.170) | (32.992) | (32.877) | (31.231) |
| new_year | 160.978** | 125.237* | 237.975*** | 109.337* | 120.093* | 143.854** |
| | (62.426) | (68.638) | (79.014) | (64.442) | (64.596) | (56.424) |
| exp_grain_3_ma_st:jan_21 | | | | | | −262.794* |
| | | | | | | (157.038) |
| exp_grain_3_ma_st:july | | | | | | −139.534*** |
| | | | | | | (26.355) |
| Constant | 1,173.760*** | 874.321** | −411.319 | 980.914*** | 1,061.923*** | 1,195.664*** |
| | (331.095) | (359.577) | (349.323) | (259.813) | (256.661) | (225.532) |
| Observations | 96 | 96 | 96 | 96 | 96 | 96 |
| $R^2$ | 0.854 | 0.819 | 0.738 | 0.818 | 0.812 | 0.861 |
| Adjusted $R^2$ | 0.835 | 0.797 | 0.711 | 0.801 | 0.797 | 0.846 |
| Residual Std. Error | 95.725 | 106.105 | 126.759 | 105.208 | 106.066 | 92.372 |
| F Statistic | 44.723*** | 38.378*** | 26.940*** | 48.726*** | 54.446*** | 59.170*** |
| *Note:* | | | | | | *p<0.1; **p<0.05; ***p<0.01 |

**Table 14:** Results. Without Export grain squared.

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Target variable | | | | | |
| | Model 1.1 | Model 2.1 | Model 3.1 | Model 4.1 | Model 5.1 | Model 6.1 |
| exp_grain_3_ma_st | 115.523*** | 132.567*** | 151.974*** | 132.572*** | 135.697*** | 175.831*** |
| | (21.885) | (16.563) | (21.607) | (18.506) | (19.458) | (36.355) |
| exp_grain_3_ma_st_sq | −41.680*** | −46.413*** | −49.419** | −46.708*** | −46.913*** | −16.328 |
| | (15.417) | (15.534) | (19.034) | (16.824) | (16.995) | (34.795) |
| p_diesel_st | 160.178*** | 162.999*** | | 164.265*** | 165.025*** | 172.494*** |
| | (27.056) | (36.240) | | (18.614) | (19.067) | (22.546) |
| exch_r | 2.751 | 5.374 | 4.626 | 5.438* | 4.339 | 3.138 |
| | (2.457) | (3.274) | (3.680) | (3.248) | (3.127) | (3.507) |
| exp_grain_east_v_ma | −0.727 | −0.358 | 1.080 | −0.414 | −0.656 | 0.008 |
| | (1.693) | (2.074) | (3.085) | (2.284) | (2.211) | (2.334) |
| exp_coal_5_fert | −0.004 | 0.011 | −0.012 | | | |
| | (0.031) | (0.030) | (0.036) | | | |
| ppi | −6.511* | 0.005 | 9.082*** | | | |
| | (3.294) | (2.384) | (2.228) | | | |
| exp_corn_3 | −0.349* | −0.481*** | −0.862*** | −0.464*** | | |
| | (0.183) | (0.173) | (0.248) | (0.143) | | |
| p_wheat | 1.358*** | | | | | |
| | (0.429) | | | | | |
| jan_21 | 91.918** | 131.499** | 81.883 | 129.445** | 142.293** | 100.690* |
| | (45.991) | (59.201) | (77.323) | (61.657) | (63.236) | (52.303) |
| harvesting | 88.869 | 38.468 | −25.580 | 38.735 | 46.403 | 88.302 |
| | (67.178) | (60.301) | (71.430) | (57.459) | (58.922) | (56.521) |
| new_year | 138.482** | 103.570 | 211.916** | 99.709 | 110.126 | 135.974 |
| | (56.975) | (94.070) | (81.790) | (96.715) | (93.577) | (108.856) |
| exp_grain_3_ma_st:jan_21 | | | | | | −250.094*** |
| | | | | | | (71.743) |
| exp_grain_3_ma_st:july | | | | | | −112.553 |
| | | | | | | (86.886) |
| Constant | 1,320.111*** | 1,065.644*** | −173.519 | 1,071.849*** | 1,151.021*** | 1,198.234*** |
| | (316.984) | (378.421) | (595.959) | (324.795) | (306.626) | (285.073) |
| *Note:* | | | | | | *p<0.1; **p<0.05; ***p<0.01 |

**Table 15:** With squared exports. Newey-West standard errors.

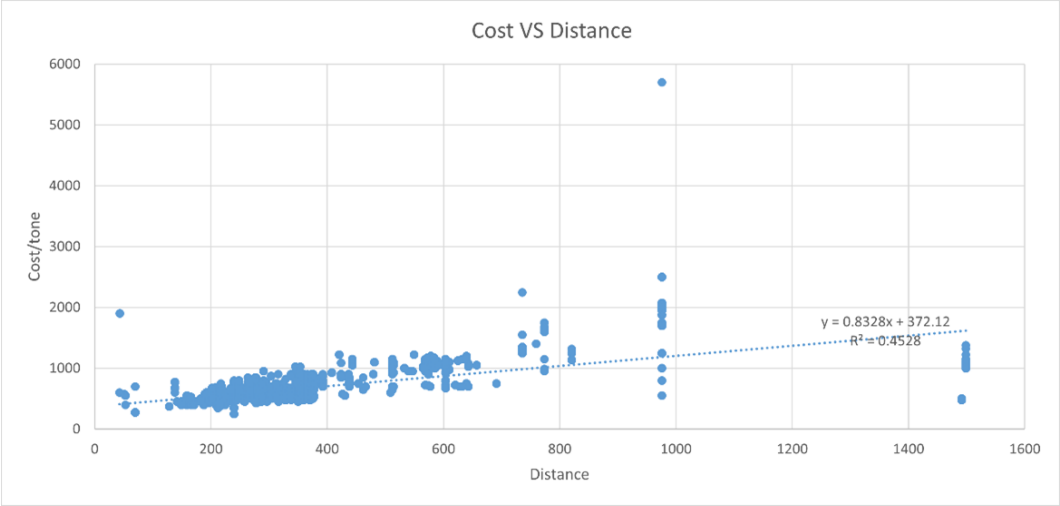
| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Target variable | | | | | |
| | Model 1.2 | Model 2.2 | Model 3.2 | Model 4.2 | Model 5.2 | Model 6.2 |
| exp_grain_3_ma_st | 114.576*** | 133.283*** | 153.266*** | 131.643*** | 134.852*** | 185.033*** |
| | (34.886) | (27.618) | (33.531) | (29.058) | (30.293) | (21.422) |
| exp_grain_east_v_ma | 0.382 | 0.928 | 2.491 | 0.652 | 0.408 | 0.352 |
| | (1.641) | (1.928) | (2.698) | (1.810) | (1.767) | (1.466) |
| p_diesel_st | 163.615*** | 167.161*** | | 179.711*** | 180.563*** | 177.106*** |
| | (31.919) | (37.722) | | (18.849) | (21.486) | (14.869) |
| exch_r | 2.424 | 5.281 | 4.506 | 5.470 | 4.340 | 2.818 |
| | (2.885) | (3.342) | (4.190) | (3.496) | (3.540) | (3.741) |
| p_wheat | 1.484*** | | | | | |
| | (0.356) | | | | | |
| ppi | −6.531** | 0.664 | 10.032*** | | | |
| | (3.256) | (2.881) | (2.704) | | | |
| exp_coal_5_fert | 0.006 | 0.024 | 0.001 | | | |
| | (0.044) | (0.040) | (0.043) | | | |
| exp_corn_3 | −0.391* | −0.542** | −0.937*** | −0.477*** | | |
| | (0.226) | (0.209) | (0.269) | (0.176) | | |
| jan_21 | 129.315*** | 177.726*** | 129.834* | 175.325*** | 188.745*** | 112.102*** |
| | (44.661) | (51.651) | (69.884) | (52.244) | (54.539) | (29.879) |
| harvesting | 75.353 | 17.985 | −49.170 | 21.892 | 29.702 | 95.245** |
| | (75.172) | (70.114) | (89.439) | (67.830) | (69.205) | (47.378) |
| new_year | 160.978*** | 125.237 | 237.975*** | 109.337 | 120.093 | 143.854 |
| | (54.781) | (106.333) | (83.844) | (119.854) | (119.400) | (123.245) |
| exp_grain_3_ma_st:jan_21 | | | | | | −262.794*** |
| | | | | | | (39.381) |
| exp_grain_3_ma_st:harvesting | | | | | | −139.534*** |
| | | | | | | (42.968) |
| Constant | 1,173.760** | 874.321* | −411.319 | 980.914*** | 1,061.923*** | 1,195.664*** |
| | (477.508) | (495.180) | (566.036) | (301.572) | (301.551) | (310.300) |
| *Note:* | | | | | | *p<0.1; **p<0.05; ***p<0.01 |

**Table 16:** Without squared exports. Newey-West standard errors.

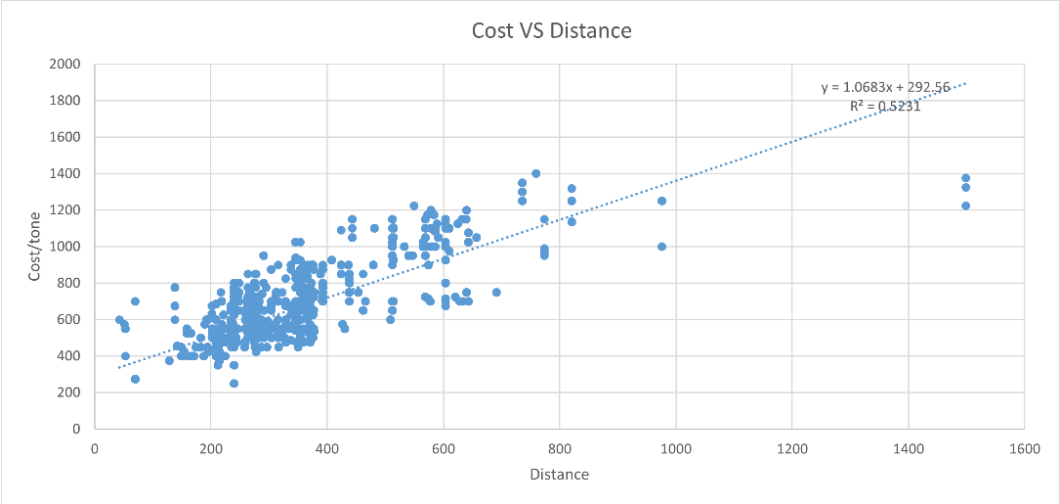
| (p, d, q) | AIC | MAE | (p, d, q) | AIC | MAE |
|---|---|---|---|---|---|
| (0,0,0) | - | - | (0,1,0) | 770 | 104 |
| (1,0,0) | 785 | 178 | (1,1,0) | 768 | 104 |
| (2,0,0) | 782 | 231 | (2,1,0) | 770 | 104 |
| (3,0,0) | 783 | 251 | (3,1,0) | 772 | 104 |
| (4,0,0) | 785 | 268 | (4,1,0) | 770 | 101 |
| (5,0,0) | 784 | 203 | (5,1,0) | 771 | 101 |
| (6,0,0) | 785 | 251 | (6,1,0) | 773 | 101 |
| (7,0,0) | 787 | 258 | (7,1,0) | 775 | 104 |
| (0,0,1) | 867 | 372 | (0,1,1) | 769 | 104 |
| (1,0,1) | 782 | 209 | (1,1,1) | 770 | 108 |
| (2,0,1) | 784 | 197 | (2,1,1) | 770 | 106 |
| (3,0,1) | 785 | 267 | (3,1,1) | 771 | 108 |
| (4,0,1) | 784 | 249 | (4,1,1) | 771 | 101 |
| (5,0,1) | 786 | 227 | (5,1,1) | 773 | 101 |
| (6,0,1) | 786 | 282 | (6,1,1) | 775 | 101 |
| (7,0,1) | 789 | 258 | (7,1,1) | 777 | 101 |
| (0,0,2) | 839 | 363 | (0,1,2) | 770 | 104 |
| (1,0,2) | 784 | 225 | (1,1,2) | 770 | 106 |
| (2,0,2) | 785 | 212 | (2,1,2) | 772 | 108 |
| (3,0,2) | 785 | 287 | (3,1,2) | 773 | 110 |
| (4,0,2) | 785 | 264 | (4,1,2) | 773 | 106 |
| (5,0,2) | 787 | 251 | (5,1,2) | 775 | 104 |
| (6,0,2) | 787 | 246 | (6,1,2) | 776 | 101 |
| (7,0,2) | 790 | 227 | (7,1,2) | 778 | 101 |
| (0,0,3) | 806 | 363 | (0,1,3) | 771 | 104 |
| (1,0,3) | 783 | 262 | (1,1,3) | 770 | 108 |
| (2,0,3) | 783 | 251 | (2,1,3) | 772 | 108 |
| (3,0,3) | 785 | 251 | (3,1,3) | 771 | 119 |
| (4,0,3) | 787 | 242 | (4,1,3) | 775 | 106 |
| (5,0,3) | 787 | 264 | (5,1,3) | 777 | 104 |
| (6,0,3) | 789 | 260 | (6,1,3) | 778 | 101 |
| (7,0,3) | 791 | 251 | (7,1,3) | 780 | 101 |

**Table 17:** ARIMA. Results

## 13.2 Appendix B. Plots. Data Analysis.

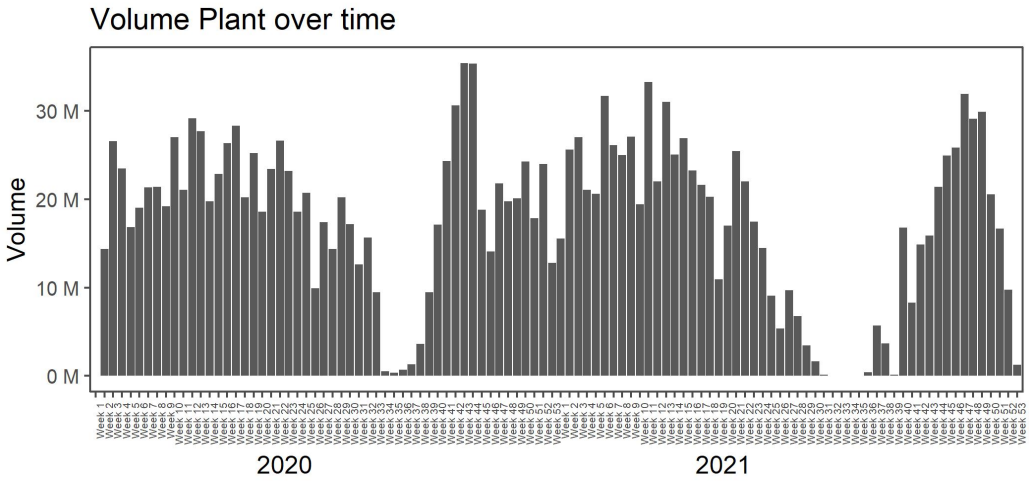



**(a)** Cost vs Distance. Before outliers removal.

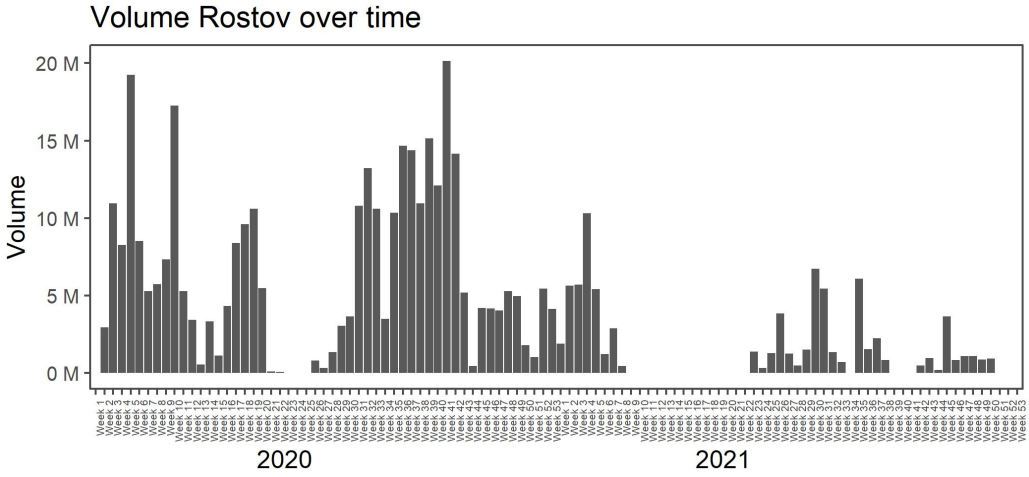



**(b)** Cost vs Distance. After outliers removal.

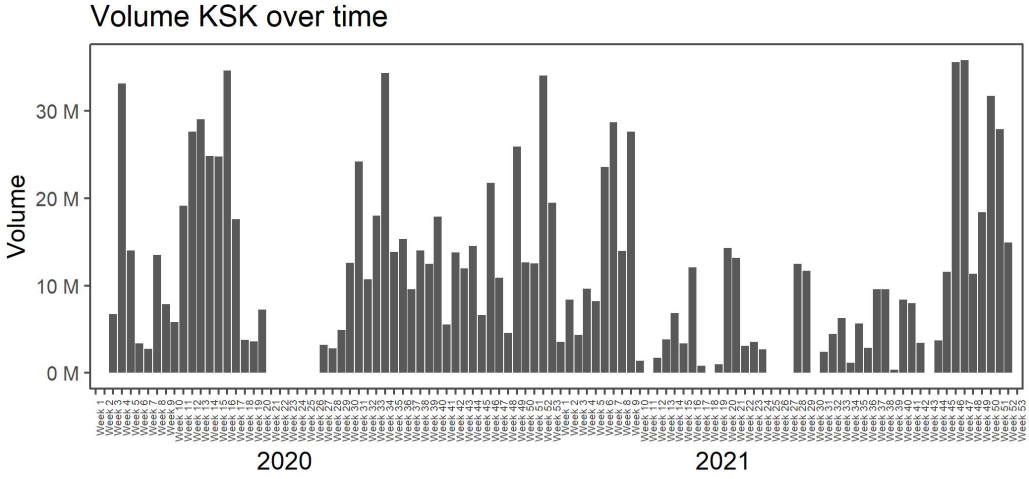
**Figure 11:** Before/after outliers removal.




**(a)** Weekly volume, Novoaninnskyy.




**(b)** Weekly volume, Rostov.




**(c)** Weekly volume, Novorossiysk.

**Figure 12:** Total weekly volume to different locations.

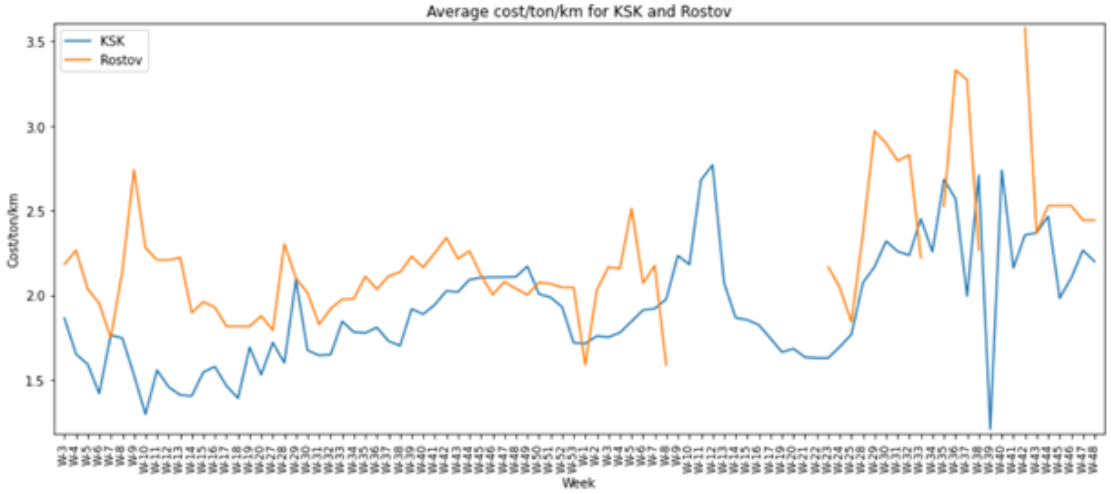



**Figure 13:** Cost/ton, plant VS ports

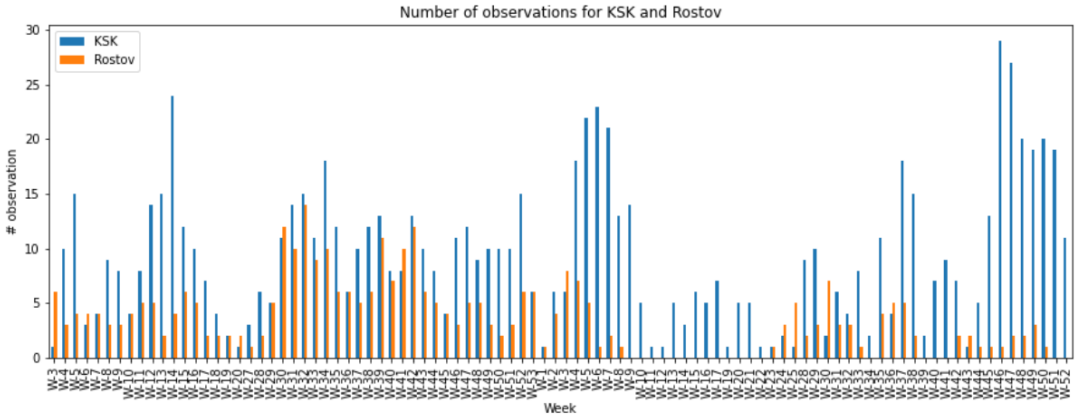



**Figure 14:** Number of observations. Novorossiysk VS Rostov

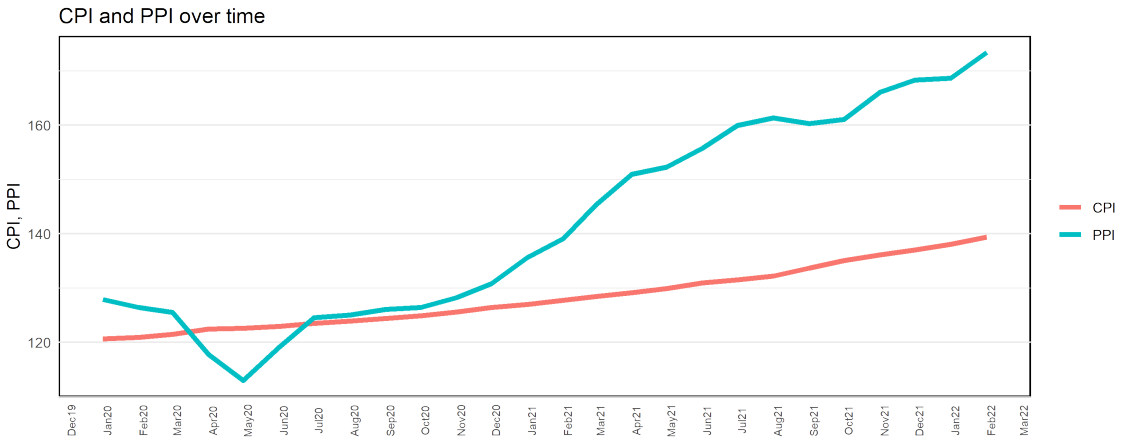



**Figure 15:** CPI and PPI over time

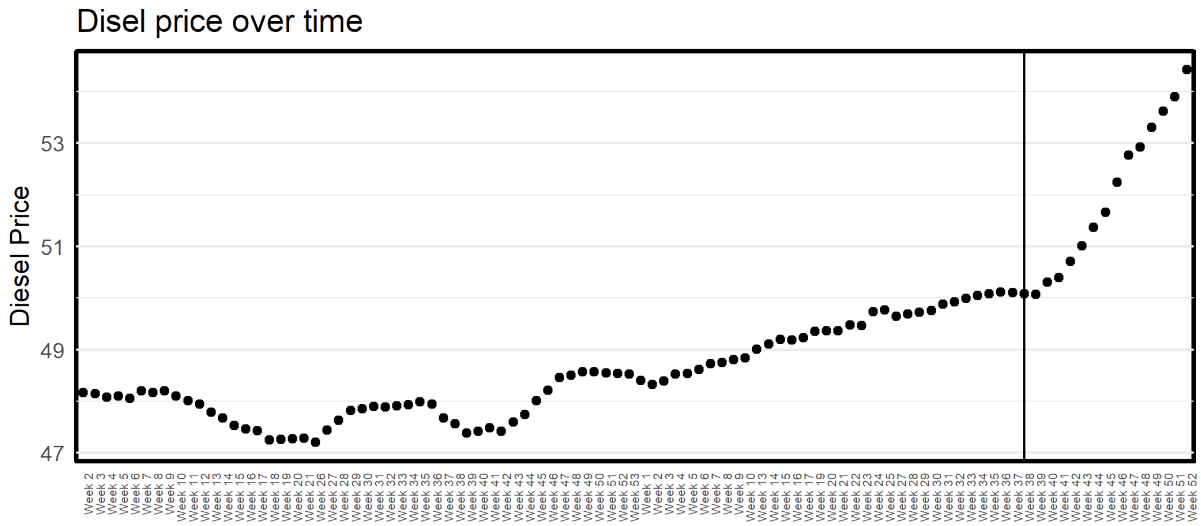



**Figure 16:** Diesel prices over time

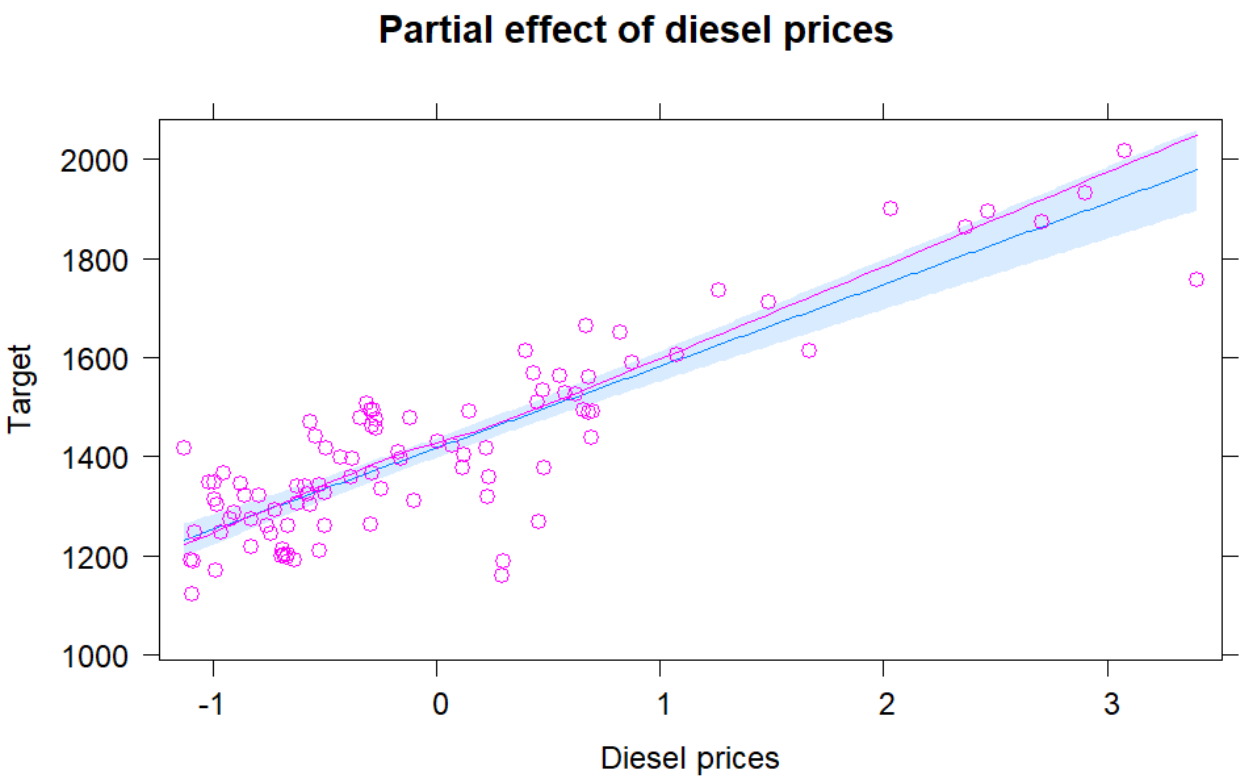



**Figure 17:** Partial effect. Price of diesel

# 14 References


1. Altay, E., Satman, M. H. (2005). Stock market forecasting: artificial neural network and linear regression comparison in an emerging market. Journal of Financial Management Analysis, 18(2), 18.

2. Ariyo, A. A., Adewumi, A. O., Ayo, C. K. (2014, March). Stock price prediction using the ARIMA model. In 2014 UKSim-AMSS 16th international conference on computer modelling and simulation (pp. 106-112). IEEE.

3. ATI.SU: рынок автомобильных грузоперевозок по итогам 2020 года вырос почти на 10%. (2021). Retrieved 12 May 2022, from https://www.tadviser.ru/index.php/Статья:Грузоперевозки_автомобильные_(рынок_России)

4. Azari, A. (2019). Bitcoin price prediction: An ARIMA approach. arXiv preprint arXiv:1904.05315.

5. Budak, A., Ustundag, A., Guloglu, B. (2017). A forecasting approach for truckload spot market pricing. Transportation Research Part A: Policy And Practice, 97, 55-68.
   doi: 10.1016/j.tra.2017.01.002

6. Bai, X., (2018). Forecasting short term trucking rates, M.S. thesis, Dept. Transp. Logistics, Massachusetts Inst. Technol., Cambridge, MA, USA.

7. Contreras, J., Espinola, R., Nogales, F., Conejo, A. (2003). ARIMA models to predict next-day electricity prices. IEEE Transactions On Power Systems, 18(3), 1014-1020. doi: 10.1109/tpwrs.2002.804943

8. De Oliveira, E., Cyrino Oliveira, F. (2018). Forecasting mid-long term electric energy consumption through bagging ARIMA and exponential smoothing methods. Energy, 144, 776-788. doi: 10.1016/j.energy.2017.12.049

9. Dhaval, B., Deshpande, A. (2020). Short-term load forecasting with using multiple linear regression. International Journal Of Electrical And Computer Engineering (IJECE), 10(4), 3911. doi: 10.11591/ijece.v10i4.pp3911-3917

10. Garlapati, A., Krishna, D. R., Garlapati, K., Rahul, U., Narayanan, G. (2021, April). Stock Price Prediction Using Facebook Prophet and Arima Models. In 2021 6th International Conference for Convergence in Technology (I2CT) (pp. 1-7). IEEE.

11. Güleryüz, D., Özden, E. (2020). The prediction of Brent crude oil trend using LSTM and Facebook Prophet. Avrupa Bilim ve Teknoloji Dergisi, (20), 1-9.

12. Ho, M., Darman, H., Musa, S. (2021). Stock Price Prediction Using ARIMA, Neural Network and LSTM Models. Journal Of Physics: Conference Series, 1988(1), 012041. doi: 10.1088/1742-6596/1988/1/012041

13. Ibrahim, Z., Rusli, D. (2007). Predicting students' academic performance: comparing artificial neural network, decision tree and linear regression. In 21st Annual SAS Malaysia Forum, 5th September.

14. Ismail, Z., Yahya, A., Shabri, A. (2009). Forecasting Gold Prices Using Multiple Linear Regression Method. American Journal Of Applied Sciences, 6(8), 1509-1514. doi: 10.3844/ajassp.2009.1509.1514

15. Jha, B. K., Pande, S. (2021, April). Time series forecasting model for supermarket sales using FB-prophet. In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 547-554). IEEE.

16. Kaninde, S., Mahajan, M., Janghale, A., Joshi, B. (2022). Stock Price Prediction using Facebook Prophet. In ITM Web of Conferences (Vol. 44, p. 03060). EDP Sciences.

17. Kay, M., Warsing, D. (2009). Estimating LTL rates using publicly available empirical data. International Journal Of Logistics Research And Applications, 12(3), 165-193.

    doi: 10.1080/13675560802392415

18. Kim, Y. S. (2008). Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size. Expert Systems with Applications, 34(2), 1227-1234.

19. Koenker, R., Hallock, K. F. (2001). Quantile regression. Journal of economic perspectives, 15(4), 143-156.

20. Kotur, D., Žarković, M. (2016, September). Neural network models for electricity prices and loads short and long-term prediction. In 2016 4th International Symposium on Environmental Friendly Energies and Applications (EFEA) (pp. 1-5). IEEE.

21. Lim, K., Nomikos, N., Yap, N. (2019). Understanding the fundamentals of freight markets volatility. Transportation Research Part E: Logistics And Transportation Review, 130, 1-15. doi: 10.1016/j.tre.2019.08.003

22. Liu, C., Hu, Z., Li, Y., Liu, S. (2017). Forecasting copper prices by decision tree learning. Resources Policy, 52, 427-434. doi: 10.1016/j.resourpol.2017.05.007

23. Melkadze, A. (2021). Topic: Transportation industry in Russia. Retrieved 9 April 2022, from https://www.statista.com/topics/7050/transportation-industry-in-russia/dossierKeyfigures

24. Miller, J. (2018). ARIMA Time Series Models for Full Truckload Transportation Prices. Forecasting, 1(1), 121-134. doi: 10.3390/forecast1010009

25. Özkaya, E., Keskinocak, P., Roshan Joseph, V., Weight, R. (2010). Estimating and benchmarking Less-than-Truckload market rates. Transportation Research Part E: Logistics And Transportation Review, 46(5), 667-682. doi: 10.1016/j.tre.2009.09.004

26. Patil, G. R., Sahu, P. K. (2016). Estimation of freight demand at Mumbai Port using regression and time series models. KSCE Journal of Civil Engineering, 20(5), 2022-2032.

27. Russia in the Global Transport and Logistics System: the Main Vectors of Development. (2022). Retrieved 3 April 2022, from https://credo-trans.com/international-logistics-business-and-freight-forwarding-in-russia/

28. Russia - Construction and Infrastructure. (2020). Retrieved 13 May 2022, from https://www.trade.gov/country-commercial-guides/russia-construction-and-infrastructure

29. Scott, A. (2015). The value of information sharing for truckload shippers. Transportation Research Part E: Logistics And Transportation Review, 81, 203-214. doi: 10.1016/j.tre.2015.07.002

30. Smith, L., Campbell, J., Mundy, R. (2007). Modeling net rates for expedited freight services. Transportation Research Part E: Logistics And Transportation Review, 43(2), 192-207. doi: 10.1016/j.tre.2005.11.001

31. Spreeuwenberg, S.J.J. (2020). Developing a forecast model for freight prices in Poland. M.S. thesis, Dept. of Economics and Econometrics, Tilburg University, Tilburg, Netherlands.

32. Swenseth, S.R., Godfrey, M.R. (1996). ESTIMATING FREIGHT RATES FOR LOGISTICS DECISIONS. Journal of Business Logistics.

33. Uras, N., Marchesi, L., Marchesi, M., Tonelli, R. (2020). Forecasting Bitcoin closing price series using linear regression and neural networks models. PeerJ Computer Science, 6, e279.

34. Velonias, P. M. (1995). Forecasting tanker freight rates (Doctoral dissertation, Massachusetts Institute of Technology).

35. Wanjawa, B. W., Muchemi, L. (2014). Ann model to predict stock prices at stock exchange markets. arXiv preprint arXiv:1502.06434

36. Xiao, W., Xu, C., Liu, H., Yang, H., Liu, X. (2020). Short-Term Truckload Spot Rates' Prediction in Consideration of Temporal and Between-Route Correlations. IEEE Access, 8, 81173-81189. doi: 10.1109/access.2020.2990751

37. Yang, J., Rahardja, S., Fränti, P. (2019, December). Outlier detection: how to threshold outlier scores?. In Proceedings of the international conference on artificial intelligence, information processing and cloud computing (pp. 1-6).